GAZE-CONTINGENT VISUAL COMMUNICATION

A Dissertation

by

ANDREW TED DUCHOWSKI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 1997

Major Subject: Computer Science

GAZE-CONTINGENT VISUAL COMMUNICATION

A Dissertation

by

ANDREW TED DUCHOWSKI

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Bruce H. McCormick (Chair of Committee)	Udo W. Pooch (Member)
John J. Leggett (Member)	Norman C. Griswold (Member)
Wayne L. Shebilske (Member)	Richard A. Volz (Head of Department)

August 1997

Major Subject: Computer Science

ABSTRACT

Gaze-Contingent Visual Communication. (August 1997)

Andrew Ted Duchowski, B.Sc., Simon Fraser University

Chair of Advisory Committee: Dr. Bruce H. McCormick

Virtual environments today lack realism. Real-time display of visually rich scenery is encumbered by the demand of rendering excessive amounts of information. This problem is especially severe in virtual reality. To minimize refresh latency, image quality is often sacrificed for speed. What's missing is knowledge of the participant's locus of visual attention, and an understanding of how an individual scans the visual field. Neurophysiological and psychophysical literature on the human visual system suggests the field of view is inspected *minutatim* through brief fixations over small regions of interest. Significant savings in scene processing can be realized if fine detail information is presented "just in time" in a *gaze-contingent* manner, delivering only as much information as required by the viewer.

An attentive model of vision is proposed where *Volumes Of Interest* (VOIs) represent fixations through time. The visual scanpath, composed of raw two-dimensional point of regard (POR) data, is analyzed over a sequence of video frames in time. Fixation locations are predicted by a piecewise auto-regressive integrated moving average (PARIMA) time series model of eye movements. PARIMA model parameters are derived from established spatio-temporal characteristics of eye movements. POR data is fitted to the PARIMA model through the application of the three-dimensional wavelet transform. Identified fixations are assembled into volumes in three-dimensional space-time, delineating dynamic foveal attention.

The attentive visual model utilizes VOIs to synthesize video sequences matching human visual acuity. Specifically, spatial resolution drops off smoothly with the degree of eccentricity from the viewer's point of gaze. Seamless degradation of individual video frames is accomplished through inhomogeneous wavelet reconstruction where the intersections of VOIs and frames constitute expected foveal regions. Peripheral degradation of video is evaluated through human subjective quality testing in a gaze-contingent environment. The proposed method of visual representation is applicable to systems with inherently intensive display requirements including teleoperator and virtual environments.

To my wife and family.

ACKNOWLEDGMENTS

I would like to thank Professor Bruce H. McCormick, my advisor and mentor, for his masterful guidance. His strict adherence to the scientific method inspired me to strive for a high level of rigor throughout the course of my investigation. The ultimate reward in maintaining this discipline is my personal satisfaction and confidence in the completion of this work. I will never forget Professor McCormick's insights and poignant prose. I am forever indebted and deeply grateful. I would like to extend thanks to my advisory committee members for their contribution and patience in the preparation of this dissertation. I thank Professor Wayne Shebilske from the Psychology Department for his invaluable help in setting up the eye tracking apparatus. His forewarning of potential difficulties in dealing with human subjects were right on the mark and I am grateful for his suggestions. I would also like to thank Dr. Shebilske for his help in my understanding of the theoretical aspects of human perception and performance and above all his encouragement. I would like to thank Professor Don House of the Visualization Department for putting together the summer workshop on wavelets. This workshop, together with Dr. House's enthusiasm, prompted me to learn this mathematical concept which forms the analytical foundation of the work found herein. Without Dr. House's interest and support I may never have braved through the initial learning curve. I would like to express my appreciation to Dr. Rick Jacoby from NASA Ames Research for his contribution to the development of the gaze-contingent video system. Rick's help with the implementation of shared memory prompted me to design the eye tracking system without which none of the human subject experiments would have been possible. I would like to extend my warm thanks to Drs. Nancy Amato and Mac Lively for their help and guidance in the preparation of my defense. I would also like to express my thanks to the support staff of the Computer Science Department at Texas A&M. They are the silent support force who enabled me to put this work together. I thank my family for seeing me through graduate school. Their moral (and of course financial) support carried me through. Finally, I owe my greatest debt of gratitude to my wife, Corey. Thanks for your patience, understanding, and above all your love.

This research was supported in part by the National Science Foundation, under Infrastructure Grant CDA-9115123 and CISE Research Instrumentation Grant CDA-9422123, and by the Texas Advanced Technology Program under Grant 999903-124.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

APPENDIX Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

> "Vision is investigated by three different schools of the scientific community. *Neurophysiologists* attempt to understand how sensory and neural mechanisms of biological systems function. *Perceptual Psychologists* try to understand the psychological issues governing the task of perception, and *Computer Vision Scientists* investigate the computational and algorithmic issues associated with image acquisition, processing, and understanding."
>
> – Trivedi and Rosenfeld (as found in [JR94, p.1]).

Computer scientists build systems. This dissertation addresses the design of a gaze-contingent system aimed at matching the requirements of human visual attention. This differs significantly from the goal of building a system to replicate human vision through computational means. The latter is reminiscent of early artificial intelligence efforts which, although ambitious, have not yet been fully realized. In contrast, the work developed here is concerned primarily with adapting visual display systems to exploit inherent limitations of human perception.

Neurophysiological and psychophysical literature on the human visual system suggests the field of view is inspected *minutatim* through brief fixations over small regions of interest. This allows perception of detail through the fovea. Foveal vision, subtending 5° (visual angle), allows fine scrutiny of 3% of the entire screen (21in monitor at ∼60cm viewing distance). Approximately 90% of viewing time is spent in fixations. When visual attention is directed to a new area, fast eye movements (saccades) reposition the fovea. Significant savings in scene processing can be realized if fine detail information is presented "just in time" in a *gaze-contingent* manner, delivering only as much information as required by the viewer.

1.1 Research Objective

The objective of this dissertation is to build a gaze-contingent (GC) visual display system to match human vision. GC systems display computer-mediated imagery (e.g., video or graphics) manipulated in a manner dependent on (the system's) knowledge of the viewer's point of regard. Application examples include, but are not limited to, video telephony, flight simulation, and virtual reality. These visual communication systems

This dissertation follows the style and format of SIAM Journal on Numerical Analysis.

display information to the viewer through imagery. The term "communication" as used here loosely refers to human-computer communication rather than to communication between computers (e.g., over networks), or between processor and monitor (e.g., real-time display). The latter type of communication is a subsidiary consideration as a result of bandwidth-limited information requirements outside the viewer's field of view. The main objective of the current work is to adaptively present video information matching the visual attention and perceptual (not performance) capacity of the human visual system.[1]

The task of a visual communication system is to display natural scenery by either:

- replicating the environment through a digitally-processed facsimile (e.g., video), in a manner indiscernible from the original, or
- synthetically modeling the environment (e.g., graphics) to evoke its perception.

Examples of the former include digital imagery where natural scenes are manipulated (encoded/decoded) in an imperceptible manner. A prominent instance of this methodology is the Joint Photographers Experts Group (JPEG) image compression standard [Wal91]. The JPEG standard quantizes digital imagery at just-perceptible degradation quality levels. An example of the latter is virtual reality, where scenery is generated to imitate reality. Virtual reality creates an illusory environment where computer generated imagery is meant to evoke the perception of natural scenes. Realism and computational efficiency are common goals of the two domains. The visual communication system proposed herein falls in the former class since the objective is a method of video degradation imperceptible to the human observer. The methods for imperceptible degradation developed in this thesis are directly applicable to the latter domain (e.g., virtual reality) where the distinction between natural and synthetic scenery is becoming increasingly blurred. The principles behind the imperceptible display methodology are applicable to both domains.

The experimental testbed developed to evaluate imperceptible video display techniques is a multithreaded program which simultaneously transfers video and tracks eye movements in real-time. Treating the eye tracker as an ordinary positional sensor, the architecture shares many similarities with traditional virtual environment system designs.

It is tacitly assumed that the measured point of regard: (1) coincides with the subject's true point of gaze, and (2) indicates the subject's focus of attention. Under this assumption, the objectives of this dissertation are three-fold: (1) the study of the dynamic aspects of eye movements, (2) the visualization of attention in space-time, and (3) the quantification of perceptual limitations of human foveal and peripheral vision.

[1]The distinction between visual perception and performance is relevant to the empirical evaluation of the peripheral image degradation technique, and is addressed in §XIII.

1.2 Specific Aims

To address the gaze-contingent requirements of the visual display, this dissertation targets three specific aims:

1. Development of eye movement analysis and modeling techniques for the spatiotemporal localization of saccades, fixations, and smooth pursuit movements.
2. Visualization of eye movements in space-time through *Volumes Of Interest* providing an aggregate graphical representation of visual attention.
3. Implementation of video resolution degradation methods matching human visual acuity.

The first and third objectives are realized using the multiresolution discrete wavelet transform through its properties of frequency localization and multiscale subsampling. A real-time eye tracker is used for eye movement data acquisition and performance of subjective quality experiments.

1.3 Dissertation Organization

The human visual system is studied from three conceptual perspectives. First, the phenomenon of visual attention is presented in Chapter II. Second, the structure and functionality of the visual system is described in Chapter III. Third, characteristics of eye movements are reviewed in Chapter IV. Chapters II through IV constitute the theoretical foundation for the development of the gaze-contingent system.

Chapter V introduces wavelet theory, which forms the mathematical framework for deterministic digital signal analysis. Chapter VI reviews time series analysis appropriate for modeling stochastic processes. These chapters provide the mathematical framework for eye movement analysis and modeling, and gaze-contingent video synthesis. Readers familiar with wavelet and time series analysis theories may omit these chapters.

Chapter VII integrates wavelet analysis in the context of time series and proposes a wavelet-based time series model of eye movements. Chapter VIII introduces a three-dimensional graphical technique for visualizing eye movements in space-time. Aggregate three-dimensional eye movement patterns constitute an attentive model of vision in the form of *Volumes Of Interest* (VOIs) representing dynamic fixations. Chapter IX develops a wavelet-based image synthesis technique matching human visual acuity. This method is used to smoothly degrade resolution in peripheral regions of an image. Chapters VII through IX resolve the three specific aims of this dissertation.

Evaluation of the proposed techniques is presented in chapters X through XIII. Chapter X presents the experimental method and apparatus, and empirical results of three experiments are given in Chapters XI, XII, and XIII.

Concluding remarks are offered in Chapter XIV and future directions are recommended in Chapter XV.

CHAPTER II

VISUAL ATTENTION

> "Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others..."
>
> –William James [Jam81, pp.381-382]

Humans are finite beings that cannot attend to all things at once. Attention is used to focus our mental capacities on selections of the sensory input so that the mind can successfully process the stimulus of interest. Our capacity for information processing is roughly bounded by the "magic" number 7 ± 2 [Mil56]. While listening to an orchestra, it is possible to concentrate on specific instruments, e.g., the flute or the oboe. The brain processes sensory input by concentrating on specific components of the entire sensory realm so that interesting sights, sounds, smells, etc., may be examined with greater attention to detail than peripheral stimuli. This is particularly true of vision. Visual scene inspection is performed *minutatim*, not *in toto*. That is, human vision is a piecemeal process relying on the perceptual integration of small regions to construct a coherent representation of the whole. In this chapter, attention is recounted from a historical perspective following the narrative found in [Van92]. The discussion focuses on attentional mechanisms involved in vision, with emphasis on two main components of visual attention, namely the "what" and the "where".

2.1 Chronological Review of Visual Attention

The phenomenon of visual attention has been studied for over a century. Early studies of attention were technologically limited to simple ocular observations and oftentimes to introspection. Since then the field has grown to an interdisciplinary subject involving the disciplines of psychophysics, cognitive neuroscience, and computer science, to name three. This section presents a qualitative historical background of visual attention.

2.1.1 Von Helmholtz's "where"

At the start of the 20^{th} century, Hermann Von Helmholtz posited visual attention as an essential mechanism of visual perception. In his *Treatise on Physiological Optics*, he notes, "We let our eyes roam continually over the visual field, because that is the only way we can see as distinctly as possible all the individual parts of the field in turn." [VonH25, p.63]. Noting that attention is concerned with a small region of space, Von Helmholtz

observed visual attention's natural tendency to wander to new things. He also remarked that attention can be controlled by a conscious and voluntary effort, allowing attention to peripheral objects without making eye movements to that object. Von Helmholtz was mainly concerned with eye movements to spatial locations, or the "where" of visual attention. In essence, although visual attention can be consciously directed at peripheral objects, eye movements reflect the will to inspect these objects at fine detail. In this sense, eye movements provide evidence of overt visual attention.

2.1.2 James' "what"

In contrast to Von Helmholtz's ideas, William James believed attention to be a more internally covert mechanism akin to imagination, anticipation, or in general, thought [Jam81, § XI]. James defined attention mainly in terms of the "what", or the identity, meaning, or expectation associated with the focus of attention. James favored the active and voluntary aspects of attention although he also recognized its passive, reflexive, non-voluntary and effortless qualities.

Both views of attention, which are not mutually exclusive, bear significantly on contemporary concepts of visual attention. The "what" and "where" of attention roughly correspond to foveal (James) and parafoveal (Von Helmholtz) aspects of visual attention, respectively.

2.1.3 Broadbent's "selective filter"

Attention, in one sense, is seen as a "selective filter" responsible for regulating sensory information to sensory channels of limited capacity. In the 1950's, Donald Broadbent performed auditory experiments designed to demonstrate the selective nature of auditory attention [Bro58]. The experiments presented a listener with information arriving simultaneously from two different channels, e.g., the spoken numerals $\{7,2,3\}$ to the left ear, $\{9,4,5\}$ to the right. Broadbent reported listeners' reproductions of either $\{7,2,3,9,4,5\}$, or $\{9,4,5,7,2,3\}$, with no interwoven (alternating channel) responses. Broadbent concluded that information enters in parallel but is then selectively filtered to sensory channels.

2.1.4 Deutsch and Deutsch's "importance weightings"

In contrast to the notion of a selective filter, J. Anthony Deutsch and Diana Deutsch proposed that all sensory messages are perceptually analyzed at the highest level, precluding a need for a selective filter [DD63]. Deutsch and Deutsch rejected the selective filter and limited capacity system theory of attention; they reasoned that the filter would need to be at least as complex as the limited capacity system itself. Instead, they proposed the existence of central structures with preset "importance weightings" which determined selec-

tion. Deutsch and Deutsch argued that it is not attention as such but the weightings of importance that have a causal role in attention. That is, attentional effects are a result of importance, or relevance, interacting with the information.

It is interesting to note that Broadbent's selective filter corresponds to Von Helmholtz's "where", while Deutsch and Deutsch's importance weightings correspond to James' expectation, or the "what". These seemingly opposing ideas were incorporated into a unified theory of attention by Anne Treisman in the 1960's (although not fully recognized until 1971). Treisman brought together the attentional models of Broadbent and Deutsch and Deutsch by specifying two components of attention: the attenuation filter followed by later (central) structures referred to as 'dictionary units'. The attenuation filter is similar to Broadbent's selective filter in that its function is selection of sensory messages. Unlike the selective filter, it does not completely block unwanted messages, but only attenuates them. The later stage dictionary units then process weakened and unweakened messages. These units contain variable thresholds tuned to importance, relevance, and context. Treisman thus brought together the complementary models of attentional unit or selective filter (the "where"), and expectation (the "what").

2.1.5 Noton and Stark's "scanpaths"

In their study of eye movements, Noton and Stark helped cast doubt on the Gestalt hypothesis that recognition is a parallel, one-step process [NS71a, NS71b]. The Gestalt view of recognition is a wholistic one which suggests that vision relies to a great extent on the tendency to group objects. Although well known visual illusions exist to support this view (e.g., subjective contours of the Kanizsa figure, see [Mar82, p.51]), results of Noton and Stark's work suggest that visual recognition is serial in nature. Noton and Stark measured eye movements over images and identified these patterns as "scanpaths". Scanpaths show that subjects tend to fixate identifiable regions of interest, or "informative details", even though the order of eye movements over these regions is quite variable. That is, given a picture of a square, subjects will fixate on the corners, although the order in which the corners are viewed differs from viewer to viewer and even differs between consecutive observations made by the same individual.[1] In contrast to the Gestalt view, Noton and Stark's work suggests that a coherent picture of the visual field is constructed piecemeal through the assembly of serially viewed regions of interest. Noton and Stark's results support James' "what" of visual attention. With respect to eye movements, the "what" corresponds to regions of interest selectively filtered by foveal vision for detailed processing.

[1] See also the work of Yarbus [Yar67].

2.1.6 Posner's "spotlight"

Contrary to the serial "what" of visual attention, the orienting, or the "where", is performed in parallel [PSD80]. Posner suggested an attentional mechanism able to move about the scene in a manner similar to a "spotlight". The spotlight, being limited in its spatial extent, seems to fit well with Noton and Stark's empirical identification of foveal regions of interest. Posner, however, dissociates the spotlight from foveal vision. Instead, the spotlight is an attentional mechanism independent of eye movements. Posner identified two aspects of visual attention: the orienting and the detecting of attention. Orienting of attention may be an entirely central (covert, or mental) aspect of attention, while the detecting aspect is context-sensitive requiring contact between the attentional beam and the input signal. The orienting of attention is not always dependent on the movement of the eyes. That is, it is possible to attend to an object while maintaining gaze elsewhere. According to Posner, the orienting of attention must be done in parallel and must precede the detecting aspect of attention.

2.1.7 Treisman's "glue"

The dissociation of attention from foveal vision is an important point. In terms of the "what" and the "where", it seems likely that the "what" relates to serial foveal vision. The "where", on the other hand, is a parallel process performed parafoveally, or peripherally, which dictates the next focus of attention.[2] Posner and Noton and Stark advanced the theory of visual attention along similar lines forged by Von Helmholtz and James (and then Broadbent and Deutsch and Deutsch). Treisman once again brought these concepts together with a feature integration theory of visual attention [TG80, Tre86]. In essence, attention provides the "glue" which integrates the separated features in a particular location so that the conjunction, i.e., the object, is perceived as a unified whole. Attention selects features from a master map of locations showing *where* all the feature boundaries are located, but not *what* those features are. That is, the master map specifies where things are, but not what they are. The feature map also encodes simple and useful properties of the scene such as color, orientation, size, and stereo distance.

2.1.8 Kosslyn's "window"

Recently, Kosslyn proposed a refined model of visual attention [Kos94]. Kosslyn describes attention as a selective aspect of perceptual processing, and proposes an attentional "window" responsible for selecting patterns in the "visual buffer". The window is needed since there is more information in the visual buffer than can be passed downstream, and hence the transmission capacity must be selectively allocated. That is,

[2] This point will be examined further when independent processing of object shape and location is discussed in §3.1.2.

some information can be passed along, but other information must be filtered out. This notion is similar to Broadbent's selective filter and Treisman's attenuation filter. The novelty of the attentional window is its ability to be adjusted incrementally, i.e., the window is scalable. Another interesting distinction of Kosslyn's model is the hypothesis of a redundant stimulus-based attention-shifting subsystem (e.g., a type of context-sensitive spotlight) in mental imagery. Mental imagery involves the formation of mental maps of objects, or of the environment in general. It is defined as "...the mental invention or recreation of an experience that in at least some respects resembles the experience of actually perceiving an object or an event, either in conjunction with, or in the absence of, direct sensory stimulation" [Fin89, p.2].

2.1.9 Chronological Review Summary

An historical account of attention is a prerequisite to forming an intuitive impression of the selective nature of perception. The singular idioms describing the selective nature of attention are the "what" and the "where". The "where" of visual attention corresponds to the visual selection of specific regions of interest from the entire visual field for detailed inspection. Notably, this selection is often carried out through the aid of peripheral vision. The "what" of visual attention corresponds to the detailed inspection of the spatial region through a perceptual channel limited in spatial extent.

2.2 Visual Search

The area of fixation is referred to as the *perceptual span*, or perhaps more appropriately, the *attentional span* [Hen92]. The shape of the attentional span may be asymmetrical due to attentional factors: readers may have a strategy for reading, anticipating oncoming words [Ore92]. For example, the window is skewed to the right for readers of English, and it is skewed to the left for readers of Hebrew text. Irwin used a moving window to limit the number of text characters presented to readers [Irw92]. A window of only the current word plus 14 characters was sufficient for reading purposes. The general shape of the attentional span is not well understood. In viewing imagery, there is no "right way" to look at a picture [Ken92]. If the picture is accompanied by text describing it, the description may present an overt strategy to the reader [Heg92, Koo88]. In either case (with or without text) there appears to be no canonical scanpath for viewing pictures [NS71a].

Posner's spotlight model and Treisman's Feature Integration Theory (FIT) are both well known theories predicting visual search to limited extents [KKP92, Ser92]. Extensions to these theories have sparked debate as to whether visual attention is carried out in sequence or in parallel. Henderson's *sequential attention model* predicts saccadic programming based on an estimated 350ms "timeout," whereby if the next fixation has not been programmed, the eyes re-fixate on the current target [Hen92]. Wolfe's Guided Search (version

2.0) extends FIT and argues that a limited number of basic visual features are analyzed in parallel across large portions of the visual field [Wol93]. The plausibility of parallel, attentional processing of the periphery is gaining acceptance. Evidence to support parallel aspects of Wolfe's theory was recently demonstrated through electrophysiological recordings of brain activity in [Luc93].

Visual search is at least partially deterministic for an individual, and not completely random [DMS93]. The next fixation location can be determined in either or both the following ways: the viewer's strategy (as in reading), and/or based on knowledge of surrounding stimuli gained through peripheral vision. The relative importance of the surrounding stimulus, however, is still under investigation, although it is known that in its absence gaze durations increase, suggesting problematic processing of the scene.

2.3 Scene Integration

Given that the attentional span is constrained spatiotemporally, an interesting problem arises: how is an individual able to "piece together" the entire scene? This is known as the *scene integration* problem. At present, no coherent theory of scene recognition exists [DeG92]. Proposed partial theories of scene integration suggest that in order to perceive an entire scene, the pieces observed foveally (during fixations) need to be spliced together [HLM92]. Extrafoveal information may contribute to scene perception as well. There may be information in the periphery that adds to the processing of the scene. This is known as *pre-attentive* (or *preview*) benefit. For the purposes of visual communication, it is clear that in order to present a coherent video sequence to the viewer, both foveal and peripheral regions must contain sufficient information for the viewer to perceive the image. In order to minimize the amount of data from the system's point of view, both foveal and peripheral imagery need to be characterized in terms of saliency. The notion that information from pixel-to-pixel is retained is almost certainly wrong—there is no "integrative buffer", "retinoid", or "stable feature frame" representation [Irw92]. This is due to the motion-sensitive, single-cell physiological organization of the human visual system (see §3.2.2.5). The "images" that are stored are probably more abstract, akin to Marr's 2 1/2D sketch [Mar82]. This may be the reason small differences in successive video frames are difficult to perceive. Thus, characterizing saliency requires determining just the right, or perhaps "just perceptible", image components.

2.4 Summary

The attentional "what" and "where" duality is relevant to display and interface design since this concept can guide informational content presentation by matching the processing capacity of foveal and peripheral vision. In visual search work, the consensus view is that a parallel, pre-attentive stage acknowledges the presence of

four basic features: color, size, orientation, and presence and/or direction of motion [Dol93, Wol93]. Doll et al. suggest that features likely to attract attention include edges, corners, but not plain surfaces [DMS93]. Todd and Kramer suggest that attention (presumably in the periphery) is captured by sudden onset stimuli, uniquely colored stimuli (to a lesser degree than sudden onset), and bright and unique stimuli [TK93]. There is doubt in the literature that human visual search can be described as an integration of independently processed features [VOD93]. Van Oden and DiVita suggest that "...any theory on visual attention must address the fundamental properties of early visual mechanisms." To attempt to quantify the visual system's processing capacity, the neural substrate of the human visual system is examined in Chapter III which surveys the relevant neurological literature.

CHAPTER III

NEUROPHYSIOLOGY

The limited information capacity of the Human Visual System (HVS) provides epistemological reasons for visual attention from a phylogenetical standpoint, and is the *raison d'être* of visual attention. The dynamics of visual attention probably evolved in harmony with (or perhaps in response to) the perceptual limitations imposed by the neurological substrate of the visual system. For this reason, the design of a visual communication system matched to human perception must consider the neural mechanisms of the HVS. The neural substrate of the human visual system is examined in this section. Emphasis is placed on differentiating the processing capability of foveal and peripheral vision.

3.1 The Brain and the Visual Pathways

The cerebral cortex is composed of numerous regions classified by their function [Zek93]. A simplified representation of cortical regions is shown in Figure 1. The human visual system is functionally described by the

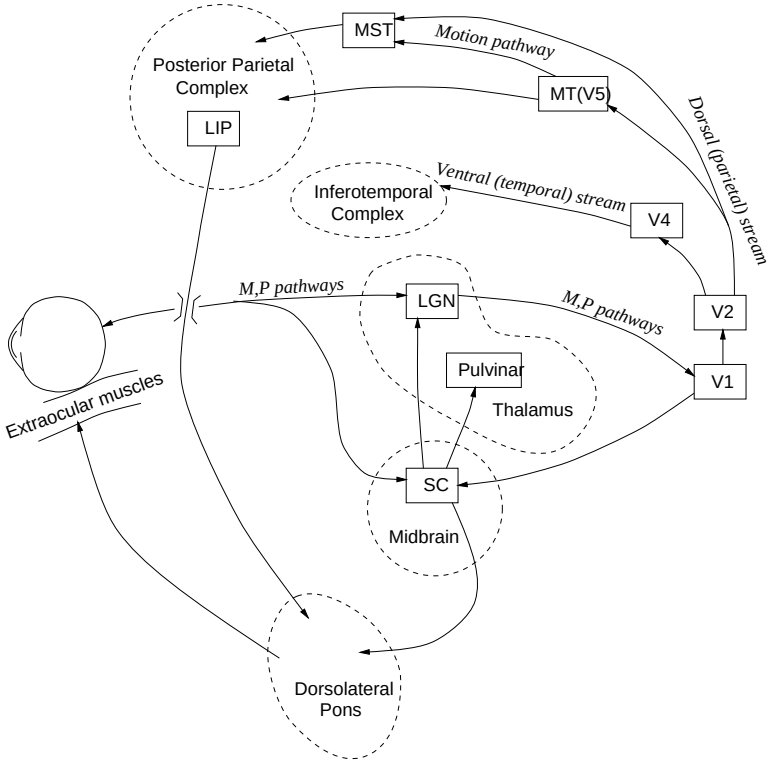


Fig. 1. The brain and the visual pathways.

connections between the retina and cortical regions, known as visual pathways. Pathways joining multiple cortical areas involved in common visual functions are referred to as streams. Figure 1 highlights regions and pathways relevant to selective visual attention. For clarity, many connections, particularly involving the Pulvinar, are omitted.

There may be as many as twenty regions topographically coding all or part of the visual field. The visual stimulus is represented (encoded) by at least two functionally distinct cognitive and sensorimotor representations [Bri95]. The cognitive representation provides visual knowledge (perception) of identities and locations of objects, while the sensorimotor encoding drives visually guided behavior. These dual vision systems do not always function independently, as will be discussed in §3.1.1.

Visual pathways nervate from the retina, through the optic chiasm and the Lateral Geniculate Nucleus (LGN). The LGN is the visual nucleus body in the Thalamus [Mum95]. The Thalamus is a collection of smaller (functionally specific and non-specific) nuclei. Essentially all input to the cortex is relayed through the Thalamus. The Pulvinar is another member of the Thalamus [OK95]. It is a large subcortical structure, itself divided into at least four smaller nuclei. Although not shown in Figure 1, it is heavily interconnected with visual cortical areas. The projection of the LGN to the visual (striate) cortex (area V1) preserves the two-dimensional retinal layout of the visual image. Thalmacortical pathways are reciprocated by feedback pathways from the cortex back to the Thalamus, forming a massive system of local loops between the Thalamus and the entire cortex. It is believed that cortical feedback gates thalamic transmission of subcortical data, i.e., feedback allows the cortex to attend to part of these data selectively. Along the retinocortical projections, pathways branch off to the Superior Colliculus (SC), a region implicated in saccadic programming. In a double-step paradigm, single unit recordings have shown that, before the first saccade, the brain encodes the second target's retinotopic location within the parietal cortex, frontal eye field, and the SC [PS95]. During execution of the first eye movement, this information is updated to represent the location of the second target as it would be found on the retina after the first saccade. The SC also mediates the function of remapping auditory space into visual coordinates (presumably) for the purpose of target foveation through saccadic eye movements. In general terms, it is thought the parietal cortex "disengages attention", the SC "moves attention", and the Pulvinar "engages attention" [OK95].

The LGN projects parallel Magno- and Parvo-cellular (M-, P-) pathways to the striate cortex. The visual cortex (area 17, or V1) is the primary visual processing center (top of the visual hierarchy). Pathways proceed from the visual cortex to higher visual centers, V2–V4, the Middle Temporal (MT) area, or area V5, and the Middle Superior Temporal (MST) cortex. Functionally, center V1 is responsible for the detection of a complete range of stimuli, e.g., color, motion, and orientation. Centers V2, V3, V3A, V4, MT, etc., are the

secondary specialized visual centers, responsible for a host of higher-level visual functions such as retinal disparity, orientation of contours, direction of motion, size, shape, and color. The MT and MST furnish a large projection to the Dorsolateral Pons, known to contribute to smooth pursuit eye movements [GN95].

Nervations emanating from V1 to the parietal and temporal cortices are referred to as the parietal (dorsal stream) and temporal (ventral stream) pathways, respectively [Nel95]. The temporal cortex is associated with cognitive aspects of vision while the parietal cortex is involved with sensorimotor functions. The ventral stream corresponds to the central visual field, and is generally identified as the "what" of visual processing. The dorsal stream is affiliated with the peripheral visual field and performs the "where" function. Specifically, the ventral stream performs precise figure synthesis and recognition of successively fixated objects and shares the output with the dorsal stream. Parietal areas along the dorsal stream perceive and learn the spatial arrangements of objects. Nelson also refers to the ventral and dorsal pathways as the object and place streams, respectively. Dorsal cells encoding peripheral information are more numerous and exchange more spatial information than ventral cells which share perceptual information regarding the fixated object of interest.

The temporal pathway contains regions in which neurons respond to specific features or properties of fixated objects [Bri95]. The pathway culminates in the inferotemporal cortex, a region composed of large receptive fields, requiring very specific and complex trigger features. The nonretinotopic representation found within the inferotemporal cortex suggests that it specializes in recognition. Neurons in this region usually include the fixation point.

The parietal pathway specializes in physical features of the visual world, e.g., motion and location. Its spatial function is further subdivided by two processing regions, the Lateral Intra-Parietal (LIP) area in the Posterior Parietal Cortex (PPC) and area 7a. The latter area is involved in ocular fixation and contains neurons which provide information sufficient for reconstruction of the fixated object's spatial position in head-centered coordinates. Responses of these neurons depend on the retinal target location and the orbital position of both eyes. The LIP area contains receptive fields which are corrected before each saccade so that they may respond to future stimuli. Changes in these receptive fields can be conceived as activity that signals candidates for planned eye movements. The LIP area directly projects to eye movement centers and is active during the programming of saccadic eye movements [MA95].

3.1.1 Independence of Visual Channels

The human visual system (HVS) has been compared to a set of independent processors in the sense that visual components such as color, motion, orientation, etc., are processed through separate, loosely coupled channels. Borrowing parallel algorithm terminology, tightly coupled machines (also referred to as multiprocessors) share a common memory, while loosely coupled machines (multicomputers) rely on an interconnection network [Akl89, p.17]. The essential difference between the two is that "multicomputers" are able to perform work independently, whereas "multiprocessors" are more dependent on each other. In this view, there is more evidence for loose coupling of visual areas than strong coupling. The neurological substrate of the HVS proffers a parallel view of computation among visual areas over a serial view of a hierarchical visual pipeline. The latter view, considering a serial 'flow of data' from areas V1 to V2 to V3, through to V5, where the sequence of visual areas successively computes and adds different features of the visual array, is no longer common since it does not recognize the known complexity of the system [Kaa89]. Kaas states that "the obvious problem with this line of reasoning is that the proposal ultimately seems unworkable because too many classes of very specific neurons would be needed at the highest level."

Evidence for loose coupling comes from two major physiological findings:

1. There is apparent high modularity of visual processing, particularly for handling color, shape, motion and location [LH88]. Each visual attribute seems to correspond to a highly specialized cortical center that responds to the attribute's presence. This physiological segregation is evident from the high degree of parallelism of magno-cellular (M) and parvo-cellular (P) channels defining four functional pathways, carrying information from X, Y, and W retinal ganglion cells [VOD93, Zek93, Kaa89]. These cell types are described below.
2. Lesion studies show that some functions remain intact while selected mechanisms are destroyed. For example, selective destruction of centers responsible for motion perception allows subjects to perform saccades to targets, but disables pursuit eye movements [And89].

The evidence for the existence of specialized visual processing centers supports the notion of loose coupling in that there is no 'common memory' for the independent processors to share. Paradoxically, evidence for the interconnections between the centers supports the idea of "re-entry" of visual information into areas V1 and V2, which may suggest strong coupling [Zek93].

There has been some opposition raised to a completely modular view of visual processing [SA93]. Stoner and Albright argue that there may be a high degree of cooperation among processors. For example, even though motion processing is considered to be specialized in the middle temporal (MT) area, it may be strongly linked to other visual centers such as those thought to be responsible for processing shape information. The cooperation of both areas may be required for shape-from-motion processing.

Another argument against total independence and specialization of visual processors is that each control zone may be able to undertake at least two operations, possibly more [Zek93]. Zeki notes that while the visual system appears functionally segregated, processors and interconnections are not hard-wired, instead they are adaptable providing "brain plasticity." This type of dynamic reconfigurability would require strong cooperation among visual centers, possibly even a centralized control mechanism to allocate processing functions (i.e., something akin to a job scheduler).

The idea of a modular, loosely coupled visual processing center seems to end in paradox—while there may be independent, multiple areas, parallel pathways, and a "deep division of labor", the visual experience is one of wholeness [Zek93]. The paradox raises the open question of how the brain integrates information processed by specialized visual areas. According to Zeki, integration is inherently non-modular in character. For a strongly coupled, integrative (shared memory) mechanism, there would have to be an area where all specialized processors project to. There is no such area, although many visual centers re-enter (back-project) to areas V1 and V2. The existence of feedback connections implies stronger connections among visual zones than purely independent processors.

The evidence for loose coupling suggests that highly specialized visual areas exist, capable of processing various visual features simultaneously. Because these areas are adaptable, capable of carrying out more than one function, and because of the fact that the brain must somehow accomplish coherent scene integration, it seems that the apparently loosely coupled areas must be highly interconnected.

3.1.2 Independent Processing of Object Shape ("what") and Location ("where")

The existence of parallel magno-cellular and parvo-cellular pathways emanating from the retina through the lateral geniculate nucleus (LGN) suggests strong evidence for the independent processing of shape and location. Hubel and Livingston report the existence of segregation of shape (form), stereo, color, and movement pathways evident in the LGN, primary visual cortex (area V1), which carries through area V2, onto the middle temporal area (MT), V4, and possibly higher visual areas [LH88]. The authors also report that lesion studies have defined "two functionally distinct divisions of visual association areas: the temporal-occipital region, necessary for learning to identify objects by their appearance, and the parieto-occipital region, needed for tasks involving the positions of objects," the familiar distinction of "what" versus "where" [LH88, p.744].

The "what" and "where" dichotomy has been attributed to the parvo- and magno-cellular (P/M) channels by Zeki [Zek93]. Selective destruction of the P layers of the LGN results in deficit in shape perception, while

selective destruction of M layers results in inhibited motion perception but does not affect shape discrimination. The “what” and “where” notions are also supported by psychophysical studies of visual attention where attention is characterized by spatial location (attention, the “where”) and by content (expectation, the “what”) [Van92].

Zeki warns, however, that the what/where hypothesis may be too simplistic a view. He states, “the precise position of an object and its relationship to other objects (spatial vision) can give the vital clue to the identity of the object, and the precise shape of an object can give the vital clue to its position. There are, in brief, far too many facts militating against the ‘what and where’ doctrine for it to be retained as a serious indicator of how the visual cortex is organized.” [Zek93, p.194] Again, the conclusion that presents itself is that although shape and location may be anatomically independent, there is probably a high degree of cooperation between these processes. Indeed, information between temporal and parietal streams is exchanged via connections between several higher cortical areas, particularly temporal area TEO and parietal area FST as well as by areas in the rostral STS [Nel95].

In summary, although a description of the neurological substrate of human vision in terms of the “what” and “were” may be somewhat of a simplified view, neurological evidence generally supports this functional segregation. The review of the cortical visual regions supports the general psychophysical description of visual attention. Specifically, relevant cortical structures have been identified as the centers providing the neuronal attentive mechanism. To gain further insight into the limitations of the visual channels, lower level neuronal pathway components are described in the next section.

3.2 Physiology of the Eye

Often called “the world’s worst camera”, the eye, shown in Figure 2, suffers from numerous optical imperfections, including spherical aberrations, chromatic aberrations, light scatter, diffraction, and imperfect focus. Dimensions of retinal features are usually described in terms of projected scene dimensions in units of degrees visual angle, defined as

$$A = 2 \arctan \frac{S}{2D},$$

where S is the size of the scene object and D is the distance to the object (see Figure 3). Common visual angles include a thumbnail at arm’s length (1.5–2°), the sun and moon (.5° or $30'$ of arc). An American quarter coin at arm’s length subtends 2°, $1'$ at 85 meters, and $1''$ at 5km.

The retina contains receptors sensitive to light (photoreceptors) which constitute the first stage of visual perception. Neural signals leading to cortical visual centers originate at these receptors. Photoreceptors are

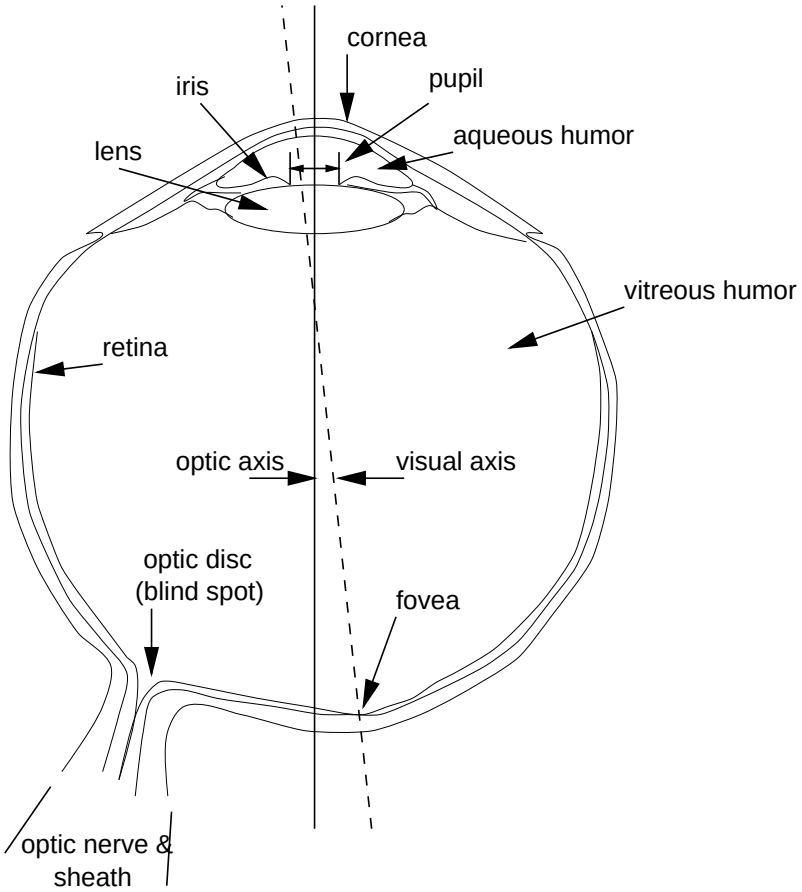


Fig. 2. The eye. Adapted from [BL88b, p.34 (Fig. 1)].

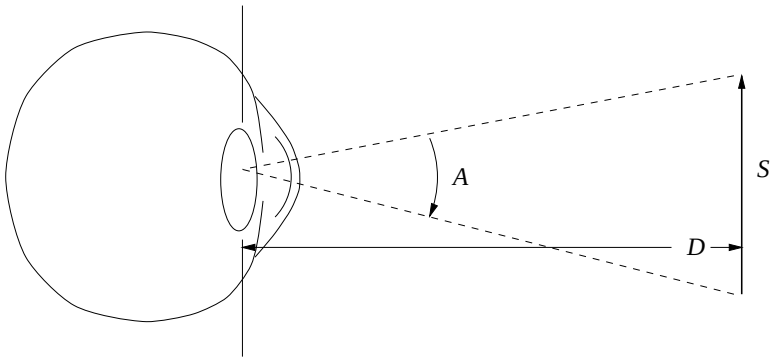


Fig. 3. Visual angle. Adapted from [HH73, p.15 (Fig. 2.7)].

functionally classified into rods and cones. Rods are sensitive to dim and achromatic light (night vision), while cones respond to brighter, chromatic light (daylight vision). The retina contains 120 million rods and 7 million cones, and is is arranged concentrically.

The innermost region is the fovea centralis (or foveola) which measures 400μm in diameter and contains 25,000 cones. The fovea proper measures 1500μm in diameter and holds 100,000 cones. The macula (or central retina) is 5000μm in diameter, and contains 650,000 cones. One degree visual angle corresponds to approximately 300μm distance on the human retina [DD88, p.48]. The foveola, measuring 400μm subtends 1.3° visual angle, while the fovea and macula subtend 5° and 16.7°, respectively (see Figure 4(a)). Figure 4(b) shows the retinal distribution of rod and cone receptors. The fovea contains 147,000 cones/mm^2

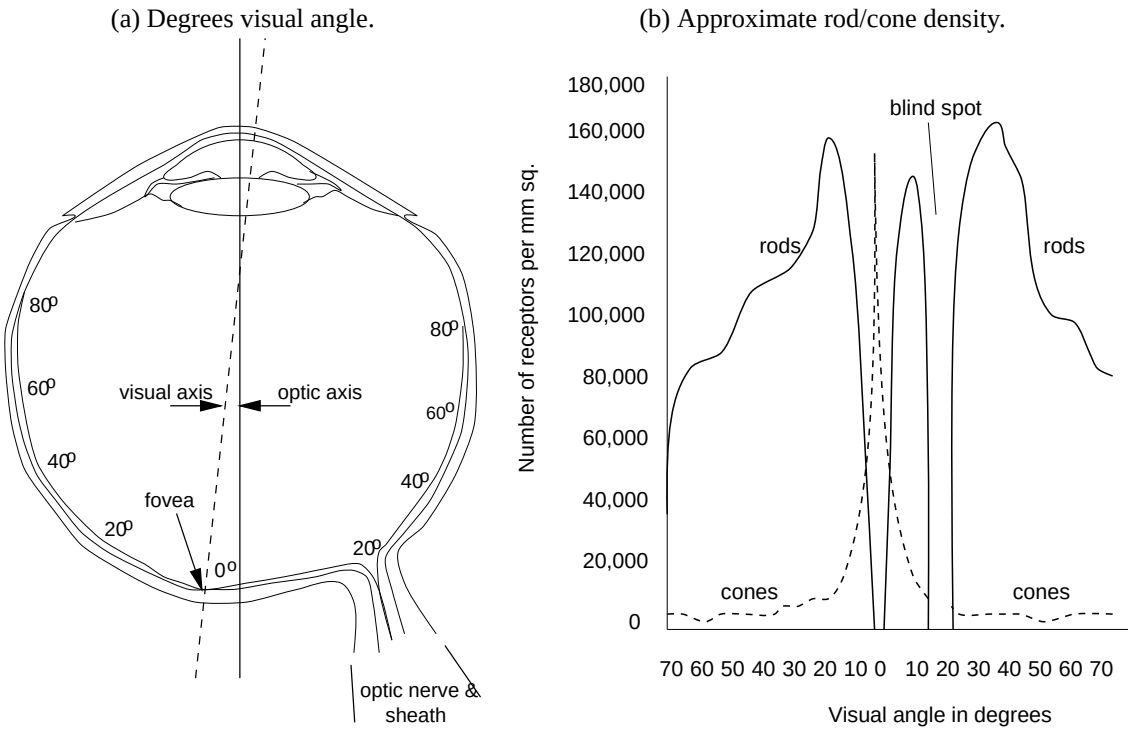


Fig. 4. Retinotopic receptor distribution. Adapted from [HH73, p.25 (Fig. 2.16)].

and a slightly smaller number of rods. At about 10° the number of cones drops sharply to less than 20,000 cones/mm^2 while at 30° the number of rods in the periphery drops to about 100,000 rods/mm^2 [HH73].

The entire visual field roughly corresponds to a 23,400 square degree area defined by an ellipsoid with the horizontal major axis subtending 180° visual angle, and the minor vertical axis subtending 130°. The diameter of the highest acuity circular region subtends 2°, the parafovea (zone of high acuity) extends to about 4°

or 5°, and acuity drops off sharply beyond. At 5°, acuity is only 50% [Irw92]. The so-called "useful" visual field extends to about 30°. The rest of the visual field has very poor resolvable power and is mostly used for perception of ambient motion. With increasing eccentricity the cones increase in size, while the rods do not [DD88]. Cones, not rods, make the largest contribution to the information going to deeper brain centers, and provide most of the fine-grained spatial resolvability of the visual system.

3.2.1 Optics and Visual Acuity

The Modulation Transfer Function (MTF) theoretically describes the spatial resolvability of retinal photoreceptors by considering the cells as a finite array of sampling units. The 400μm-diameter rod-free foveola contains 25,000 cones. Using the area of a circle, $25000 = \pi r^2$, approximately $2\sqrt{25000/\pi} = 178.41$ cones occupy a 400μm linear cross-section of the foveola with an estimated average linear inter-cone spacing of 2.24μm. Cones in this region measure about 1μm in diameter. Since one degree visual angle corresponds to approximately 300μm distance on the human retina, roughly 133 cones are packed per degree visual angle in the foveola. By the sampling theorem, this suggests a resolvable spatial Nyquist frequency of 66 c/deg. Subjective resolution has in fact been measured at about 60 c/deg [DD88, pp.46-53]. In the fovea, a similar estimate based on the foveal diameter of 1500μm and a 100,000 cone population, gives an approximate linear cone distribution of $2\sqrt{100000/\pi} = 356.82$ cones per 1500μm. The average linear inter-cone spacing is then 71 cones/deg suggesting a maximum resolvable frequency of 35 cycles/deg, roughly half the resolvability within the foveola. This is somewhat of an underestimate since cone diameters increase two-fold by the edge of the fovea suggesting a slightly milder acuity degradation. These one-dimensional approximations are not fully generalizable to the two-dimensional photoreceptor array although they provide insight into the theoretic resolution limits of the eye. Effective relative visual acuity measures are usually obtained through psychophysical experimentation.

At photopic light levels (day, or cone vision), foveal acuity is fairly constant within the central 2°, and drops approximately linearly from there to the 5° foveal border. Beyond the 5°, acuity drops sharply (approximately exponentially). At scotopic light levels (night, or rod-vision), acuity is poor at all eccentricities. Figure 5 shows the variation of visual acuity at various eccentricities and light intensity levels. Intensity is shown varying from 9.0 to 4.6 log micromicrolamberts, denoted by $\log\ \mathrm{mmL}$ (9.0 log micromicrolamberts = 10^9 micromicrolamberts = 1 mL, see [Dav80, p.311]). The correspondence between foveal receptor spacing and optical limits generally holds in foveal regions of the retina, but not necessarily in the periphery. In contrast to the approximate 60 c/deg resolvability of foveal cones, the highest spatial frequencies resolvable by rods are on the order of 5 c/deg, suggesting poor resolvability in the relatively cone-free periphery. Visual acuity beyond the foveal 5° is discussed in §9.2.1. Although visual acuity correlates fairly well with cone

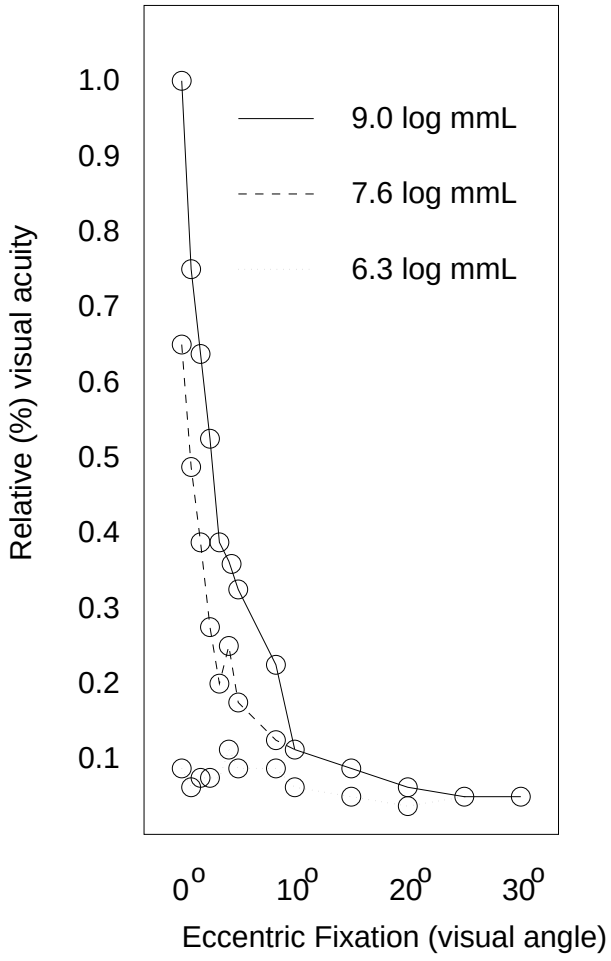


Fig. 5. Visual acuity at various eccentricities and light levels. Adapted from [Dav80, p.311 (Fig. 13.1)].

distribution density, it is important to note that synaptic organization and later neural elements (e.g., ganglion cells concentrated in the central retina) are also contributing factors in determining visual acuity. Retinogeniculo-cortical anatomy and physiology is discussed in the following sections.

3.2.2 Retinogeniculate Anatomy and Physiology

The retina is composed of multiple layers of different cell types [DD88]. Surprisingly, the "inverted" retina is constructed in such a way that photoreceptors are found at the bottom layer. This construction is somewhat counterintuitive since rods and cones are furthest away from incoming light, buried beneath a layer of cells. The retina resembles a three-layer cell sandwich, with connection bundles between each layer. These connectional layers are called plexiform or synaptic layers.

The retinogeniculate organization is schematically depicted in Figure 6. The outermost layer (w.r.t. incoming

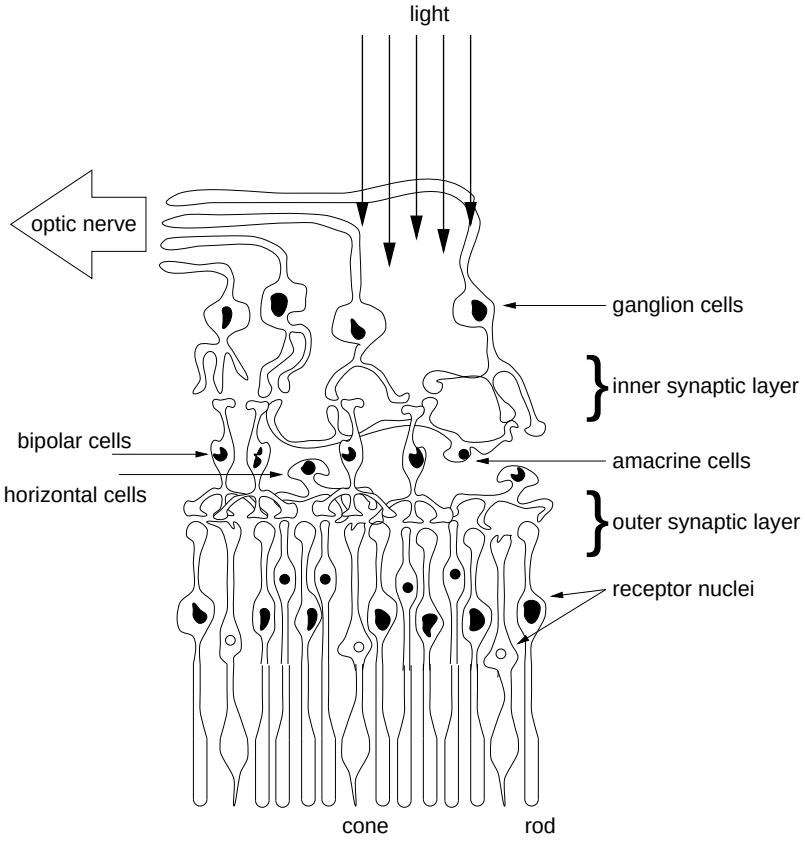


Fig. 6. The retina. Adapted from [HH73, p.24 (Fig. 2.15)].

light) is the outer nuclear layer which contains the photoreceptor (rod/cone) cells. The first connectional

layer is the outer plexiform layer which houses connections between receptor and bipolar nuclei. The next outer layer of cells is the inner nuclear layer containing bipolar (amacrine, bipolar, horizontal) cells. The next plexiform layer is the inner plexiform layer where connections between inner nuclei cells and ganglion cells are formed. The top layer, or the ganglion layer, is composed of ganglion cells.

The fovea's photoreceptors are special types of neurons—the nervous system's basic elements (see Figure 7). Retinal rods and cones are specific types of dendrites. In general, individual neurons can connect to as many

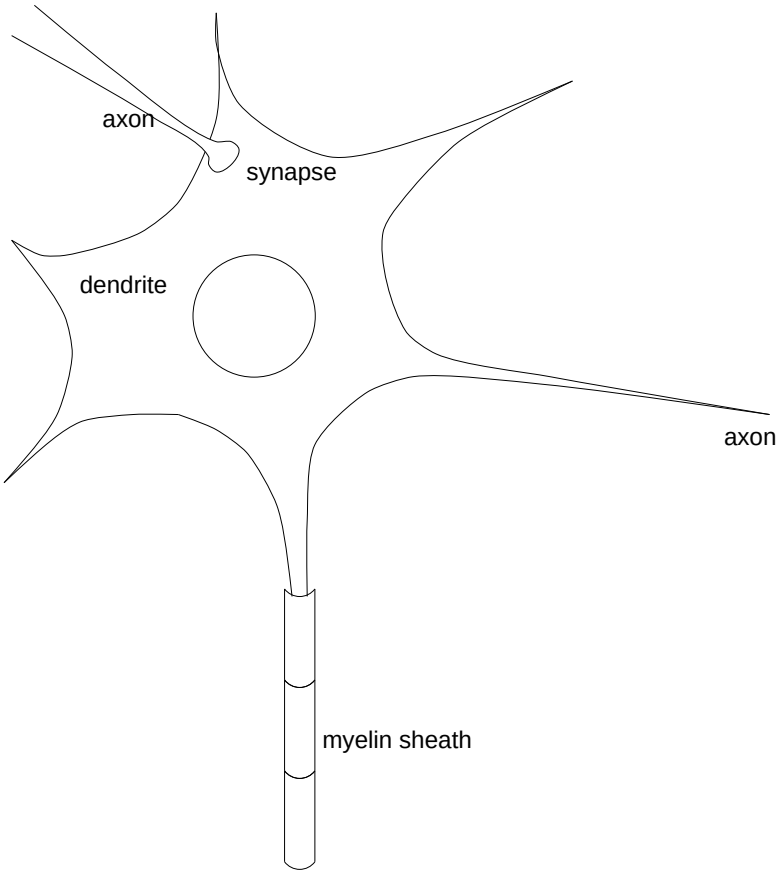


Fig. 7. Schematic of the neuron. Adapted from [BL88a, pp.31-32 (Fig. 2.1, Fig. 2.2)].

as 10,000 other neurons. Comprised of such interconnected building blocks, as a whole, the nervous system behaves like a large neural circuit.

Certain neurons (e.g., ganglion cells) resemble a "digital gate", sending a signal (firing) when the cell's activation level exceeds a threshold. Ganglion cell activation comes as an effect of the potassium-sodium pump. Signals propagate through the axon in a wave of depolarization—the action potential. The action poten-

tial (lasting less than 1ms) occurs as sodium ions (Na+) permeate the depolarized neuronal membrane. As sodium ions flow in, potassium (K+) ions flow out restoring resting potential. Inhibitory signals allow inflow of chloride ions (Cl-) preventing depolarization. The myelin sheath is an axonal cover providing insulation which speeds up conduction of impulses. Unmyelinated axons of the ganglion cells converge to the optic disk (an opaque myelin sheath would block light). Axons are myelinated at optic disk, and connect to the Lateral Geniculate Nuclei (LGN) and the Superior Colliculus (SC).

3.2.2.1 The Outer Layer

Rods and cones of the outer retinal layer respond to incoming light. A simplified account of the function of these cells is that rods provide monochromatic, scotopic (night) vision, and cones provide trichromatic, photopic (day) vision. Both types of cells are partially sensitive to mesopic (twighlight) light levels.

3.2.2.2 The Inner Nuclear Layer

Outer receptor cells are laterally connected to the horizontal cells. In the fovea, each horizontal cell is connected to about 6 cones, and in the periphery to about 30-40 cones. Centrally, the cone bipolar cells contact one cone directly, and several cones indirectly through horizontal or receptor-receptor coupling. Peripherally, cone bipolar cells directly contact several cones. The number of receptors increases eccentrically. The rod bipolar cells contact a considerably larger number of receptors than cone bipolars. There are two main types of bipolar cells, ones that depolarize to increments of light (+), and others that depolarize to decrements of light (-). The signal profile (cross-section) of bipolar receptive fields is a “Mexican hat”, or center-surround, with an on-center, or off-center signature.

3.2.2.3 The Ganglion Layer

Ganglion cells form an “active contrast-enhancing system,” not a camera-like plate. Centrally, ganglion cells directly contact one bipolar. Peripherally, ganglion cells directly contact several bipolars. The receptive fields of ganglion cells are similar to those of bipolar cells (center-surround, on-center, off-center). Ganglion cells are distinguished by their morphological and functional characteristics. Morphologically, there are two types of ganglion cells, the α and β cells. Approximately 10% of retinal ganglion cells are α cells possessing large cell bodies and dendrites, and about 80% are β cells with small bodies and dendrites [LWL95]. The α cells project to the magnocellular (M-) layers of LGN while the β cells project to the parvocellular (P-) layers. A third channel of input relays through narrow, cell-sparse laminae between the main M- and P-layers of the LGN. Its origin in the retina is not yet known. Functionally, ganglion cells fall into three classes, the X, Y, and W cells [DD88, Kap91]. X cells respond to sustained stimulus, location and fine detail,

TABLE 1
Functional characteristics of ganglionic projections.

Characteristics	Magnocellular	Parvocellular
ganglion size	large	small
transmission time	fast	slow
receptive fields	large	small
sensitivity to small objects	poor	good
sensitivity to change in light levels	large	small
sensitivity to contrast	low	high
sensitivity to motion	high	low
color discrimination	no	yes

and nervate along both M- and P-projections. Y cells nervate only along the M-projection, and respond to transient stimulus, coarse features, and motion. W cells respond to coarse features, and motion, and project to the superior colliculus. The functional characteristics of ganglionic projections are summarized in Table 1.

3.2.2.4 Cells in the Striate Cortex

Thalamic axons from the M- and P-layers of the LGN terminate mainly in the lower and upper halves (β, α divisions, respectively) of layer 4C in middle depth of area V1 [LWL95]. Cell receptive field size and contrast sensitivity signatures are distinctly different in the M- and P- inputs of the LGN, and vary continuously through the depth of layer 4C. Unlike the center-surround receptive fields of retinal ganglion and LGN cells, cortical cells respond to orientation-specific stimulus [Hub88, pp.67-71]. Cortical cells are distinguished by two classes: *simple* and *complex*.

The size of a simple cell's receptive field depends on its retinal position, relative to the fovea. The smallest fields are in and near the fovea, with sizes of about 1/4 $\times$ 1/4 degree. This is about the size of the smallest diameters of the smallest receptive field centers of retinal ganglion or LGN cells. In the far periphery, simple cell receptive field sizes are about 1 $\times$ 1 degree. Simple cells fire only when a line or edge of preferred orientation falls within a particular location of the cell's receptive field. Complex cells fire wherever such a stimulus falls into the cell's receptive field [LWL95]. The optimum stimulus width for either cell type is, in the fovea, about 2 minutes of arc. The resolving power (acuity) of both cell types is the same.

About 10-20% of complex cells in the upper layers of the striate cortex show marked directional selectivity [Hub88, p.77]. Directional selectivity (DS) refers to the cell's response to a particular direction of movement. Cortical directional selectivity (CDS) contributes to motion perception and to the control of eye movements [GN95]. CDS cells establish a motion pathway from V1 projecting to MT and V2 (which also projects to MT) and to MST. In contrast, there is no evidence that retinal directional selectivity (RDS)

contributes to motion perception. RDS contributes to oculomotor responses [GSA95]. In vertebrates, it is involved in optokinetic nystagmus, a type of eye movement discussed in §IV.

3.2.2.5 Significance of Motion-Sensitive Single-Cell Physiology for Perception

There are two somewhat counterintuitive implications of the visual system's motion-sensitive single-cell organization for perception. First, due to motion sensitive cells, fixations are never perfectly still but make constant tiny movements called *microsaccades* [Hub88, p.81]. These movements are more or less spatially random varying over 1 to 2 minutes of arc in amplitude. The counterintuitive fact regarding fixations is that if an image is artificially stabilized on the retina, vision fades away within about a second and the scene becomes blank. Fixations and saccades are further discussed in §IV. Second, due to the response characteristics of single (cortical) cells, the "retinal buffer" representation of natural images is much more abstract than what intuition suggests. An object in the visual field stimulates only a tiny fraction of the cells on whose receptive field it falls [Hub88, pp.85-87]. Perception of the object depends mostly on the response of (orientation-specific) cells to the object's borders. For example, the homogeneously shaded interior of an arbitrary form (e.g., a kidney bean) does not stimulate cells of the visual system. Awareness of the interior shade or hue depends on only cells sensitive to the borders of the object. In Hubel's words, "...our perception of the interior as black, white, gray, or green has nothing to do with cells whose fields are in the interior—hard as that may be to swallow...What happens at the borders is the only information you need to know: the interior is boring." [Hub88, p.87]

3.3 Implications for Attentional Visual Display Design

From the above discussion, both the structure and functionality of human visual system components place constraints on the design parameters of a visual communication system. In particular, the design of the gaze-contingent system must distinguish the characteristics of foveal and peripheral vision.

The parvocellular pathway in general responds to signals possessing the following attributes: high contrast (the parvocellular pathway is less sensitive to luminance), chromaticity, low temporal frequency, and high spatial frequency (due to the small receptive fields). Conversely, the magnocellular pathway can be characterized by sensitivity to the following signals: low contrast (the magnocellular pathway is more sensitive to luminance), achromaticity, moderate-to-high temporal frequency (sudden onset stimuli), and low spatial frequency (due to the large receptive fields). In terms of motion responsiveness, Koenderink et al. provide support that the foveal region is more receptive to slower motion than the periphery, although motion is perceived uniformly across the visual field [KDG85].

M and P ganglion cells in the retina connect to M and P channels, respectively. Zeki suggests the existence of four functional pathways defined by the M and P channels [Zek93]: motion, dynamic form, color, and form (size and shape). Furthermore, it is thought that fibers reaching the superior colliculus represent retinal receptive fields in rod-rich peripheral zones, while the fibers reaching the LGN represent cone-rich areas of high acuity [BL88a]. It seems likely that the M ganglion cells correspond to rods, mainly found in the periphery, and the P cells correspond to cones, which are chromatic cells concentrated mainly in the foveal region. A *visuotopic* representation model for imagery based on these observations is proposed:

1. **Spatial Resolution** should remain high within the foveal region and smoothly degrade within the peripheral, matching human visual acuity.
2. **Temporal Resolution** must be available in the periphery. Sudden onset events are potential attentional attractors.
3. **Luminance** should be coded for high visibility in the peripheral areas since the periphery is sensitive to dim objects.
4. **Chrominance** should be coded for high exposure almost exclusively in the foveal region, with chromaticity decreasing sharply into the periphery. This requirement is a direct consequence of the high density of cones and parvocellular ganglion cells in the fovea.
5. **Contrast** sensitivity should be high in the periphery, corresponding to the sensitivity of the magnocellular ganglion cells found mainly outside the fovea.
6. **Spatial frequency**: High frequency components must be more pronounced in foveal regions, than in the periphery. High spatial frequency features in the periphery must be made visible “just in time” to anticipate gaze-contingent fixation changes.

Special consideration should be given to sudden onset, luminous, high frequency objects (i.e., suddenly appearing bright edges).

A gaze-contingent visual system faces an implementational difficulty not yet addressed: matching the dynamics of human eye movement. Any system designed to incorporate, for example, an eye-slaved high resolution of interest must deal with the inherent delay imposed by the processing required to track and process real-time eye tracking data. To consider the temporal constraints that need to be met by such systems, the dynamics of human eye movements are studied in the following chapter.

CHAPTER IV

EYE MOVEMENTS

Perception of the environment is achieved through the integration of small high-resolution "spotlights" projected onto the fovea. Portions of the scene are repositioned for foveal inspection through rapid movement of the eyes. Eye movement measurement is an essential component of gaze-contingent applications. Successful systems rely on an accurate measurement device and sound analysis techniques based on knowledge of the underlying physical processes. A primary goal is the proper identification of expected characteristic patterns in the recorded signal. This section starts with a brief description of the device used to measure eye movements, namely the eye tracker. Then an outline of the neuronal substrate of eye movements is presented, followed by an itemized description of characteristic eye movement patterns.

4.1 Eye Trackers

The measurement device most commonly used for measuring eye movements is an eye tracker. The first method for objective eye measurements using corneal reflection was reported in 1901. To improve accuracy, techniques using a contact lens were developed in the 1950's. Devices attached to the contact lens ranged from small mirrors to coils of wire. For a short review of early eye tracking methods, see [Rob68, §II]. Measurement devices relying on physical contact with the eyeball generally provide very sensitive measurements. The obvious drawback of these devices is their invasive requirement of wearing the contact lens. So-called non-invasive (sometimes called remote) eye trackers typically rely on the measurement of the corneal reflection of a closely positioned directed light source. Inexpensive devices available today utilize a video camera to observe the eyeball and calculate the point of regard in real-time. The corneal reflection of the light source (typically infra-red) is measured relative to the location of the pupil center. Corneal reflections are known as the Purkinje reflections, or Purkinje images [Cra94]. Due to the construction of the eye, four Purkinje reflections are formed, as shown in Figure 8. Video-based eye trackers typically locate the first Purkinje image. With appropriate calibration procedures, these eye trackers are capable of measuring a viewer's point of regard (POR) on a suitably positioned (perpendicularly planar) surface on which calibration points are displayed. Two points of reference are needed to separate eye movements from head movements. Positional difference between the pupil center and corneal reflection changes with pure eye rotation, but remains relatively constant with minor head movements. Approximate relative positions of pupil and first Purkinje reflections are graphically shown in Figure 9, as the left eye rotates to fixate 9 correspondingly placed calibration points. The Purkinje reflection is shown as a small white circle in close proximity to the pupil, represented by a black circle. Since the infra-red light source is usually placed at some fixed position rel-

PR, Purkinje reflections: 1, reflection from front surface of the cornea; 2, reflection from rear surface of the cornea; 3, reflection from front surface of the lens; 4, reflection from rear surface of the lens—almost the same size and formed in the same plane as the first Purkinje image, but due to change in index of refraction at rear of lens, intensity is less than 1% of that of the first Purkinje image; IL, incoming light; A, aqueous humor; C, cornea; S, sclera; V, vitreous humor; I, iris; L, lens; CR, center of rotation; EA, eye axis; $a \approx 6$mm; $b \approx 12.5$mm; $c \approx 13.5$mm; $d \approx 24$mm; $r \approx 7.8$mm [Cra94].

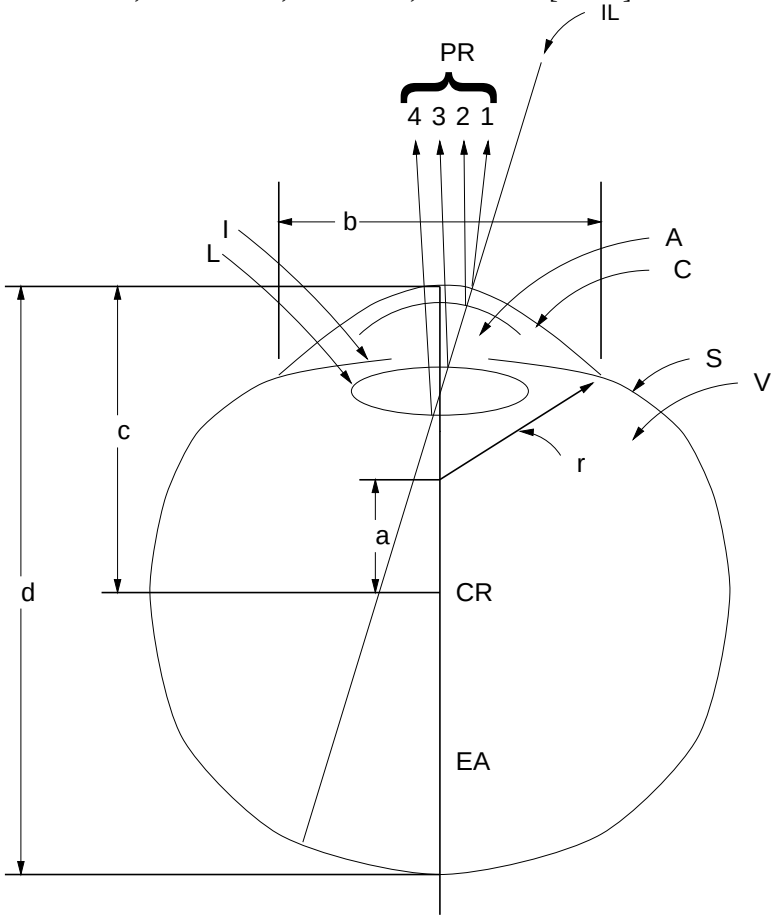


Fig. 8. Purkinje images. Adapted from [Cra94, p.19 (Fig. 1)].

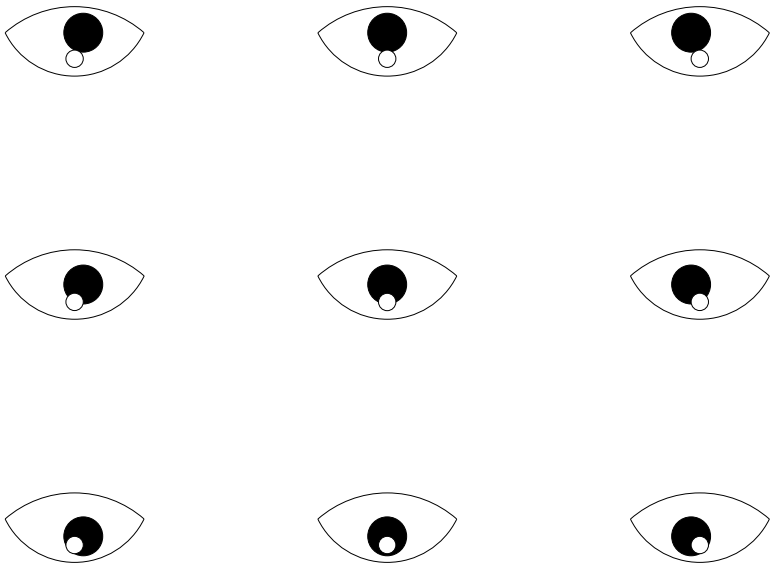


Fig. 9. Relative positions of pupil and first Purkinje images as seen by the eye tracker's camera.

ative to the eye, the Purkinje image is relatively stable while the eyeball, and hence the pupil, rotates in its orbit. So-called generation-V eye trackers also measure the fourth Purkinje image [CS85]. By measuring the first and fourth Purkinje reflections, these dual-Purkinje image (DPI) eye trackers separate translational and rotational eye movements. Both reflections move together through exactly the same distance upon eye translation but the images move through different distances, thus changing their separation, upon eye rotation.

4.2 The Oculomotor System

In general, the eyes move within six degrees of freedom: three translations within the socket, and three rotations. There are six muscles responsible for movement of the eyeball: the *medial* and *lateral recti* (sideways movements), the *superior* and *inferior recti* (up/down movements), and the *superior* and *inferior obliques* (twist) [Dav80]. These are shown in Figure 10. The neural system involved in generating eye movements is known as the oculomotor plant. The general plant structure and connections are shown in Figure 11 and described in [Rob68]. Eye movement control signals emanate from several functionally distinct regions. Areas 17, 18, 19, and 22 are areas in the occipital cortex thought to be responsible for high-level visual functions such as recognition. The superior colliculus bears afferents emanating directly from the retina, particularly from peripheral regions conveyed through the magno-cellular pathway. The semicircular canals react to head movements in three-dimensional space. All three areas, i.e., the occipital cortex, the superior colliculus, and the semicircular canals convey efferents to the eye muscles through the mesencephalic and pontine reticular formations. Classification of observed eye movement signals relies in part on the known functional characteristics of these cortical regions.

Left (view from above): 1, superior rectus; 2, levator palbebrae superioris; 3, lateral rectus; 4, medial rectus; 5, superior oblique; 6, reflected tendon of the superior oblique; 7, annulus of Zinn. *Right (lateral view):* 8, inferior rectus; 9, inferior oblique.

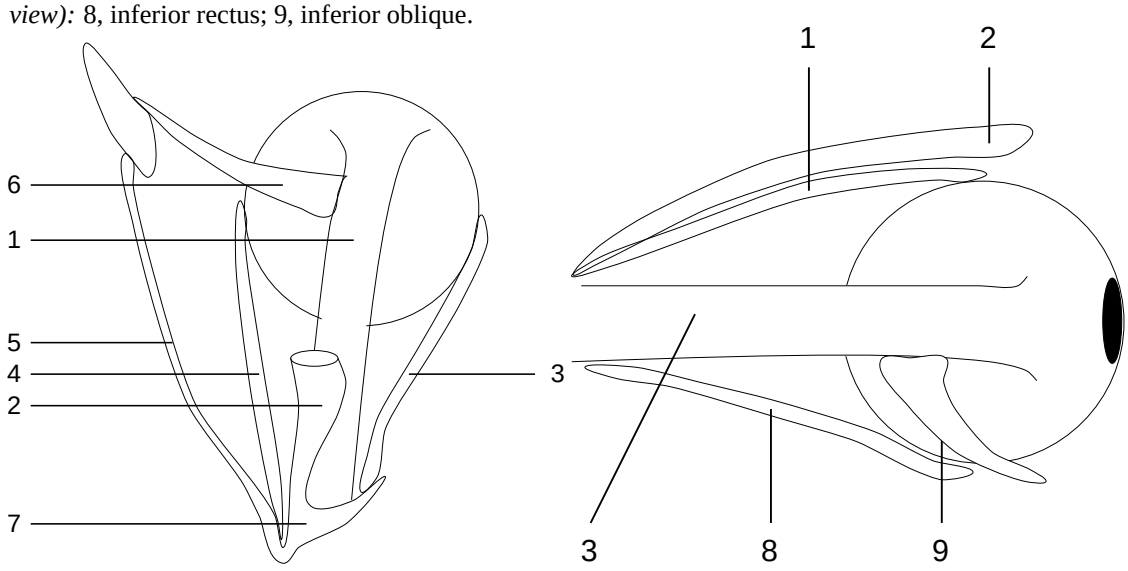


Fig. 10. Extrinsic muscles of the eye. Adapted from [Dav80, p.385 (Fig. 16.2), p.386 (Fig. 16.3)].

Two pertinent observations regarding eye movements can be drawn from the oculomotor plant's organization:

1. The eye movement system is, to a large extent, a feedback circuit.
2. Signals controlling eye movement emanate from cortical regions which can be functionally categorized as voluntary (occipital cortex), involuntary (superior colliculus), and reflexive (semicircular canals).

The feedback-like circuitry is utilized mainly in the types of eye movements requiring stabilization of the eye. Orbital equilibrium is necessitated for the steady retinal projection of an object, concomitant with the object's motion and movements of the head. Stability is maintained by a neuronal control system.

4.3 Taxonomy and Models of Eye Movements

Almost all normal primate eye movements used to reposition the fovea result as combinations of five basic types: saccadic, smooth pursuit, vergence, vestibular, and physiological nystagmus (miniature movements associated with fixations) [Rob68, p.1033]. Vergence movements are used to focus the pair of eyes over a distant target (depth perception). Other movements such as adaptation and accommodation refer to non-positional aspects of eye movements (i.e., pupil dilation, lens focusing). With respect to visual display design, positional eye movements are of primary importance.

CBT, corticobular tract; CER, cerebellum; ICTT, internal corticotectal tract; LG, lateral geniculate body; MLF, medial longitudinal fasciculus; MRF, mesencephalic and pontine reticular formations; PT, pretectal nuclei; SA, stretch afferents from extraocular muscles; SC, superior colliculi; SCC, semicircular canals; T, tegmental nuclei; VN, vestibular nuclei; II, optic nerve; III, IV, and VI, the oculomotor, trochlear, and abducens nuclei and nerves; 17, 18, 19, 22, primary and association visual areas, occipital and parietal (Brodmann); 8, the frontal eye fields.

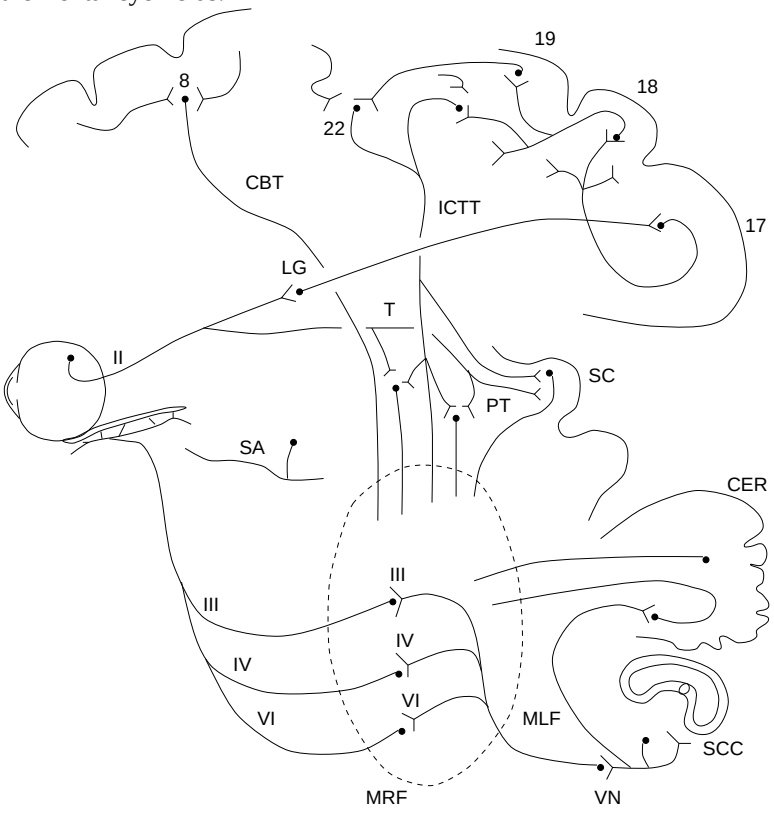


Fig. 11. Schematic of the major known elements of the oculomotor system. Adapted from [Rob68, p.1035 (Fig. 2)].

4.3.1 Saccades

Saccades are rapid eye movements used in repositioning the fovea to a new location in the visual environment. The term comes from an old French word meaning “flick of a sail” [Gre90, p.64]. Saccadic movements are both voluntary and reflexive. The movements can be voluntarily executed or they can be invoked as a corrective optokinetic or vestibular measure (see below). Saccades range in duration from 10ms to 100ms, which is a sufficiently short duration to render the executor effectively blind during the transition [SF83]. There is some debate over the underlying neuronal system driving saccades. On the one hand, saccades have been deemed ballistic and stereotyped. The term stereotyped refers to the observation that particular movement patterns can be evoked repeatedly. The term ballistic refers to the presumption that saccade destinations are pre-programmed. That is, once the saccadic movement to the next desired fixation location has been calculated (programming latencies of about 200ms have been reported), saccades cannot be altered. One reason behind this presumption is that during saccades there is insufficient time for visual feedback to guide the eye to its final position [Car77, p.57]. One the other hand, a saccadic feedback system is plausible if it is assumed that instead of visual feedback, an internal copy of head, eye, and target position is used to guide the eyes during a saccade [LR86, FKS85]. Due to their fast velocities, saccades may only appear to be ballistic [ZOCR+76, p.251].

Various models for saccadic programming have been proposed [Fin92]. These models, with the exception of ones including “center-of-gravity” coding (see for example [HK89]), may inadequately predict unchangeable saccade paths. Instead, saccadic feedback systems based on an internal representation of target position may be more plausible since they tend to correctly predict the so-called double-step experimental paradigm. The double-step paradigm is an experiment where target position is changed during a saccade in mid-flight. Scudder et al. proposed a refinement of Robinson’s feedback model which is based on a signal provided by the superior colliculus and a local feedback loop [FKS85]. The local loop generates feedback in the form of motor error produced by subtracting eye position from a mental target-in-space position. Sparks and Mays cite compelling evidence that intermediate and deep layers of the SC contain neurons that are critical components of the neural circuitry initiating and controlling saccadic movements [SM90]. These layers of the SC receive inputs from cortical regions involved in the analysis of sensory (visual, auditory, and somatosensory) signals used to guide saccades. The authors also rely on implications of Listing’s and Donder’s Laws which specify an essentially null torsion component in eye movements, requiring virtually only two degrees of freedom for saccadic eye motions [Dav80, SM90]. According to these laws, motions can be resolved into rotations about the horizontal x- and vertical y-axes.

Models of saccadic generation attempt to provide an explanation of the underlying mechanism responsible for generating the signals sent to the motor neurons. Although there is some debate as to the source of the

saccadic program, the observed signal resembles a pulse/step function [SM90, p.315]. The pulse/step function refers to a dual velocity and position command to the extraocular muscles [LZ91, p.180]. A possible simple representation of a saccadic step signal is a differentiation filter. Carpenter suggests such a possible filter arrangement for generating saccades coupled with an integrator [Car77, p.288]. The integrating filter is in place to model the necessary conversion of velocity-coded information to position-coded signals [LZ91, p.182]. A perfect neural integrator converts a pulse signal to a step function. An imperfect integrator (called leaky) will generate a signal resembling a decaying exponential function. The principle of this type of neural integration applies to all types of conjugate eye movements. Neural circuits connecting structures in the brain stem and the cerebellum exist to perform integration of coupled eye movements including saccades, smooth pursuits, and vestibular and optokinetic nystagmus (see below) [LZ91, p.183].

A differentiation filter can be modeled by a linear moving average filter as shown in Figure 12. In the time

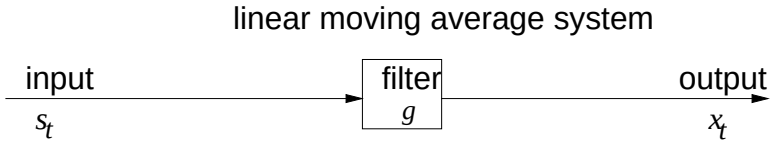


Fig. 12. Block diagram of a simple linear moving average system modeling saccadic movements.

domain, the moving average filter is modeled by the following equation

$$\begin{aligned} x_t &= g_0 s_t + g_1 s_{t-1} + \cdots \\ &= \sum_{k=0}^{\infty} g_k s_{t-k}, \end{aligned}$$

where s_t is the input (pulse), x_t is the output (step), and b_k are the moving average filter coefficients. To ensure differentiation, the filter coefficients typically must satisfy properties which approximate mathematical differentiation. An example of such a filter is the Haar filter with coefficients $\{1,-1\}$. Under the z-transform (see §VI) the transfer function $X(z)/S(z)$ of this linear filter is

$$\begin{aligned} x_t &= g_0 s_t + g_1 s_{t-1} \\ x_t &= (1)s_t + (-1)s_{t-1} \\ x_t &= (1)s_t + (-1)z s_t \\ x_t &= (1-z)s_t \\ X(z) &= (1-z)S(z) \\ \frac{X(z)}{S(z)} &= 1-z. \end{aligned}$$

The Haar filter is a length-2 filter which approximates the first derivate between successive pairs of inputs.

4.3.2 Smooth Pursuits

Pursuit movements are involved when visually tracking a moving target. Depending on the range of target motion, the eyes are capable of matching the velocity of the moving target. Pursuit movements provide an example of a control system with built-in negative feedback [Car77, p.41]. A simple closed-loop feedback loop used to model pursuit movements is shown in Figure 13, where s_t is the target position, x_t is the (desired) eye position, and h is the (linear, time-invariant) filter, or gain of the system [Car77, LZ91]. Tracing the loop

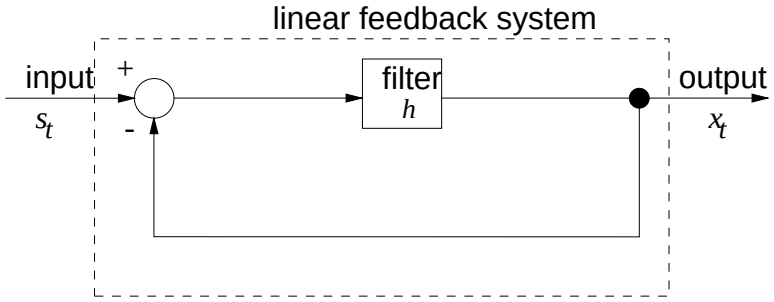


Fig. 13. Block diagram of a simple linear feedback system modeling smooth pursuit movements.

from the feedback start point gives the following equation in the time domain

$$h(s_t - x_t) = x_{t+1}.$$

Under the z-transform (see §VI) the transfer function $X(z)/S(z)$ of this linear system is

$$\begin{aligned} H(z)(S(z) - X(z)) &= X(z) \\ H(z)S(z) &= X(z)(1 + H(z)) \\ \frac{H(z)}{1 + H(z)} &= \frac{X(z)}{S(z)}. \end{aligned}$$

Signals from visual receptors constitute the error signal indicating needed compensation to match the target's retinal image motion.

4.3.3 Fixations

Fixations are eye movements which stabilize the retina over a stationary object of interest. It seems intuitive that fixations should be generated by the same neuronal circuit controlling smooth pursuits with fixations

being a special case of a target moving at zero velocity. This is probably incorrect [LZ91, pp.139-140]. Fixations, instead, are characterized by the miniature eye movements: tremor, drift, and microsaccades. These eye movements are considered noise present in the control system (possibly distinct from the smooth pursuit circuit) attempting to hold gaze steady. This noise appears as a random fluctuation about the area of fixation, typically no larger than 5° visual angle [Car77, p.105]. Although the classification of miniature movements as noise may be an oversimplification of the underlying natural process, it allows the signal to be modeled by a feedback system similar to the one shown in Figure 13. The additive noise in Figure 13 is represented by $e_t = s_t - x_t$, where the (desired) eye position x_t is subtracted from the steady fixation position s_t at the summing junction. In this model, the error signal stimulates the fixation system in a manner similar to the smooth pursuit system, except that here e_t is an error-position signal instead of an error-velocity signal (see [LZ91, p.150]). The feedback system modeling fixations, using the noisy "data reduction" method, is in fact simpler than the pursuit model since it implicitly assumes a stationary stochastic process [Car77, p.107]. Stationarity in the statistical sense refers to a process with constant mean. Other relevant statistical measures of fixations include their duration range of 150ms to 600ms, and the observation that 90% of viewing time is devoted to fixations [Irw92].

4.3.4 Nystagmus

Nystagmus eye movements are conjugate eye movements characterized by a sawtooth-like time course (time series signal) pattern. Optokinetic nystagmus is a smooth pursuit movement interspersed with saccades invoked to compensate for the retinal movement of the target. The smooth pursuit component of optokinetic nystagmus appears in the slow phase of the signal [Rob68]. Vestibular nystagmus is a similar type of eye movement compensating for the movement of the head. The time course of vestibular nystagmus is virtually indistinguishable from its optokinetic counterpart [Car77].

4.4 Implications for Eye Movement Analysis

From the above discussion, two significant observations relevant to eye movement analysis can be made. First, based on the functionality of eye movements, only three types of movements need be modeled to gain insight into the overt localization of visual attention. These types of eye movements are fixations, smooth pursuits, and saccades. Second, based on signal characteristics and plausible underlying neural circuitry, all three types of eye movements may be approximated by a linear, time-invariant (LTI) system, i.e., a linear filter.

The primary requirement of eye movement analysis, in the context of gaze-contingent system design, is the identification of fixations, saccades, and smooth pursuits. It is assumed that these movements provide

evidence of voluntary, overt visual attention. This assumption does not preclude the plausible involuntary utility of these movements, or conversely, the covert non-use of these eye movements (e.g., as in the case of parafoveal attention). Fixations naturally correspond to the desire to maintain one's gaze on an object of interest. Similarly, pursuits are used in the same manner for objects in smooth motion. Saccades are considered manifestations of the desire to voluntary change the focus of attention.

Eye movement signals can be approximated by linear filters. Fixations and pursuits are driven by a relatively simple neuronal feedback system. In the case of fixations, the neuronal control system is responsible for minimizing fixation error. For pursuit movements, the error is similarly measured as distance off the target, but in this case the target is non-stationary. Fixations and pursuits may be detected by a simple linear model based on linear summation.

The linear approach to eye movement modeling is an operational simplification of the underlying nonlinear natural processes [Car77, p.44]. The linear model assumes that position and velocity is processed by the same neuronal mechanism. The visual system processes these quantities in different ways. The position of a target is signaled by the activation of specific retinal receptors. The velocity of the target, on the other hand, is registered by the firing rate (amplitude) of the firing receptors. Furthermore, nonlinearities are expected in most types of eye movements. Accelerational and decelerational considerations alone suggest the inadequacy of the linear assumption. Nevertheless, from a signal processing standpoint, linear filter analysis is sufficient for the localization of distinct features in eye movement signals. Although this approach is a poor estimate of the underlying system, it nonetheless establishes a useful approximation of the signal in the sense of pattern recognition.

4.5 Implications for Pre-Attentional Visual Display Design

In estimating parameters in the design of gaze-contingent displays, peripheral regions present a curious problem: if the human visual system's processing capacity is peripherally limited, why not simply eliminate peripheral information? The reason for providing peripheral information is to assure preview benefit. Without peripheral information, the human visual system would not be able to select future fixation regions. Conversely, peripheral information designed to attract attention may be induced artificially if the objective is to draw (or cue or direct) visual attention. The manipulation of peripheral information should cater to the characteristics of peripheral vision.

The neural center thought to be responsible for directing attention to peripheral visual stimuli (the "foveation hypothesis" [Dav80]) is the superior colliculus (SC). From Figure 11, connections from the SC to the MRF

suggest that the SC has direct influence on eye movement. Since connections from higher visual areas (areas 17,18,19,22 and the frontal eye fields) also terminate in the MRF, it appears that voluntary eye movement control emanates from higher cortical areas while peripherally stimulated movement originates within the superior colliculus. Thus one possible view of the peripheral regions of the visual system is that they are responsible in part for the orienting of attention. Consequently, from the standpoint of visual display design, to attract one's attention, displays should trigger peripheral receptors. Conversely, in order to build unobtrusive displays, no distractory stimulus should be shown peripherally.

An obtrusive system designed to attract attention should do so by displaying appropriate stimulus in the periphery. Sudden onset events are particularly suitable candidates. For example, bright, blinking stimuli in the periphery are likely to attract attention. Presentation of such stimuli allows degradation of spatially distant objects once gaze has been diverted. It should be noted that such attentional factors may not be limited to visual cues. Virtual reality systems incorporating three-dimensional audio may also use aural cues to attract attention.

An unobtrusive system matching visual perceptual capacity should provide a foveal region of interest in the periphery "just in time" to meet the participant's change of fixation. This *anticipatory* strategy meets foveal vision instead of reacting to it. Unfortunately this design requires the nontrivial ability of predicting visual patterns (scanpaths). Prediction of eye movements is complicated by the system's need for an internal representation (model) of the underlying imagery in terms of potential visual regions of interest. Automatic identification of these regions is currently an open problem in computer vision.

In either system design, care must be exercised in manipulating peripheral imagery so that performance, if not perception, is not impaired. In this dissertation, peripheral degradation effects are studied in terms of perceptual impairment (see §X). The problem of eye movement prediction is handled in two ways: (1) a visual tracking paradigm is used to effectively eliminate scanpath variability, and (2) potential regions of interest in video are identified by analyzing multiple viewers' eye movement patterns. Both strategies are utilized to empirically test methods of peripheral spatial resolution degradation.

CHAPTER V

INTRODUCTION TO WAVELETS

Wavelets have generated tremendous interest in both theoretical and applied areas, especially within the latter half of this decade. Wavelet theory can be viewed as a synthesis of ideas originating in engineering (sub-band coding in part with quadrature mirror filters), physics (coherent states, renormalization group), and pure and applied mathematics (Calderón-Zygmund operators) [Dau92]. Historically, fundamental mathematical concepts of wavelet theory can be traced back to Fourier's work of 1807, Haar's algorithmic structure of 1909, with most of the development more directly related to wavelet theory occurring in the 1930s and 1960s [Mey93]. Wavelet theory has recently received wide acclaim due to the amalgamation of the diverse yet related analytical techniques into one elegant, coherent framework. Numerous researchers contributed to this effort. Grossmann, Morlet, Daubechies, Meyer, and Chui [Dau88, Mey93, Chu92] have greatly influenced the development of the mathematical theory, while Mallat and Meyer are credited with the introduction of multiresolution analysis in the wavelet context [Mal89a, Mal89b, Mey93]. Most of the analytical techniques developed herein for image/video processing and eye movement modeling are based on Mallat's multiresolution results, as they relate to pyramidal image processing techniques pioneered by Burt [Bur81, BA83b], and Mallat et al.'s singularity detection theory in the wavelet domain [MH91, MH92, MZ92a]. This section starts with the review of fundamental concepts of wavelet theory closely following Chui's derivations, then presents the theory of multiresolution analysis, wavelet filters and the discrete wavelet transform, and concludes with three applications of wavelet analysis, namely:

1. multiscale sharp variation (edge) detection in spatiotemporal data,
2. anisotropic multidimensional discrete wavelet analysis, and
3. multiresolution image representation through MIP mapping.

5.1 Fundamentals

The central idea behind wavelet analysis is the use of compactly supported basis functions which are used to approximate arbitrary signals. In essence, a wavelet basis is a generalized, functional extension of a vector basis. This section examines the fundamental concept of a basis and its use in the expression of arbitrary vectors and functions. The idea of a basis is studied in the domains of linear algebra, Fourier series, and wavelet series. Conventions for mathematical expressions are shown in Table 2.

TABLE 2
Notational conventions.

$c_1, c_2, \ldots, c_n$	Constants, scalars.
$\mathbf{v}$	Vector.
$(v_1, v_2, \ldots, v_n)$	Vector components.
$\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$	Vector set.
$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{k=1}^{n} u_k v_k = u_1 v_1 + \cdots + u_n v_n$	Inner product of vectors $\mathbf{u}, \mathbf{v}$.
$\|\mathbf{v}\| = \langle \mathbf{v}, \mathbf{v} \rangle^{1/2} = \sqrt{v_1^2 + \cdots + v_n^2}$	Vector norm.
V_n	Vector space (n-dimensional), i.e., set of all n-dimensional vectors.
$\mathbf{Z}$	Set of integers.
$\mathbf{R}$	Set of real numbers.
$\delta_{j,l} = \begin{cases} 1 & \text{for } j = l; \\ 0 & \text{for } j \neq l, \quad j, l \in \mathbf{Z} \end{cases}$	Kronecker delta.
E^n	Euclidean (n-dimensional) space.
$L^2(0, 2\pi)$	Vector space of 2π-periodic square-integrable one-dimensional functions $f(x)$.
$L^2(\mathbf{R})$	Vector space of measurable, square-integrable one-dimensional functions $f(x)$.
$L^2(\mathbf{R}^2)$	Vector space of measurable, square-integrable two-dimensional functions $f(x, y)$.
$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) \overline{g(x)}\, dx$	Inner product of $f, g \in L^2(\mathbf{R})$.
$\langle f, g \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \overline{g(x, y)}\, dx\, dy$	Inner product of $f, g \in L^2(\mathbf{R}^2)$.
$\|f\| = \langle f, f \rangle^{1/2}$	Norm of $f \in L^2(\mathbf{R})$ or $f \in L^2(\mathbf{R}^2)$, where the corresponding inner product is assumed.
f^j, g^l	Representations of functions f, g at j^{th} and l^{th} levels of resolution, respectively.
$\{h_k\}, \{g_m\}$	Digital filter sequences.
H, G	Digital filters.
$\mathbf{H}, \mathbf{G}$	Matrices (usually representing multidimensional digital filters).

5.1.1 Linear Algebra and Vector Spaces

Recall, from linear algebra, a vector $\mathbf{v}$ in n-dimensional Euclidean space E^n is defined as an ordered n-tuple $(v_1, v_2, \ldots, v_n)$ of real numbers, recognized as *vector components*. Vectors are *linearly independent* when the equation

$$\sum_{k=1}^{n} c_k \mathbf{v}_k = 0, \quad k \in \mathbf{Z}$$

can only hold if $c_1 = c_2 = \cdots = c_n = 0$. A vector *basis* in n-dimensional space is defined as any set of linearly independent vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Linear independence guarantees that any vector $\mathbf{v}$ in n-space can be expressed uniquely as a linear combination of the basis vectors, i.e.,

$$\begin{aligned} \mathbf{v} &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n \\ &= \sum_{k=1}^{n} c_k \mathbf{v}_k, \quad k \in \mathbf{Z}. \end{aligned}$$

The vectors are *orthogonal* (perpendicular) if

$$\langle \mathbf{v}_l, \mathbf{v}_m \rangle = \sum_{k=1}^{n} \mathbf{v}_{l_k} \mathbf{v}_{m_k} = v_{l_1} v_{m_1} + v_{l_2} v_{m_2} + \cdots + v_{l_n} v_{m_n} = 0, \quad \forall l, m,\ l \neq m,\ k, l, m \in \mathbf{Z},$$

where $\langle \mathbf{v}_l, \mathbf{v}_m \rangle$ denotes the vector *inner* (or scalar or dot) product. Every orthogonal system of n vectors forms a basis for the set of all vectors in n-space, V^n, although the orthogonality condition is not strictly necessary.[1] The orthogonal system of n vectors is *orthonormal* if each of the vectors has unit norm,

$$\|\mathbf{v}_k\| = \langle \mathbf{v}_k, \mathbf{v}_k \rangle^{1/2} = 1.$$

The theory of a vector space of infinite dimension is closely related to the theories of a function space, Fourier and wavelet series. The binding thread among these theories is the expression of an arbitrary function $f(x)$ by a series expansion using a set of basis functions $\{\psi_k(x)\}$ such that $f(x) = \sum_k c_k \psi_k(x)$.

5.1.2 Function Spaces

Wavelet theory is concerned with a particular functional vector space, namely the space $L^2(\mathbf{R})$ of all real, (Lebesgue) measurable, square integrable functions defined on the real line $\mathbf{R}$. The space $L^2(\mathbf{R})$ is a vector (Hilbert) space in which wavelet functions are typically constructed to serve as basis functions. Due to its pertinence to wavelet theory, the definition and properties of $L^2(\mathbf{R})$ are briefly discussed here. The present outline closely follows the very readable text by Holland [Hol90] where the reader is referred for further clarifications. References to particular sections and pages are given where appropriate.

[1] Given n linearly independent vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, it is always possible to construct an orthogonal system of n vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, each of which is a linear combination of $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ (Gram-Schmidt orthogonalization process) [Kap84, p.55].

The terms *functional vector space*, *vector space of functions*, and *function space*, refer to a set of functions possessing the same formal properties as a vector space of *n*-tuples in linear algebra, i.e., closure under sum and closure under scalar multiplication. That is, a set V of functions forms a vector space if for any functions f,g in V, $f+g$, and cf are also in V. The set W is a (functional) subspace of V if W is a vector space in its own right [Hol90, pp.22-23]. Note that the linear algebra concepts of linear dependence and independence carry over seamlessly to functional vector spaces. The function space V is called an *inner product space* if a scalar-valued expression called the *inner product*, denoted $\langle f,g\rangle$, can be defined for all $f,g \in V$, satisfying the following three conditions:

1. linear in the first variable, conjugate linear in the second (conjugate bilinear):

$$\langle af+bg,h\rangle = a\langle f,h\rangle + b\langle g,h\rangle, \text{ and } \langle f,ag+bh\rangle = \overline{a}\langle f,g\rangle + \overline{b}\langle f,h\rangle; \tag{5.1}$$

2. Hermitian symmetric:

$$\langle f,g\rangle = \overline{\langle g,f\rangle} \;\; \forall f,g \in V; \tag{5.2}$$

3. positive definite:

$$\langle f,f\rangle \geq 0 \;\; \forall f \in V, \text{ and } \langle f,f\rangle = 0 \text{ implies } f = 0, \tag{5.3}$$

where the symbol ($\bar{\cdot}$) denotes complex conjugation. In the case of real-valued functions and real scalars, complex conjugation has no effect, i.e., $\overline{a} = a$ for all scalars a. In this case, condition (5.1) states that the inner product is linear in each variable separately, condition (5.2) states that $\langle f,g\rangle = \langle g,f\rangle, \;\; \forall f,g \in V$, and condition (5.3) states that f,g are "essentially" equal if, in the sense of the inner product, $\langle f,g\rangle = 0$, and f is "essentially" zero if $\langle f,f\rangle = 0$. The last condition is somewhat subtle in that for $f(x)$ to be "essentially" zero (zero "almost everywhere") does not necessarily mean that $f(x)$ has to be zero at every point, only that whatever is used to define the inner product (e.g., an integral) evaluates to zero [Hol90, §3.6].

The above conditions define the abstract scalar value of an inner product axiomatically. In the case of the real-valued functional space $L^2(\mathbf{R})$, the inner product is specified as:

$$\langle f,g\rangle = \int_{-\infty}^{\infty} f(x)\,\overline{g(x)}\,dx. \tag{5.4}$$

Noting that conjugation has no effect on real-valued functions, the inner product satisfies the above properties (see [Hol90, pp.108-111]).

1. Bilinearity:

$$\begin{aligned}
\langle af+bg,h\rangle &= \int_{-\infty}^{\infty} (af+bg)h\,dx \\
&= \int_{-\infty}^{\infty} afh+bgh\,dx \\
&= a\int_{-\infty}^{\infty} fh\,dx + b\int_{-\infty}^{\infty} gh\,dx \\
&= a\langle f,h\rangle + b\langle g,h\rangle,
\end{aligned}$$

with a similar argument for the second variable.

2. Hermitian symmetry:

$$\langle f,g\rangle = \int_{-\infty}^{\infty} f\,g\,dx \;=\; \int_{-\infty}^{\infty} g\,f\,dx \;=\; \langle g,f\rangle.$$

3. Positive definiteness:

$$\langle f,f\rangle = \int_{-\infty}^{\infty} (f(x))^2\,dx \geq 0,$$

because $(f(x))^2 \geq 0$ since $f(x)$ is real, and

$$\langle f,f\rangle = \int_{-\infty}^{\infty} (f(x))^2\,dx \;=\; 0$$

implies $f(x)$ is "essentially" zero, meaning that $f(x)$ integrates to zero over the real line $\mathbf{R}$. The subtlety of this property can be illustrated by the function

$$f(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise,} \end{cases}$$

which has $\langle f,f\rangle = \int_{-\infty}^{\infty} |f(x)|^2\,dx = 0$ although $f(x)$ is not zero at every point. In fact, any function $f(x)$ that is zero except at a finite number of points has a zero integral.

The positive definite condition provides a definition of function equality: using the principle that absolute convergence implies convergence (so permitting the inspection of the integral of $|f(x)\overline{g(x)}|$ instead of $f(x)\overline{g(x)}$) and the fact that $|\overline{z}| = |z|$ for any complex number z, functions f and g are "essentially" equal if $\int_{-\infty}^{\infty} |f(x)-g(x)|\,dx = 0$ and f is "essentially" zero if $\int_{-\infty}^{\infty} |f(x)|^2\,dx = 0$.

Any function f is said to belong to the space $L^2(\mathbf{R})$ if it satisfies:

$$\int_{-\infty}^{\infty} |f(x)|^2\,dx < \infty, \tag{5.5}$$

where the integral is not required to have any particular value, only that it be finite [Hol90, pp.138-139].[2] The main issue in testing whether a given $f(x)$ belongs to $L^2(\mathbf{R})$ is to consider whether or not the integral in (5.5) is convergent or divergent. If it is not divergent, then $f(x)$ is said to be in $L^2(\mathbf{R})$. With the above definitions

[2]Technically, attention is usually restricted to Lebesgue measurable functions where all integrals should be interpreted as Lebesgue integrals. Under Lebesgue theory of integration, $f(x)$ is differentiable "almost everywhere" meaning that $f(x)$ is the Lebesgue integral of its derivative $f'(x)$, i.e.,

$$f(x) = \int_a^x f'(t)\,dt + f(a).$$

Such functions are called *absolutely continuous*. Although Lebesgue theory is beyond the scope of the present discussion (see [Hol90, p.165,p.253] for an introduction), it is tacitly assumed, without loss of generality, that all functions in $L^2(\mathbf{R})$ are absolutely continuous. This assumption only excludes functions where integration by parts may fail, i.e., functions of the Cantor-Lebesgue type, and practically restricts the discussion to functions that are piecewise continuous [Chu92, p.1].

of the space $L^2(\mathbf{R})$, and of the inner product for $L^2(\mathbf{R})$ in (5.4), it is clear that $L^2(\mathbf{R})$ is an inner product vector space (see [Hol90, pp.141-143]).

Since $L^2(\mathbf{R})$ is an inner product space, all properties associated with an inner product hold in $L^2(\mathbf{R})$. The properties particularly applicable to wavelet theory are listed below (see also [Chu92, p.4]):

- *Norm*:
$$\|f\| = \langle f, f\rangle^{1/2} = \left[\int_{-\infty}^{\infty} f(x)\overline{f(x)}\,dx\right]^{1/2} = \left[\int_{-\infty}^{\infty} |f(x)|^2\,dx\right]^{1/2}$$
- *Mean-square (distance) metric*:
$$\|f-g\| = \left[\int_{-\infty}^{\infty} |f(x)-g(x)|^2\,dx\right]^{1/2}$$
- *Schwarz' inequality*:
$$|\langle f,g\rangle| \le \|f\|\,\|g\| \Rightarrow \left|\int_{-\infty}^{\infty} f(x)\overline{g(x)}\,dx\right| \le \left[\int_{-\infty}^{\infty} |f(x)|^2\,dx\right]^{1/2}\left[\int_{-\infty}^{\infty} |g(x)|^2\,dx\right]^{1/2}$$

The above definition of $L^2(\mathbf{R})$ spaces naturally extends to the general class of inner product spaces, the L^2 spaces. L^2 spaces are denoted by
$$L^2(a,b), \quad -\infty \le a < b \le +\infty$$
where any function f is said to belong to the space $L^2(a,b)$ if it satisfies:
$$\int_a^b |f(x)|^2\,dx < \infty.$$
The inner product space $L^2(\mathbf{R})$ is a particular instance of the class of L^2 spaces with $a = -\infty$, $b = \infty$. The formal properties of a vector space and inner product, as well as the properties associated with the inner product, defined for $L^2(\mathbf{R})$ above, extend analogously to the general class of L^2 spaces. All L^2 spaces are Hilbert spaces since they satisfy the property of metric completeness [Hol90, p.143]. Although the notion of a Hilbert space is somewhat superfluous in the context of wavelet theory, some of the concepts and definitions of a Hilbert space do apply and are worth mentioning. In particular, the properties of *denseness*, *separability*, and *completeness*, required for the definition of the Hilbert space, are given below. The reader is referred to [Hol90, §3.10] for the complete account.

- *Denseness*: A subspace W of an abstract inner product space V is *dense* if, given any $v \in V$ there exists an $w \in W$ such that $\|v-w\| < \varepsilon$, for any small ε. This property states that W is dense in V if elements of W can be found as close as desired to any element of V.
- *Separability*: An inner product space V is *separable* if it contains a sequence of elements $w_1, w_2, \ldots$ that span a dense subspace of V. All finite-dimensional vector spaces V are separable since a basis can be found for V where the subspace spanned by the basis is V itself. Note that V is trivially dense in itself. All L^2 spaces are separable. As a consequence of this property, any separable inner product space has an orthogonal basis.

- *Completeness*: The notion of completeness is defined in terms of what is called a Cauchy sequence. A sequence of elements $v_k,\ k = 1,2,\ldots$ in an inner product space is a *Cauchy sequence* if: given any small positive number ε, an integer N (generally dependent on ε) can be found such that $\|v_k - v_l\| < \varepsilon$ whenever both $k,l \geq N$. Loosely speaking, a Cauchy sequence is one whose terms eventually cluster together. An inner product space V is said to be *complete* if, given any Cauchy sequence $v_k,\ k = 1,2,\ldots \in V$, there exists a $v \in V$ such that the sequence v_k converges to v. Roughly, the notion of completeness states that every Cauchy sequence in V must converge to an element of V.
- *Hilbert space*: The *Hilbert space* is a separable real (or complex) inner product space that is complete in the metric derived from its inner product.

Although concepts such as denseness and separability appear throughout the wavelet literature, arguably the most useful property (or at least notation) of wavelet theory is the idea of the inner product $\langle \cdot, \cdot \rangle$ of $L^2(\mathbf{R})$, and by analogous extension, of n-dimensional L^2 spaces denoted by $L^2(\mathbf{R}^n)$. This is due to the fact that wavelets are functions generating a basis in $L^2(\mathbf{R}^n)$ which, as explained below, is defined in terms of the inner product.

5.1.3 Fourier Series

Shifting from n-dimensional Euclidean space E^n to the space of 2π-periodic square-integrable functions $L^2(0,2\pi)$, a measurable function f is defined on the interval $(0,2\pi)$ as

$$\int_0^{2\pi} |f(x)|^2\, dx < \infty.$$

It may be assumed that f is a piecewise continuous function, extended periodically to the real line $\mathbf{R} = (-\infty,\infty)$ in $L^2(0,2\pi)$ by $f(x) = f(x-2\pi), \forall x$ [Chu92, p.1]. Any f in $L^2(0,2\pi)$ has a Fourier series expansion:

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx},$$

where $i = \sqrt{-1}$ and the Fourier coefficients c_k of f are defined by

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx.$$

Omitting the convergence considerations of Fourier series (detailed in [Kap84, Chu92]), the important feature to note here is that the function e^{ix} forms an orthonormal basis of $L^2(0,2\pi)$ which itself is a vector space. The original (2π-periodic square-integrable) function f is decomposed into (infinitely many) mutually orthogonal components $c_k e^{ikx}$ by the generation of the orthonormal basis $\{w_k\}$ from the *dilation* of the basis function $w(x) = e^{ix}$, i.e., $w_k(x) = w(kx)$, over all integers k.[3] Since the *sinusoidal wave* e^{ix} is the only function required to generate all 2π-periodic square-summable functions, every function in $L^2(0,2\pi)$ is composed of waves of various frequencies.

[3]The Fourier basis is often referred to as a basis of sines and cosines due to the identity: $e^{ix} = cosx + isinx$.

5.1.4 Wavelet Series

Considering the space $L^2(\mathbf{R})$ of (Lebesgue) measurable functions f defined on the real line $\mathbf{R}$, again a single basis function is sought which can be made to express any function f in $L^2(\mathbf{R})$. The function space $L^2(\mathbf{R})$ differs from $L^2(0,2\pi)$ in that the local average values of every function must attenuate to zero at $\pm\infty$. The sinusoidal wave functions $w_k(x)$ do not belong to $L^2(\mathbf{R})$, and cannot be used directly to generate a basis in $L^2(\mathbf{R})$. Instead, "short-term" (quickly decaying) waves, known as *wavelets*, are required. In order to cover the entire space, these *compactly supported* wavelet functions must be shifted (translated) in space. The region where the function is nonzero is said to be its *support*, hence wavelets are nonzero in limited (compact) regions. The set of wavelet functions are formed by dilations and translations of a single function $\psi(x)$ called the "mother wavelet", "basic wavelet", or "analyzing wavelet" [RBC+92]. The term *wavelet* refers to wavelet functions of the form

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi(\frac{x-b}{a}), \quad a>0, b\in\mathbf{R}, \tag{5.6}$$

where dilations and translations are governed by parameters a, b, respectively. Every wavelet ψ generates a series representation of $f \in L^2(\mathbf{R})$:

$$f(x) = \sum_{a,b=-\infty}^{\infty} c_{a,b}\psi_{a,b}, \quad a>0, b\in\mathbf{R},$$

with wavelet coefficients $\{c_{a,b}\}$ given by the integral transform W_ψ:

$$\begin{aligned} c_{a,b} &= \{W_\psi f(x)\}(a,b) \\ &= \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(x)\overline{\psi(\frac{x-b}{a})}dx, \quad f\in L^2(\mathbf{R}),\ a>0, b\in\mathbf{R}, \\ &= \langle f, \psi_{a,b}\rangle. \end{aligned} \tag{5.7}$$

The linear transformation W_ψ is called the *integral wavelet transform*, or simply *wavelet transform*, relative to ψ.[4]

For reasons concerning sampling theory and computational efficiency, the parameters a, b are chosen so that frequency space is partitioned into consecutive frequency bands (or "octaves") by a *binary* dilation, and space is covered by a *dyadic* translation, i.e.,

$$a = 2^{-j};\ b = \frac{k}{2^j}, \quad j,k\in\mathbf{Z}.$$

[4]The resemblance of the wavelet function to the *ket* of quantum mechanics is not accidental. Both the wavelet and the ket are used to represent vector bases. The vector basis *bra*, associated with the ket, corresponds to the scaling function ϕ in the wavelet context, an essential component of multiresolution analysis, described in §5.4. Further similarities between the two domains include the inner product $\langle\cdot,\cdot\rangle$ which is used in place of the Dirac notation $\langle\cdot|\cdot\rangle$. The projection operator $P_\psi = |\psi\rangle\langle\psi|$ used in quantum mechanics is not expressly used in the wavelet literature although the wavelet transform itself defines the projection of an arbitrary function onto the wavelet basis generated by ψ. The tensor product of two vector bases defined in quantum mechanics as $|\phi\rangle\otimes|\psi\rangle$ is significant in the wavelet domain insofar as it is used to construct multidimensional wavelet bases (see §5.5,§5.7) [CDL77].

The *dyadic wavelet* can now be expressed in terms of dilation and translation parameters j,k,

$$\psi_{j,k}(x) \quad = \quad 2^{j/2}\psi(2^j x - k), \quad j,k \in \mathbf{Z}. \tag{5.8}$$

Henceforth expression (5.8) is used to define both *dyadic wavelets* and *wavelets* although it should be noted that a dyadic wavelet is technically distinguished from the basic wavelet, as defined in (5.6), not only by the binary dilation and dyadic translation, but also by a "stability condition" imposed on the basic wavelet (see [Chu92, p.11]).

The dyadic wavelet $\psi_{j,k}$ generates a dyadic series representation of $f \in L^2(\mathbf{R})$:

$$f(x) = \sum_{j,k=-\infty}^{\infty} c_{j,k}\psi_{j,k}, \quad j,k \in \mathbf{Z}, \tag{5.9}$$

with wavelet coefficients $\{c_{j,k}\}$ given by the integral transform:

$$\begin{aligned} c_{j,k} &= \{W_\psi f(x)\}(j,k) \\ &= 2^{j/2}\int_{-\infty}^{\infty} f(x)\overline{\psi(2^j x - k)}dx, \quad f \in L^2(\mathbf{R}),\ j,k \in \mathbf{Z}, \\ &= \langle f, \psi_{j,k}\rangle. \end{aligned} \tag{5.10}$$

That is, the $(j,k)^{th}$ wavelet coefficient of f is given by the integral wavelet transformation of f at dyadic position $b = k/2^j$ with binary dilation $a = 2^{-j}$. Provided the wavelet ψ is orthogonal, the same wavelet ψ is used to generate the wavelet series (5.9) and to define the integral wavelet transform (5.10) [Chu92, p.5]. Orthogonal and other classifications of wavelets are discussed in §5.2.

5.1.5 From Vectors to Wavelets

The common goal among the above three methodologies is the representation of an arbitrary function (or generalized vector) $f(x)$ by a linear combination of basis elements, i.e.,

$$f(x) = \sum_{k=-\infty}^{\infty} c_k \psi_k(x).$$

For vectors, a set of basis vectors is used, i.e., $\{\psi_k(x)\} = \{\mathbf{v}_k\}$. In the Fourier domain, the linear combination is formed from the frequency dilation of the basis function $\{\psi_k(x)\} = \{e^{ikx}\}$. In the wavelet domain, the linear combination is formed from the frequency dilation and spatial translation of the basis function $\{\psi_k(x)\} = \{\psi_{j,k}(x)\}$.[5] Note that for wavelets, j,k are the dilation and translation parameters, respectively,

[5]The precise representation of $f(x)$ is an l^2-linear combination [Chu92, p.3], where l^2 denotes the space of all square-summable bi-infinite sequences; that is, $\{c_k\} \in l^2$ if and only if

$$\sum_{k=-\infty}^{\infty} |c_k|^2 < \infty.$$

whereas in the the Fourier domain these parameters are reversed, that is, k is the frequency dilation parameter, and there is no explicit translation parameter, i.e., $j = 1$, since the Fourier basis functions are infinite in extent [GB92]. The dilation of the basis function in both cases generates the representation of $f(x)$ at multiple frequencies.

The primary distinction between the Fourier and wavelet representations is that the Fourier series uses one basis function at multiple frequencies over all space. The wavelet representation also uses one basis function at multiple frequencies, but to cover all space, it uses many shifted versions of the compactly supported basis function (the mother wavelet), each over a limited spatial region. That is, each wavelet is *localized* in space. Furthermore, the wavelet basis $\{\psi_{j,k}(x)\}$ analyzes a function over a consecutive distribution of frequency bands governed by the parameter j. This frequency distribution results in a hierarchical partitioning of the function by a flexible *space-frequency window* which automatically narrows at high frequencies and widens at low frequencies.[6] The dimensions of the window on the space-frequency grid are governed by j,k with constant area $4\Delta_x\Delta_\omega$, where Δ_x denotes the spatial extent and Δ_ω the frequency extent. The Heisenberg uncertainty principle states that the area of the space-frequency windows (also called *Heisenberg boxes*) can be no greater than 2. The automatic dilation of the wavelet Heisenberg boxes maintains constant area, but as the boxes shrink in space they stretch over frequency. This characteristic of the wavelet representation is known as the wavelets' *zooming property* and it is the wavelets' paramount advantage over traditional Fourier techniques for signal analysis. For further detail and precise definition of the wavelet space-frequency window see [JS94a, Chu92].

The benefit gained by the wavelets' flexible space-frequency window can be illustrated by the following abstract example of burst signal detection. Because the Fourier representation integrates the basis function over all space, i.e., over the entire signal, the frequency content is recorded over all space. If a high-frequency burst is present in the signal, Fourier analysis will only report that such a high frequency component is present *somewhere* in the signal. The wavelet representation, on the other hand, due to its hierarchical frequency partitioning, enables the detection *and* localization of transient signal components such as bursts. In an effort to provide similar functionality, short-term Fourier approaches (such as the Short-Time Fourier Transform, or STFT) use spatially localized basis functions, but still suffer from a fixed space-frequency window (see [Chu92, §1.2] for details). A schematic of the STFT and wavelet space-frequency tiling is shown in Figure 14.

[6]The space-frequency window is also known as the time-frequency window. There is no real distinction between space and time except within the context of the analysis. Typically time refers to time-varying signals such as speech, whereas space may refer to the spatial (x,y) location of an image pixel.

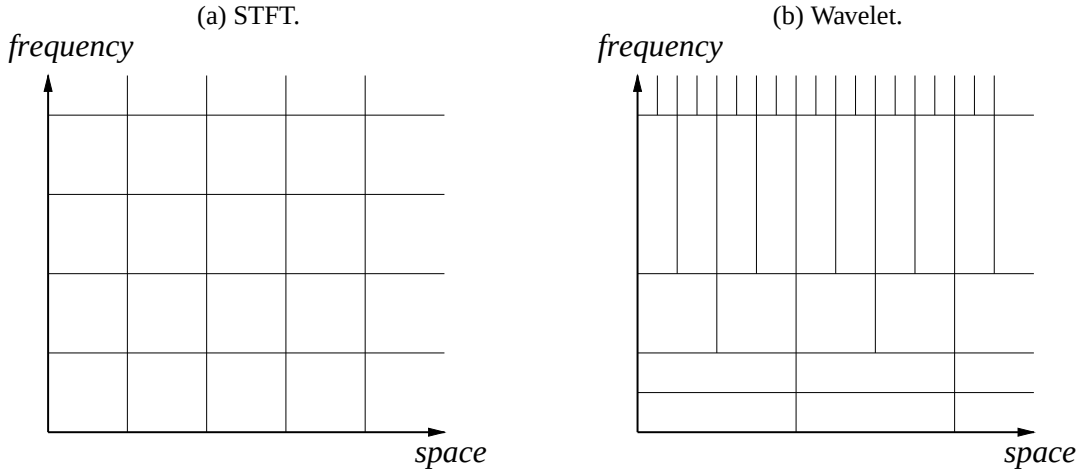


Fig. 14. Space-frequency tiling of the STFT and Wavelet representations. Adapted from [Bar94, p.9 (Fig. 2.1)].

5.2 Wavelet Functions

Denoting the closure of the linear span of $\{\psi_{j,k} : k \in \mathbf{Z}\}$ by W_j for each $j \in \mathbf{Z}$, i.e.,

$$W_j = clos_{L^2(\mathbf{R})}\langle \psi_{j,k} : k \in \mathbf{Z}\rangle,$$

$L^2(\mathbf{R})$ can be decomposed as a *direct sum* of the spaces W_j:

$$L^2(\mathbf{R}) = \sum_{j\in\mathbf{Z}}^{\bullet} W_j = \cdots \dot{+} W_{-1} \dot{+} W_0 \dot{+} W_1 \dot{+} \cdots, \tag{5.11}$$

where $\dot{+}$ indicates "direct sum", in the sense that every function $f \in L^2(\mathbf{R})$ has a unique decomposition:

$$f(x) = \cdots + g^{-1}(x) + g^0(x) + g^1(x) + \cdots, \tag{5.12}$$

where $g^j \in W_j$ and $g^l \in W_l$. Any wavelet generates a direct sum decomposition of $L^2(\mathbf{R})$.

5.2.1 Bi-orthogonal Wavelets

Every wavelet $\psi \in L^2(\mathbf{R})$, as defined by (5.8), has a *dual* $\widetilde{\psi} \in L^2(\mathbf{R})$ defined by

$$\widetilde{\psi}_{l,m}(x) = 2^{l/2}\widetilde{\psi}(2^l x - m), \quad l, m \in \mathbf{Z}.$$

If the bases $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{l,m}\}$ generated by the dual wavelets ψ and $\widetilde{\psi}$ satisfy

$$\langle \psi_{j,k}, \widetilde{\psi}_{l,m}\rangle = \delta_{j,l} \cdot \delta_{k,m}, \quad j,k,l,m \in \mathbf{Z}, \tag{5.13}$$

i.e., the bases establish inter-scale ($\delta_{j,l}$) and intra-scale ($\delta_{k,m}$) orthogonality, then $(\psi, \widetilde{\psi})$ form a pair of *bi-orthogonal wavelets*, and every $f \in L^2(\mathbf{R})$ can be written as a wavelet series

$$f(x) \quad = \quad \sum_{j,k\in\mathbf{Z}} d_{j,k}\widetilde{\psi}_{j,k}(x) \tag{5.14}$$

$$= \sum_{j,k\in\mathbf{Z}} d_{j,k}\psi_{j,k}(x), \tag{5.15}$$

where, analogous to Fourier coefficients, wavelet coefficients are given by

$$\begin{aligned} d_{j,k} &= \langle f, \psi_{j,k}\rangle \quad \text{in (5.14);} \\ &= \langle f, \widetilde{\psi}_{j,k}\rangle \quad \text{in (5.15).} \end{aligned}$$

If ψ and $\widetilde{\psi}$ constitute a bi-orthogonal wavelet pair, then they generate two subspaces $\{W_j\}, \{\widetilde{W}_j\}$ of $L^2(\mathbf{R})$ where the subspaces are not generally mutually orthogonal,

$$W_j \not\perp W_l, \text{ and } \widetilde{W}_j \not\perp \widetilde{W}_l, \quad j \neq l, \tag{5.16}$$

but instead are orthogonal in the dual sense,

$$W_j \perp \widetilde{W}_l, \quad j \neq l. \tag{5.17}$$

Since both $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{l,m}\}$ are bases of $L^2(\mathbf{R})$, the space can be decomposed by either basis, i.e.,

$$\begin{aligned} L^2(\mathbf{R}) &= \dot{\sum_{j\in\mathbf{Z}}} W_j = \cdots \dot{+} W_{-1} \dot{+} W_0 \dot{+} W_1 \dot{+} \cdots \\ &= \dot{\sum_{j\in\mathbf{Z}}} \widetilde{W}_j = \cdots \dot{+} \widetilde{W}_{-1} \dot{+} \widetilde{W}_0 \dot{+} \widetilde{W}_1 \dot{+} \cdots. \end{aligned} \tag{5.18}$$

Equations (5.14) and (5.15) effectively state that any function in $L^2(\mathbf{R})$ projected onto one basis can be recovered by expansion in the other. In contrast to an orthogonal wavelet basis (see below), the bi-orthogonal system permits greater freedom in the construction of wavelet filters (see §5.6.4). For details pertaining to the convergence of the series, see [Chu92, p.5 and §3.6].

5.2.2 Orthogonal Wavelets

A function $\psi \in L^2(\mathbf{R})$ is called an *orthogonal wavelet* if the family $\{\psi_{j,k}\}$ forms an orthonormal basis of $L^2(\mathbf{R})$,

$$\langle \psi_{j,k}, \psi_{l,m}\rangle = \delta_{j,l} \cdot \delta_{k,m}, \quad j,k,l,m \in \mathbf{Z}. \tag{5.19}$$

An orthogonal wavelet is self-dual with $\psi = \widetilde{\psi}$ generating $\{\psi_{j,k}\}$ so that every $f \in L^2(\mathbf{R})$ can be represented by the wavelet series

$$f(x) = \sum_{j,k\in\mathbf{Z}} d_{j,k}\psi_{j,k}(x),$$

with wavelet coefficients

$$d_{j,k} = \langle f, \psi_{j,k}\rangle.$$

Given an orthogonal wavelet ψ, the subspaces $\{W_j\}$ of $L^2(\mathbf{R})$ generated by the wavelet are mutually orthogonal,

$$W_j \perp W_l, \quad j \neq l. \tag{5.20}$$

The function decompositions, as given by (5.12), are also orthogonal, i.e.,

$$\langle g^j, g^l \rangle = 0, \quad j \neq l,$$

and the direct sum of subspaces (5.11) becomes an *orthogonal sum*:

$$L^2(\mathbf{R}) = \bigoplus_{j \in \mathbf{Z}} W_j = \cdots \oplus W_{-1} \oplus W_0 \oplus W_1 \oplus \cdots, \tag{5.21}$$

where $\oplus$ indicates "orthogonal sum" (see [Chu92, pp.14-15] for details).

5.2.3 Semi-orthogonal Wavelets

A function $\psi \in L^2(\mathbf{R})$ is called a *semi-orthogonal wavelet* if the generated basis $\{\psi_{j,k}\}$ satisfies

$$\langle \psi_{j,k}, \psi_{l,m} \rangle = 0, \quad j \neq l, \quad j,k,l,m \in \mathbf{Z}, \tag{5.22}$$

where $\langle \psi_{j,k}, \psi_{j,m} \rangle$ may be non-zero. That is, the wavelet does not necessarily provide intra-scale orthogonality. The distinction between orthogonal and semi-orthogonal wavelets is analogous to orthogonal and orthonormal vector bases. Semi-orthogonal wavelets, also known as "pre-wavelets", can produce orthogonal wavelets through an orthogonalization procedure [RBC+92, p.8]. Every semi-orthogonal wavelet generates an (inter-scale) orthogonal decomposition (but not necessarily an orthonormal one), and every orthogonal wavelet is also a semi-orthogonal wavelet since (5.19) guarantees (5.22).

Although semi-orthogonal wavelets are not generally fully orthonormal, the subspaces they generate are mutually orthogonal, i.e., the condition expressed by (5.20) holds, and the function decompositions, as given by (5.12), are also orthogonal, i.e.,

$$\langle g^j, g^l \rangle = 0, \quad j \neq l, \quad j,l \in \mathbf{Z}.$$

5.2.4 Non-orthogonal Wavelets

A wavelet ψ is called *non-orthogonal* if it is not a semi-orthogonal wavelet. Bi-orthogonal wavelets are generally non-orthogonal, meaning that the resulting bases typically lack both inter-scale and intra-scale orthogonality [RBC+92, p.8].

5.3 Wavelet Maxima and Multiscale Edges

A critical consideration in almost any signal analysis task is the detection of sharp variation points. The wavelet transform is closely related to multiscale edge detection employed in computer vision [MZ92a]. Mallat et al. have developed an adaptive sampling technique to locate signal sharp variation points by detecting local maxima of the wavelet transform modulus. The method is equivalent to the Canny edge detector [Can86]. Several papers and book chapters by Mallat can be found on this topic, including [MZ92a, MZ92b, MH92, FM92, Mal91]. Because this edge detection technique is directly applicable to eye movement modeling and video analysis it is summarized here, closely following Mallat's derivations. Where appropriate, the reference to the relevant source is provided.

The wavelet transform of f at scale j and position x, given in (5.10), defines the convolution product

$$\{W_\psi f(x)\}(j) = f * \psi_j(x),$$

where the translation parameter k is made implicit. The *dyadic wavelet transform* is defined as the sequence of functions

$$\mathbf{W}f = \left[\{W_\psi f(x)\}(j)\right]_{j\in\mathbf{Z}},$$

where $\mathbf{W}$ is the dyadic wavelet transform operator. Assuming a twice-differentiable smoothing function $\theta(x)$ exists, whose integral is equal to 1 and that converges to 0 at infinity, e.g., a Gaussian, define the first- and second-order derivatives of $\theta(x)$:

$$\psi'(x) = \frac{d\theta(x)}{dx}, \text{ and, } \psi''(x) = \frac{d^2\theta(x)}{dx^2}.$$

The functions ψ' and ψ'' are by definition wavelets since their integral is equal to 0. Denoting the wavelet transforms of $f(x)$ relative to ψ', ψ'' as,

$$\{W_{\psi'} f(x)\}(j) = f * \psi_j'(x), \text{ and } \{W_{\psi''} f(x)\}(j) = f * \psi_j''(x),$$

$\{W_{\psi'} f(x)\}(j)$, $\{W_{\psi''} f(x)\}(j)$ are the first and second derivative of the signal smoothed at scale j [MZ92a]:

$$\{W_{\psi'} f(x)\}(j) = f * (j\frac{d\theta_j}{dx})(x) = j\frac{d}{dx}(f * \theta_j)(x), \text{ and}$$
$$\{W_{\psi''} f(x)\}(j) = f * (j^2\frac{d^2\theta_j}{dx^2})(x) = j^2\frac{d^2}{dx^2}(f * \theta_j)(x).$$

The local extrema of $\{W_{\psi'} f(x)\}(j)$ correspond to the zero crossings of $\{W_{\psi''} f(x)\}(j)$ and to the inflection points of $f * \theta_j(x)$. In the particular case where $\theta(x)$ is a Gaussian, the zero-crossing detection is equivalent to a Marr-Hildreth edge detection [Mar80], and the extrema detection corresponds to Canny edge detection [Can86]. The wavelet approach follows the latter, relying on $\{W_{\psi'} f(x)\}(j)$ to distinguish between sharp and slow variation points of $f * \theta_j(x)$, which is often difficult using a second derivative operator.

Sharp variation points are detected by finding the local maxima of the modulus $|\{W_{\psi'} f(x)\}(j)|$. At each scale j, local modulus maxima are located by finding the points where $|\{W_{\psi'} f(x)\}(j)|$ is larger than its two closest neighbor values, and strictly larger than at least one of them [MH92]. That is, a modulus maxima is located at scale j and location (x_0) if:

$$|\{W_{\psi'} f(x_0-1)\}(j)| \le |\{W_{\psi'} f(x_0)\}(j)| \ge |\{W_{\psi'} f(x_0+1)\}(j)|, \text{ and} \tag{5.23}$$

$$\begin{cases} |\{W_{\psi'} f(x_0)\}(j)| > |\{W_{\psi'} f(x_0-1)\}(j)|, & \text{or} \\ |\{W_{\psi'} f(x_0)\}(j)| > |\{W_{\psi'} f(x_0+1)\}(j)|. \end{cases} \tag{5.24}$$

The modulus maxima of the wavelet transform at scale j and location (x_0) is a strict local maxima of the modulus on the right or the left of location x_0.

The local maxima detection is extendible to multiple dimensions if there exists a smoothing function θ, which converges to 0 at infinity yet totally integrates to 1. In two dimensions, the image function $f(x,y)$ is smoothed at different scales j by convolution with the two-dimensional smoothing function $\theta_j(x,y)$. Computing the gradient vector, $\nabla(f * \theta_j)(x,y)$, edges are defined as points (x_0,y_0) where the modulus of the gradient vector is maximum in the direction of the gradient in the image plane. Introducing two 2D wavelet functions,

$$\psi_x(x,y) = \frac{\partial\theta(x,y)}{\partial x} \text{ and } \psi_y(x,y) = \frac{\partial\theta(x,y)}{\partial y},$$

two components of the wavelet transform of $f(x,y) \in L^2(\mathbf{R}^2)$ at scale j are defined with implicit translation parameter k:

$$\{Wf(x,y)\}_x(j) = \{W_{\psi_x} f(x,y)\}(j) = f * \psi_{x_j}(x,y), \text{ and}$$
$$\{Wf(x,y)\}_y(j) = \{W_{\psi_y} f(x,y)\}(j) = f * \psi_{y_j}(x,y).$$

Edge points can be located from the two components $\{Wf(x,y)\}_x(j)$, $\{Wf(x,y)\}_y(j)$ since

$$\begin{pmatrix} \{Wf(x,y)\}_x(j) \\ \{Wf(x,y)\}_y(j) \end{pmatrix} = j \begin{pmatrix} \frac{\partial}{\partial x}(f * \theta_j)(x,y) \\ \frac{\partial}{\partial y}(f * \theta_j)(x,y) \end{pmatrix} = j\nabla(f * \theta_j)(x,y).$$

Sharp variation points are detected analogously to the 1D case, where the modulus at scale j and position (x,y), denoted by $\{Mf(x,y)\}(j)$, is proportional to:

$$|\{Wf(x,y)\}(j)| \propto \{Mf(x,y)\}(j) = \sqrt{|\{Wf(x,y)\}_x(j)|^2 + |\{Wf(x,y)\}_y(j)|^2}. \tag{5.25}$$

At each scale j, the local modulus maxima are again found by comparing the point $\{Mf(x,y)\}(j)$ with its two neighbors as in (5.23) and (5.24), except now neighboring points must be examined along the direction

of the gradient vector. The angle of the gradient vector with horizontal direction at scale j and position (x,y), denoted by $\{Af(x,y)\}(j)$ is given by [MZ92b]:

$$\begin{aligned} \{Af(x,y)\}(j) &= \arg(\{Wf(x,y)\}_x(j) + i\{Wf(x,y)\}_y(j)) \\ &= \tan^{-1}\left(\frac{\{Wf(x,y)\}_y(j)}{\{Wf(x,y)\}_x(j)}\right). \end{aligned} \tag{5.26}$$

In three dimensions, the volume function $f(x,y,t)$ is smoothed at different scales j by convolution with the three-dimensional smoothing function $\theta_j(x,y,t)$. Computing the gradient vector, $\nabla(f * \theta_j)(x,y,t)$, edges are defined as points (x_0,y_0,t_0) where the modulus of the gradient vector is maximum in the direction of the gradient in the volume. Introducing three 3D wavelet functions,

$$\psi_x(x,y,t) = \frac{\partial\theta(x,y,t)}{\partial x},\ \psi_y(x,y,t) = \frac{\partial\theta(x,y,t)}{\partial y},\ \text{and}\ \psi_t(x,y,t) = \frac{\partial\theta(x,y,t)}{\partial t},$$

three components of the wavelet transform of $f(x,y,t) \in L^2(\mathbf{R}^3)$ at scale j are defined with implicit translation parameter k:

$$\begin{aligned} \{Wf(x,y,t)\}_x(j) &= \{W_{\psi_x} f(x,y,t)\}(j) = f * \psi_{x\,j}(x,y,t), \\ \{Wf(x,y,t)\}_y(j) &= \{W_{\psi_y} f(x,y,t)\}(j) = f * \psi_{y\,j}(x,y,t),\ \text{and} \\ \{Wf(x,y,t)\}_t(j) &= \{W_{\psi_t} f(x,y,t)\}(j) = f * \psi_{t\,j}(x,y,t). \end{aligned}$$

Edge points can be located from the above three components since

$$\begin{pmatrix} \{Wf(x,y,t)\}_x(j) \\ \{Wf(x,y,t)\}_y(j) \\ \{Wf(x,y,t)\}_t(j) \end{pmatrix} = j \begin{pmatrix} \frac{\partial}{\partial x}(f*\theta_j)(x,y,t) \\ \frac{\partial}{\partial y}(f*\theta_j)(x,y,t) \\ \frac{\partial}{\partial t}(f*\theta_j)(x,y,t) \end{pmatrix} = j\nabla(f*\theta_j)(x,y,t).$$

Sharp variation points are detected analogously to the 1D case, where the modulus at scale j and position (x,y,t), denoted by $\{Mf(x,y,t)\}(j)$, is proportional to:

$$\begin{aligned} &|\{Wf(x,y,t)\}(j)| \propto \{Mf(x,y,t)\}(j) = \\ &\sqrt{|\{Wf(x,y,t)\}_x(j)|^2 + |\{Wf(x,y,t)\}_y(j)|^2 + |\{Wf(x,y,t)\}_t(j)|^2}. \end{aligned} \tag{5.27}$$

At each scale j, the local modulus maxima are again found by comparing the point $\{Mf(x,y,t)\}(j)$ with its two neighbors along the direction of the gradient vector. The angle of the gradient vector is now determined by three planar angles at scale j and position (x,y,t), denoted by $\{Af(x,y,t)\}_{\angle ab}(j)$ where $\angle ab$ specifies the directional plane, given by:

$$\{Af(x,y,t)\}_{\angle xy}(j) = \tan^{-1}\left(\frac{\{Wf(x,y,t)\}_y(j)}{\{Wf(x,y,t)\}_x(j)}\right), \tag{5.28}$$

$$\{Af(x,y,t)\}_{\angle xt}(j) = \tan^{-1}\left(\frac{\{Wf(x,y,t)\}_t(j)}{\{Wf(x,y,t)\}_x(j)}\right), \tag{5.29}$$

$$\{Af(x,y,t)\}_{\angle yt}(j) = \tan^{-1}\left(\frac{\{Wf(x,y,t)\}_y(j)}{\{Wf(x,y,t)\}_t(j)}\right). \tag{5.30}$$

5.4 Multiresolution Analysis

Multiresolution analysis (MRA), introduced by Meyer and Mallat [Mal89a], is an algorithmic framework for representing functions at hierarchical levels of scale (or resolution). The wavelet basis described above analyzes the underlying signal in terms of spatially-localized frequency components. Using the wavelet basis alone, reconstruction of the signal may be problematic. In order to reconstruct the original signal from its wavelet representation, the wavelet dual $\widetilde{\psi}$ is used as the reconstruction kernel function in the inversion formula defined by Equation (5.14). In general, however, $\widetilde{\psi}$ does not exist [Chu92, p.13]. Multiresolution analysis addresses this reconstruction problem by maintaining a scaled version of the signal at consecutive levels of resolution. The original signal can be faithfully reconstructed by successively combining the scaled signal with the wavelet coefficients at each level of resolution.

5.4.1 Scaling Functions

At the heart of multiresolution analysis is the notion of a *scaling function*, denoted by $\phi(x)$. The scaling function is very similar in nature to the wavelet in that it also generates a basis of $L^2(\mathbf{R})$. The scaling function is also a compactly supported function, defined as

$$\phi_{a,b}(x) = \frac{1}{\sqrt{a}}\phi(\frac{x-b}{a}), \quad a > 0, b \in \mathbf{R},$$

where again a, b are the dilation and translation parameters. As for the wavelet function, integral powers of 2 are used where the scaling function is obtained by a binary dilation (dilation by 2^j), and a dyadic translation (translation of $k/2^j$) of a single function ϕ. That is, a, b are chosen as for the wavelet function, and the scaling function becomes

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \quad j, k \in \mathbf{Z}.$$

5.4.2 Scale Subspaces

Since the scaling function generates a basis of $L^2(\mathbf{R})$, it also generates subspaces $\{V_j\}$, just as the subspaces $\{W_j\}$ are generated by ψ above, i.e.,

$$V_j = clos_{L^2(\mathbf{R})}\langle \phi_{j,k} : k \in \mathbf{Z} \rangle,$$

with ϕ generating a reference subspace V_0, i.e.,

$$V_0 = clos_{L^2(\mathbf{R})}\langle \phi_{0,k} : k \in \mathbf{Z} \rangle.$$

In contrast to the sequence of orthogonal subspaces $\{W_j\}$ generated by an orthogonal ψ satisfying (5.20), the *nested* sequence of closed subspaces $\{V_j\}$ generated by the scaling function possess the following properties:

(5.31) 1. $\cdots \subset V_{-1} \subset V_0 \subset V_1 \cdots$ (*containment*);

(5.32) 2. $\overline{\cup_{j\in\mathbf{Z}} V_j} = L^2(\mathbf{R})$ (*completeness*);

(5.33) 3. $\cap_{j\in\mathbf{Z}} V_j = \{0\}$ (*uniqueness*);

(5.34) 4. $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}, j \in \mathbf{Z}$ (*scalability*).

Property (5.31) states that the sequence of subspaces is nested; property (5.32) states that every function f in $L^2(\mathbf{R})$ can be approximated as closely as desired by its projections in V_j; property (5.33), on the other hand, states that, by decreasing j, the projections could have arbitrarily small energy [Chu92, p.16]; and property (5.34) is the multiresolution condition which states that as j increases, the spaces V_j correspond to "finer resolution": if the function f is in the basic multiresolution space V_0, then the finer resolution function $f(2^j\cdot) : x \mapsto f(2^j x)$ is in the space indexed by j [Fou95, p.43]. The scaling function is said to generate a multiresolution analysis if it generates a nested sequence of subspaces $\{V_j\}$ satisfying the above properties such that $\{\phi_{0,k}\}$ forms a basis of V_0.

5.4.3 Bi-orthogonal Multiresolution

Given a pair of scaling and wavelet functions (ϕ, ψ), neither of which necessarily forms an orthogonal basis, the goal is to specify dual functions $(\widetilde{\phi}, \widetilde{\psi})$ so that the original function f in $L^2(\mathbf{R})$ can be perfectly reconstructed. Recall that wavelets are bi-orthogonal if they satisfy condition (5.13) and generate dually orthogonal subspaces as expressed by Equations (5.16) and (5.17). Assuming that the scaling functions $(\phi, \widetilde{\phi})$ are dual as per (5.13), and imposing the following intra-scale orthogonality conditions:

(5.35) $$\langle \phi_{j,k}, \widetilde{\psi}_{j,l} \rangle = 0, \text{ and } \langle \widetilde{\phi}_{j,k}, \psi_{j,l} \rangle = 0, \quad j,k,l \in \mathbf{Z},$$

the double multiresolution generated by $(\phi, \widetilde{\phi})$ with two sequences of subspaces $\{V_j\}, \{\widetilde{V}_j\}$ then satisfies

$$V_j \perp \widetilde{W}_j \text{ and } \widetilde{V}_j \perp W_j, \quad j \in \mathbf{Z},$$

and $L^2(\mathbf{R})$ is decomposed as in (5.18), with

$$V_{j+1} = V_j \dot{+} W_j \text{ and } \widetilde{V}_{j+1} = \widetilde{V}_j \dot{+} \widetilde{W}_j, \quad j \in \mathbf{Z}.$$

The pairs (ϕ, ψ) and $(\widetilde{\phi}, \widetilde{\psi})$ are interchangeable in the sense that only one of the pairs needs to be specified. The second pair is derived from the first with ϕ connected to $\widetilde{\psi}$ and ψ connected to $\widetilde{\phi}$ (see §5.6.4) [Fou95, §II].

5.4.4 Orthogonal Multiresolution

If the scaling function ϕ can be chosen so that the set of translates $\{\phi_{0k}\} = \{\phi(x-k)\}$ forms an orthonormal basis and generates a set of multiresolution subspaces $\{V_j\}$, then an orthonormal wavelet basis can be

constructed from ϕ [RBC+92]. Defining W_j as $V_j^{\perp}$, where the orthogonal complement is taken in V_{j+1}, so that

$$V_{j+1} = V_j \oplus W_j \text{ and } V_j \perp W_j,$$

$L^2(\mathbf{R})$ is decomposed as in (5.21). A function ψ is sought so that $\{\psi_{j,k}\}$ forms an orthonormal basis for $L^2(\mathbf{R})$, and subsequently $\{\psi_{j,k}\}$ is an orthonormal basis for W_j. Assuming that integer translates of ϕ generate an orthonormal basis for V_0 and there exist c_k such that

$$\phi(x) \quad = \quad \sum_{k\in\mathbf{Z}} c_k \phi(2x-k),$$

then $\psi(x)$ is given by

$$\psi(x) \quad = \quad \sum_{k\in\mathbf{Z}} (-1)^k c_{k+1} \phi(2x+k). \tag{5.36}$$

By the above construction and orthonormality of ϕ,

$$\langle \phi_{j,k}, \psi_{j,l} \rangle = 0, \quad j,k,l \in \mathbf{Z},$$

and ϕ, ψ are each self-dual, satisfying (5.35). That is, orthogonal multiresolution is a special case of bi-orthogonal multiresolution where $\phi = \widetilde{\phi}$ and $\psi = \widetilde{\psi}$.

5.5 Wavelet Decomposition and Reconstruction

Given the multiresolution framework, wavelet decomposition and reconstruction algorithms can be derived for any f in $L^2(\mathbf{R})$. Since $\widetilde{\phi} \in L^2(\mathbf{R})$ generates $\{\widetilde{V}_j\}$ and $\widetilde{\psi} \in L^2(\mathbf{R})$ generates $\{\widetilde{W}_j\}$, and by multiresolution property (5.32) above, every function f in $L^2(\mathbf{R})$ can be approximated by an $f^N \in \widetilde{V}_N$, for some $N \in \mathbf{Z}$. Consider $\widetilde{V}_N$ as the "sample space" and f^N the "data" (or measurement) of f on $\widetilde{V}_N$. Since

$$\begin{aligned} \widetilde{V}_N &= \widetilde{W}_{N-1} \dot{+} \widetilde{V}_{N-1} \\ &= \widetilde{W}_{N-1} \dot{+} \cdots \dot{+} \widetilde{W}_{N-M} \dot{+} \widetilde{V}_{N-M}, \end{aligned}$$

for any positive integer M, f^N has a unique decomposition:

$$f^N(x) \quad = \quad f^{N-1}(x) + g^{N-1}(x),$$

where $f^{N-1} \in \widetilde{V}_{N-1}$ and $g^{N-1} \in \widetilde{W}_{N-1}$. Recursively,

$$f^N(x) = g^{N-1}(x) + g^{N-2}(x) + \cdots + g^{N-M}(x) + f^{N-M}(x), \tag{5.37}$$

where

$$f^j(x) \quad = \quad \sum_k c_{j,k} \widetilde{\phi}(2^j x - k) \in \widetilde{V}_j \ : \ \mathbf{c}^j = \{c_{j,k}\}, \quad k \in \mathbf{Z}; \tag{5.38}$$

$$g^j(x) \quad = \quad \sum_k d_{j,k} \widetilde{\psi}(2^j x - k) \in \widetilde{W}_j \ : \ \mathbf{d}^j = \{d_{j,k}\}, \quad k \in \mathbf{Z}; \tag{5.39}$$

and

$$f^{N-M}(x) \in \widetilde{V}_{N-M}, \quad j = N-M, N-M+1, \ldots, N-1,$$

with the normalization factor $2^{j/2}$ folded into the series coefficients and M chosen so that f^{N-M} is sufficiently decomposed. The decomposition in (5.37) is uniquely determined by the sequences $\mathbf{c}^j$ and $\mathbf{d}^j$, in (5.38) and (5.39) [Chu92, pp.156-157], which are the scale and wavelet coefficients obtained from the multiresolution projection of f^j onto subspaces V_j, W_j as generated by ϕ, ψ, respectively:

$$c_{j,k} = \langle f^j, \phi_{j,k} \rangle, \quad d_{j,k} = \langle f^j, \psi_{j,k} \rangle. \tag{5.40}$$

Note that here, contrary to convention, ψ, not $\widetilde{\psi}$, is used as the analyzing wavelet, although by the *duality principle* [Chu92, p.156], the pairs $(\phi, \psi), (\widetilde{\phi}, \widetilde{\psi})$ are interchangeable for decomposition and reconstruction purposes. That is, incorporating Equations (5.38) and (5.39) into (5.37), the function f^{j+1} can be obtained from either combination of dual pairs by the following (bi-orthogonal) inversion formula [GB92, p.634]:

$$f^{j+1}(x) = \sum_{j,k \in \mathbf{Z}} c_{j,k} \widetilde{\phi}_{j,k}(x) + \sum_{j,k \in \mathbf{Z}} d_{j,k} \widetilde{\psi}_{j,k}(x) \tag{5.41}$$

$$= \sum_{j,k \in \mathbf{Z}} c_{j,k} \phi_{j,k}(x) + \sum_{j,k \in \mathbf{Z}} d_{j,k} \psi_{j,k}(x), \tag{5.42}$$

with scale and wavelet coefficients

$$\begin{cases} c_{j,k} = \langle f^j, \phi_{j,k} \rangle, \quad d_{j,k} = \langle f^j, \psi_{j,k} \rangle & \text{in (5.41);} \\ c_{j,k} = \langle f^j, \widetilde{\phi}_{j,k} \rangle, \quad d_{j,k} = \langle f^j, \widetilde{\psi}_{j,k} \rangle & \text{in (5.42).} \end{cases}$$

In orthogonal MRA, with self-dual ϕ and ψ functions, Equations (5.41) and (5.42) condense into one (orthogonal) inversion formula:

$$f^{j+1}(x) = \sum_{j,k \in \mathbf{Z}} c_{j,k} \phi_{j,k}(x) + \sum_{j,k \in \mathbf{Z}} d_{j,k} \psi_{j,k}(x),$$

with coefficients as given by (5.40).

The most important property of the subspaces $\{V_j\}$ and $\{W_j\}$ (or $\{\widetilde{V}_j\}$ and $\{\widetilde{W}_j\}$, depending on which dual pair is used for decomposition), and hence multiresolution analysis in general, is that as $j \to -\infty$, more and more "variations" of the analyzed function are removed at each "rate of variation", or frequency band, j, and stored in W_j. The remaining coarser approximations to the function remain in V_j. The crux of the recursive nature of MRA is the decomposition of the coarse function at level j into the function's coarser approximation and stripped "variation" at level $j-1$, as projected onto V_{j-1} and W_{j-1}, respectively.[7] The

[7] Note that some authors use a convention of increasing subspaces [RBC+92]. Roughly speaking, in the Meyer convention (adopted here) the functions in V_j scale like 2^{-j}, whereas in the Daubechies convention they scale like 2^j. That is, in the Meyer convention, the decomposition level j is commensurate with the resolution of the function under study, i.e., level $j = 0$ represents the coarsest resolution. In the Daubechies

algorithmic approach for decomposing and reconstructing the function f^j between resolution levels is accomplished through the use of discrete sequences which approximate the scaling and wavelet functions ϕ, $\widetilde{\phi}$, ψ, $\widetilde{\psi}$.

Since both $\widetilde{\phi} \in \widetilde{V}_0$ and $\widetilde{\psi} \in \widetilde{W}_0$ are in $\widetilde{V}_1$, and since $\widetilde{V}_1$ is generated by $\widetilde{\phi}_{1,k}(x) = 2^{1/2}\widetilde{\phi}(2x-k),\ k \in \mathbf{Z}$, there exist two sequences denoted by $\{\widetilde{p}_k\}$ and $\{\widetilde{q}_k\}$ such that

$$\widetilde{\phi}(x) = \sum_k \widetilde{p}_k \widetilde{\phi}(2x-k); \tag{5.43}$$

$$\widetilde{\psi}(x) = \sum_k \widetilde{q}_k \widetilde{\phi}(2x-k), \tag{5.44}$$

for all $x \in \mathbf{R}$. These are the *two-scale*, *dilation*, or *refinement* relations of the scaling and wavelet functions, respectively. These relations imply that $\widetilde{\phi}(x)$ and $\widetilde{\psi}(x)$ must be generated by the finer scale functions $\widetilde{\phi}(2x-k)$, and lead to the decomposition algorithm.

Conversely, since both $\phi(2x)$ and $\phi(2x-1)$ are in V_1 and $V_1 = V_0 \dot{+} W_0$, there are two sequences denoted by $\{p_k\}$ and $\{q_k\}$, k in $\mathbf{Z}$, such that

$$\phi(2x-l) \quad = \quad \sum_k [p_{l-2k}\phi(x-k) + q_{l-2k}\psi(x-k)], \quad l \in \mathbf{Z}. \tag{5.45}$$

This is called the *decomposition relation* of ϕ and ψ. Mathematically, the decomposition relation roughly states that the function under analysis at a given resolution level (scale) can be decomposed into a coarser resolution approximation plus the stripped-off detail. Computationally, perhaps somewhat counterintuitively, the decomposition leads to the reconstruction algorithm. The two pairs of sequences $(\{\widetilde{p}_k\}, \{\widetilde{q}_k\})$ and $(\{p_k\}, \{q_k\})$, are unique once the normalization of ϕ is fixed (see [Chu92, §1.6] for details).

Representing f^j and g^j from (5.38) and (5.39) by the "digital" sequences $\mathbf{c}^j$ and $\mathbf{d}^j$, the following generalized (bi-orthogonal) decomposition and reconstruction algorithms emerge:

Decomposition:

$$c_k^{j-1} = \sum_l p_{l-2k} c_l^j; \quad d_k^{j-1} = \sum_l q_{l-2k} c_l^j, \tag{5.46}$$

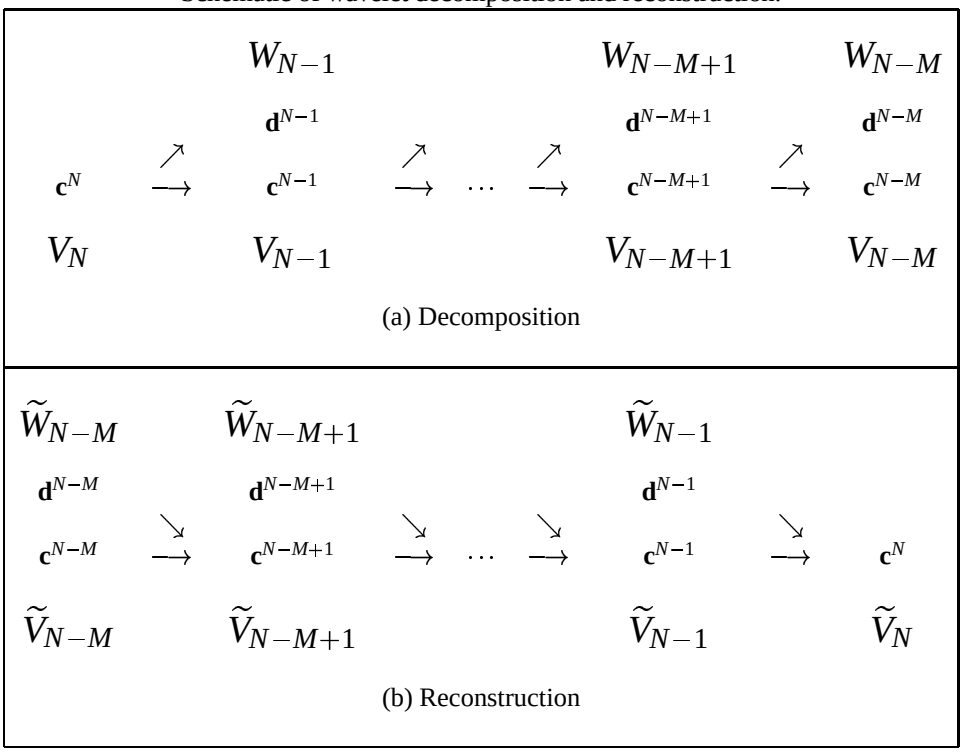
Reconstruction:

$$c_k^j = \sum_l [\widetilde{p}_{k-2l} c_l^{j-1} + \widetilde{q}_{k-2l} d_l^{j-1}], \tag{5.47}$$

convention, the decomposition level j pertains to the number of decompositions applied to the function under study, i.e., level $j = 0$ represents the finest resolution since no decompositions have been applied to the function. Both conventions are equally informative since in the former the "current" resolution level can be used directly in estimating the extent of the function (i.e., the number of samples present in the scaled signal—this is particularly useful when dealing with images). The latter convention provides information in terms of number of decompositions applied to the function, which can be a valuable measure in a recursive implementation.

where $\{p_k\}$ and $\{q_k\}$ are *decomposition sequences*, while $\{\widetilde{p}_k\}$ and $\{\widetilde{q}_k\}$ are *reconstruction sequences*. These sequences correspond to digital filters in signal analysis.[8] Note that in the case of orthogonal MRA, the filters coincide, i.e., $\{p_k\} = \{\widetilde{p}_k\}$ and $\{q_k\} = \{\widetilde{q}_k\}$. The decomposition and reconstruction algorithms are shown schematically in Table 3.

TABLE 3
Schematic of wavelet decomposition and reconstruction.



The wavelet transform generalizes to multiple dimensions provided the scaling functions and wavelets generate multidimensional bases. In the particular two-dimensional case, there are two ways in which the 1D transform can be generalized, namely through the *standard* and *non-standard* decompositions.

The standard decomposition of a typical 2D function, i.e., an image, $f(x,y)$, is obtained by first applying the 1D wavelet transform to each row of (pixel) values, giving average (smoothed) values with detail coefficients for each row. The transformed rows are treated as 1D functions themselves and the 1D wavelet transform is

[8] Some authors prefer to concentrate on reconstruction filters as the "nice" filters and denote decomposition sequences by a special symbol. Because the decomposition is more pertinent to signal analysis, here the opposite convention is used where the "nice" filters are associated with decomposition and the distinguishing symbol ($\widetilde{\ }$) denotes reconstruction filters.

applied again on each column. The standard decomposition gives coefficients for a basis formed by the *standard construction* of wavelet basis functions, consisting of all possible tensor products of the one-dimensional basis functions,

$$\phi(x) \otimes \phi(x),\ \phi(x) \otimes \psi(x),\ \psi(x) \otimes \phi(x),\ \psi(x) \otimes \psi(x),$$

where $\phi(x) \otimes \phi(x)$ is the 2D scaling function and the rest are wavelets (see [Fou95, p.20] for details and examples).

The non-standard decomposition of a 2D function alternates between operations on rows and columns. That is, the decomposition is obtained by first applying the 1D wavelet transform to each row of (pixel) values *at one resolution level*, giving average (smoothed) values with detail coefficients for each row. The transformed rows are again treated as 1D functions and one level the 1D wavelet transform is applied again on each column. To complete the transform, the process is repeated recursively on the quadrant containing both row and column averages. The *non-standard construction* of a two-dimensional basis is similar to the standard construction, except that the tensor products are obtained using transposed versions of the 1D scaling and wavelet functions. That is, the two-dimensional scaling function is defined as

$$\phi\phi(x,y) = \phi(x) \otimes \phi^T(x),$$

and the three wavelet functions are:

$$\begin{aligned} \phi\psi(x,y) &= \phi(x) \otimes \psi^T(x), \\ \psi\phi(x,y) &= \psi(x) \otimes \phi^T(x), \\ \psi\psi(x,y) &= \psi(x) \otimes \psi^T(x). \end{aligned}$$

Both constructions will generate orthogonal 2D bases given orthogonal 1D functions [Fou95]. Examples of the non-standard decomposition and some of its properties are given in §5.7.

In three dimensions, the wavelet transform depends on three-dimensional scaling and wavelet bases functions. The standard decomposition of a typical 3D function, e.g., a video frame sequence, $f(x,y,t)$, is obtained by first applying the 1D wavelet transform on inter-frame pixels between two successive video frames at each resolution level. This gives the temporal decomposition of the video frames, analogous to the wavelet transform of one-dimensional signals. The first transformed frame contains the overall temporal average value, while the last frame contains the overall temporal difference of the original frames. The transformed frames are then treated as 2D functions and the standard 2D wavelet decomposition is applied to all frames.

The non-standard decomposition is obtained by first applying the 1D wavelet transform on each pixel between each of two successive video frames in the sequence. One of the two transformed frames contains the

temporal average values, while the other frame contains the temporal difference of the two original frames. The transformed frames are then treated as 2D functions and the non-standard wavelet decomposition is applied to both frames. Provided there were four frames to begin with, the process is repeated recursively on the two quadrants containing both temporal and spatial averages which are contained in the two temporal average frames. The non-standard construction of a three-dimensional basis is similar to the two-dimensional case except that the temporal basis is obtained first. That is, the three-dimensional scaling function is defined as:

$$\phi\phi\phi(x,y,t) = \phi(x) \otimes \phi(x) \otimes \phi^T(x),$$

and the seven wavelet functions are:

$$\begin{aligned}
\phi\phi\psi(x,y,t) &= \phi(x) \otimes \phi(x) \otimes \psi^T(x),\\
\phi\psi\phi(x,y,t) &= \phi(x) \otimes \psi(x) \otimes \phi^T(x),\\
\phi\psi\psi(x,y,t) &= \phi(x) \otimes \psi(x) \otimes \psi^T(x),\\
\psi\phi\phi(x,y,t) &= \psi(x) \otimes \phi(x) \otimes \psi^T(x),\\
\psi\phi\psi(x,y,t) &= \phi(x) \otimes \phi(x) \otimes \psi^T(x),\\
\psi\psi\phi(x,y,t) &= \phi(x) \otimes \psi(x) \otimes \phi^T(x),\\
\psi\psi\psi(x,y,t) &= \phi(x) \otimes \psi(x) \otimes \psi^T(x).
\end{aligned}$$

The constructions will generate orthogonal 3D bases given orthogonal 1D functions. Examples of the non-standard decomposition are given in §5.7.

5.6 Wavelet Filters

The multiresolution wavelet decomposition and reconstruction, depicted in Table 3, can be implemented by a two-band filter bank, as shown in Figure 15. To maintain consistency with signal processing convention, the discrete sequences $\{p_k\}$, $\{q_k\}$, $\{\widetilde{p}_k\}$, $\{\widetilde{q}_k\}$ are replaced by the digital filters $H, G, \widetilde{H}, \widetilde{G}$ represented by discrete sequences $\{h_k\}$, $\{g_k\}$, $\{\widetilde{h}_k\}$, $\{\widetilde{g}_k\}$, respectively. Figure 15 displays decomposition and reconstruction of the signal f at one resolution level. The symbols $\downarrow 2$ and $\uparrow 2$ within circles represent dyadic downsampling and upsampling, respectively.

Multiresolution analysis at multiple levels resembles a nonuniform, tree-structured filter bank. The nonuniform qualification refers to the flexible tiling of the space-frequency grid generated by wavelet analysis (see §5.1.5, Figure 14) [Vai93]. Multiresolution decomposition and reconstruction at three levels is shown in Figures 16. In general, the digital implementation of multiresolution analysis, as described in §5.4, is often

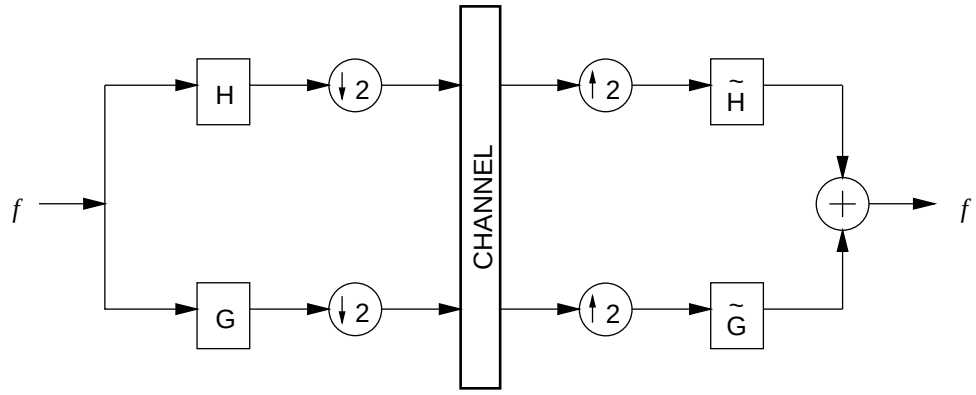


Fig. 15. One-level wavelet decomposition and reconstruction implemented by a two-band filter bank.

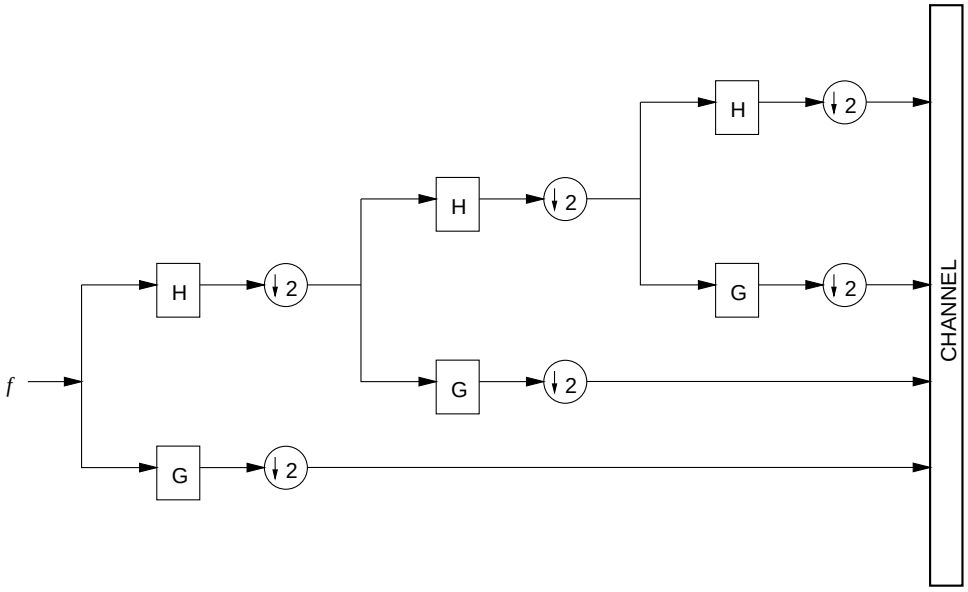
referred to as the Discrete Wavelet Transform (DWT).[9]

In practice, the filters H and G are chosen as lowpass and and highpass (in general, bandpass) filters, respectively. The lowpass filter corresponds to the scaling function ϕ by subsampling the signal at decreasing levels of resolution. The highpass (or bandpass) filter corresponds to the wavelet function ψ decomposing the signal by projections onto consecutive frequency bands. The dual filters $\widetilde{H}, \widetilde{G}$ are derived from H, G subject to desired orthogonality constraints between filters. These constraints are delineated by the four wavelet classes discussed in §5.2 resulting in the consonant families of filters, namely *orthogonal*, *bi-orthogonal*, *semi-orthogonal*, and *non-orthogonal*.

The Discrete Wavelet Transform can be represented in matrix form [PTVF92]. At a given scale j, the finite, discrete function f, represented by the sequence $\mathbf{c}^j$, is transformed into the sequences $\mathbf{c}^{j-1}$ and $\mathbf{d}^{j-1}$ by the square matrix $\mathbf{M}^j$ consisting of null (zero) elements, and elements of the scaling and wavelet filters $\{h_k\}$, $\{g_k\}$. The transformed sequences $\mathbf{c}^{j-1}$, $\mathbf{d}^{j-1}$ are each half the length of $\mathbf{c}^j$ due to downsampling. For example, using scaling and wavelet filters $\{h_k\}$ and $\{g_k\}$, each of length 4, the decomposition of the sequence $\mathbf{c}^j$ of

[9]Strictly speaking, the term *Wavelet Transform* generally refers to the *Integral Wavelet Transform*, relative to the basic wavelet ψ, defined in Equation (5.7), and *Discrete Wavelet Transform* refers to the wavelet series expansion of f, relative to ψ. The transform is *dyadic* when a and b are chosen such that the wavelet basis is obtained by a binary dilation and dyadic translation of a single function ψ. In the signal processing domain, and especially in image and video processing applications, the term *Wavelet Transform*, or *Discrete Wavelet Transform (DWT)*, has come to mean a multiresolution analysis of the underlying signal. Although not technically accurate, this terminology is adopted here meaning that *Discrete Wavelet Transform* and the abbreviation *DWT* should be interpreted as "discrete, dyadic multiresolution analysis". The term *Inverse Discrete Wavelet Transform (IDWT)* should be interpreted as "discrete, dyadic multiresolution synthesis".

(a) Decomposition



(b) Reconstruction

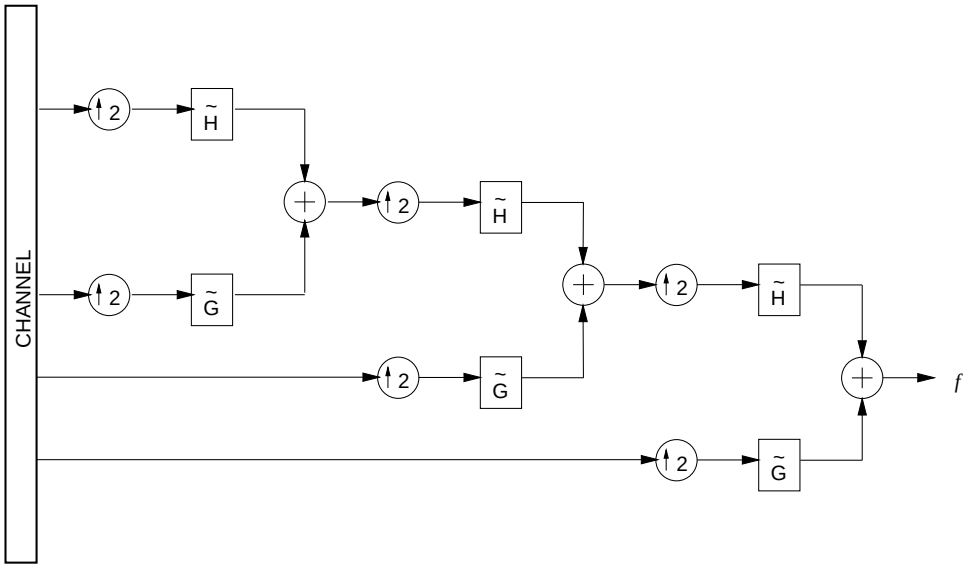


Fig. 16. Discrete Wavelet Transform implemented by a nonuniform, tree-structured, two-band filter bank.

length 8 is given by:

$$\begin{bmatrix} c_0^{j-1} \\ c_1^{j-1} \\ c_2^{j-1} \\ c_3^{j-1} \\ d_0^{j-1} \\ d_1^{j-1} \\ d_2^{j-1} \\ d_3^{j-1} \end{bmatrix} = \begin{bmatrix} h_0 & h_1 & h_2 & h_3 & & & & \\ & & h_0 & h_1 & h_2 & h_3 & & \\ & & & & h_0 & h_1 & h_2 & h_3 \\ h_2 & h_3 & & & & & h_0 & h_1 \\ g_0 & g_1 & g_2 & g_3 & & & & \\ & & g_0 & g_1 & g_2 & g_3 & & \\ & & & & g_0 & g_1 & g_2 & g_3 \\ g_2 & g_3 & & & & & g_0 & g_1 \end{bmatrix} \begin{bmatrix} c_0^j \\ c_1^j \\ c_2^j \\ c_3^j \\ c_4^j \\ c_5^j \\ c_6^j \\ c_7^j \end{bmatrix}$$

$$\mathbf{f}^{j-1} = \mathbf{M}^j \mathbf{f}^j \tag{5.48}$$

where $\mathbf{f}$ denotes the finite, discrete function f, and null elements of the matrix $\mathbf{M}^j$ are shown as empty spaces. The original function f can be perfectly reconstructed if the inverse matrix $(\mathbf{M}^j)^{-1}$ can be found and the dual filters $\{\widetilde{h}_k\}$, $\{\widetilde{g}_k\}$ exist. Construction of the dual filters depends on the chosen class of wavelets. Reconstruction is represented by a similar matrix operation where the reconstruction matrix resembles $\mathbf{M}^j$ except that the reconstruction filters $\{\widetilde{h}_k\}$ and $\{\widetilde{g}_k\}$ replace the decomposition filters $\{h_k\}$, $\{g_k\}$, e.g.,

$$\mathbf{f}^j = (\mathbf{M}^j)^{-1} \mathbf{f}^{j-1}.$$

Considering the filters $\{h_k\}, \{g_k\}$ as convolution kernels, notice that the above matrix operation incorporates the subsampling step by performing dyadic translations of the kernels. In some signal processing implementations, convolution is carried out through monadic translation of the convolution filter, relying on the subsampling step to drop every other element. In the above matrix representation, however, the subsampling step is made implicit by dyadic translation precluding the need for explicit subsampling and supersampling. In the discussion on filters, below, dyadic kernel translation is assumed.

5.6.1 Orthogonal Filters

The orthogonality condition for the wavelet ψ, initially given in §5.2.2 by Equation (5.19), is restated here with respect to the analysis filter $\{g_k\}$: if ψ is an orthogonal wavelet, then the filter G forms an intra-scale orthonormal basis of $L^2(\mathbf{R})$,

$$\langle g_k, g_m \rangle = \delta_{k,m}, \quad k, m \in \mathbf{Z}. \tag{5.49}$$

Provided the filter $\{g_k\}$ also satisfies inter-scale orthogonality, the subspaces of $L^2(\mathbf{R})$ generated by G are mutually orthogonal as in Equation (5.20). Condition (5.49) effectively states that the translated wavelet function ψ, and hence highpass filter $\{g_k\}$, does not overlap, or if it does, the overlapped segments sum to zero in the sense of the $L^2(\mathbf{R})$ inner product.

As outlined in §5.4, $\{g_k\}$ can be obtained from the orthonormal lowpass filter $\{h_k\}$, corresponding to the scaling function ϕ, as per Equation (5.36), in which case H and G are called *quadrature mirror filters* [Mal89a]. Equation (5.36) can be rewritten more compactly in terms of the filters $\{h_k\}, \{g_k\}$ by:

$$g_k \quad = \quad (-1)^k h_{1-k}, \quad k \in \mathbf{Z}, \tag{5.50}$$

so that the following intra-scale conditions hold:

$$\langle h_k, h_m \rangle = \langle g_k, g_m \rangle = \delta_{k,m}, \quad k, m \in \mathbf{Z} \quad \text{(orthonormal filters);} \tag{5.51}$$

$$\langle h_k, g_m \rangle = 0, \quad k, m \in \mathbf{Z} \qquad \text{(orthogonal subspaces } V_j \perp W_j\text{).} \tag{5.52}$$

Under this construction, the matrix $\mathbf{M}^j$ is orthogonal in the sense that the reconstruction matrix $(\mathbf{M}^j)^{-1}$ is the transpose of $\mathbf{M}^j$, i.e., $(\mathbf{M}^j)^{-1} = (\mathbf{M}^j)^T$, and the filters $\{h_k\}, \{g_k\}$ are self-dual, i.e., $\{\widetilde{h}_k\} = \{h_k\}$ and $\{\widetilde{g}_k\} = \{g_k\}$.

Referring to the above matrix decomposition example with filters $\{h_k\}$ and $\{g_k\}$ of length 4, the reconstruction $\mathbf{f}^j = (\mathbf{M}^j)^{-1}\mathbf{f}^{j-1}$ is given by:

$$\begin{bmatrix} c_0^j \\ c_1^j \\ c_2^j \\ c_3^j \\ c_4^j \\ c_5^j \\ c_6^j \\ c_7^j \end{bmatrix} = \begin{bmatrix} \widetilde{h}_0 & \widetilde{g}_0 & & & & & \widetilde{h}_2 & \widetilde{g}_2 \\ \widetilde{h}_1 & \widetilde{g}_1 & & & & & \widetilde{h}_3 & \widetilde{g}_3 \\ \widetilde{h}_2 & \widetilde{g}_2 & \widetilde{h}_0 & \widetilde{g}_0 & & & & \\ \widetilde{h}_3 & \widetilde{g}_3 & \widetilde{h}_1 & \widetilde{g}_1 & & & & \\ & & \widetilde{h}_2 & \widetilde{g}_2 & \widetilde{h}_0 & \widetilde{g}_0 & & \\ & & \widetilde{h}_3 & \widetilde{g}_3 & \widetilde{h}_1 & \widetilde{g}_1 & & \\ & & & & \widetilde{h}_2 & \widetilde{g}_2 & \widetilde{h}_0 & \widetilde{g}_0 \\ & & & & \widetilde{h}_3 & \widetilde{g}_3 & \widetilde{h}_1 & \widetilde{g}_1 \end{bmatrix} \begin{bmatrix} c_0^{j-1} \\ d_0^{j-1} \\ c_1^{j-1} \\ d_1^{j-1} \\ c_2^{j-1} \\ d_2^{j-1} \\ c_3^{j-1} \\ d_3^{j-1} \end{bmatrix} \tag{5.53}$$

Substituting $\{\widetilde{h}_k\}$ by $\{h_k\}$ and $\{\widetilde{g}_k\}$ by $\{g_k\}$, where $\{g_k\}$ is obtained as in (5.50), and permuting rows of $\mathbf{f}^{j-1}$ and columns of $(\mathbf{M}^j)^{-1}$, Equation (5.53) is rewritten as:

$$\begin{bmatrix} c_0^j \\ c_1^j \\ c_2^j \\ c_3^j \\ c_4^j \\ c_5^j \\ c_6^j \\ c_7^j \end{bmatrix} = \begin{bmatrix} h_0 & h_3 & & & & & h_2 & h_1 \\ h_1 & -h_2 & & & & & h_3 & -h_0 \\ h_2 & h_1 & h_0 & h_3 & & & & \\ h_3 & -h_0 & h_1 & -h_2 & & & & \\ & & h_2 & h_1 & h_0 & h_3 & & \\ & & h_3 & -h_0 & h_1 & -h_2 & & \\ & & & & h_2 & h_1 & h_0 & h_3 \\ & & & & h_3 & -h_0 & h_1 & -h_2 \end{bmatrix} \begin{bmatrix} c_0^{j-1} \\ d_0^{j-1} \\ c_1^{j-1} \\ d_1^{j-1} \\ c_2^{j-1} \\ d_2^{j-1} \\ c_3^{j-1} \\ d_3^{j-1} \end{bmatrix} \tag{5.54}$$

In this example, $(\mathbf{M}^j)^T$ is the inverse of $\mathbf{M}^j$ if and only if

$$h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1, \text{ and,} \tag{5.55}$$

$$h_0 h_2 + h_1 h_3 = 0. \tag{5.56}$$

Equations (5.55) and (5.56) in combination express the intra-scale orthonormality condition (5.51). If condition (5.56) is not evident from the invertible matrix requirement, consider the intra-scale orthogonality of

the subspace $\{V_j\}$, covered by the scaling function, which can be exemplified by two vectors formed by the spatial translation of the lowpass filter, h_k, h_{k+1}:

$$\begin{bmatrix} \dots & h_0 & h_1 & h_2 & h_3 & 0 & 0 & \dots \end{bmatrix}$$
$$\begin{bmatrix} \dots & 0 & 0 & h_0 & h_1 & h_2 & h_3 & \dots \end{bmatrix}$$

where the inner product $\langle h_k, h_{k+1} \rangle = h_0 h_2 + h_1 h_3$. These are precisely the terms required to sum to 0 in Equation (5.56). In general, for any different translations $k, m,\ k \neq m$, the inner product must sum to zero. In other words, $\langle h_k, h_m \rangle = 0,\ k \neq m$ so that the scaling function ϕ generates an orthogonal basis. Equations (5.55) and (5.56), along with two additional relations, were recognized and solved by Daubechies, while coefficients for filters of length 2 were first given by Haar. Coefficients of both filters are given in Table 4.

TABLE 4
Orthonormal filters.

(a) Haar.

k	$\sqrt{2}(h_k)$	$\sqrt{2}(g_k)$
0	1	1
1	1	-1

(b) Daubechies-4.

k	$4\sqrt{2}(h_k)$	$4\sqrt{2}(g_k)$
0	$1+\sqrt{3}$	$1-\sqrt{3}$
1	$3+\sqrt{3}$	$-3+\sqrt{3}$
2	$3-\sqrt{3}$	$3+\sqrt{3}$
3	$1-\sqrt{3}$	$-1-\sqrt{3}$

Orthogonal wavelets guarantee perfect reconstruction and generally facilitate implementation. In practice, however, orthogonal wavelets are not always easily constructed and may lack desirable properties such as symmetry or continuity. Filter symmetry is incompatible with exact reconstruction, if the same FIR filters are used for decomposition and reconstruction. Except for the Haar basis, *all* compactly supported, real orthonormal wavelet bases are asymmetric [Dau92, p.252,p.253,p.259]. The Haar wavelet is the only real-valued wavelet that is compactly supported, symmetric and orthogonal [JS94a]. Non-orthogonal, or overlapping filters, relax the orthogonality condition and subsequently are considered more flexible.

5.6.2 Semi-Orthogonal Filters

Recall that a function $\psi \in L^2(\mathbf{R})$ is semi-orthogonal if the generated basis $\{\psi_{j,k}\}$ is orthogonal, as expressed by (5.22). This condition suggests that the corresponding filters need not be fully orthonormal, only orthogonal, generating mutually orthogonal subspaces. In effect, the intra-scale orthonormality condition (5.51) is relaxed so that

$$\langle h_k, h_m \rangle = \langle g_k, g_m \rangle = 0, \quad k \neq m, \quad k, m \in \mathbf{Z} \quad \text{(orthogonal filters)};$$

while condition (5.52) remains:

$$\langle h_k, g_m \rangle = 0, \quad k, m \in \mathbf{Z} \quad \text{(orthogonal subspaces } V_j \perp W_j\text{)}.$$

Semi-orthogonal filters can produce orthogonal filters through an orthogonalization procedure, as mentioned in §5.2.3.

5.6.3 Non-Orthogonal Filters

Non-orthogonal filters are filters that are not semi-orthogonal. That is, non-orthogonal filters do not generate mutually orthogonal subspaces. Effectively, they are overlapping filters. In general, non-orthogonal filters require their duals to guarantee perfect reconstruction.

5.6.4 Bi-orthogonal Filters

Following §5.4.3, given a pair of lowpass and highpass filters $\{h_k\}, \{g_k\}$, neither necessarily being orthogonal, dual filters $\{\widetilde{h}_k\}, \{\widetilde{g}_k\}$ are required to guarantee perfect reconstruction. In particular, by (5.35), the following relations must hold:

$$\langle h_k, \widetilde{g}_m \rangle = 0, \quad k, m \in \mathbf{Z}; \tag{5.57}$$

$$\langle g_k, \widetilde{h}_m \rangle = 0, \quad k, m \in \mathbf{Z}. \tag{5.58}$$

Note that orthogonal filters satisfy these requirements through the stringent condition of orthonormality placed on $\{h_k\}$ and subsequently on $\{g_k\}$. Since $H = \widetilde{H}$, and H is orthonormal, i.e., $\langle h_k, h_m \rangle = \delta_{k,m}$, then $\langle \widetilde{h}_k, \widetilde{h}_m \rangle = \delta_{k,m}$ also holds. Moreover, since $\{g_k\}$ is the quadrature mirror of $\{h_k\}$, $\{g_k\}$ and $\{\widetilde{g}_k\}$ are also orthonormal. The construction of biorthogonal filters, on the other hand, is based on the relaxation of the orthonormality condition, so that in general, $H \neq \widetilde{H}$. The requirement of bi-orthogonal dual bases remains. That is, the intra-scale orthonormality condition, contained in (5.13), is rewritten in terms of the two sets of filters $H, \widetilde{H}, G, \widetilde{G}$ as:

$$\langle h_k, \widetilde{h}_m \rangle = \delta_{k,m}, \quad k, m \in \mathbf{Z}; \tag{5.59}$$

$$\langle g_k, \widetilde{g}_m \rangle = \delta_{k,m}, \quad k, m \in \mathbf{Z}. \tag{5.60}$$

In general, the relaxation of the orthonormality condition and the use of dual filters provides greater flexibility in the construction of filters. Specifically, symmetric filters can be constructed. The construction of bi-orthogonal wavelets is typically performed by specifying the decomposing (or reconstructing) pair of functions (ϕ, ψ), then deriving their duals such that the above bi-orthogonal conditions are satisfied. One method, as suggested in §5.4.3, is to derive $\{\widetilde{h}_k\}$ from $\{g_k\}$ and $\{\widetilde{g}_k\}$ from $\{h_k\}$ by the quadrature mirror construction (see [Dau92, §8.3]):

$$\widetilde{g}_k = (-1)^k h_{1-k}, \quad k \in \mathbf{Z},$$

$$\widetilde{h}_k = (-1)^k g_{1-k}, \quad k \in \mathbf{Z}.$$

Various authors have constructed symmetric, bi-orthogonal filters. Most constructions rely on filter bank theory [GB92] or on multiresolution derivations usually using the spline family of functions which provides continuity as well as symmetry. Ueda and Lodha provide an excellent introduction into B-spline wavelets including derivations of linear, quadratic, and cubic B-spline wavelet filters [UL95]. Well known bi-orthogonal filters have been designed by Cohen, Daubechies, and Feaveau [Bar94]. Chui has developed a family of spline wavelets based on cardinal B-spline functions [Chu92, §4]. Barlaud derived near-orthonormal dual spline wavelets constructed from the popular Laplacian pyramid filter introduced by Burt and Adelson [BA83b], which itself is a near-orthonormal wavelet filter [ABMD92]. The Laplacian filters are in turn very similar to the orthonormal *coiflet* basis developed by Coifman. Mallat et al. have developed quadratic spline wavelets which are particularly suitable for singularity detection [MZ92a]. Coefficients of the Mallat, Chui (multiplicity-2), Barlaud, and Burt and Adelson filters are given in §A.

Unfortunately, although one set of filters may possess many desirable properties, the dual filters are, in general, not compactly supported [JS94a]. This may cause significant implementational problems. For example, Chui's (multiplicity-2) decomposition filters are of length 41, while Mallat's filters require special normalization operations at various levels of reconstruction. Furthermore, if the filters are not separable then implementation of multi-dimensional wavelet transforms becomes even more problematic.

5.7 Discrete Wavelet Transform

The one-dimensional Discrete Wavelet Transform (DWT) is characterized by the decomposition and reconstruction Equations (5.46) and (5.47) described in §5.5. The implementation of the 1D DWT follows the general digital filter representation portrayed by Figures 15 and 16, and an example of the 1D decomposition through convolution was given by the matrix representation (5.48) in §5.6.

Given an n-length discrete function at the j^{th} level of resolution,

$$f^j(x) = f^j_\phi(1), f^j_\phi(2), \ldots, f^j_\phi(n), \tag{5.61}$$

the decomposition relations of the function are:

$$f^{j-1}_\phi(x) = \sum_k h_k f^j_\phi(2x+k), \tag{5.62}$$

$$f^{j-1}_\psi(x) = \sum_k g_k f^j_\phi(2x+k), \tag{5.63}$$

where $\{h_k\}, \{g_k\}$ are the one-dimensional low- and high-pass filters. This gives the discrete wavelet transform:

$$\{Wf(x)\}(j-1) = f^{j-1}_\phi(1), f^{j-1}_\psi(2), \ldots, f^{j-1}_\phi(n-1), f^{j-1}_\psi(n). \tag{5.64}$$

Permuting the terms so that the first $n/2$ elements are the low-pass (scale) coefficients, i.e.,

$$\{Wf(x)\}(j-1) = \\ f_\phi^{j-1}(1), f_\phi^{j-1}(3), \ldots, f_\phi^{j-1}(n-1), f_\psi^{j-1}(2), f_\psi^{j-1}(4), \ldots, f_\psi^{j-1}(n), \tag{5.65}$$

and relabeling the indices,

$$\{Wf(x)\}(j-1) = f_\phi^{j-1}(1), \ldots, f_\phi^{j-1}(n/2-1), f_\psi^{j-1}(n/2), \ldots, f_\psi^{j-1}(n) \tag{5.66}$$

the smooth (or averaged) elements (the first $n/2$ elements) are recursively decomposed. The fully transformed function contains the global average as the first element, the next 2^j elements contain the detail (or difference) information at each resolution level j. Except for the average value, the transformed elements comprise the so-called wavelet coefficients of the function.

To reconstruct the function, the terms at each resolution level are repermuted so that the average and wavelet coefficients are interleaved, as per Equation (5.64). Introducing the $\bowtie$ operator denoting element interleave, the $j-1$ level coefficients can be arranged into an intermediate representation for reconstruction at level j:

$$f_{\phi\bowtie\psi}^{j-1}(2x+p) \quad = \quad (1-p)f^{j-1}(x) + (p)f^{j-1}(x),$$

for $p \in \{0,1\}$. Reconstruction at level j, with $p \in \{0,1\}$ is then written as:

$$f_\phi^j(2x+p) = (1-p)\sum_k \widetilde{h}_k f_{\phi\bowtie\psi}^{j-1}(x-k) + (p)\sum_k \widetilde{g}_k f_{\phi\bowtie\psi}^{j-1}(x-k)$$

which gives the original function $f^j(x)$ in (5.61). Note that the variable p is used as a selection variable, that is, in the dyadic wavelet reconstruction, the element at position $2x+p$ is the result of filtering the lower level elements with either $\{\widetilde{h}_k\}$ or $\{\widetilde{g}_k\}$. This is a convenient substitute for writing two equations:

$$f_\phi^j(2x) = \sum_k \widetilde{h}_k f_{\phi\bowtie\psi}^{j-1}(x-k);$$
$$f_\phi^j(2x+1) = \sum_k \widetilde{g}_k f_{\phi\bowtie\psi}^{j-1}(x-k).$$

The permutation function $f_{\phi\bowtie\psi}^j$ serves as an alternate method of reconstruction used instead of traditional supersampling. In contrast, without permutating the average and wavelet coefficients, there would be two sequences f_ϕ^{j-1}, f_ψ^{j-1}, each of length $n/2$, where n is the length of the sequence at level j. In this case, reconstruction is given as:

$$f_\phi^j(2x-l) = \sum_k \widetilde{h}_{l-2k} f_\phi^{j-1}(x-k) + \sum_k \widetilde{g}_{l-2k} f_\psi^{j-1}(x-k), \tag{5.67}$$

which follows from the decomposition relation given in (5.45). The reconstruction in (5.67) is equivalent to the one given by (5.67), however its implementation is obviously different. In all following (multidimensional) wavelet transform discussions the permutation approach is adopted. This is consistent with the 1D

reconstruction example of (5.54) corresponding to the decomposition of (5.48) in §5.6. A numerical example of orthogonal decomposition and reconstruction using Haar filters is given in Table 5, where the symbol $\bowtie$ denotes element permutation.

TABLE 5
Numerical 1D DWT example.

Decomposition								
$\mathbf{f}^2 = \mathbf{c}^2$:	1		4		0		-2	
$\mathbf{d}^1$:		$\frac{-3}{\sqrt{2}}$				$\frac{2}{\sqrt{2}}$		
$\mathbf{c}^1$:		$\frac{5}{\sqrt{2}}$				$\frac{-2}{\sqrt{2}}$		
$\mathbf{d}^0$:				$\frac{7}{2}$				
$\mathbf{c}^0$:				$\frac{3}{2}$				
$\mathbf{W}f$:	$\frac{3}{2}$		$\frac{7}{2}$		$\frac{-3}{\sqrt{2}}$		$\frac{2}{\sqrt{2}}$	
Reconstruction								
$\mathbf{c}^0 \bowtie \mathbf{d}^0$:		$\frac{3}{2}$				$\frac{7}{2}$		
$\mathbf{c}^1$:				$\frac{5}{\sqrt{2}}, \frac{-2}{\sqrt{2}}$				
$\mathbf{c}^1 \bowtie \mathbf{d}^1$:	$\frac{5}{\sqrt{2}}$		$\frac{-3}{\sqrt{2}}$		$\frac{-2}{\sqrt{2}}$		$\frac{2}{\sqrt{2}}$	
$\mathbf{c}^2$:		1,4				0,-2		
$\mathbf{c}^2 = \mathbf{f}^2$:	1		4		0		-2	

Multidimensional extensions of the DWT rely on the use of multidimensional bases, described in §5.5, which are constructed by obtaining the tensor product of unidimensional bases. The matrix tensor product operation is reviewed in §B. The computational realization of the 2D DWT described here is an instance of the well-known pyramidal multiresolution representational framework first proposed by Tanimoto and Pavlidis (see [TK80, §2, pp.31-56] and [JR94, p.3]). The pyramidal model stipulates a hierarchical processing paradigm known as the *coarse-to-fine* (resolution) strategy. The pyramidal wavelet transform in particular is related to the Laplacian pyramid introduced by Burt and Adelson for image coding [BA83b]. In two dimensions, a $\log_2 N$ level pyramid is constructed from an $N \times N$ image, where the bottom level of the pyramid

(level $j = \log_2 N - 1$) contains the finest resolution, and the top level ($j = 0$) contains the coarsest. Note that the original image is considered the finest level of resolution (level $j = \log_2 N$), however it is not contained within the pyramid itself.

Given 1D scaling and wavelet filters H, G associated with ϕ, ψ, respectively, the 2D filters corresponding to the 2D wavelet bases $\phi\phi$, $\phi\psi$, $\psi\phi$, $\psi\psi$, as described in §5.5, are generated by the non-standard 2D wavelet basis construction using tensor products:

$$\begin{aligned} HH &= H \otimes H^T, \\ HG &= H \otimes G^T, \\ GH &= G \otimes H^T, \\ GG &= G \otimes G^T. \end{aligned}$$

As an example consider the Haar filters given in Table 4. Their two-dimensional extensions are derived below:

$$\begin{aligned} HH = \phi \otimes \phi^T &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} & \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}; \end{aligned}$$

$$\begin{aligned} HG = \phi \otimes \psi^T &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} & \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}; \end{aligned}$$

$$\begin{aligned} GH = \psi \otimes \phi^T &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} & -\frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}; \end{aligned}$$

$$\begin{aligned} GG = \psi \otimes \psi^T &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} & -\frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}. \end{aligned}$$

Multiplying each 2D filter by the dyadic normalization factor $2^{j/2}$ as suggested by Equations (5.8) and (5.10), the 2D filters become:

$$HH = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}; \; HG = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix};$$
$$GH = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}; \; GG = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

At the next level of resolution, the filters are derived by:

$$\begin{aligned} HHHH &= (\phi \otimes \phi^T) \otimes (\phi \otimes \phi^T); \\ HGHH &= (\phi \otimes \psi^T) \otimes (\phi \otimes \phi^T); \\ GHHH &= (\psi \otimes \phi^T) \otimes (\phi \otimes \phi^T); \\ GGHH &= (\psi \otimes \psi^T) \otimes (\phi \otimes \phi^T). \end{aligned}$$

Taking dyadic filter translation into consideration, the 2D filters are clearly mutually orthogonal in *x*- and *y*-directions, and in this example orthonormal. In general, multidimensional orthogonal filters are also *separable*, i.e., satisfying

$$h(k,m) = h(k)h(m),$$

due to their tensor product construction. Note that the above example illustrates one of the drawbacks of the DWT, namely, depending on the choice of bases, the DWT is neither necessarily translationally nor rotationally invariant.

The multidimensional tensor product filters are useful for visualizing spatiotemporal properties of the multidimensional wavelet transform, however direct implementation with multidimensional filters is inefficient. Instead, relying on the separability of the filters, the wavelet transform can be implemented by processing each dimension separately. The decomposition relations describing the non-standard decomposition of the spatial average image at level j are:

$$\begin{aligned} f_{\phi_r}^{j-1}(x,y) &= \textstyle\sum_k h_k f_{\phi\phi}^{j}(x,2y+k) & f_{\psi_r}^{j-1}(x,y) &= \textstyle\sum_k g_k f_{\phi\phi}^{j}(x,2y+k) \\ f_{\phi\phi}^{j-1}(x,y) &= \textstyle\sum_k h_k f_{\phi_r}^{j-1}(2x+k,y) & f_{\phi\psi}^{j-1}(x,y) &= \textstyle\sum_k h_k f_{\psi_r}^{j-1}(2x+k,y) \\ f_{\psi\phi}^{j-1}(x,y) &= \textstyle\sum_k g_k f_{\phi_r}^{j-1}(2x+k,y) & f_{\psi\psi}^{j-1}(x,y) &= \textstyle\sum_k g_k f_{\psi_r}^{j-1}(2x+k,y) \end{aligned}$$

where $\{h_k\}, \{g_k\}$ are the one-dimensional low- and high-pass filters. The non-standard DWT first involves subsampling the rows of the lower resolution level spatial average image (denoted by $f_{\phi\phi}^{j}$) to generate the temporary upper resolution level images $f_{\phi_r}^{j-1}$ and $f_{\psi_r}^{j-1}$. Due to dyadic downsampling of rows, these images

are half the width of the image at the lower resolution level. The columns of these subimages are then subsampled to generate the four subimages denoted by the subscripts $\phi\phi$, $\phi\psi$, $\psi\phi$, and $\psi\phi$. The above equations are written verbosely in order to facilitate implementation. Rewriting the equations concisely,

$$f_{\phi\phi}^{j-1}(x,y) = \sum_{k,m}(h_k \otimes h_m) f_{\phi\phi}^{j}(2x+k,2y+m) \tag{5.68}$$

$$f_{\psi\phi}^{j-1}(x,y) = \sum_{k,m}(g_k \otimes h_m) f_{\phi\phi}^{j}(2x+k,2y+m) \tag{5.69}$$

$$f_{\phi\psi}^{j-1}(x,y) = \sum_{k,m}(h_k \otimes g_m) f_{\phi\phi}^{j}(2x+k,2y+m) \tag{5.70}$$

$$f_{\psi\psi}^{j-1}(x,y) = \sum_{k,m}(g_k \otimes g_m) f_{\phi\phi}^{j}(2x+k,2y+m) \tag{5.71}$$

it is clear that the decomposition algorithm follows the two-scale relations (5.43) and (5.44). The smooth (or averaged) subimage $f_{\phi\phi}^{j}$ is recursively subsampled at each stage of the decomposition. The transformed image contains the global average at the top of the pyramid, the lower layers contain the detail (or difference) information at each pyramid level. These lower layers comprise the so-called wavelet coefficients of the transformed image. The decomposition is shown schematically in Figure 17 where the image matrix f is subsampled with low- and high-pass filters h,g. The subscripts r,c represent the subsampling operation

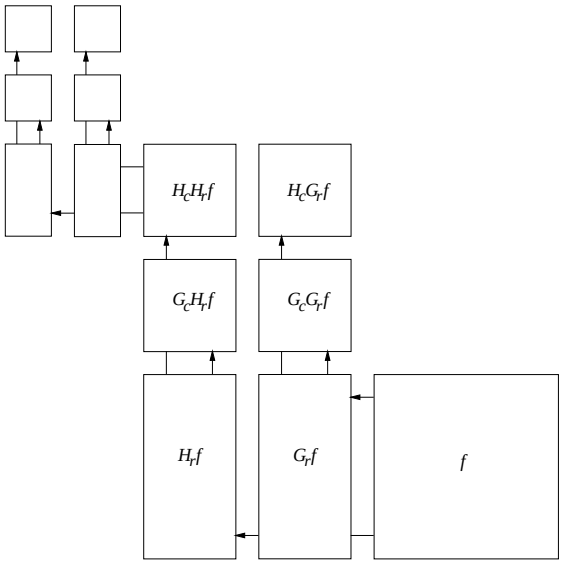


Fig. 17. Non-standard 2D pyramidal decomposition.

performed on rows and columns, i.e., $f_{\phi\phi}^{j-1} = H_c H_r f^j$.

In practice, depending on the length of filters $\{h_k\},\{g_k\}$, boundary conditions require special consideration. There are generally two strategies used to handle this problem: extending the image by padding with zero

values, or periodic extension of the image (i.e., tiling copies of the image). In the present implementation the latter strategy is implemented by applying modulo r,c to the indices at each resolution level. This generates a *wraparound* at the image borders when the filters extend beyond the image boundary. All references to image locations (x,y) are extended beyond image boundaries by the indices $((r+x) \bmod r, (c+y) \bmod c)$ where r,c are the dimensions of $f^j_{\phi\phi}$. This strategy allows the use of negative indices (required during reconstruction) and the processing of non-square images. An example of the 2D DWT applied to an image is shown in Figure 18. The transformed image has been processed for display purposes.

(a) Original *cnn* image. Reprinted with permission from Turner Broadcasting System, Inc. (see §F).

(b) 2-level DWT (processed by histogram equalization with subsampled image inset).

Fig. 18. Non-standard 2D DWT.

In traditional pyramidal approaches, where the pyramid contains only smoothed multiscale versions of the original image (e.g., texture-mapping applications), subimages at each level provide the pixel intensity values for reconstruction usually involving interpolation (cf. §5.10). In the wavelet transform, the image is synthesized by a recursive process of adding detail information to the average (smoothed) subimages in order to reconstruct the next level's average subimage. Rows and columns are interleaved prior to filtering instead of the traditional null row and column padding (see [PTVF92] for an example of the interleave operation, and [Cas96] for padding examples). Generally, row and column padding (supersampling) is used under monadic convolution. Dyadic convolution precludes the need for padding, but instead requires that rows and columns be interleaved prior to reconstruction filtering. Introducing the $\bowtie_r$ and $\bowtie_c$ operators denoting row and column interleave, respectively, the reconstruction is obtained with the use of the following intermediate relations:

$$f^j_{\phi\phi \bowtie_r \psi\phi}(2x+p,y) \quad = \quad (1-p)\, f^{j-1}_{\phi\phi}(x,y) + (p)\, f^{j-1}_{\psi\phi}(x,y), \tag{5.72}$$

$$f^j_{\phi\psi\bowtie_r\psi\psi}(2x+p,y) = (1-p)\,f^{j-1}_{\phi\psi}(x,y)+(p)\,f^{j-1}_{\psi\psi}(x,y), \tag{5.73}$$

$$f^j_{\phi_r\bowtie_c\psi_r}(x,2y+q) = (1-q)\,f^j_{\phi_r}(x,y)+(q)\,f^j_{\psi_r}(x,y), \tag{5.74}$$

where $p,q\in\{0,1\}$, $x,y,k\in\mathbf{Z}$ and

$$\begin{aligned} f^j_{\phi_r}(2x+p,y) &= (1-p)\sum_k \widetilde{h}_k f^j_{\phi\phi\bowtie_r\psi\phi}(2x-k,y)+(p)\sum_k \widetilde{g}_k f^j_{\phi\phi\bowtie_r\psi\phi}(2x-k,y),\\ f^j_{\psi_r}(2x+p,y) &= (1-p)\sum_k \widetilde{h}_k f^j_{\phi\psi\bowtie_r\psi\psi}(2x-k,y)+(p)\sum_k \widetilde{g}_k f^j_{\phi\psi\bowtie_r\psi\psi}(2x-k,y), \end{aligned}$$

so that

$$f^j_{\phi\phi}(x,2y+q) = (1-q)\sum_k \widetilde{h}_k f^j_{\phi_r\bowtie_c\psi_r}(x,2y-k)+(q)\sum_k \widetilde{g}_k f^j_{\phi_r\bowtie_c\psi_r}(x,2y-k).$$

The above reconstruction relations in two dimensions can be rewritten succinctly following the decomposition relation given in (5.45):

$$\begin{aligned} f^j_{\phi\phi}(2x+p,2y+q) = (1-q)\Bigg[&(1-p)\sum_{k,m}(\widetilde{h}_k\otimes\widetilde{h}_m)\,f^{j-1}_{\phi\phi}(x-k,y-m)+\\ &p\sum_{k,m}(\widetilde{h}_k\otimes\widetilde{g}_m)\,f^{j-1}_{\psi\phi}(x-k,y-m)\Bigg]+\\ (q)\Bigg[&(1-p)\sum_{k,m}(\widetilde{g}_k\otimes\widetilde{h}_m)\,f^{j-1}_{\phi\psi}(x-k,y-m)+\\ &p\sum_{k,m}(\widetilde{g}_k\otimes\widetilde{g}_m)\,f^{j-1}_{\psi\psi}(x-k,y-m)\Bigg], \end{aligned} \tag{5.75}$$

where $x,y,k,m\in\mathbf{Z}$, wraparound indices in the reverse direction, written as $(x-k)$, are assumed, and $p,q\in\{0,1\}$ are used as selection variables analogously as in the one-dimensional reconstruction.

To show that Equation (5.75) is derived from the above relations, expand $f^j_{\phi\phi}(x,2y+q)$:

$$\begin{aligned} f^j_{\phi\phi}(x,2y+q) &= (1-q)\sum_k \widetilde{h}_k f^j_{\phi_r\bowtie_c\psi_r}(x,2y-k)+(q)\sum_k \widetilde{g}_k f^j_{\phi_r\bowtie_c\psi_r}(x,2y-k)\\ &= (1-q)\sum_k \widetilde{h}_k\Big[(1-q)\,f^j_{\phi_r}(x,y-k)+(q)\,f^j_{\psi_r}(x,y-k)\Big]+\\ &\qquad (q)\sum_k \widetilde{g}_k\Big[(1-q)\,f^j_{\phi_r}(x,y-k)+(q)\,f^j_{\psi_r}(x,y-k)\Big]. \end{aligned} \tag{5.76}$$

Since the (q), $(1-q)$ terms are symbolic for binary selection, multiple like terms can be combined into one, i.e., $(1-q)^k=(1-q)(1-q)\cdots(1-q)=(1-q)$, and $(q)^k=(q)(q)\cdots(q)=(q)$ for all $q\in\{0,1\},k\in\mathbf{Z}$. Conversely, unlike terms cancel, simplifying Equation (5.76) to:

$$f^j_{\phi\phi}(x,2y+q) = (1-q)\sum_k \widetilde{h}_k f^j_{\phi_r}(x,y-k)+(q)\sum_k \widetilde{g}_k f^j_{\psi_r}(x,y-k).$$

Substituting appropriately for $f^j_{\phi_r}$, $f^j_{\psi_r}$ by changing the resolution of x to $2x+p$ and taking care to disambiguate summation indices,

$$\begin{aligned} f^j_{\phi\phi}(2x+p,2y+q) &= (1-q)\sum_m \widetilde{h}_m \Big[(1-p)\sum_k \widetilde{h}_k f^j_{\phi\phi\bowtie_r\psi\phi}(2x-k,y-m)+ \\ &\qquad (p)\sum_k \widetilde{g}_k f^j_{\phi\phi\bowtie_r\psi\phi}(2x-k,y-m)\Big]+ \\ &\quad (q)\sum_m \widetilde{g}_m \Big[(1-p)\sum_k \widetilde{h}_k f^j_{\phi\psi\bowtie_r\psi\psi}(2x-k,y-m)+ \\ &\qquad (p)\sum_k \widetilde{g}_k f^j_{\phi\psi\bowtie_r\psi\psi}(2x-k,y-m)\Big]. \end{aligned} \tag{5.77}$$

Noting that the filter summation terms apply to rows and columns separately, i.e., $\sum_m \widetilde{h}_m \sum_k \widetilde{g}_k$ refers to the tensor product of $\widetilde{H}$ and $\widetilde{G}$ since $\sum_k \widetilde{g}_k$ applies to the columns of f^j, the double summation terms can be collected and represented by the single summation term $\sum_{k,m}(\widetilde{h}_k \otimes \widetilde{g}_m)$. Equation (5.77) is then rewritten as:

$$\begin{aligned} f^j_{\phi\phi}(2x+p,2y+q) &= (1-q)\Big[(1-p)\sum_{k,m}(\widetilde{h}_k \otimes \widetilde{h_m})\, f^j_{\phi\phi\bowtie_r\psi\phi}(2x-k,y-m)+ \\ &\qquad (p)\sum_{k,m}(\widetilde{h}_k \otimes \widetilde{g}_m)\, f^j_{\phi\phi\bowtie_r\psi\phi}(2x-k,y-m)\Big]+ \\ &\quad (q)\Big[(1-p)\sum_{k,m}(\widetilde{g}_k \otimes \widetilde{h}_m)\, f^j_{\phi\psi\bowtie_r\psi\psi}(2x-k,y-l)+ \\ &\qquad (p)\sum_{k,m}(\widetilde{g}_k \otimes \widetilde{g}_m)\, f^j_{\phi\psi\bowtie_r\psi\psi}(2x-k,y-m)\Big]. \end{aligned} \tag{5.78}$$

Substituting Equations (5.72) and (5.73) into (5.78) gives:

$$\begin{aligned} f^j_{\phi\phi}(2x+p,2y+q) &= (1-q)\Big[(1-p)\sum_{k,m}(\widetilde{h}_k \otimes \widetilde{h}_m)\Big\{(1-p)\, f^{j-1}_{\phi\phi}(x-k,y-m)+ \\ &\qquad (p)\, f^{j-1}_{\psi\phi}(x-k,y-m)\Big\}+ \\ &\qquad (p)\sum_{k,m}(\widetilde{h}_k \otimes \widetilde{g}_m)\Big\{(1-p)\, f^{j-1}_{\phi\phi}(x-k,y-m)+ \\ &\qquad (p)\, f^{j-1}_{\psi\phi}(x-k,y-m)\Big\}\Big]+ \\ &\quad (q)\Big[(1-p)\sum_{k,m}(\widetilde{g}_k \otimes \widetilde{h}_m)\Big\{(1-p)\, f^{j-1}_{\phi\psi}(x-k,y-m)+ \\ &\qquad (p)\, f^{j-1}_{\psi\psi}(x-k,y-m)\Big\}+ \\ &\qquad (p)\sum_{k,m}(\widetilde{g}_k \otimes \widetilde{g}_m)\Big\{(1-p)\, f^{j-1}_{\phi\psi}(x-k,y-m)+ \\ &\qquad (p)\, f^{j-1}_{\psi\psi}(x-k,y-m)\Big\}\Big]. \end{aligned} \tag{5.79}$$

Using similar arguments for like $(1-p)$, (p) terms in Equation (5.79), the reconstruction algorithm simplifies to:

$$\begin{aligned} f^j_{\phi\phi}(2x+p,2y+q) \quad = \quad & (1-q)\left[(1-p)\sum_{k,m}(\widetilde{h}_k\otimes\widetilde{h}_m)\,f^{j-1}_{\phi\phi}(x-k,y-m)+\right. \\ & \left.(p)\sum_{k,m}(\widetilde{h}_k\otimes\widetilde{g}_m)\,f^{j-1}_{\psi\phi}(x-k,y-m)\right]+ \\ & (q)\left[(1-p)\sum_{k,m}(\widetilde{g}_k\otimes\widetilde{h}_m)\,f^{j-1}_{\phi\psi}(x-k,y-m)+\right. \\ & \left.(p)\sum_{k,m}(\widetilde{g}_k\otimes\widetilde{g}_m)\,f^{j-1}_{\psi\psi}(x-k,y-m)\right], \end{aligned} \tag{5.80}$$

which is equivalent to (5.75). Equations (5.75) and (5.80) roughly state that $f^j_{\phi\phi}$ is reconstructed from the expansion of the f^{j-1} functions at resolution level $j-1$ by the dual (bi-orthogonal) filters $\{\widetilde{h}_k\},\{\widetilde{g}_k\}$ representing scale and wavelet bases generated by $\widetilde{\phi}$, $\widetilde{\psi}$. Since $f^j_{\phi\phi}$ was originally projected by the 2D bases functions $\phi\phi$, $\phi\psi$, $\psi\phi$, and $\psi\psi$, represented by the tensor products of $\{h_k\}$ and $\{g_k\}$, producing the four lower resolution level functions $f^{j-1}_{\phi\phi}$, $f^{j-1}_{\phi\psi}$, $f^{j-1}_{\psi\phi}$, and $f^{j-1}_{\psi\psi}$, $f^j_{\phi\phi}$ is faithfully reconstructed provided $(\widetilde{\phi},\widetilde{\psi})$, (ϕ,ψ) satisfy (bi-)orthogonality conditions as specified in §5.2 and §5.4.3. Equation (5.75) also makes intuitive sense when Equations (5.68)–(5.71) are considered in matrix form:

$$\begin{aligned} \mathbf{f}^{j-1}_{\phi\phi} &= (\mathbf{H}\otimes\mathbf{H})\mathbf{f}^j_{\phi\phi}, \quad \mathbf{f}^{j-1}_{\psi\phi} = (\mathbf{G}\otimes\mathbf{H})\mathbf{f}^j_{\phi\phi}, \\ \mathbf{f}^{j-1}_{\phi\psi} &= (\mathbf{H}\otimes\mathbf{G})\mathbf{f}^j_{\phi\phi}, \quad \mathbf{f}^{j-1}_{\psi\psi} = (\mathbf{G}\otimes\mathbf{G})\mathbf{f}^j_{\phi\phi}. \end{aligned}$$

To reconstruct $\mathbf{f}^j$, both sides of each equation should be left-multiplied by some tensor product combination of $\widetilde{\mathbf{H}}$ and $\widetilde{\mathbf{G}}$. If the decomposition matrices are orthogonal, then each reconstruction matrix is the transpose of the decomposition matrix. That is,

$$\begin{aligned} \mathbf{f}^{j-1}_{\phi\phi} &= (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{H}})^T\mathbf{f}^j_{\phi\phi}, \quad \mathbf{f}^{j-1}_{\phi\phi} = (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{H}})^T\mathbf{f}^j_{\psi\phi}, \\ \mathbf{f}^{j-1}_{\phi\phi} &= (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{G}})^T\mathbf{f}^j_{\phi\psi}, \quad \mathbf{f}^{j-1}_{\phi\phi} = (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{G}})^T\mathbf{f}^j_{\psi\psi}. \end{aligned}$$

In the orthogonal case, $\mathbf{H}=\widetilde{\mathbf{H}}$ and $\mathbf{G}=\widetilde{\mathbf{G}}$, and

$$\begin{aligned} (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{H}})^T &= (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{H}}), \quad (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{H}})^T = (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{G}}), \\ (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{G}})^T &= (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{H}}), \quad (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{G}})^T = (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{G}}), \end{aligned}$$

giving

$$\begin{aligned} \mathbf{f}^j_{\phi\phi} &= (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{H}})\mathbf{f}^{j-1}_{\phi\phi}, \quad \mathbf{f}^j_{\phi\phi} = (\widetilde{\mathbf{H}}\otimes\widetilde{\mathbf{G}})\mathbf{f}^{j-1}_{\psi\phi}, \\ \mathbf{f}^j_{\phi\phi} &= (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{H}})\mathbf{f}^{j-1}_{\phi\psi}, \quad \mathbf{f}^j_{\phi\phi} = (\widetilde{\mathbf{G}}\otimes\widetilde{\mathbf{G}})\mathbf{f}^{j-1}_{\psi\psi}, \end{aligned}$$

which coincides with the order of the tensor product terms in (5.75). Note that each tensor product transpose is not the tensor product's inverse. That is, the individual matrix product components above do not each

guarantee perfect reconstruction. However, the appropriate row- and column-permutation of the combination of the tensor product components will create the appropriate perfect (orthogonal) reconstruction matrix. To illustrate, consider the matrix $(\mathbf{M}^j)^{-1}$ defined by:

$$(\mathbf{M}^j)^{-1} = ((\widetilde{\mathbf{H}} \otimes \widetilde{\mathbf{H}}) \bowtie_r (\widetilde{\mathbf{G}} \otimes \widetilde{\mathbf{H}})) \bowtie_c ((\widetilde{\mathbf{H}} \otimes \widetilde{\mathbf{G}}) \bowtie_r (\widetilde{\mathbf{G}} \otimes \widetilde{\mathbf{G}})).$$

Provided the appropriate (bi-)orthogonality conditions hold, matrix $(\mathbf{M}^j)^{-1}$ is symmetric-orthogonal, being the transpose and inverse of the decomposition matrix $\mathbf{M}^j$ (cf.§5.6). In this case, the reason for the tensor product terms in the reconstruction Equation (5.75) being the transpose of the tensor products in the decomposition Equation (5.68) can be seen as a constraint on the orthogonality of $\mathbf{M}^j$. This is clearly seen in the case of the Haar filters where $\mathbf{M}^j$ is given below:

$$\mathbf{M}^j = \left[\begin{array}{c|c} HH & HG \\ \hline GH & GG \end{array}\right] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= ((\widetilde{\mathbf{H}} \otimes \widetilde{\mathbf{H}}) \bowtie_r (\widetilde{\mathbf{G}} \otimes \widetilde{\mathbf{H}})) \bowtie_c ((\widetilde{\mathbf{H}} \otimes \widetilde{\mathbf{G}}) \bowtie_r (\widetilde{\mathbf{G}} \otimes \widetilde{\mathbf{G}})) = (\mathbf{M}^j)^{-1}.$$

The matrix $\mathbf{M}^j$ is its own inverse save for the resolution scale factor 2^j. Reconstruction of the image f, built by applying reconstruction filters $\widetilde{h}, \widetilde{g}$, is shown schematically in Figure 19. Row- and column-interleave

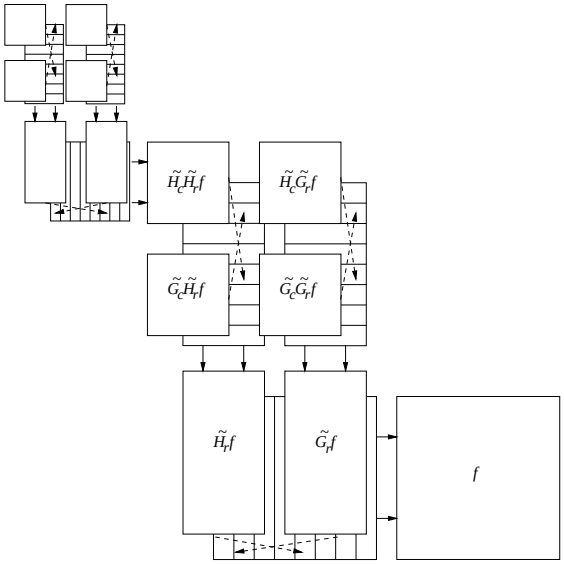


Fig. 19. Non-standard 2D pyramidal reconstruction.

operations are denoted by dashed arrows.

The 3D filters corresponding to the 3D wavelet bases described in §5.5, are generated by the non-standard 3D wavelet basis construction using tensor products:

$$\begin{aligned} HHH &= H \otimes H \otimes H^T, \\ HHG &= H \otimes H \otimes G^T, \\ HGH &= H \otimes G \otimes H^T, \\ HGG &= H \otimes G \otimes G^T, \\ GHH &= G \otimes H \otimes H^T, \\ GHG &= G \otimes H \otimes G^T, \\ GGH &= G \otimes G \otimes H^T, \\ GGG &= G \otimes G \otimes G^T. \end{aligned}$$

For example, the three-dimensional Haar filter GHG is a 2×4 filter derived below:

$$\begin{aligned} GHG = \psi \otimes \phi \otimes \psi^T &= \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \otimes \left(\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \right) \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} & \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} & -\frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix}. \end{aligned}$$

Multiplying the filter by the dyadic normalization factor $2^{j/2}$ in each dimension, that is by $2^{2j/2}$ or 2 at one resolution level, the filter becomes:

$$GHG = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Considering the application of the filter on two consecutive video frames, the filter can be represented by the two 2×2 spatial templates:

$$GHG = \begin{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} & -\frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \end{bmatrix}$$

$$\rightsquigarrow \quad \left[\begin{array}{cc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{array}\right]_{G_t}, \left[\begin{array}{cc} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array}\right]_{G_{t+1}},$$

where the normalization factor is implicit and the subscripts G_t, G_{t+1} denote the application of the appropriate temporal element of G. To help visualize the temporal filter application, Figure 20 shows the correspondence between filter elements. This representation is introduced only for convenience since it facilitates the representation of the three-dimensional gradient components. The remaining seven filters are derived in a similar manner.

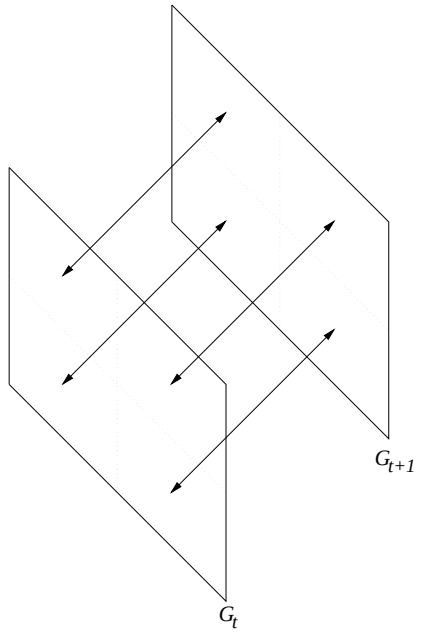


Fig. 20. Visualization of temporal filter element application.

The decomposition relations describing the non-standard decomposition of a 3D digital signal represented by a sequence of 2D images are:

$$f_{\phi}^{j-1}(x,y,t) = \sum_k h_k f_{\phi\phi\phi}^{j}(x,y,2t+k) \quad f_{\psi}^{j-1}(x,y,t) = \sum_k g_k f_{\phi\phi\phi}^{j}(x,y,2t+k)$$
$$f_{\phi_r\phi}^{j-1}(x,y,t) = \sum_k h_k f_{\phi}^{j-1}(x,2y+k,t) \quad f_{\phi_r\psi}^{j-1}(x,y,t) = \sum_k h_k f_{\psi}^{j-1}(x,2y+k,t)$$
$$f_{\psi_r\phi}^{j-1}(x,y,t) = \sum_k g_k f_{\phi}^{j-1}(x,2y+k,t) \quad f_{\psi_r\psi}^{j-1}(x,y,t) = \sum_k g_k f_{\psi}^{j-1}(x,2y+k,t)$$
$$f_{\phi\phi\phi}^{j-1}(x,y,t) = \sum_k h_k f_{\phi_r\phi}^{j-1}(2x+k,y,t) \quad f_{\phi\phi\psi}^{j-1}(x,y,t) = \sum_k h_k f_{\phi_r\psi}^{j-1}(2x+k,y,t)$$
$$f_{\psi\phi\phi}^{j-1}(x,y,t) = \sum_k g_k f_{\phi_r\phi}^{j-1}(2x+k,y,t) \quad f_{\psi\phi\psi}^{j-1}(x,y,t) = \sum_k g_k f_{\phi_r\psi}^{j-1}(2x+k,y,t)$$

$$f^{j-1}_{\phi\psi\phi}(x,y,t) = \sum_k h_k f^{j-1}_{\psi_r\phi}(2x+k,y,t) \quad f^{j-1}_{\phi\psi\psi}(x,y,t) = \sum_k h_k f^{j-1}_{\psi_r\psi}(2x+k,y,t)$$
$$f^{j-1}_{\psi\psi\phi}(x,y,t) = \sum_k g_k f^{j-1}_{\psi_r\phi}(2x+k,y,t) \quad f^{j-1}_{\psi\psi\psi}(x,y,t) = \sum_k g_k f^{j-1}_{\psi_r\psi}(2x+k,y,t)$$

where $\{h_k\}, \{g_k\}$ are the one-dimensional low- and high-pass filters. The non-standard DWT first temporally subsamples the sequence images of the lower (finer) resolution level's spatial average sequence (denoted by $f^j_{\phi\phi\phi}$) to generate the temporary upper resolution level images f^{j-1}_ϕ and f^{j-1}_ψ. These images will be of the same dimension as the images at the lower resolution level, but half of them will contain averaged temporal values, the other half will contain temporal difference values. Each temporal average and difference image is then decomposed at one level by the 2D DWT. The lower resolution level sequence is transformed into $n/2$ temporal average frames $f^{j-1}_{\cdot\cdot\phi}$, and $n/2$ temporal difference frames $f^{j-1}_{\cdot\cdot\psi}$. These frames are then permuted as in the 1D decomposition using the organization given in (5.65) and (5.66):

$$\{Wf(x)\}(j-1) = f^{j-1}_{\cdot\cdot\phi}(1), \ldots, f^{j-1}_{\cdot\cdot\psi}(n/2), \ldots, f^{j-1}_{\cdot\cdot\psi}(n) \tag{5.81}$$

The spatiotemporal smooth (or averaged) subimages $f^j_{\phi\phi\phi}$ are recursively processed at each stage of the decomposition. The transformed image sequence contains the global spatiotemporal average at the top of the (volume) pyramid. The decomposition is shown schematically in Figure 21, where the resultant wavelet transform volume is represented by a segmented cube resembling a nonuniform octtree data structure. The subscript t represents the temporal subsampling operation performed on sequence images, i.e., $f^{j-1}_\phi = H_t * f^j$. A pictorial example is shown in Figure 22.

The video sequence is synthesized by a recursive process of adding detail information to the average (smoothed) subimages in order to reconstruct the next level's temporal average subimages. At each level of reconstruction, frames must be interleaved in the temporal dimension prior to the spatial 2D wavelet reconstruction:

$$f^{j-1}_{\cdot\cdot\phi\bowtie\cdot\cdot\psi}(x,y,2t+p) \quad = \quad (1-p)\, f^{j-1}(x,y,t) + (p)\, f^{j-1}(x,y,t).$$

After permutation, each frame is spatially reconstructed by the 2D DWT, following (5.75), giving the two temporally subsampled frames $f^{j-1}_\phi(x,y,t)$ and $f^{j-1}_\psi(x,y,t)$. Each such pair of frames is then reconstructed as in the 1D DWT case:

$$f^j_{\phi\phi\phi}(x,y,2t+p) = (1-p)\sum_k \widetilde{h}_k f^{j-1}_{\phi\bowtie\psi}(x,y,2t-k) + (p)\sum_k \widetilde{g}_k f^{j-1}_{\phi\bowtie\psi}(x,y,2t-k).$$

The above description of the DWT assumes that decomposition in each dimension is possible to the same extent, i.e., the above multidimensional transform is *isotropic* with respect to spatiotemporal dimensions. The maximum level of isotropic signal decomposition coincides with the dimension of *lowest* magnitude. To illustrate, consider a video sequence of 4 frames, each frame of size 640×480. The sequence can only be decomposed isotropically to 2 levels since there is an insufficient number of sequence frames to decompose the signal any further in the temporal dimension. Conversely, given 1024 frames, a temporal decomposition

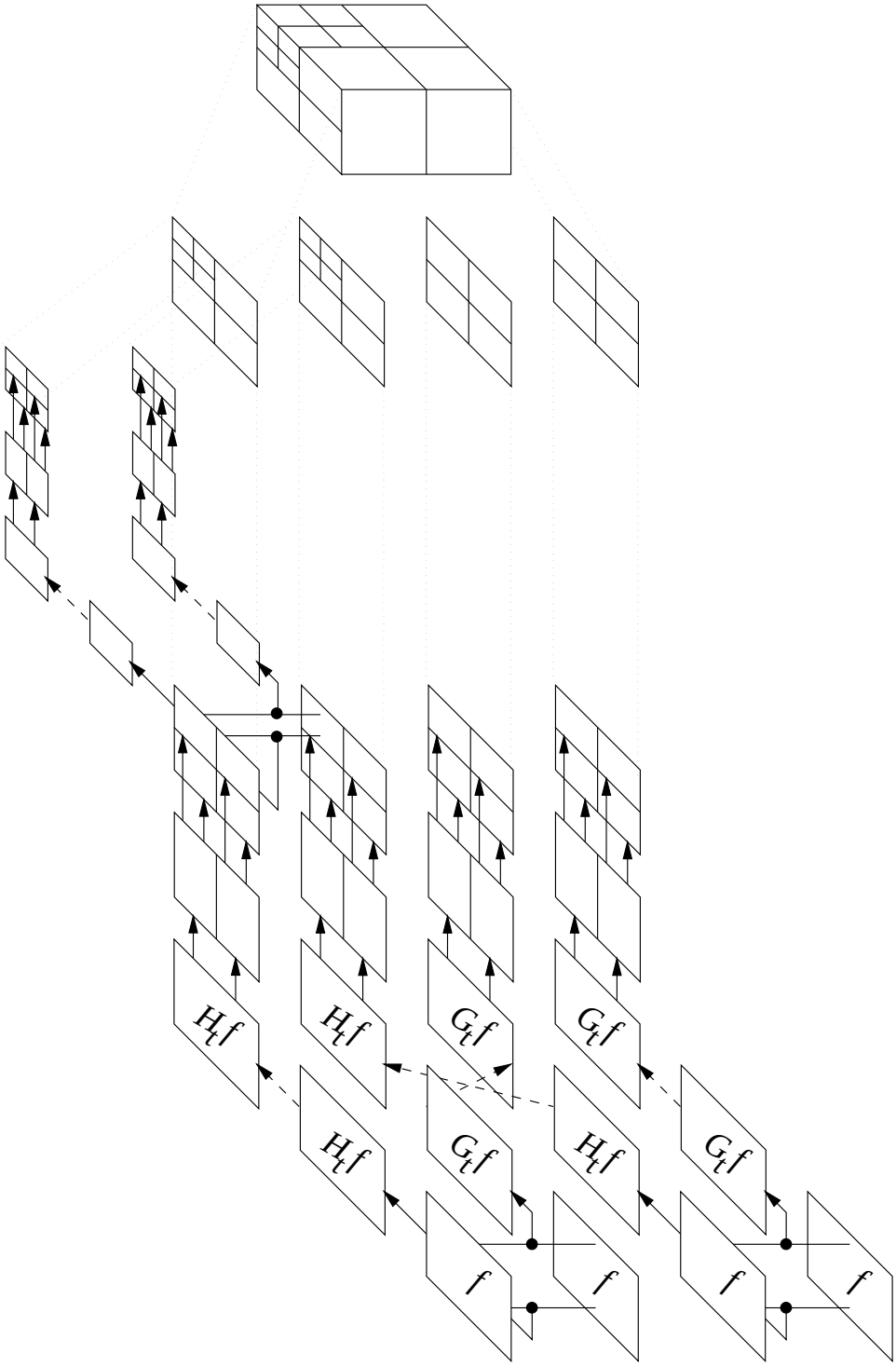
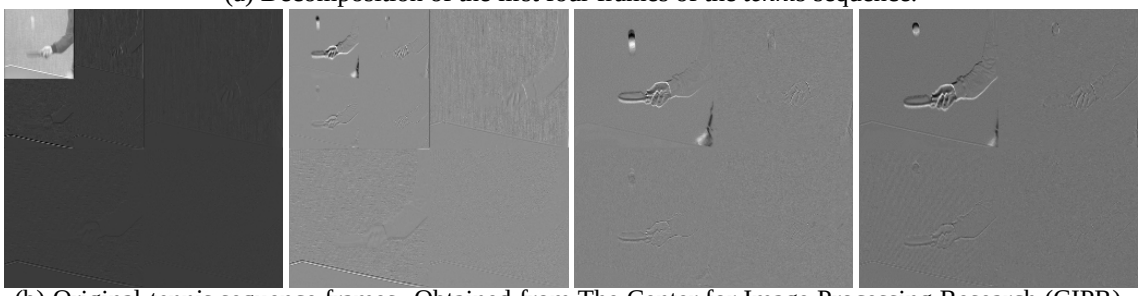


Fig. 21. Schematic non-standard 3D pyramidal wavelet decomposition.

(a) Decomposition of the first four frames of the *tennis* sequence.



(b) Original *tennis* sequence frames. Obtained from The Center for Image Processing Research (CIPR), an Internet public domain archive (ftp://ipl.rpi.edu/pub/image/sequence/tennis/).

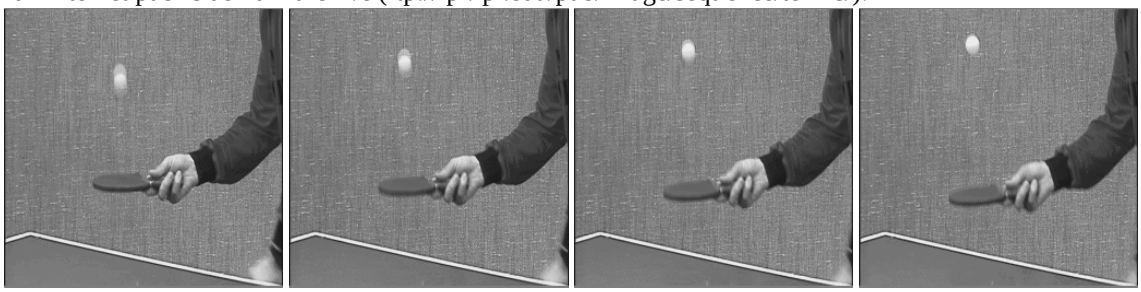


Fig. 22. Non-standard 3D pyramidal discrete wavelet decomposition.

of 10 levels is not possible since there is insufficient information in the x-dimension beyond the 8^{th} level of decomposition (the x-dimension is reduced to 1 pixel). Alternatively, it is possible to decompose the 4-frame sequence *anisotropically* to 2 levels in the temporal dimension, then decompose each frame spatially to 8 levels but in this case the decomposition is incomplete in terms of partial derivative information (see §5.9). The complete spatiotemporal decomposition governed by the dimension of least magnitude gives at least a full first-order representation of the signal. That is, depending on the choice of wavelet, all three partial derivatives $\partial/\partial x$, $\partial/\partial y$, $\partial/\partial t$ are available. This enables localization of multiscale, three-dimensional edges in video (see §5.8). This is a significantly powerful method of video analysis since it extends inspection of the temporal domain over a longer temporal duration than just two frames. Two-frame motion detection has been extensively studied in the context of motion-compensated video encoding, but temporal analysis over many frames has not been widely utilized. Furthermore, the complete 3D DWT offers second-order information, i.e., the partial derivatives $\partial/\partial x\partial y$, $\partial/\partial x\partial t$, $\partial/\partial y\partial t$, and $\partial/\partial x\partial y\partial t$ are all present in the transformed signal, although it is not clear at this point how this information can be used to enhance video analysis applications relying on first-order derivative information. For example, three-dimensional edges can be located by examining either the set of first-order partials, or the set of second-order partial derivatives (detection of zero-crossings), in which case the set not used appears redundant.

5.8 Multidimensional Multiscale Edge Detection

Edges, or sharp variation points in general, can be located in signals through the use of Mallat's multiscale edge detection algorithm. If ψ is chosen such that it approximates the first derivative of a smoothing function θ, as described in §5.3, then the 2D filters HG, GH constitute the gradient components $\partial\theta(x,y)/\partial x$, $\partial\theta(x,y)/\partial y$, respectively. This is due to the tensor product construction of the filters. The elements of HG are two rotated (transposed) copies of G multiplied by scalar elements of H. Similarly, the elements of GH are copies of G scaled by elements of H. Considering the 1D filter G as the 1D vector $[\partial\theta/\partial x]$, the transpose of G is the derivative of θ in the y-direction, i.e., $[\partial\theta/\partial y] = [\partial\theta/\partial x]^T$. The scalar multiplier cancels if the original filters H, G are orthonormal, and the filters HG, GH are left with components of the gradient:

$$HG = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial x} \end{bmatrix}; \; GH = \begin{bmatrix} \frac{\partial f}{\partial y} \\ \\ \frac{\partial f}{\partial y} \end{bmatrix}.$$

Considering the Haar wavelet, which is a first derivative operator, the second resolution level filters are:

$$HGHH = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}; \; GHHH = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Save for scale factors, these resemble (directionally, at least) the well known Sobel or Prewitt gradient operators used in edge detection (for comparison, see the discussions on gradient operators in [GW87, pp.336-338], [Sch89, pp.146-150], and [Jai89, pp.348-349]). The locations of the gradient components in the 2D-DWT of an image frame are schematically shown Figure 23(a). Due to the finite sampling of a 2D image $f(x,y)$, the angle defined by (5.26) can be quantized to octants specifying 8 possible neighbors, relative to the point at scale j and location (x_0, y_0), namely $\{Mf(x_0+\Delta x, y_0+\Delta y)\}(j)$, where $(\Delta x, \Delta y)$ are arranged as:

(-1,-1)	(0,-1)	(1,-1)
(-1, 0)	(0, 0)	(1, 0)
(-1, 1)	(0, 1)	(1, 1)

defining 4 planar directions corresponding to the compass directions N-S, NE-SW, E-W, and SE-NW. In the two-dimensional case, Equation (5.26) directly specifies one of the 4 directions used to identify pixel neighbor for determination of the modulus maxima.

Extending Mallat's algorithm to three dimensions, gradient components $\partial\theta(x,y,t)/\partial x$, $\partial\theta(x,y,t)/\partial y$, and $\partial\theta(x,y,t)/\partial t$, are represented by the 3D filters HHG, HGH and GHH, respectively:

$$HHG = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{G_t}, \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{G_{t+1}},$$

$$HGH = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{G_t}, \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{G_{t+1}},$$

$$GHH = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}_{G_t}, \begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{G_{t+1}}.$$

The conceptual volumetric representation of the 3D-DWT is shown in Figure 23(b). The locations of the gradient components in the two resolution level 3D-DWT of four frames are schematically shown in Figure 24. Quantizing the angle defined by Equations (5.28)–(5.30), 26 possible neighbors are specified relative to the point at scale j and location (x_0, y_0, t_0), namely $\{Mf(x_0+\Delta x, y_0+\Delta y, t_0+\Delta t)\}(j)$, where $(\Delta x, \Delta y, \Delta t)$ are arranged as:

(-1,-1,-1)	(0,-1,-1)	(1,-1,-1)
(-1, 0,-1)	(0, 0,-1)	(1, 0,-1)
(-1, 1,-1)	(0, 1,-1)	(1, 1,-1)

(-1,-1, 0)	(0,-1, 0)	(1,-1, 0)
(-1, 0, 0)	(0, 0, 0)	(1, 0, 0)
(-1, 1, 0)	(0, 1, 0)	(1, 1, 0)

(-1,-1, 1)	(0,-1, 1)	(1,-1, 1)
(-1, 0, 1)	(0, 0, 1)	(1, 0, 1)
(-1, 1, 1)	(0, 1, 1)	(1, 1, 1)

along 13 cubic (voxel) directions, depicted in Figure 25.

In the three-dimensional case, Equations (5.28)–(5.30) do not readily specify the 13 (voxel) neighbor directions. Instead, the values of the first-order partial derivatives must be examined to first determine the relevant angular plane. This is accomplished by inspecting the three first-order partials for zero (or near-zero) values. Since there are only three values, there are $2^3 = 8$ possibilities for zero-value combinations. Labeling non-zero values as 1 for ease of notation, Table 6 lists the possible directions given the 8 possible gradient value combinations. Referring to each case associated by its binary value, the 0^{th} case identifies a uniform region. This is a common property shared by gradient operators [Jai89, p.349]. Case 1 trivially specifies direction 1. In the 2^{nd}, 4^{th}, and 6^{th} cases, Equation (5.26) can be used directly to determine the relevant direction in the xy-plane. In case 3, only $\partial/\partial x$ is zero which suggests the gradient is in the yt-plane. Equation (5.30) can be used to determine whether the gradient falls along direction 4 or 8. Case 5 is similar to case 3 in that only $\partial/\partial y$ is zero so that Equation (5.29) can be used to determine gradient direction 2 or 6, both in the xt-plane. Case 7 is the most complicated since the gradient direction has components in both xt- and yt-planes. One of the four directions is found by projecting the gradient onto each plane in turn.

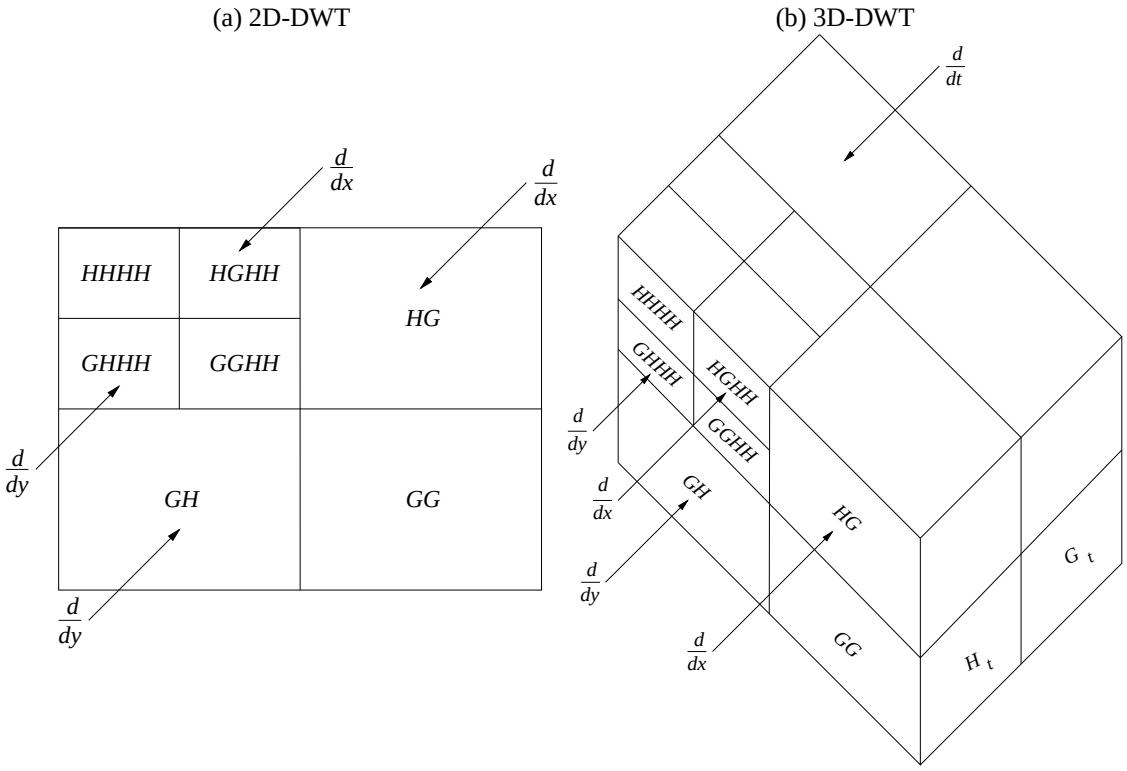


Fig. 23. 2D- and 3D-DWT multiresolution quadrants and octants, with gradient components.

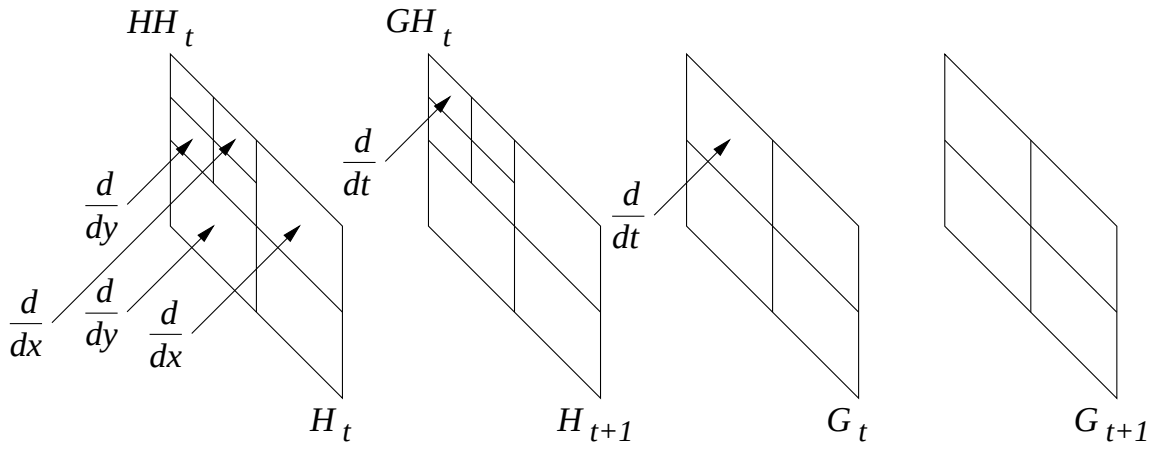


Fig. 24. Schematic 3D-DWT with gradient components.

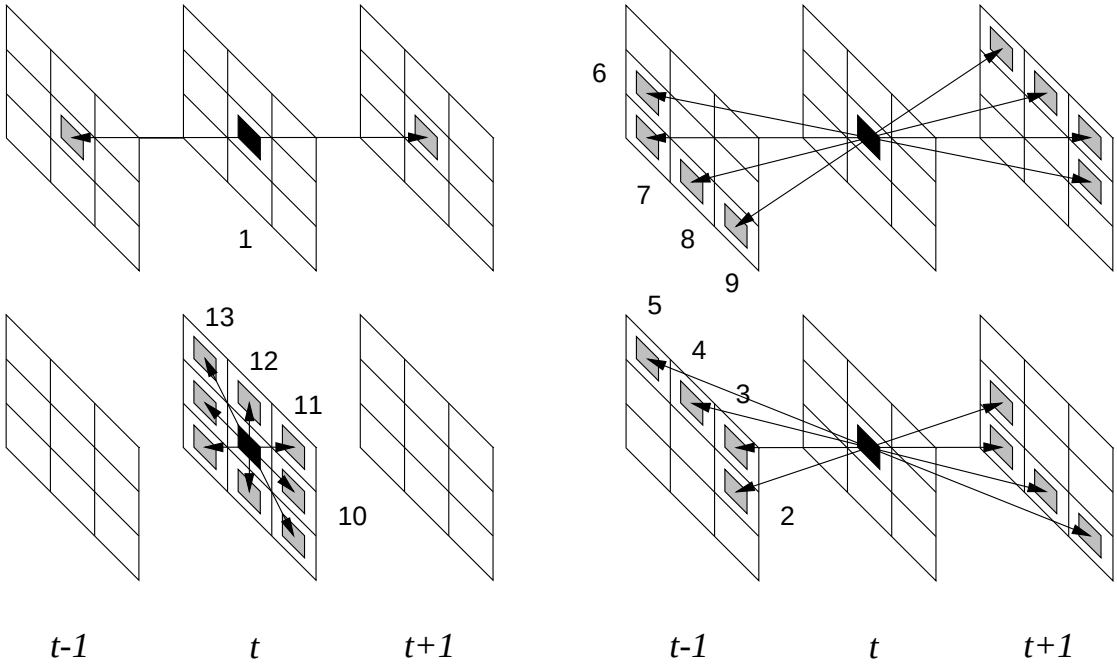


Fig. 25. Modula maxima planar (pixel) and cubic (voxel) neighbors.

TABLE 6
Three-dimensional direction identification based on first-order partial derivatives.

$\partial/\partial x$	$\partial/\partial y$	$\partial/\partial t$	Direction
0	0	0	$\{\}$
0	0	1	$\{1\}$
0	1	0	$\{12\}$
0	1	1	$\{4,8\}$
1	0	0	$\{10\}$
1	0	1	$\{2,6\}$
1	1	0	$\{10,11,12,13\}$
1	1	1	$\{3,5,7,9\}$

Given the modulus values as calculated by Equations (5.25) and (5.27) in 2D and 3D, respectively, along with the directional neighbor locations defined above, modulus maxima are located according to Equations (5.23) and (5.24) in both two- and three-dimensions. If the modulus satisfies both of these equations, then a record at the specific location is stored by the value $\{mf\}(j)$ set to the value of the modulus if the modulus is a local maxima, or 0. That is, in two dimensions, defining $\max\{Mf(x_0,y_0)\}(j)$ as:

$$\{Mf(x_0+\Delta_l x, y_0+\Delta_l y)\} \le \{Mf(x_0,y_0)\} \ge \{Mf(x_0+\Delta_r x, y_0+\Delta_r y)\}, \text{ and}$$

$$\begin{cases} \{Mf(x_0,y_0)\} > \{Mf(x_0+\Delta_l x, y_0+\Delta_l y)\}, & \text{or} \\ \{Mf(x_0,y_0)\} > \{Mf(x_0+\Delta_r x, y_0+\Delta_r y)\}. \end{cases}$$

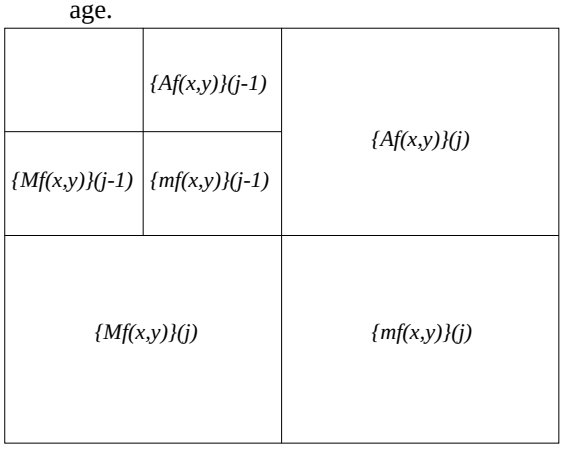
with the resolution level index made implicit, then $\{mf(x,y)\}(j)$ is defined as:

$$\{mf(x,y)\}(j) = \begin{cases} \{Mf(x,y)\}(j) & \text{if } \max\{Mf(x_0,y_0)\}(j) \\ 0 & \text{otherwise,} \end{cases}$$

where the left and right neighbors $\{Mf(x+\Delta_l x, y+\Delta_l y)\}(j)$, $\{Mf(x+\Delta_r x, y+\Delta_r y)\}(j)$, are identified by one of the four directions as given above. In three dimensions, $\max\{Mf(x_0,y_0,t_0)\}(j)$ is defined similarly with the left and right neighbors $\{Mf(x+\Delta_l x, y+\Delta_l y, t+\Delta_l t)\}(j)$, $\{Mf(x+\Delta_r x, y+\Delta_r y, t+\Delta_r t)\}(j)$, identified by one of the above specified thirteen directions.

During modulus maxima detection, a 2D data structure is created to hold values $\{Mf(x,y)\}(j)$, $\{Af(x,y)\}(j)\}$, and $\{mf(x,y)\}(j)$, as shown schematically in Figure 26(a). An example of modula maxima detection in two dimensions is shown in Figures 26(b), where the image has been globally normalized and gamma-corrected ($\gamma = 3$) to facilitate display. In the 2D case, the unmodified 2D wavelet transform is needed for perfect

(a) 2D-DWT modula maxima information storage.



(b) Modula maxima in 2-level decomposition of *cnn* image.

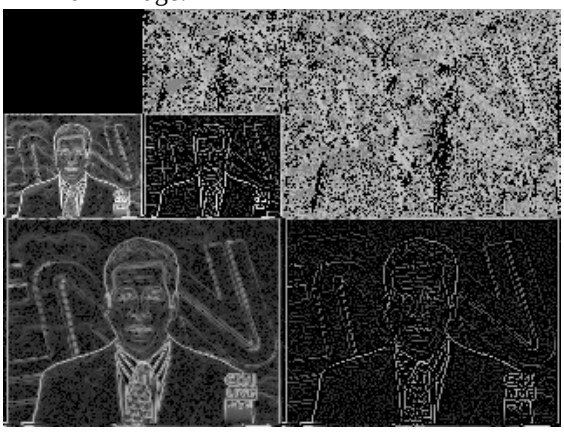
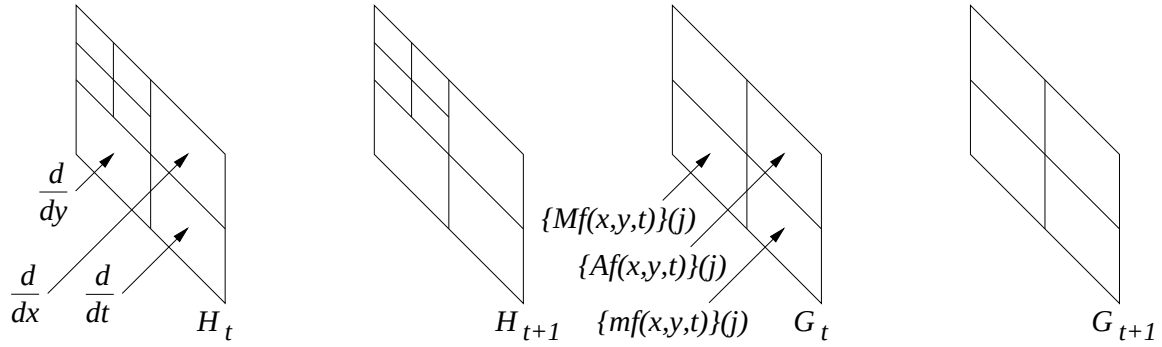


Fig. 26. 2D modula maxima detection.

reconstruction of the image. The image cannot be readily reconstructed from the maxima modulus information, although Mallat has developed an iterative algorithm that almost achieves perfect reconstruction in most cases [MZ92a]. (Meyer gives a counterexample to Mallat's conjecture in [Mey93, §8].) Typically a second image matrix array is allocated for this purpose.

In three dimensions, the 3D data structure is rearranged to hold values $\{Mf(x,y)\}(j)$, $\{Af(x,y)\}(j)\}$, and $\{mf(x,y)\}(j)$, as well as $\partial/\partial x$, $\partial/\partial y$, $\partial/\partial t$. This organization is shown schematically in Figure 27(a). An example of temporal modulus maxima detection (in a three-dimensional video sequence) is shown in Figure 27(b), where individual frames have been globally normalized and gamma-corrected ($\gamma = 3$) to facilitate display. Analogous to the 2D case, unmodified 3D wavelet transform information is needed for perfect re-

(a) Schematic 3D-DWT modula maxima information storage.



(b) Modula maxima in 2-level decomposition of the first four frames of the *tennis* sequence.

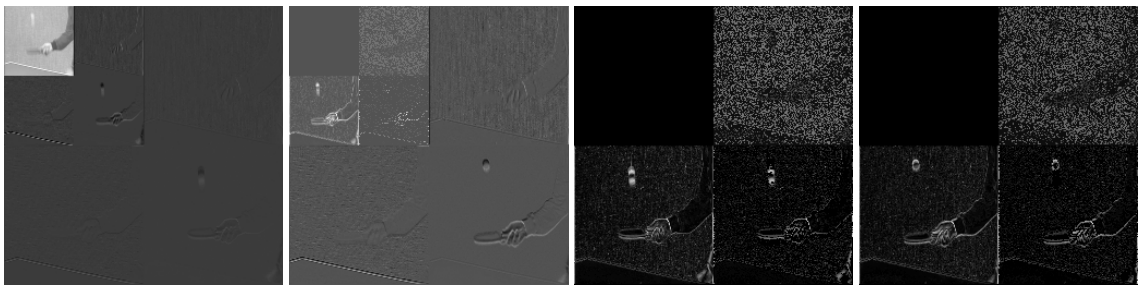


Fig. 27. 3D modula maxima detection.

construction. Best results are obtained if wavelet transform coefficients can be stored in memory with double precision. It is desirable to store both modulus maxima and wavelet coefficient information in memory. In the case of a 2D static image, a copy of the image, for the purposes of modulus maxima information storage, can generally be made due to sufficient memory availability. Unlike the two-dimensional image, duplicat-

ing an entire sequence in memory may be problematic due to memory capacity limitations. Fortunately, second-order information is not needed for modula maxima detection.[10] This allows the data quadrants to be rearranged such that the values needed for modula maxima detection can be accumulated in place of the second-order information. Note that this organization prevents straightforward reconstruction of the video sequence, but provides storage of both modulus maxima and first-order derivative information.

5.9 Anisotropic Multidimensional Discrete Wavelet Transform

The Discrete Wavelet Transform as described in §5.7 performed equal extent decomposition in each dimension. That is, the preceding discussion centered on an *isotropic* wavelet decomposition of a multidimensional signal. In general, signals may be transformed maximally in each dimension if each dimension is considered individually. In this sense, the transformation is *anisotropic* since the decomposition varies along each dimension. In this section, anisotropic decomposition of video sequences is discussed where decomposition in the spatial dimension is dissociated from the temporal dimension.

In dealing with video, the sequence may be considered as a one-dimensional temporal signal composed of two-dimensional elements. Accordingly, the sequence may first be fully decomposed in the temporal dimension on a per-pixel basis following Equations (5.62) and (5.63), i.e.,

$$\begin{aligned} f_\phi^{j-1}(x,y,t) &= \sum_k h_k f_\phi^j(x,y,2t+k), \\ f_\psi^{j-1}(x,y,t) &= \sum_k g_k f_\phi^j(x,y,2t+k), \end{aligned}$$

giving the DWT in the temporal direction:

$$\begin{aligned} &\{Wf(x,y,t)\}_t(j-1) = \\ &\quad f_{\phi_t}^{j-1}(x,y,1), f_{\psi_t}^{j-1}(x,y,2), \ldots, f_{\phi_t}^{j-1}(x,y,n-1), f_{\psi_t}^{j-1}(x,y,n). \end{aligned} \tag{5.82}$$

As in the one-dimensional case, the sequence frames are permuted so that the first $n/2$ frames, containing low-pass coefficients, are decomposed recursively. The fully transformed video sequence contains the global

[10] As an alternative to first-order derivative edge detection, directional information of edges can be obtained by searching zero-crossings of the second-order derivative along r for each direction θ, since

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r} = f_x \cos\theta + f_y \sin\theta,$$

where r is the gradient direction vector, and

$$\frac{\partial^2 f}{\partial r^2} = \frac{\partial f_x}{\partial r}\cos\theta + \frac{\partial f_y}{\partial r}\sin\theta = \frac{\partial^2 f}{\partial x^2}\cos^2\theta + 2\frac{\partial^2 f}{\partial x \partial y}\sin\theta\cos\theta + \frac{\partial^2 f}{\partial y^2}\sin^2\theta.$$

See [Jai89, p.348,p.353].

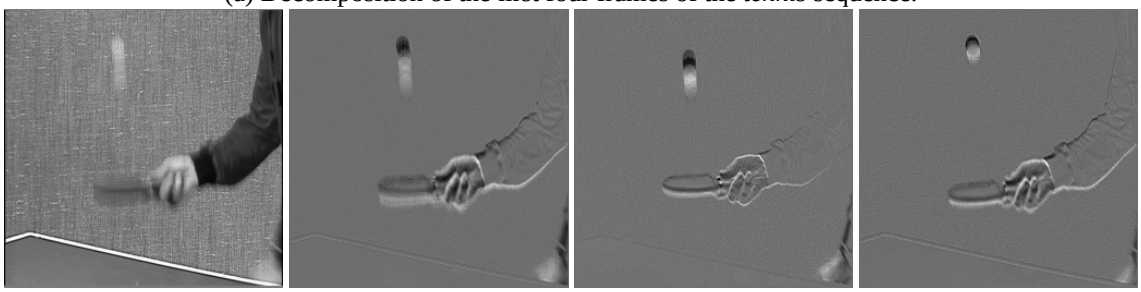
average in the first frame, with the next 2^j frames containing detail information at each resolution level j. Note that the first frame will contain an image displaying average motion information over the entire sequence. The frames containing difference information will contain motion differences over a specific interval of frames thereby indicating presence of motion over specific frequencies relative to the sampling rate of the video sequence. For example, consider a video sequence recorded at a sampling rate of 18ms, i.e., 55.5Hz, or 55.5 frames per second (fps).[11] The (dyadic) DWT of such a video sequence will contain temporal information of 55.5Hz at 1 level of decomposition (resolution level $j = 1$), 18.5Hz at level $j = 2$, 7.9Hz at level $j = 3$, etc. In general, at the j^{th} decomposition level, the dyadic DWT will contain temporal information at frequency $1/(2^j - 1)s_r$ where s_r is the sampling rate. In other words, $(2^j - 1)s_r$ gives the temporal difference information at level j with units matching the sampling rate (e.g., milliseconds). Considering the sequence captured at the 18ms sampling rate, then at level $j = 1$, the DWT will contain information over 18ms, at level $j = 2$ over 54ms, at level $j = 3$ over 126ms, etc. This is due to the fact that each temporal (dyadic) decomposition obtains difference information over $2^j - 1$ frame intervals, where each interval corresponds to the sampling rate. Such temporal information is extremely valuable in applications such as motion detection.

Having applied the 1D temporal DWT over a sequence of frames, the next natural task is motion detection. This involves detection of motion within sequence frames containing wavelet coefficients at specific (dyadic) frequencies of interest. Using the 55.5Hz video sequence as an example, if fast motion over 18Hz is sought, then sequence frames containing temporal differences at resolution level $j = 2$ must be inspected. Slower motion artifacts under 18.5Hz will be found at higher resolution levels. The problem of locating these artifacts is essentially a two-dimensional problem. In this respect, the 2D DWT (including 2D edge detection) is well suited for analysis. Temporal difference frames should be treated as two-dimensional image frames and the 2D DWT is applied exactly as in §5.7. Two-dimensional edge detection (as described in §5.8) is again applicable on a per-frame basis. The result of such an analysis locates features in video at specific frequencies. Four frames of the *tennis* sequence are shown under 2 levels of temporal decomposition in Figure 28(a). Note that the frames have been interleaved as described in §5.7 so that the first frame contains the global temporal average. Figure 28(b) shows two-dimensional edge detection carried out on each frame of the transformed sequence (compare with Figure 27 in §5.8).

Reconstruction of the sequence is carried out in the reverse order where each frame is processed by the 2D IDWT. The entire sequence is then treated as a one-dimensional signal and the 1D IDWT is applied on a per-pixel basis taking care to properly interleave whole image frames as required (see Equation (5.82)). Using

[11] The standard (NTSC) video rate is 30 fps, 30Hz, or in other words, video sampled at a rate of 33.3ms. This should not be confused with the NTSC *field* rate of 60Hz where a field corresponds to only half a given video *frame*, e.g., even or odd lines of a frame.

(a) Decomposition of the first four frames of the *tennis* sequence.



(b) 2D edge (modulus maxima) detection.

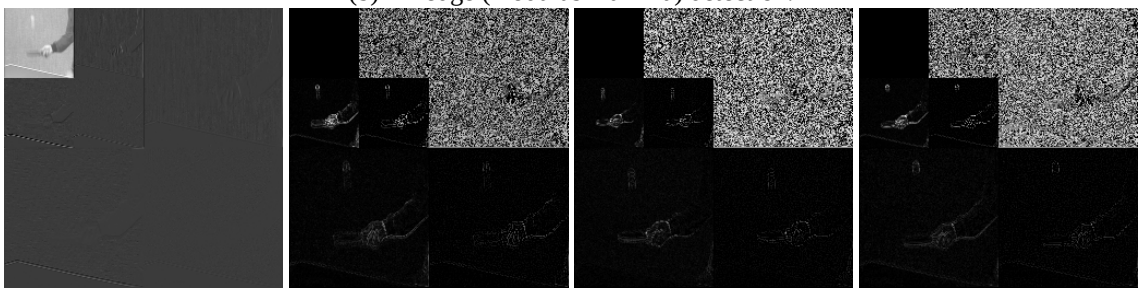


Fig. 28. Anisotropic non-standard 3D pyramidal discrete wavelet decomposition and 2D edge detection.

the interleave operator $\bowtie$, image frames are arranged for reconstruction at level j by:

$$f^{j-1}_{\phi\bowtie\psi}(x,y,2t+p) \quad = \quad (1-p)\,f^{j-1}(x,y,t) + (p)\,f^{j-1}(x,y,t),$$

for $p \in \{0,1\}$. Reconstruction is then written as:

$$f^{j}_{\phi}(x,y,2t+p) = (1-p)\sum_k \widetilde{h}_k f^{j-1}_{\phi\bowtie\psi}(x,y,t-k) + (p)\sum_k \widetilde{g}_k f^{j-1}_{\phi\bowtie\psi}(x,y,t-k),$$

giving the original function $f^{j}(x,y,t)$.

5.10 Wavelet Interpolation

The Discrete Wavelet Transform can be used to texture map images by selectively scaling wavelet coefficients. Provided appropriate wavelet filters can be found, reconstruction exactly matches linear *MIP-mapping*. MIP-mapping is a well known texture mapping algorithm used extensively in computer graphics [Wil83, WW92, §4.7].[12] MIP-mapping involves preprocessing an image at several resolution levels (decomposition) in order to texture map (reconstruct) an image at variable resolution. In this sense, MIP-mapping falls under the classical pyramid framework for early vision [JR94]. In the present context the relevant feature of MIP-mapping is the multiresolution reconstruction of the image. Specifically, gaze-contingent visual representation of digital imagery, discussed in §IX, relies on wavelet coefficient scaling developed here. In this

[12]The acronym MIP, introduced by Williams, is from the Latin phrase *multum in parvo* meaning "many things in a small place".

section the MIP-mapping approach is briefly described, then a set of filters, termed *MIP-wavelets*, is derived to match the box filter frequently used in MIP-mapping. Using these filters, it is shown that linear interpolation of scaled wavelet coefficients is equivalent to linear interpolation under MIP-mapping. Although the present discussion is limited to two-dimensional images, the wavelet coefficient scaling method is applicable to multidimensional signals for multiresolution representation.

5.10.1 MIP-Mapping

Given an $N \times N$ image, assuming N is a power of 2 with $n = \log_2 N$, the original image $f^n(x,y)$ is subsampled and smoothed into $n+1$ subimages (or *maps*),

$$f^j(\left\lfloor \frac{x}{M} \right\rfloor, \left\lfloor \frac{y}{M} \right\rfloor) = \frac{1}{M^2} \sum_{k=0}^{M-1} \sum_{m=0}^{M-1} f^n(x+k, y+m), \quad 0 \le j \le n, \tag{5.83}$$

where M is a smoothing filter of size 2^{n-j}, and j is the resolution level. Equation (5.83) generates projections of the original image onto $n+1$ scaled subspaces equivalent to the subspaces generated by the scaling function of the DWT. The subspaces in this instance are scaled analogously to the DWT with resolution level $j = 0$ corresponding to the coarsest resolution level.[13] Equation (5.83) is a slightly different representation from the classical recursive pyramidal approach since each subimage is subsampled directly from the original image f^n, not from the image at the next finer resolution level f^{j+1}. In general, the recursive form of Equation (5.83) is given by:

$$f^{j-1}(x,y) = \sum_{k=0}^{M} \sum_{m=0}^{M} h(k,m) f^j(2x+k, 2y+m), \quad 0 < j < n, \tag{5.84}$$

where $M+1$ is the (constant) width of the convolution kernel. If $h(k,m) = 1/\sqrt{2}$, $\forall k,m$ with $M = 1$, then $\{h(k,m)\}$ is the Haar smoothing filter. Equation (5.84) corresponds to the two-scale relation of the scaling function ϕ, given by (5.43), and is equivalent to the two-dimensional lowpass decomposition relation (5.68) where the scaling filter is the tensor product of the one-dimensional lowpass filter $\{h_k\}$, i.e., $h(k,m) = h_k \otimes h_m$. In general, the smoothing filter should satisfy the following constraints [JR94]:

1. normalization

$$\sum_k \sum_m h(k,m) = 1;$$

2. symmetry

$$h(k,m) = h(M+1-k,m) = h(k,M+1-m) = h(M+1-k,M+1-m) \quad \forall k,m;$$

[13] To scale subspaces in the opposite direction (e.g., the Daubechies convention), each subimage is generated by

$$f^j(\left\lfloor \frac{x}{M} \right\rfloor, \left\lfloor \frac{y}{M} \right\rfloor) = \frac{1}{M^2} \sum_{k=0}^{M-1} \sum_{m=0}^{M-1} f^0(x+k, y+m), \quad 0 \le j \le n,$$

where the filter M is of size 2^j and resolution level $j = 0$ corresponds to the highest resolution.

3. unimodality
$$0 \le h(j,k) \le h(m,n) \text{ for } j \le m < \frac{M}{2} \text{ and } k \le n < \frac{M}{2};$$
4. equal contribution (all pixels of f contribute the same total weight to each pixel of f^j);
5. separability
$$h(k,m) = h(k)h(m).$$

The averaging box filter is often chosen as the smoothing filter with $h(k,m) = 1/M^2, \forall k,m$, where $M = 2^{n-j}$ defines the filter size as well as the filter coefficients. For example, the following subimages are obtained from a 4×4 image:

$$f^1(\frac{x}{2},\frac{y}{2}) = \sum_{k=0}^{1}\sum_{m=0}^{1}\frac{f^2(x+k,y+m)}{4},\ f^0(\frac{x}{4},\frac{y}{4}) = \sum_{k=0}^{3}\sum_{m=0}^{3}\frac{f^2(x+k,y+m)}{16},$$

where $f^2(x,y)$ is the original image. Using the normalized box filter, example subsampled images are shown in Figure 29. The MIP-mapping pyramid is formed by the union of the original image and the set of subsampled images.

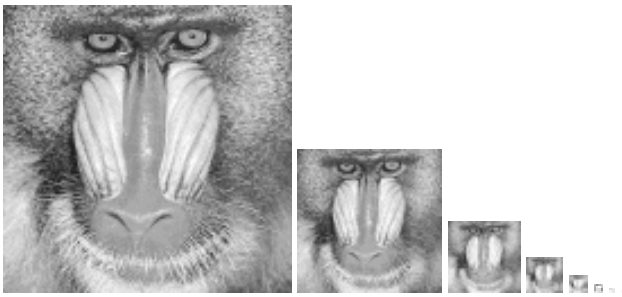


Fig. 29. MIP-map subimages, processed by normalized box filter. Obtained from The Center for Image Processing Research (CIPR), an Internet public domain archive (ftp://ipl.rpi.edu/pub/image/still/usc/bgr/-baboon).

Reconstruction of the image at a given pixel location (x,y) depends on the desired resolution of the pixel. The desired resolution level is bandlimited to the number of decomposed resolution levels (typically the decomposition is dyadic in nature) bounded by the two closest resolution subimages f^{j-1} and f^j. The final pixel value at location (x,y) is calculated as a linear combination of pixel intensities in the pyramid:

$$f(x,y) = (1-p)f^{j-1}(\left\lfloor \frac{x}{2^{n-(j-1)}} \right\rfloor, \left\lfloor \frac{y}{2^{n-(j-1)}} \right\rfloor) + (p)f^j(\left\lfloor \frac{x}{2^{n-j}} \right\rfloor, \left\lfloor \frac{y}{2^{n-j}} \right\rfloor). \tag{5.85}$$

Equation (5.85) represents linear *inter-map* interpolation, shown schematically for an 8×8 image in Figure 30. In cases where the image is not being mapped onto a flat surface, *intra-map* interpolation may also be used to prevent aliasing artifacts (see [WW92, §4.7.1]). The combination of inter- and intra-map interpolation

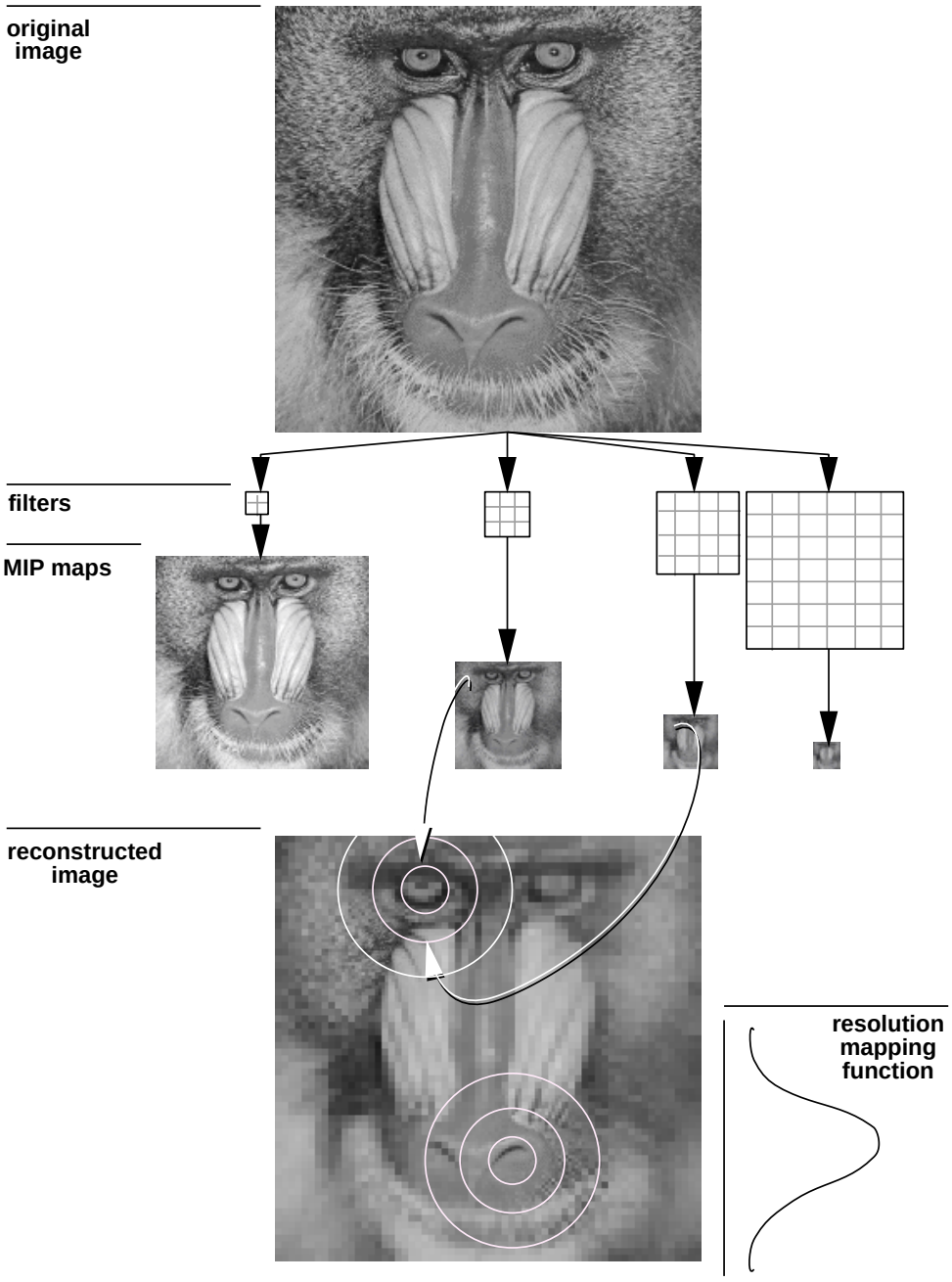


Fig. 30. Depiction of MIP-mapping algorithm.

is called bi-linear interpolation. The wavelet technique equivalent to Equation (5.85) discussed below does not consider bi-linear interpolation.

5.10.2 MIP-Wavelets

To generate wavelet multiscale representations of a given image matching the MIP-map decomposition using the normalized box filter, the lowpass wavelet filter $\{h_k\}$ is set to $\{1/2, 1/2\}$. The detail filter is then the quadrature mirror of $\{h_k\}$, i.e., $\{g_k\} = \{1/2, -1/2\}$. To guarantee perfect reconstruction, MIP-wavelet dual filters are required so that conditions (5.58) and (5.59) are simultaneously satisfied. For filters of length 2, the following equations must hold:

$$h_0\widetilde{h}_0 + h_1\widetilde{h}_1 = 1 \tag{5.86}$$

$$g_0\widetilde{h}_0 + g_1\widetilde{h}_1 = 0. \tag{5.87}$$

Given $(\{h_k\}, \{g_k\})$, $\{\widetilde{h}_k\}$ is derived with:

$$\frac{1}{2}\widetilde{h}_0 + \frac{1}{2}\widetilde{h}_1 = 1, \quad \text{from (5.86), and,} \tag{5.88}$$

$$\frac{1}{2}\widetilde{h}_0 - \frac{1}{2}\widetilde{h}_1 = 0, \quad \text{from (5.87).} \tag{5.89}$$

From (5.89), $\widetilde{h}_0 = \widetilde{h}_1$, and substituting into (5.88),

$$\widetilde{h}_0 = \widetilde{h}_1 = 1.$$

The dual detail filter coefficients are derived from conditions (5.57) and (5.60), which generate equations:

$$g_0\widetilde{g}_0 + g_1\widetilde{g}_1 = 1 \tag{5.90}$$

$$h_0\widetilde{g}_0 + h_1\widetilde{g}_1 = 0. \tag{5.91}$$

Using derived filters $(\{h_k\}, \{g_k\})$, $\{\widetilde{g}_k\}$ is found by:

$$\frac{1}{2}\widetilde{g}_0 + \frac{1}{2}\widetilde{g}_1 = 0, \quad \text{from (5.91), and,} \tag{5.92}$$

$$\frac{1}{2}\widetilde{g}_0 - \frac{1}{2}\widetilde{g}_1 = 1, \quad \text{from (5.90).} \tag{5.93}$$

From (5.92), $\widetilde{g}_0 = -\widetilde{g}_1$, and substituting into (5.93),

$$\widetilde{g}_0 = -\widetilde{g}_1 = 1.$$

MIP-wavelets, with coefficients given in Table 7, are unnormalized versions of the Haar filters. That is, MIP-wavelets are semi-orthogonal Haar wavelets (or pre-wavelets). In fact, normalized Haar filters will generate the same texture mapping at dyadic resolution boundaries, but will lose luminance information between

TABLE 7
MIP-wavelet filters.

k	$2(h_k)$	$2(g_k)$	$2(\tilde{h}_k)$	$2(\tilde{g}_k)$
0	1	1	2	2
1	1	-1	2	-2

boundaries where linear interpolation is required. The benefit of the semi-orthogonal MIP-wavelets is that correct luminance values will be generated at any desired resolution level. Note that the filter coefficients of the lowpass filter $\{h_k\}$ match the averaging box filter above exactly. This can easily be verified by obtaining the tensor product of the scaling filter at any resolution level. For example, at one level of resolution $(j = 1)$ the effective sampling filter is a 2×2 filter with cells equal to $1/4$. At level $j = 0$, the filter is a 4×4 filter with cells equal to $1/16$. Note that under the DWT, the zeroth $(j = 2)$ resolution level (i.e., the original image) is not present in the pyramidal transformation. Because MIP-wavelets generate identical subsampled lowpass images to subimages generated by averaged box filters, in terms of resolution, both decompositions are identical. That is, in the case of monochromatic images, the same luminance resolution information is present in the scaled subimages of both approaches. In other words, the MIP-wavelets derived above serve as the basis functions for the multiresolution averaged box filter.

5.10.3 Variable Resolution Reconstruction with MIP-wavelets

Reconstruction with MIP-wavelets is lossless due to the orthogonality of the filters. To obtain interpolation results identical to MIP-mapping, an intuitive approach would be to maintain reconstructed scaled subimages produced by successive steps of the IDWT, then to perform the interpolation step as given by Equation (5.85). Although this approach would yield identical results due to the equivalence of subsampling filters and the DWT's perfect reconstruction (due to the orthogonality of the MIP-wavelets), it is memory-intensive. What is perhaps not obvious is that identical interpolation results can be obtained by scaling wavelet coefficients prior to reconstruction. Scaling of the wavelet coefficients prior to reconstruction results in the attenuation of the signal with respect to the average (low-pass) signal. Full decimation of the coefficients (scaling by 0) results in a lossy, subsampled reproduction of the original. Conversely, scaling wavelet coefficients by 1 preserves all detail information producing lossless reconstruction. Selectively scaling the coefficients by a value in the range $[0,1]$ at appropriate levels of the wavelet pyramid produces a variable resolution image upon reconstruction. This approach is equivalent to MIP-mapping reconstruction with linear interpolation of pixel values.

In MIP-mapping, the value of the interpolant p is determined by some mapping function which specifies the desired resolution level l. The two closest pyramid resolution levels are then determined by rounding down

and up to find subimage levels $j-1$ and j. The interpolant value is obtained by the relation:

$$p = l - \lfloor l \rfloor .$$

Note that the slope of the mapping function should match the resolution hierarchy of the pyramid, i.e., if resolution decreases eccentrically from some reference point, the parameter l should also decrease eccentrically. If it does not, its value may be reversed by subtracting from the number of resolution levels, i.e., $n-l$. To scale wavelet coefficients, p is set to either 0, 1, or the interpolant value at particular subbands according to the following relations dependent on l:

$$p = \begin{cases} 1, & j \leq \lfloor l \rfloor; \\ l - \lfloor l \rfloor, & j = \lceil l \rceil; \\ 0, & j > \lceil l \rceil. \end{cases} \tag{5.94}$$

For example, if at some particular pixel location (x,y), $l = 1.5$, then wavelet coefficients would be preserved (scaled by 1) at levels $j \leq 1$, scaled by .5 at level $j = 2$, and decimated (scaled by 0) at levels $j > 2$ at the appropriate pixel location in the subimages.

Theorem 1 *Wavelet coefficient scaling is equivalent to linear pixel interpolation under MIP mapping.*

Proof: Consider the interpolation step in the latter,

$$f = (1-p) f^{j-1} + (p) f^{j},$$

which is equivalent to Equation (5.85) with implicit pixel coordinates. When there is no need for interpolation, i.e., the desired resolution level at a pixel falls on a mapped resolution boundary, $l - \lfloor l \rfloor = 0$, and $j = \lceil l \rceil = l$, then

$$f = \begin{cases} f^{j-1} = f^{l-1} & \text{if } p = 0; \\ f^{j} = f^{l} & \text{if } p = 1; \end{cases}$$

or simply $f = f^l$ meaning that the resolution at the given pixel will match the resolution level of the subimage at the l^{th} pyramid level. Simplifying the IDWT reconstruction Equation (5.75) and expanding,

$$f^{j}_{\phi\phi} = (p f^{j-1}_{\psi} + \ldots + (p f^{2}_{\psi} + (p f^{1}_{\psi} + (p f^{0}_{\psi} + f^{0}_{\phi\phi}))) \ldots), \tag{5.95}$$

where f^{j}_{ψ} collectively represents subimages containing wavelet coefficient, $f^{j}_{\psi\phi}$, $f^{j}_{\phi\psi}$, $f^{j}_{\psi\psi}$, with implied pixel coordinates. Note that in Equation (5.95) the symbol p is now the interpolant and not the binary selection variable as used in Equation (5.75). In the case when $l - \lfloor l \rfloor = 0$,

$$f = (0 \ldots + (f^{l}_{\psi} \ldots + (f^{2}_{\psi} + (f^{1}_{\psi} + (f^{0}_{\psi} + f^{0}_{\phi\phi}))) \ldots) \ldots).$$

Since $f^{j+1}_{\phi\phi} = f^{j}_{\psi} + f^{j}_{\phi\phi}$ at each level of reconstruction, the resultant image will contain resolution matching the contents of subimage f^{l+1}. The subimage f^{l+1} contains the average (lowpass) component at level $l+1$, or equivalently, it contains the entire reproduced image at level l. In this case, the reconstructed image will

contain resolution no finer than that found in subimage f^l, i.e., $f = f^l$, as with MIP-mapping.

When $l - \lfloor l \rfloor \neq 0$, i.e., interpolation is required due to the mapping falling between resolution boundaries, MIP mapping reproduces resolution at each pixel location according to Equation (5.85). In the IDWT, with $\lfloor l \rfloor < l < \lceil l \rceil$,

$$f = (0 \ldots + (0 + (p f_\psi^{\lceil l \rceil} + (f_\psi^{\lfloor l \rfloor} \ldots + (f_\psi^2 + (f_\psi^1 + (f_\psi^0 + f_{\phi\phi}^0))) \ldots))) \ldots).$$

That is, resolution at the given pixel location is no finer than what is reproduced at subimage $f_{\phi\phi}^{\lceil l \rceil + 1}$, since

$$\begin{aligned} f_{\phi\phi}^{\lceil l \rceil + 1} &= p f_\psi^{\lceil l \rceil} + f_{\phi\phi}^{\lfloor l \rfloor + 1} \\ &= p f_\psi^j + f_{\phi\phi}^j \\ &= f_{\phi\phi}^{j+1} \\ &= f^j, \end{aligned} \tag{5.96}$$

where $\lfloor l \rfloor + 1 = \lceil l \rceil$, and $j = \lceil l \rceil$ from Equation (5.94). Equation (5.96) shows that the reconstructed resolution will be no finer than the resolution contained in f^l matching the finest resolution level produced by MIP-mapping as specified by Equation (5.85).

What remains to be shown is that the resolution gain obtained by scaling wavelet coefficients (at level j) is equivalent to the gain obtained by the MIP-mapping interpolation. That is, the scaling operation $p f_\psi^j + f_{\phi\phi}^j$ is equivalent (in terms of resolution gain) to the interpolation $(1-p) f^{j-1} + (p) f^j$. To prove this, recall the reconstruction relation,

$$\begin{aligned} f_{\phi\phi}^j &= f_\psi^{j-1} + f_{\phi\phi}^{j-1} \text{ or,} \\ f_{\phi\phi}^{j+1} &= f_\psi^j + f_{\phi\phi}^j \text{ giving,} \\ f_\psi^j &= f_{\phi\phi}^{j+1} - f_{\phi\phi}^j. \end{aligned} \tag{5.97}$$

Substituting Equation (5.97) into the wavelet scaling equation,

$$\begin{aligned} p f_\psi^j + f_{\phi\phi}^j &= f_{\phi\phi}^j + p f_\psi^j \\ &= f_{\phi\phi}^j + p(f_{\phi\phi}^{j+1} - f_{\phi\phi}^j) \\ &= f_{\phi\phi}^j - p f_{\phi\phi}^j + p f_{\phi\phi}^{j+1} \\ &= (1-p) f_{\phi\phi}^j + p f_{\phi\phi}^{j+1} \\ &= (1-p) f_{\phi\phi}^{j-1} + p f_{\phi\phi}^j \\ &= (1-p) f^{j-1} + p f^j \end{aligned}$$

completes the proof showing that scaling wavelet coefficients by p at pyramid level j is equivalent to linearly interpolating pixel values of scaled subimages f^{j-1}, f^j through MIP-mapping. □

CHAPTER VI

TIME SERIES ANALYSIS

Time series analysis is a powerful framework for modeling autoregressive temporal data, that is, a signal that is to some extent dependent on its historical component. Time series analysis (TSA) has been used in the social sciences for modeling temporally periodic data such as monthly stock quotes, airline passenger data, and behavioral patterns, although in general, TSA techniques are related to analysis with digital filters and even to wavelet theory. TSA differs from traditional statistical approaches in one significant aspect: time series analysis assumes dependence of observed time series values on previous values in the series whereas traditional ordinary least squares (OLS) techniques assume sample independence. It is this aspect of TSA that is particularly suitable for modeling signals where temporal samples at time t are correlated with samples at $t - k$, at least for some duration k.

This section starts with a brief review of linear systems theory, closely following Robinson and Silvia's presentation [RS79], including feedback and feedforward systems. The discussion continues with the Box-Jenkins time series models of random processes, including concepts of stationarity and invertibility pertinent to modeling autoregressive integrated moving average (ARIMA) processes. The section concludes by the introduction of *piecewise* autoregressive integrated moving average (PARIMA) processes and wavelet methods suitable for detection of PARIMA demarcations (interventions).

6.1 Fundamentals

Time series analysis of a stochastic (random) process is in general based on the representation of the stochastic process by a linear system. Feedback and feedforward components of the system model the autoregressive characteristics of the process. A simplified view of a stochastic time series, for example, is that the series is the output of a linear (feedback/feedforward) system with a pure random signal (e.g., white noise) as its input. The prevalent method of time series modeling (the Box-Jenkins method), in essence, is concerned with finding an appropriate linear filter so that a given time series s_t passed through the filter resembles white noise x_t on output. If such a filter can be found then the reverse filter will generate the observed time series given white noise as its input (the Slutzky effect). In this case the filter completely describes the observed time series. In this section the fundamental properties of a linear filter are reviewed, including the relevant concepts of convolution and the z-transform.

6.1.1 Classification of Discrete Time Series

A discrete time series is represented by a sequence of observations x_t which are assumed to be equally spaced in time (for convenience $\Delta t = 1$). In this section, the discussion is restricted to *deterministic* time series, i.e., the observations x_t are fixed quantities, and not random variables. The simplest time series is the delta function δ_t (not to be confused with the Kronecker delta, $\delta_{j,l}$) defined by:

$$\delta_t = \begin{cases} 1, & t = 0, \\ 0, & t \neq 0, \ t \in \mathbf{Z}. \end{cases}$$

The delta function is often referred to as the *unit impulse* or *unit spike*. Any arbitrary, deterministic, time series x_t can be represented by a sum of scaled delayed and advanced unit impulses by the expression [RS79, §4.1]:

$$\begin{aligned} x_t &= \cdots + x_{-2}\delta_{t+2} + x_{-1}\delta_{t+1} + x_0\delta_t + x_1\delta_{t-1} + x_2\delta_{t-2} + \cdots \\ &= \sum_{k=-\infty}^{\infty} x_k \delta_{t-k} \quad k \in \mathbf{Z}. \end{aligned}$$

A *finite-length* time series is classified by nonzero values lying in some finite interval, i.e., $x_t \neq 0$ for t in some interval $[\alpha, \beta]$. The *total energy* of a time series is defined by:

$$E = \sum_{t=-\infty}^{\infty} x_t \overline{x_t} = \sum_{t=-\infty}^{\infty} |x_t|^2,$$

where the symbol ($\bar{\cdot}$) denotes complex conjugation. Defining the *average power*, $\dot{P}$, of a finite-length time series as:

$$\dot{P} = \frac{1}{2t+1} \sum_{t=-k}^{k} |x_t|^2,$$

a *transient* time series is one of finite energy, whereas a continually oscillatory time series is classified as *non-transient* satisfying the condition:

$$0 < \lim_{t\to\infty} \dot{P} = \lim_{t\to\infty} \left[\frac{1}{2t+1} \sum_{t=-k}^{k} |x_t|^2 \right] < \infty.$$

The unit impulse is an example of a transient time series, while a sinusoidal time series is an example of a non-transient time series. A transient time series has zero power while a non-transient series has infinite energy. Discrete time series can in general be classified as one or the other or neither, but not both. Due to its finite energy, a wavelet is a transient time series. If the wavelet satisfies the condition that $x_t = 0$ for $t < 0$ then it is also *causal*. It is important to note that time series (and wavelets) need not be of finite length to be transient, only that they are of finite energy.

6.1.2 Linear Systems

In practice, discrete time series may be recorded through observation of some natural process. A central problem in time series analysis is modeling the response of a system to a specified input or time series.

There are various methods in which such a model can be built including the formal solution of the difference equation characterizing the system, use of transform techniques, and analysis based on the response of a system to an elementary time series using the principle of superposition (convolution). The latter approach is limited to linear systems. A *linear system* which defines a filter prescribing a transformation of the input series s_t into the output series x_t, is often depicted as a "black box" with s_t as the input and x_t as the output generated by the black box, i.e., $s_t \mapsto x_t$. The system is *linear* if and only if it possesses the same formal properties as a (linear) vector space in linear algebra, i.e.,

(6.1) 1. if $u_t \mapsto v_t$ and $w_t \mapsto x_t$ then $u_t + w_t \mapsto v_t + x_t$ (*closure under sum*);

(6.2) 2. if $s_t \mapsto x_t$ then $cs_t \mapsto cx_t$ (*closure under scalar multiplication*).

A system which does not satisfy both properties (6.1) and (6.2) is called *non-linear*. A linear system or filter is also termed time-invariant if the transformation of s_t to x_t is independent of a time origin, i.e.,

$$\text{if } s_t \mapsto x_t \text{ then } s_{t-k} \mapsto x_{t-k}.$$

The output of a linear system or filter is expressed in terms of the input and the system's *unit impulse response*. This concept is known as convolution. The impulse response of the system is defined as the response of a linear, time-invariant system to the unit impulse, δ_t. The system response is often denoted by h_t, i.e., $\delta_t \mapsto h_t$. Convolution of a time series s_t with h_t generates the output x_t by the convolution sum equation:

$$x_t = \sum_{k=-\infty}^{\infty} s_k h_{t-k}, \quad k \in \mathbf{Z}. \tag{6.3}$$

Equation (6.3) is derived from the properties of the linear system, i.e., given that the linear system transforms the unit impulse,

$$\delta_t \mapsto h_t$$

and since the system is time-invariant,

$$\delta_{t-k} \mapsto h_{t-k},$$

then by the multiplicative property (6.2),

$$s_k \delta_{t-k} \mapsto s_k h_{t-k},$$

and by the additive property (6.1),

$$\sum_{k=-\infty}^{\infty} s_k \delta_{t-k} \quad \mapsto \quad \sum_{k=-\infty}^{\infty} s_k h_{t-k}, \quad k \in \mathbf{Z}.$$

Since any arbitrary time series can be resolved into a sum of unit impulses, i.e.,

$$s_t = \sum_{k=-\infty}^{\infty} s_k \delta_{t-k} = \cdots + s_{-2}\delta_{t+2} + s_{-1}\delta_{t+1} + s_0\delta_t + s_1\delta_{t-1} + s_2\delta_{t-2} + \cdots,$$

it follows that

$$s_t = \sum_{k=-\infty}^{\infty} s_k \delta_{t-k} \mapsto \sum_{k=-\infty}^{\infty} s_k h_{t-k} = x_t,$$

or simply,

$$s_t \mapsto x_t$$

where x_t denotes the output of the linear system expressed by Equation (6.3) [RS79, pp.172-173]. Equation (6.3) is often abbreviated as

$$x_t = s_k * h_{t-k},$$

where the asterisk denotes the convolution operation.

A particularly useful transformation of a deterministic discrete time series, common to digital filter design, is the z-transform defined by:

$$X(z) = \sum_{t=-\infty}^{\infty} x_t z^t,$$

where the coefficient for z^t is the value of x_t of the time series at time t. Given a causal time series (i.e., $x_t = 0$ for $t < 0$), the z-transform becomes the power series:

$$X(z) = \sum_{t=0}^{\infty} x_t z^t,$$

and in the case of a finite-length time series of length n, the z-transform becomes an n-degree polynomial $X(z) = x_0 + x_1 z + \cdots + x_n z^n$. Given two z-transforms

$$S(z) = \sum_{t=-\infty}^{\infty} s_t z^t, \quad H(z) = \sum_{t=-\infty}^{\infty} h_t z^t,$$

the multiplication of $S(z)H(z)$ results in the z-transform $X(Z)$:

$$X(z) = S(z)H(z) = \sum_{t=-\infty}^{\infty} x_t z^t, \tag{6.4}$$

where the coefficients x_t of $X(z)$ are related to the coefficients s_t and h_t by:

$$x_t = \sum_{k=-\infty}^{\infty} s_k h_{t-k}, \quad k \in \mathbf{Z},$$

which is recognized as the convolution of s_t with h_t. Convolution in time domain (6.3) is equivalent to multiplication of z-transforms (6.4) [RS79, p.177]. Convolution is commutative and associative under multiplication, i.e.,

$$S(z)H(z) = H(z)S(z), \quad S(z)[H(z)X(z)] = [S(z)H(z)]X(z),$$

and distributive under addition, i.e.,

$$S(z)[H(z) + X(z)] = S(z)H(z) + S(z)X(z).$$

6.1.3 Autoregressive-Moving Average (ARMA) Linear Systems

The input-output relationship of a causal linear time-invariant filter may be represented in a number of ways. The convolution sum is one representation, from (6.3),

$$x_t = \sum_{k=0}^{\infty} h_k s_{t-k}, \quad k \in \mathbf{Z}, \tag{6.5}$$

where s_t is the input and x_t is the output of a causal filter with impulse response $h_t = (h_0, h_1, \ldots)$. Another representation is the difference equation:

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = b_0 s_t + b_1 s_{t-1} + \cdots + b_q s_{t-q}, \tag{6.6}$$

where the output time series x_t at time t is dependent only on the present and past values s_t, s_{t-1}, …, s_{t-q} of the input time series and past values x_{t-1}, x_{t-2}, …, x_{t-p} of the output time series. The coefficients $\{a_0 = 1, a_1, a_2, \ldots, a_p\}$ are known as the *feedback* or *autoregressive* (AR) coefficients and the coefficients $\{b_0, b_1, \ldots, b_q\}$ are known as the *feedforward* or *moving average* (MA) coefficients. The representation in (6.6) is commonly referred to as an autoregressive-moving average (ARMA) digital filter. Both representations yield identical results [RS79, p.201,p.202]. Rewriting (6.6) as

$$\sum_{k=0}^{p} a_k x_{t-k} = \sum_{k=0}^{q} b_k s_{t-k}, \quad a_0 = 1,\ k \in \mathbf{Z},$$

under the z-transform becomes

$$A(z)X(z) = B(z)S(z). \tag{6.7}$$

Equation (6.5) is obtained if the infinite linear filter $\{h_k\}$ can be replaced by a rational function in z composed of the finite filters $\{a_k\}$ and $\{b_k\}$,

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_0 + b_1 z + b_2 z^2 + \cdots + b_q z^q}{a_0 + a_1 z + a_2 z^2 + \cdots + a_p z^p} = h_0 + h_1 z + h_2 z^2 + \cdots.$$

The impulse response coefficients are related to the feedback and feedforward coefficients by the relation $A(z)H(Z) = B(z)$, or in the time domain,

$$\sum_{k=0}^{p} a_k h_{t-k} = \begin{cases} b_t, & 0 \le t \le q, \\ 0, & t > q, \quad k \in \mathbf{Z}. \end{cases} \tag{6.8}$$

The causal filter with impulse response h_t is completely described by the z-transform

$$H(z) = \sum_{t=0}^{\infty} h_t z^t,$$

defined as the *transfer function* of the causal filter. In (6.5), the causal filter is described by the impulse response $h_t = (h_0, h_1, \ldots)$ or the transfer function; whereas in (6.6), the causal filter is characterized by a finite number of feedback and feedforward coefficients. Since the same linear system can be described by both

representations, the fact that the ARMA representation requires a finite set of $p+q+1$ coefficients (assuming $a_0 = 1$), instead of the infinite number required by the impulse response h_t representation, is the chief advantage of the ARMA model. The generally smaller number of parameters required by ARMA models was popularized by Box and Jenkins [BJ76] who referred to this efficiency as *parsimony*. ARMA models describe systems with an infinitely long impulse response by a finite-parameter model, denoted by ARMA(p,q).

6.1.4 Moving Average (MA) and Autoregressive (AR) Linear Systems

Considering the case when $A(z) = 1$, the transfer function of the linear system is

$$H(z) = B(z) = b_0 + b_1 z + b_2 z^2 + \cdots + b_q z^q.$$

The linear system is commonly referred to as a feedforward system characterized by the moving average model of order q, or MA(q) [RS79, p.257]. From (6.7), the MA(q) model is written as

$$X(z) = B(z)S(z),$$

or in the time domain,

$$x_t = b_0 s_t + b_1 s_{t-1} + \cdots + b_q s_{t-q}.$$

Considering the case when $B(z) = 1$, the transfer function of the linear system is

$$H(z) = \frac{1}{A(z)} = \frac{1}{a_0 + a_1 z + a_2 z^2 + \cdots + a_p z^p}.$$

The linear system is referred to as a (pure) feedback system characterized by the autoregressive model of order p, or AR(p). From (6.7), the AR(p) model is written as

$$A(z)X(z) = S(z),$$

or in the time domain,

$$a_0 x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = s_t. \tag{6.9}$$

A simple feedback system is shown in Figure 31 where, assuming $a_0 = 1$, the summation term is

$$s_t - \sum_{k=1}^{p} a_t x_{t-k} = x_t,$$

which coincides with Equation (6.9) [RS79, p.210].

In general, MA(q) coefficients specify a Finite Impulse Response (FIR) filter provided the number of coefficients is finite. The AR(p) coefficients specify an Infinite Impulse Response (IIR) filter since the transfer function $A(z)^{-1}$ is, in general, composed of an infinite number of coefficients [GW94, pp.12-13]. Even with an initially finite number of autoregressive coefficients $\{a_0, a_1, \ldots, a_p\}$, this can be verified by performing the long division $1/(a_0 + a_1 z + \cdots + a_p z^p)$.

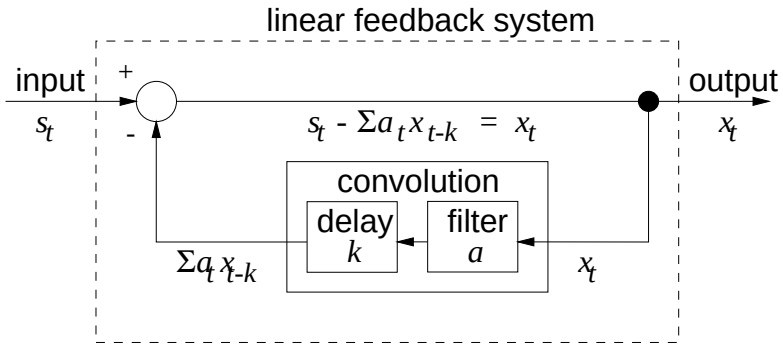


Fig. 31. Block diagram of a linear feedback system.

6.2 Nondeterministic (Stochastic) Time Series Models

The above discussion of linear systems dealt with deterministic time series. In the following section, stochastic time series will be considered where observations are assumed to contain a random component.

A stochastic or random process is characterized by its mean μ, variance σ^2, autocovariances $\{\gamma_k\}$, autocorrelations $\{\zeta_k\}$, and partial autocorrelations $\{\rho_{k,k}\}$. The mean represents the expected value of an observation at time t,

$$\mu_t = E[x_t],$$

where $E[\cdot]$ denotes expectation. The variance σ^2 refers to the squared deviation of each observation from its expected value, i.e.,

$$\sigma_t^2 = E[(x_t - E[x_t])^2].$$

In general, $E[(x_t - E[x_t])^n]$ is defined as the n^{th} *moment* of x_t about its mean, where the variance is known as the second moment. The autocovariance function,

$$\gamma_k = E[(x_t - E[x_t])(x_{t+k} - E[x_{t+k}])],$$

measures the deviation of two observations from their means. The prefix *auto* refers to both observations coming from the same time series separated by temporal *lag* k. The autocorrelation function is defined as

$$\zeta_k = \frac{E[(x_t - E[x_t])(x_{t+k} - E[x_{t+k}])]}{E[(x_t - E[x_t])(x_t - E[x_t])]} = \frac{E[(x_t - E[x_t])(x_{t+k} - E[x_{t+k}])]}{E[(x_t - E[x_t])^2]},$$

which, for x_t, x_{t+k} of strictly stationary processes with first two moments finite, depends only on the time lag k. In the special case of a zero mean, the autocorrelation function is the autocovariance function normalized so that $\zeta_0 = 1$. The autocorrelation function is an even function, where $\zeta_k = \zeta_{-k}$, meaning that it is symmetric about the origin $k = 0$. The partial autocorrelation function $\rho_{k,k}$ measures the correlation between x_t and x_{t+k}

when the $k-1$ intervening values $x_{t+1}, x_{t+2}, \ldots, x_{t+k-1}$ are fixed [RS79, p.280].

Stochastic time series are time series where any observation x_t of a time series is a random variable with a probability distribution generally dependent on time t. Time series models popularized by Box and Jenkins (referred to as Box-Jenkins time series models) assume that the time series is generated from a series of uncorrelated random variables with zero mean and constant variance, e.g., white noise $\varepsilon_t \sim N(0, \sigma^2)$. The observed time series x_t is generally modeled as the output of a causal linear filter with white noise (innovations or random shocks) ε_t as input. With the filter impulse response $h_t = (h_0, h_1, \ldots)$, the time series x_t is represented as

$$\begin{aligned} x_t &= \mu_t + h_t * \varepsilon_t \\ &= \mu_t + \sum_{k=0}^{\infty} h_k \varepsilon_{t-k}, \quad k \in \mathbf{Z}, \quad \text{or,} \\ x_t &= \mu_t + \eta_t, \end{aligned} \tag{6.10}$$

where $\eta_t = \sum_{k=0}^{\infty} h_k \varepsilon_{t-k}$ is a zero mean random process [RS79, p.275]. In this view, the time series is generally composed of of two parts, namely the deterministic component μ_t and the random or stochastic component η_t.

The stochastic time series x_t is a linear transformation of the white noise series ε_t. This can be shown by borrowing z and for the moment using it as a backward shift operator defined by

$$z x_t = x_{t-1},$$

which delays the time series x_t by one time unit. In general,

$$z^k x_t = x_{t-k},$$

where $k \in \mathbf{Z}$ represents a delay of k time units. Rewriting Equation (6.10) in terms of z,

$$\begin{aligned} x_t &= \mu_t + h_0\varepsilon_t + h_1\varepsilon_{t-1} + h_2\varepsilon_{t-2} + \cdots \\ &= \mu_t + h_0\varepsilon_t + h_1 z\varepsilon_t + h_2 z^2\varepsilon_t + \cdots \\ &= \mu_t + (h_0 + h_1 z + h_2 z^2 + \cdots)\varepsilon_t. \end{aligned} \tag{6.11}$$

Reclaiming the definition of z as defined by the z-transform, the quantity inside the parentheses in (6.11) is the z-transform of the impulse response $h_t = (h_0, h_1, h_2, \ldots)$. Since $H(z) = h_0 + h_1 z + h_2 z^2 + \cdots$ defines the transfer function of the linear filter h_t, Equation (6.11) is rewritten as

$$x_t = \mu_t + H(z)\varepsilon_t \tag{6.12}$$

showing that the time series x_t is a linear transformation of the white noise series ε_t.

6.2.1 Stationarity of Stochastic Time Series

The concept of *stationarity* of time series refers to the statistical behavior of time series and is relevant to the choice of appropriate modeling strategy. In general, ARMA models are restricted to stationary time series. A time series is stationary (covariance-stationary or second-order stationary) if both of the following conditions are satisfied:

$$E[x_t] = \mu_t = \mu \text{ constant mean for all } t, \tag{6.13}$$

$$\begin{aligned} E[x_t, x_{t+k}] &= E[(x_t - E[x_t])(x_{t+k} - E[x_{t+k}])], \text{ for } E[x_t] = 0, \\ &= \gamma_k = \text{cov}(x_t, x_{t+k}) \text{ dependent only on time difference } k. \end{aligned} \tag{6.14}$$

If either condition (6.13) or (6.14) is not satisfied, the time series is *non-stationary* [RS79, p.277]. With respect to the causal linear filter model (6.12), the time series x_t is stationary provided that:

$$\begin{aligned} \mu_t &= \mu \text{ (constant mean for all } t), \\ E &= \sum_{t=0}^{\infty} h_t \overline{h_t} = \sum_{t=0}^{\infty} |h_t|^2 < \infty \text{ (impulse response finite energy).} \end{aligned}$$

The white noise series, ε_t, is a fundamental example of a stationary time series. The white noise series is a sequence of random variables each with zero mean and constant variance, i.e.,

$$E[\varepsilon_t] = 0, \quad E[\varepsilon_t^2] = \sigma^2, \quad \forall t,\ t \in \mathbf{Z},$$

and such that any two different variables are uncorrelated, i.e.,

$$E[\varepsilon_t, \varepsilon_{t+k}] = E[\varepsilon_t \varepsilon_{t+k}] = 0, \quad k \neq 0,\ k \in \mathbf{Z},$$

giving covariance

$$\gamma_k(\varepsilon_t) = \begin{cases} E[\varepsilon_t^2] = \sigma^2, & k = 0; \\ E[\varepsilon_t \varepsilon_{t+k}], & k \neq 0, \quad k \in \mathbf{Z}. \end{cases} \tag{6.15}$$

For stationary processes in general, where the mean is constant for each observation, the autocorrelation function ζ_k is the autocovariance function normalized so that $\zeta_0 = 1$. The autocorrelation and partial autocorrelation functions (ACF, PACF) are instrumental in determining a suitable stochastic time series model.

6.2.2 Autoregressive-Moving Average (ARMA) Processes

Section 6.1.3 discussed deterministic ARMA linear systems. Given white noise as the input to the ARMA linear system, the output is a signal with a random component. Deterministic and stochastic signals are distinguished by referring to the deterministic ARMA system as a *model* of the stochastic ARMA *process*. Equation (6.10) represents the time series x_t as the output of the ARMA linear filter with impulse response h_t and transfer function $H(z)$. The filter is usually one of infinite length. If the filter can be replaced by one

whose transfer function is a rational function in z, then the stochastic ARMA process can be represented by the following finite-parameter model:

$$\begin{aligned} x_t &= \mu_t + \sum_{k=0}^{\infty} h_k \varepsilon_{t-k} \\ &= \mu_t + (h_0 + h_1 z + h_2 z^2 + \cdots)\varepsilon_t \\ &= \mu_t + H(z)\varepsilon_t \\ &= \mu_t + \frac{B(z)}{A(z)}\varepsilon_t \\ &= \mu_t + \left(\frac{b_0 + b_1 z + b_2 z^2 + \cdots + b_q z^q}{a_0 + a_1 z + a_2 z^2 + \cdots + a_p z^p} \right) \varepsilon_t . \end{aligned} \tag{6.16}$$

Transposing terms and expanding,

$$\begin{aligned} A(z)(x_t - \mu_t) &= B(z)\varepsilon_t \\ \sum_{k=0}^{p} a_k (x_{t-k} - \mu_{t-k}) &= \sum_{k=0}^{q} b_k \varepsilon_{t-k} . \end{aligned}$$

Defining the time series in terms of deviations from the deterministic mean, μ_t, by $\tilde{x} = x_t - \mu_t$, the ARMA process is modeled by

$$\begin{aligned} A(z)\tilde{x}_t &= B(z)\varepsilon_t \\ \sum_{k=0}^{p} a_k \tilde{x}_{t-k} &= \sum_{k=0}^{q} b_k \varepsilon_{t-k} \\ a_0\tilde{x}_t + a_1\tilde{x}_{t-1} + a_2\tilde{x}_{t-2} + \cdots + a_p\tilde{x}_{t-p} &= b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots + b_p\varepsilon_{t-q} . \end{aligned} \tag{6.17}$$

Equation (6.17) is the time-domain representation of Equations (6.12) and (6.16).

If the time series is stationary then its statistical properties are time-invariant, i.e., unaffected by a shift of the time origin [RS79, p.277]. That is, the statistical properties of the series x_t are identical to those of the delayed time series x_{t-k}, with mean

$$\begin{aligned} E[x_t] &= E[\mu + H(z)\varepsilon_t] \\ &= E[\mu] + E[\sum_{k=0}^{\infty} h_t \varepsilon_{t-k}] \\ &= \mu, \end{aligned}$$

since $E[\varepsilon_t] = 0 \ \forall t$, and variance

$$\begin{aligned} E[(x_t - \bar{x})^2] &= E[(x_t - E[x_t])^2] \\ &= E[(x_t - \mu)^2] = E[(\mu_t + \sum_{k=0}^{\infty} h_k \varepsilon_{t-k} - \mu_t)^2] \\ &= E[(\sum_{k=0}^{\infty} h_k \varepsilon_{t-k})^2]. \end{aligned} \tag{6.18}$$

Expanding (6.18),

$$
\begin{aligned}
E[(\sum_{k=0}^{\infty} h_k\varepsilon_{t-k})^2] &= E[(h_0\varepsilon_t + h_1\varepsilon_{t-1} + \cdots)(h_0\varepsilon_t + h_1\varepsilon_{t-1} + \cdots)] \\
&= E[\ h_0\varepsilon_t h_0\varepsilon_t + h_0\varepsilon_t h_1\varepsilon_{t-1} + h_0\varepsilon_t h_2\varepsilon_{t-2} + \cdots \\
&\qquad h_1\varepsilon_{t-1}h_0\varepsilon_t + h_1\varepsilon_{t-1}h_1\varepsilon_{t-1} + h_1\varepsilon_{t-1}h_2\varepsilon_{t-2} + \cdots \\
&\qquad h_2\varepsilon_{t-2}h_0\varepsilon_t + h_2\varepsilon_{t-2}h_1\varepsilon_{t-1} + h_2\varepsilon_{t-2}h_2\varepsilon_{t-2} + \cdots\ ] \\
&= E[\ h_0^2\varepsilon_t^2 + h_0\varepsilon_t h_1\varepsilon_{t-1} + h_0\varepsilon_t h_2\varepsilon_{t-2} + \cdots \\
&\qquad h_1\varepsilon_{t-1}h_0\varepsilon_t + h_1^2\varepsilon_{t-1}^2 + h_1\varepsilon_{t-1}h_2\varepsilon_{t-2} + \cdots \\
&\qquad h_2\varepsilon_{t-2}h_0\varepsilon_t + h_2\varepsilon_{t-2}h_1\varepsilon_{t-1} + h_2^2\varepsilon_{t-2}^2 + \cdots\ ] \\
&= E[\sum_{k=0}^{\infty} h_k^2\varepsilon_{t-k}^2 + \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h_j\varepsilon_{t-j}h_k\varepsilon_{t-k}], \quad j \neq k \\
&= E[\sum_{k=0}^{\infty} h_k^2\varepsilon_{t-k}^2 + \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h_j h_k\varepsilon_{t-j}\varepsilon_{t-k}], \quad j \neq k \\
&= E[\sum_{k=0}^{\infty} h_k^2\varepsilon_{t-k}^2] + E[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h_j h_k\varepsilon_{t-j}\varepsilon_{t-k}], \quad j \neq k \\
&= \sum_{k=0}^{\infty} h_k^2 E[\varepsilon_{t-k}^2] + \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h_j h_k E[\varepsilon_{t-j}\varepsilon_{t-k}], \quad j \neq k \\
&= \sigma^2\sum_{k=0}^{\infty} h_k^2,
\end{aligned}
$$

since $E[\varepsilon_k^2] = \sigma^2\ \forall k$, and $E[\varepsilon_j\varepsilon_k] = 0\ \forall j,k,\ j \neq k$. Note that the variance of the time series exists if and only if $\sum_{k=0}^{\infty}|h_k|^2 < \infty$, i.e., the linear filter $\{h_k\}$ has finite energy.

The autocovariance of x_t at lag k is defined as

$$
\begin{aligned}
\gamma_k(x_t) &= E[(x_t - \bar{x})(x_{t+k} - \bar{x})] \\
&= E[(x_t - E[x_t])(x_{t+k} - E[x_t])] \\
&= E[(\mu_t + \sum_{l=0}^{\infty} h_l\varepsilon_{t-l} - \mu_t)(\mu_{t+k} + \sum_{j=0}^{\infty} h_j\varepsilon_{t+k-j} - \mu_{t+k})] \\
&= E[(\sum_{l=0}^{\infty} h_l\varepsilon_{t-l})(\sum_{j=0}^{\infty} h_j\varepsilon_{t+k-j})] \qquad (6.19) \\
&= E[(h_0\varepsilon_t + h_1\varepsilon_{t-1} + \cdots)(h_0\varepsilon_{t+k} + h_1\varepsilon_{t+k-1} + \cdots)] \\
&= E[\ h_0\varepsilon_t h_0\varepsilon_{t+k} + h_0\varepsilon_t h_1\varepsilon_{t+k-1} + h_0\varepsilon_t h_2\varepsilon_{t+k-2} + \cdots \\
&\qquad h_1\varepsilon_{t-1}h_0\varepsilon_{t+k} + h_1\varepsilon_{t-1}h_1\varepsilon_{t+k-1} + h_1\varepsilon_{t-1}h_2\varepsilon_{t+k-2} + \cdots \\
&\qquad h_2\varepsilon_{t-2}h_0\varepsilon_{t+k} + h_2\varepsilon_{t-2}h_1\varepsilon_{t+k-1} + h_2\varepsilon_{t-2}h_2\varepsilon_{t+k-2} + \cdots\ ] \\
&= E[\ h_0h_0\varepsilon_t\varepsilon_{t+k} + h_0h_1\varepsilon_t\varepsilon_{t+k-1} + h_0h_2\varepsilon_t\varepsilon_{t+k-2} + \cdots \\
&\qquad h_1h_0\varepsilon_{t-1}\varepsilon_{t+k} + h_1h_1\varepsilon_{t-1}\varepsilon_{t+k-1} + h_1h_2\varepsilon_{t-1}\varepsilon_{t+k-2} + \cdots \\
&\qquad h_2h_0\varepsilon_{t-2}\varepsilon_{t+k} + h_2h_1\varepsilon_{t-2}\varepsilon_{t+k-1} + h_2h_2\varepsilon_{t-2}\varepsilon_{t+k-2} + \cdots\ ]. \qquad (6.20)
\end{aligned}
$$

The summation term of the stationary time series covariance given by Equation (6.20) is represented in matrix topology where the matrix can be thought of as being the result of vector multiplication where the summation multiplicands of (6.19) are seen as vectors. Since the autocovariance of the white noise process, given by

relation (6.15), is such that $E[\varepsilon_{t+j}\varepsilon_{t+k}] = 0$ for $j \neq k$, only diagonal terms in the "matrix" in (6.20) are non-zero for any given k. That is, for $k = 0$, only the main diagonal entries are non-zero. For $k = 1$, the first upper diagonal contains non-zero entries, for $k = -1$, it is the the first lower diagonal, for $k = 2$, the second upper diagonal, and so on. These positive and negative k^{th} matrix diagonals (w.r.t. the main diagonal where $k = 0$) are denoted by the j^{th} row and $(j+k)^{th}$ column elements for any row j. Correspondingly, the stationary time series covariance is rewritten as

$$\begin{aligned} \gamma_k(x_t) &= E[\sum_{j=0}^{\infty} h_j h_{j+k} \varepsilon_{t-j} \varepsilon_{t+k-(j+k)}] \\ &= E[\sum_{j=0}^{\infty} h_j h_{j+k} \varepsilon_{t-j} \varepsilon_{t-j}], \quad \forall k. \end{aligned}$$

Since $E[\varepsilon_{t+j}\varepsilon_{t+k}] = \sigma^2$ for $j = k$,

$$\begin{aligned} \gamma_k(x_t) &= \sum_{j=0}^{\infty} h_j h_{j+k} E[\varepsilon_{t-j}\varepsilon_{t-j}] \\ &= \sigma^2 \sum_{j=0}^{\infty} h_j h_{j+k}, \quad \forall k. \end{aligned} \tag{6.21}$$

Equation (6.21) states that the autocovariance of the time series x_t is proportional to the autocorrelation of the impulse response h_t [RS79, pp.278-280]. In fact, Equation (6.21) can be represented by the convolution of the autocovariances of the impulse response and the white noise series, i.e., $\gamma_k(x_t) = \gamma_k(h_t) * \gamma_k(\varepsilon_t)$.

As discussed in §6.2, the autocorrelation function of the time series x_t at lag k is defined as

$$\zeta_k = \frac{E[(x_t - E[x_t])(x_{t+k} - E[x_{t+k}])]}{E[(x_t - E[x_t])^2]}.$$

For the stationary time series, the autocorrelation becomes

$$\zeta_k = \frac{\gamma_k}{\gamma_0},$$

which refers to the correlation between any two observations x_t and x_{t+k} and does not depend on the units of measurement [RS79, p.279]. The autocorrelation function is the normalized autocovariance function, i.e., $\zeta_0 = 1$ and $-1 \leq \zeta_k \leq 1$ for $k \in \mathbf{Z}$.

6.2.3 The Autoregressive (AR) Process

The autoregressive process AR(p) of order p is modeled as the output of a linear filter with transfer function $A(z)^{-1}$ whose input is the white noise time series ε_t. The AR(p) process is the output of a feedback system modeled by

$$\begin{aligned} A(z)\tilde{x}_t &= \varepsilon_t \\ \sum_{k=0}^{p} a_k \tilde{x}_{t-k} &= \varepsilon_t. \end{aligned} \tag{6.22}$$

where $\tilde{x}_t = x_t - \mu_t$. Expanding (6.22) gives:

$$a_0\tilde{x}_t + a_1\tilde{x}_{t-1} + a_2\tilde{x}_{t-2} + \cdots + a_p\tilde{x}_{t-p} \quad = \quad \varepsilon_t, \tag{6.23}$$

which is the standard time-domain representation of the AR(p) process. The AR(p) process is characterized by the expected autocorrelation function. Letting $\alpha_k = -a_k$ and assuming $a_0 = 1$, Equation (6.23) is rewritten as

$$\tilde{x}_t \quad = \quad \alpha_1\tilde{x}_{t-1} + \alpha_2\tilde{x}_{t-2} + \cdots + \alpha_p\tilde{x}_{t-p} + \varepsilon_t. \tag{6.24}$$

In order to first obtain the autocovariance function, $\gamma_k(\tilde{x}_t) = E[\tilde{x}_k\tilde{x}_{t-k}]$, Equation (6.24) is multiplied by $\tilde{x}_{t-k}$ on both sides producing:

$$\tilde{x}_t\tilde{x}_{t-k} \quad = \quad \alpha_1\tilde{x}_{t-1}\tilde{x}_{t-k} + \alpha_2\tilde{x}_{t-2}\tilde{x}_{t-k} + \cdots + \alpha_p\tilde{x}_{t-p}\tilde{x}_{t-k} + \varepsilon_t\tilde{x}_{t-k}.$$

Taking expected values gives

$$E[\tilde{x}_t\tilde{x}_{t-k}] = \tag{6.25}$$
$$E[\alpha_1\tilde{x}_{t-1}\tilde{x}_{t-k}] + E[\alpha_2\tilde{x}_{t-2}\tilde{x}_{t-k}] + \cdots + E[\alpha_p\tilde{x}_{t-p}\tilde{x}_{t-k}] + E[\varepsilon_t\tilde{x}_{t-k}].$$

Noting that for stationary processes $E[\varepsilon_t\tilde{x}_{t-k}] = 0$ for $k > 0$ and that the autocovariance is dependent only on the time difference k, i.e.,

$$\begin{aligned}
\gamma_k(\tilde{x}_t) &= E[x_t x_{t-k}] \\
\gamma_{k-1}(\tilde{x}_t) &= E[x_t x_{t-k-1}] = E[x_{t-1}x_{t-k}] \\
\gamma_{k-2}(\tilde{x}_t) &= E[x_t x_{t-k-2}] = E[x_{t-2}x_{t-k}] \\
&\cdots \\
\gamma_{k-p}(\tilde{x}_t) &= E[x_t x_{t-k-p}] = E[x_{t-p}x_{t-k}], \qquad (6.26)
\end{aligned}$$

Equation (6.25) becomes:

$$\gamma_k(\tilde{x}_t) \quad = \quad \alpha_1\gamma_{k-1}(\tilde{x}_t) + \alpha_2\gamma_{k-2}(\tilde{x}_t) + \cdots + \alpha_p\gamma_{k-p}(\tilde{x}_t) + 0, \quad k > 0.$$

Dividing through by $\gamma_0(\tilde{x}_t)$ gives the recursive representation of the autocorrelation function ζ_k:

$$\zeta_k = \alpha_1\zeta_{k-1} + \alpha_2\zeta_{k-2} + \alpha_3\zeta_{k-3} + \cdots + \alpha_p\zeta_{k-p}, \quad k > 0. \tag{6.27}$$

Using z once again as the backward shift operator, i.e.,

$$\zeta_k = \alpha_1 z\zeta_k + \alpha_2 z^2\zeta_k + \cdots + \alpha_p z^p\zeta_k,$$

the autocorrelation function is determined by the following difference equation for $k > 0$:

$$(1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p)\zeta_k = 0. \tag{6.28}$$

Equation (6.28) is known as the *autocorrelation generating function* [Got81, p.159]. Since any polynomial can be expressed as a product of its (possibly complex) roots, the polynomial in z of degree p in Equation (6.28) can be factored into

$$\prod_{j=1}^{m}(1-r_j z)^{d_j} = (1-r_1 z)^{d_1}(1-r_2 z)^{d_2}(1-r_3 z)^{d_3}\cdots(1-r_m z)^{d_m}, \tag{6.29}$$

where $(1-r_j z)^{d_j}$ is a (possibly complex) root of multiplicity d_j such that $\sum_{j=1}^{m} d_j = p$ [Wei90, pp.44-46]. For the process to be stationary, $|r_j^{-1}| > 1$ and $|r_j| < 1$ in (6.29) must hold. As a consequence, the autocorrelation function (ACF) tails off as a mixture of exponential decays and/or damped sine waves depending on the roots of (6.29) (damped sine waves will appear if some of the roots are complex). Furthermore, by its construction and dependence on the autocorrelation function, the partial autocorrelation function (PACF) vanishes after lag p. These properties provide a means of estimating the order of the AR(p) process by examining the structure of the sample autocorrelation and partial autocorrelation functions.

6.2.4 The Moving Average (MA) Process

The moving average MA(q) processes of order q is a time series $\tilde{x}_t$ linearly dependent on a finite number of previous white noise samples (random shocks).[1] The MA(q) process is the output of a feedforward system with input ε_t and transfer function $B(z) = (b_0 + b_1 z + b_2 z^2 + \cdots + b_q z^q)$ and is modeled by

$$\begin{aligned} \tilde{x}_t &= B(z)\varepsilon_t \\ \tilde{x}_t &= \sum_{k=0}^{q} b_k \varepsilon_{t-k}, \end{aligned} \tag{6.30}$$

where $\tilde{x}_t = x_t - \mu_t$. Expanding (6.30) gives:

$$\tilde{x}_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots + b_p\varepsilon_{t-q}, \tag{6.31}$$

which is the standard time-domain representation of the MA(q) process. Because the linear filter $\{b_k\}$ is composed of a finite number of coefficients, the impulse response of the linear filter $(b_0, b_1, \ldots, b_q)$ has finite energy, i.e.,

$$\sum_{k=0}^{\infty} |b_k|^2 = \sum_{k=0}^{q} |b_k|^2 < \infty.$$

The finite property of the MA(q) filter guarantees the stationarity of the MA(q) process. Noting that the autocovariance function does not depend on absolute time t, $E[\tilde{x}_t] = 0$ with finite and constant variance $\gamma_0 = E[(\tilde{x}_t - E[\tilde{x}_t])^2]$. Since the autocovariance of any linear filter time series model can be expressed in terms

[1] The name *moving average* is somewhat misleading since the weights $\{b_k\}$ need not be positive and need not sum to unity.

of the impulse response $(h_0, h_1, \ldots)$ of the filter, as shown by Equation (6.21) in §6.2.2, the autocovariance of the MA(q) process at lag k is given by

$$\gamma_k(\tilde{x}_t) = \sigma^2 \sum_{j=0}^{\infty} b_j b_{j+k} = \sigma^2 \sum_{j=0}^{q} b_j b_{j+k}. \tag{6.32}$$

Because $b_j = 0$ for $j > q$, the summation in (6.32) need only collect the first $(0, 1, 2, \ldots, q-k-1, q-k)$ terms for given lag k. Also since $\gamma_k(\tilde{x}_t) = \gamma_{-k}(\tilde{x}_t)$, the autocovariance function can be rewritten as the even function:

$$\gamma_k(\tilde{x}_t) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|k|} b_j b_{j+|k|} & \text{for } k = 0, \pm 1, \pm 2, \ldots, \pm q; \\ 0 & \text{otherwise.} \end{cases}$$

The autocorrelation of the MA(q) process then becomes

$$\zeta_k(\tilde{x}_t) = \frac{\gamma_k(\tilde{x}_t)}{\gamma_0(\tilde{x}_t)} = \begin{cases} \frac{\sum_{j=0}^{q-|k|} b_j b_{j+|k|}}{\sum_{j=0}^{q} b_j^2} & \text{for } k = 0, \pm 1, \pm 2, \ldots, \pm q; \\ 0 & \text{otherwise,} \end{cases}$$

which cuts off after lag q [Wei90, p.53]. The partial autocorrelation function (PACF) of the MA(q) process, on the other hand, tails off as a mixture of exponential decays and/or damped sine waves depending on the roots of equation

$$(1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_q z^q) = 0, \tag{6.33}$$

which is the transfer function of the MA(q) filter with $\beta_k = -b_k$ and $b_0 = 1$. As with the autocorrelation function (ACF) of the AR(p) process, damped sine waves will appear if some of the roots of Equation (6.33) are complex. As with the AR(p) process, these properties provide a means of estimating the order of the MA(q) process by examining the structure of the sample autocorrelation and partial autocorrelation functions.

6.2.5 Autoregressive (AR) Moving Average (MA) Process Duality

In general, a finite order stationary AR(p) process corresponds to an infinite order MA process. The converse is also true if the finite MA(q) process is *invertible*. This AR/MA *duality* can be expressed in terms of the z-transforms of the corresponding (finite) filters. From (6.23), the AR(p) process may be written as:

$$(1 + a_1 z + a_2 z^2 + \cdots + a_p z^p)\tilde{x}_t = \varepsilon_t, \tag{6.34}$$

with $a_0 = 1$. Rewriting (6.34) with $\alpha_k = -a_k$ for $k = 1, 2, \ldots, p$ gives

$$\tilde{x}_t = \frac{1}{(1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p)} \varepsilon_t. \tag{6.35}$$

Using (6.29) to represent the roots of $(1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p)$, if the AR($p$) process is stationary, then $|r_j^{-1}| > 1$ and $|r_j| < 1$ and by long division, Equation (6.35) becomes

$$\tilde{x}_t = (1 + b_1 z + b_2 z^2 + \cdots)\varepsilon_t,$$

which is an infinite-order moving average process, or MA(∞). In this case the AR(p) process is said to be both stationary and invertible. Conversely, from (6.31), the MA(q) process may be written as:

$$\tilde{x}_t = (1 + b_1 z + b_2 z^2 + \cdots + b_q z^q)\varepsilon_t, \tag{6.36}$$

with $b_0 = 1$. Rewriting (6.36) with $\beta_k = -b_k$ for $k = 1, 2, \ldots, q$ gives

$$\frac{1}{(1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_q z^q)}\tilde{x}_t = \varepsilon_t. \tag{6.37}$$

Using (6.29) again to this time represent the roots of $(1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_q z^q)$, if the MA($q$) process is invertible, then $|r_j^{-1}| > 1$ and $|r_j| < 1$ and by long division, Equation (6.37) becomes

$$(1 + a_1 z + a_2 z^2 + \cdots)\tilde{x}_t = \varepsilon_t,$$

which is an infinite-order autoregressive process, or AR(∞). In this case the MA(p) process is said to be invertible (it is also inherently stationary due to its finite-length filter).

The AR/MA process duality is exhibited in the structures of the ACF and PACF of the AR(p) and MA(q) processes as mentioned in §6.2.3 and §6.2.4. This relationship is summarized in Table 8.

TABLE 8
AR(p)/MA(q) duality.

	ACF	PACF
AR(p)	decays	vanishes after lag p
MA(q)	vanishes after lag q	decays

6.2.6 The Mixed Autoregressive Moving Average (ARMA) Process

As seen in §6.2.2, the mixed autoregressive moving average process can be modeled by (6.17), repeated here for convenience,

$$\begin{aligned} A(z)\tilde{x}_t &= B(z)\varepsilon_t \\ \sum_{k=0}^{p} a_k \tilde{x}_{t-k} &= \sum_{k=0}^{q} b_k \varepsilon_{t-k} \\ a_0\tilde{x}_t + a_1\tilde{x}_{t-1} + a_2\tilde{x}_{t-2} + \cdots + a_p\tilde{x}_{t-p} &= b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots + b_p\varepsilon_{t-q}. \end{aligned} \tag{6.38}$$

In §6.2.2 the statistical properties of the process were considered in terms of the infinite filter h_t which implicitly modeled the ARMA process. In this section the characterizing autocorrelation and partial autocorrelation functions are derived explicitly in terms of the finite-parameter autoregressive moving average model, or

ARMA(p,q).

In order to first obtain the autocovariance function, $\gamma_k(\tilde{x}_t) = E[\tilde{x}_k\tilde{x}_{t-k}]$, Equation (6.38) is multiplied by $\tilde{x}_{t-k}$ on both sides producing:

$$a_0\tilde{x}_t\tilde{x}_{t-k} + a_1\tilde{x}_{t-1}\tilde{x}_{t-k} + a_p\tilde{x}_{t-p}\tilde{x}_{t-k} \quad = \quad b_0\varepsilon_t\tilde{x}_{t-k} + b_1\varepsilon_{t-1}\tilde{x}_{t-k} + b_p\varepsilon_{t-q}\tilde{x}_{t-k}.$$

Taking expected values gives

$$E[a_0\tilde{x}_t\tilde{x}_{t-k}] + E[a_1\tilde{x}_{t-1}\tilde{x}_{t-k}] + E[a_p\tilde{x}_{t-p}\tilde{x}_{t-k}] = \tag{6.39}$$
$$E[b_0\varepsilon_t\tilde{x}_{t-k}] + E[b_1\varepsilon_{t-1}\tilde{x}_{t-k}] + E[b_p\varepsilon_{t-q}\tilde{x}_{t-k}].$$

Noting that for stationary processes the autocovariance is dependent only on the time difference k, as shown in (6.26), Equation (6.39) becomes:

$$a_0\gamma_k(\tilde{x}_t) + a_1\gamma_{k-1}(\tilde{x}_t) + a_2\gamma_{k-2}(\tilde{x}_t) + \cdots + a_p\gamma_{k-p}(\tilde{x}_t) \quad = \tag{6.40}$$
$$\sum_{j=0}^{q} b_j E[\varepsilon_{t-j}\tilde{x}_{t-k}],$$

To evaluate $\sum_{j=0}^{q} b_j E[\varepsilon_{t-j}\tilde{x}_{t-k}]$, $E[\varepsilon_{t-j}\tilde{x}_{t-k}]$ is derived by first considering $E[\varepsilon_t\tilde{x}_{t-k}]$, noting that $\tilde{x}_t$ can be represented using (6.16) with $\tilde{x}_t = x_t - \mu_t$, i.e.,

$$\tilde{x}_t = \sum_{j=0}^{\infty} h_j\varepsilon_{t-j} = h_0\varepsilon_t + h_1\varepsilon_{t-1} + h_2\varepsilon_{t-2} + \cdots, \tag{6.41}$$

where impulse response coefficients h_t are related to the feedback and feedforward coefficients by the relation $A(z)H(Z) = B(z)$ as shown in the time domain by (6.8), repeated here for convenience using index j instead of k,

$$\sum_{j=0}^{p} a_j h_{t-j} = \begin{cases} b_t, & 0 \le t \le q, \\ 0, & t > q. \end{cases} \tag{6.42}$$

Due to the time-invariant property of the linear system, from (6.41),

$$\tilde{x}_{t-k} = \sum_{j=0}^{\infty} h_j\varepsilon_{t-k-j} = h_0\varepsilon_{t-k} + h_1\varepsilon_{t-k-1} + h_2\varepsilon_{t-k-2} + \cdots. \tag{6.43}$$

Multiplying both sides of (6.43) by ε_t and taking expectations gives

$$E[\varepsilon_t\tilde{x}_{t-k}] = E[\sum_{j=0}^{\infty} h_j\varepsilon_t\varepsilon_{t-k-j}] = \tag{6.44}$$
$$E[h_0\varepsilon_t\varepsilon_{t-k}] + E[h_1\varepsilon_t\varepsilon_{t-k-1}] + E[h_2\varepsilon_t\varepsilon_{t-k-2}] + \cdots.$$

Since $E[\varepsilon_t\varepsilon_{t-k}] = \sigma^2$ for $k > 0$ and $E[\varepsilon_t\varepsilon_{t-k}] = 0$ for $k \neq 0$, Equation (6.44) becomes

$$E[\varepsilon_t\tilde{x}_{t-k}] = \begin{cases} \sigma^2 h_{-k} & k \le 0 \\ 0 & k > 0. \end{cases} \tag{6.45}$$

To evaluate $E[\varepsilon_{t-j}\tilde{x}_{t-k}]$, the index variables in (6.45) are changed by letting $s = t - j$ and substituting $k - j$ for k, giving

$$E[\varepsilon_{t-j}\tilde{x}_{t-k}] = E[\varepsilon_s\tilde{x}_{s-(k-j)}] = \begin{cases} \sigma^2 h_{j-k} & k \le j \\ 0 & k > j. \end{cases} \tag{6.46}$$

Substituting (6.46) into (6.40) yields

$$a_0\gamma_k(\tilde{x}_t) + a_1\gamma_{k-1}(\tilde{x}_t) + a_2\gamma_{k-2}(\tilde{x}_t) + \cdots + a_p\gamma_{k-p}(\tilde{x}_t) = \begin{cases} \sigma^2 \sum_{j=0}^{q} b_j h_{j-k} & \text{for } k \le q \\ 0 & \text{for } k > q, \end{cases} \tag{6.47}$$

where the impulse response coefficients h_t are related to the feedback and feedforward coefficients a_t and b_t by (6.42). Equation (6.47) and (6.42) completely describe the autocovariance of an ARMA(p,q) process in terms of the feedback and feedforward coefficients a_t, b_t [RS79, p.308]. Dividing through by $\gamma_0(\tilde{x}_t)$ gives the recursive representation of the autocorrelation function ζ_k:

$$\zeta_k + a_1\zeta_{k-1} + a_2\zeta_{k-2} + \cdots + a_p\zeta_{k-p} = \begin{cases} \sigma^2 \sum_{j=0}^{q} b_j h_{j-k} & \text{for } k \le q \\ 0 & \text{for } k > q, \end{cases} \tag{6.48}$$

where $a_0 = 1$. Equation (6.48) states that the ACF of an ARMA(p,q) model tails off after lag q (like an AR(p) process) since after lag q the moving average (feedforward) component has no effect. The first q autocorrelations, however, depend on both autoregressive and moving average parameters [Wei90, p.57]. The PACF of the ARMA(p,q) process will also contain a mixture of exponential and/or damped sine waves due to the MA component. That is, for identification purposes, the fact that both ACF and PACF tail off (instead of either of them cutting off) suggests a mixed ARMA model [Wei90, p.59]. In general, the number of "anomalous" terms in the ACF (terms that do not follow the decay pattern) equals $(p-q)$, and the number of anomalous terms in the PACF equals $(q-p)$. The ACF for an ARMA(p,q) process is a mixture of decaying exponentials and damped sine waves after the first $(q-p)$ lags, and the PACF is a mixture of decaying exponentials and damped sine waves after the first $(p-q)$ lags [Got81, p.258].

6.3 Stochastic Process Sample Statistics

Time series modeling relies on the calculation of the sample autocorrelation and partial autocorrelation functions. As discussed in §6.2, a stochastic or random process is characterized by its mean μ, variance σ^2, autocovariances $\{\gamma_k\}$, autocorrelations $\{\zeta_k\}$, and partial autocorrelations $\{\rho_{k,k}\}$. Theoretically, calculation of the the exact values of these parameters requires an ensemble of all possible realizations of the random events. In most applications available observations constitute only a single realization making the calculation of the ensemble average impossible. The theoretical statistical parameters are instead approximated by sample parameters $\bar{x}$, s^2, $\hat{\gamma}$, $\hat{\zeta}$, and $\hat{\rho}_{k,k}$, respectively. Formulations for each parameter are given below with the understanding that the formulations are approximations based on measurements of the sample, not the

population.

The sample mean, used to approximate $E[x_t]$, is defined as:

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t,$$

which is the average of n observations. The variance of n measurements, approximating $E[(x_t - \bar{x})^2]$, is:

$$s^2 = \frac{1}{n-1}\sum_{t=1}^{n} [x_t - \bar{x}]^2,$$

which is the sum of the squared deviations of each observation divided by $n-1$ to give an unbiased estimator of σ^2 (in some cases the denominator n is used but for large samples would lead to an underestimation of σ^2). In general, the n^{th} moment of x_t, approximating $E[(x_t - \bar{x})^n]$, is obtained by

$$s^n = \frac{1}{n-1}\sum_{t=1}^{n} [x_t - \bar{x}]^n.$$

The sample autocovariance and autocorrelation functions, approximating $E[(x_t - \bar{x})(x_{t+k} - \bar{x})]$, $E[(x_t - \bar{x})(x_{t+k} - \bar{x})]/E[(x_t - \bar{x})^2]$, respectively, are defined as:

$$\begin{aligned} \hat{\gamma}_k &= \frac{1}{n}\sum_{t=1}^{n-k} [x_t - \bar{x}][x_{t+k} - \bar{x}], \\ \hat{\zeta}_k &= \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n-k} [x_t - \bar{x}][x_{t+k} - \bar{x}]}{\sum_{t=1}^{n} [x_t - \bar{x}]^2}, \quad k = 0, 1, 2, \ldots, \end{aligned}$$

which, for stationary processes, depends only on the time difference k. For a zero mean process, the autocorrelation function is the autocovariance function normalized so that $\hat{\zeta}_0 = 1$. A plot of $\hat{\zeta}$ vs. k is sometimes called a sample correlogram. The sample autocorrelation function (ACF) is an even function, where $\hat{\zeta}_k = \hat{\zeta}_{-k}$, and is symmetric about the origin $k = 0$ with

$$\hat{\zeta}_0 = \frac{\hat{\gamma}_0}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{n} [x_t - \bar{x}][x_t - \bar{x}]}{\sum_{t=1}^{n} [x_t - \bar{x}]^2} = 1,$$

which is the (trivial) case of lag 0 where each x_t is perfectly correlated with itself. The sample partial autocorrelation function (PACF) $\hat{\rho}_{k,k}$ is given by the iterative formulas:

$$\hat{\rho}_{k+1,k+1} = \frac{\hat{\zeta}_{k+1} - \sum_{j=1}^{k} \hat{\rho}_{k,j}\hat{\zeta}_{k+1-j}}{1 - \sum_{j=1}^{k} \hat{\rho}_{k,j}\hat{\zeta}_j} \tag{6.49}$$

and

$$\hat{\rho}_{k+1,j} = \hat{\rho}_{k,j} - \hat{\rho}_{k+1,k+1}\hat{\rho}_{k,k+1-j}, \quad j = 1, \ldots, k, \tag{6.50}$$

with $\hat{\rho}_{0,0} = \hat{\zeta}_0 = 1$, and $\hat{\rho}_{1,1} = \hat{\zeta}_1$. Note that the only PACF terms of interest are $\hat{\rho}_{k,k}$, sometimes abbreviated to $\hat{\rho}_k$. To calculate $\hat{\rho}_{k,k}$, obtain $\hat{\rho}_{k+1,k+1}$ for $k = -1, 0, 1, \ldots, n$. Starting at $k = -1$ Equation (6.49) gives:

$$\hat{\rho}_{0,0} = \frac{\hat{\zeta}_0 - \sum_{j=1}^{-1} \hat{\rho}_{-1,j}\hat{\zeta}_{-j}}{1 - \sum_{j=1}^{-1} \hat{\rho}_{-1,j}\hat{\zeta}_j}$$

which simplifies to

$$\hat{\rho}_{0,0} = \hat{\zeta}_0 = 1$$

since the summation terms are not involved. For $k = 0$,

$$\hat{\rho}_{1,1} = \frac{\hat{\zeta}_1 - \sum_{j=1}^{0} \hat{\rho}_{0,j}\hat{\zeta}_{1-j}}{1 - \sum_{j=1}^{0} \hat{\rho}_{0,j}\hat{\zeta}_j}$$

again the summation terms are ignored leaving

$$\hat{\rho}_{1,1} = \hat{\zeta}_1.$$

For $k > 0$, Equations (6.49) and (6.50) are used where Equation (6.50) calculates the intermediate values of $\hat{\rho}_{k,j}$ for $1 \leq j \leq k$. These values are used in the next iterative calculation of $\hat{\rho}_{k+1,k+1}$. At the end of the calculation, the intermediate values may be discarded since usually only the values of $\hat{\rho}_{k,k}$ are used in time series analysis.

6.4 Stationary Time Series Modeling

Developing a time series model of an unknown stochastic process involves an iterative sequence of process identification, model specification, estimation of model parameters, and performance of consistency (diagnostic) checks. Some of these steps may involve repetition, as depicted in Figure 32. The identification step is crucial to building an adequate model of the stochastic process. Unfortunately, a fair amount of guesswork, intuition and luck may be required since the true nature of the process is usually unknown. A priori knowledge of the system under investigation is often invaluable. Once an adequate model has been identified and tested, predictions can be attempted in an effort to forecast future trends.

6.4.1 Process Identification

The first identification step (first block in Figure 32) in modeling time series requires a test for stationarity. By definition of stationarity, the sample mean and variances are calculated and checked for trends. If the mean and variances are constant then the process under investigation satisfies the stationary criteria. Note that this does not mean that the process is stationary, only that the observations exhibit stationary characteristics.

If the observed data is stationary, then the ACF and PACF are examined. If the plots of the ACF/PACF exhibit characteristics as suggested in Table 8, then it may be possible to model the time series by either an AR(p) or MA(q) model. If not, a mixed ARMA(p, q) model may be required.

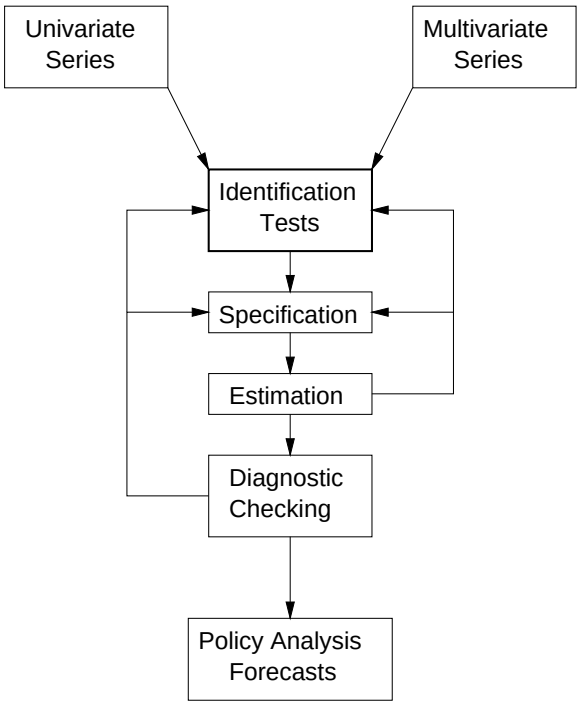


Fig. 32. Time series modeling approach. Adopted from [CHLT94, p.2 (Fig. 1.1)].

6.4.2 Model Specification

Model specification involves the determination of the order of the identified process. For example, if the observed data is found to be stationary, then an ARMA(p,q) model may be suitable. Specification of the model requires the choice of the order of the model, i.e., specifying the model parameters p and q. Once a candidate model has been specified, the model parameters (coefficients) can be calculated and tested.

6.4.3 Parameter Estimation

Coefficients of the AR(p) model can be estimated from the sample autocorrelation function. Noting that $\zeta_k = \zeta_{-k}$ and substituting $k = 1, 2, \ldots, p$ in (6.27) gives

$$
\begin{array}{rcccccccccc}
\zeta_1 & = & \alpha_1 & + & \alpha_2\zeta_1 & + & \alpha_3\zeta_2 & + & \cdots & + & \alpha_p\zeta_{p-1} \\
\zeta_2 & = & \alpha_1\zeta_1 & + & \alpha_2 & + & \alpha_3\zeta_1 & + & \cdots & + & \alpha_p\zeta_{p-2} \\
\zeta_3 & = & \alpha_1\zeta_2 & + & \alpha_2\zeta_1 & + & \alpha_3 & + & \cdots & + & \alpha_p\zeta_{p-3} \\
\vdots & & & & & & & & & & \\
\zeta_p & = & \alpha_1\zeta_{p-1} & + & \alpha_2\zeta_{p-2} & + & \alpha_3\zeta_{p-3} & + & \cdots & + & \alpha_p.
\end{array}
\tag{6.51}
$$

The linear system of Equations (6.51) is known as the set of *Yule-Walker equations* whose solution with $\hat{\zeta}$ replacing ζ gives estimates $\hat{\alpha}_k$ for the parameters α_k. The Yule-Walker equations can be used in a similar manner to estimate parameters β_k of the MA(q) model.

6.4.4 Diagnostic Checks

Diagnostic checks involve measuring deviations of estimated model parameters from observed data. In this sense, time series analysis is similar to traditional ordinary least-squares (OLS) methods. Among various diagnostic checks, a prevalent approach, identical to the popular OLD method, is checking residuals. If a given time series can be transformed to a series resembling white noise, then the filter(s) used in the transformation provide a sufficient model of the stochastic process. This is due to the *Slutzky effect*, named after its discoverer, which states that white noise is a series from which any other series can be constructed [Got81, p.29]. Consequently, if the specified filter can transform a given series into white noise, it should also be reversible and generate the given series by filtering white noise. This reverse process was discovered by Yule. If the filtered series' residuals indicate a pattern other than white noise, the model (filter) is inappropriate and the one or more of the modeling steps should be repeated.

6.5 Non-stationary (Linear) Time Series Modeling

In general, stationary time series can be modeled as mixed autoregressive moving-average (ARMA) processes. Non-stationary series with non-constant mean and/or variance cannot. Although the general model building strategy as shown in Figure 32 is also used for modeling non-stationary time series, the specified model must account for the non-stationary characteristics of the process.[2] Often the non-stationary model is composed of some transformation of the time series where the transformed series is adequately described by the stationary ARMA model. In this way the non-stationary model is composed of a stationary component (e.g., an ARMA model) and some transformation. A popular example of such a model is the autoregressive integrated moving-average (ARIMA) system.

Autoregressive integrated moving-average models rely on a variable-degree differencing transformation of the data, popularized by Box and Jenkins, in order to transform a non-stationary time series into a stationary one [CLT94, BJ76, p.10]. The differenced data and the process x_t is referred to as an *integrated* process of order d where d refers to the number of levels of differencing required to achieve stationarity. The ARIMA model incorporates the three model components of auto-regression, integration, and moving averages, each of order p,d,q, respectively. The complete model is denoted as ARIMA(p,d,q) and is specified by the relation

$$A(z)(1-z)^d x_t = \mu_0 + B(z)\varepsilon_t, \quad d \geq 0, \tag{6.52}$$

where for $d > 0$, μ_0 is called the deterministic trend component and is usually assumed to be 0. For $d =$

[2] The non-stationary models discussed herein are restricted to linear functions of past observations although it should be noted that it is also possible to specify non-linear models of past observations, e.g., exponential autoregressive models. Non-linear models are outside the scope of the present context of linear systems.

0, the original process is stationary with μ_0 related to the mean of the process, i.e., $\mu_0 = (1 + a_1 z + \cdots + a_p z^p)\mu$ [Wei90, §4.1.2]. Rewriting (6.52) in the time domain with $\mu_0 = 0$ and $b_0 = 1$,

$$(1-z)^d(x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p}) = \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q}. \tag{6.53}$$

With $d = 1$, for example, (6.53) becomes:

$$\begin{aligned}((1-z)x_t + a_1(1-z)x_{t-1} + \cdots + a_p(1-z)x_{t-p}) &= \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q} \\ ((x_t - x_{t-1}) + a_1(x_{t-1} - x_{t-2}) + \cdots + a_p(x_{t-p} - x_{t-p-1})) &= \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q}.\end{aligned}$$

The autoregressive transfer function $A(z)$ in (6.52) models stationary autoregressive processes and so the ARIMA model represents a non-stationary process by reducing it to a stationary one through the difference operation $(1-z)^d$. This can be seen by assuming that the d^{th} backward difference of the non-stationary process x_t gives a stationary process w_t defined by

$$w_t = (1-z)^d x_t = \tilde{w}_t + \mu, \tag{6.54}$$

where μ is the constant mean of w_t and $\tilde{w}_t = w_t - \mu$, $\forall t$. The stationary, zero-mean process $\tilde{w}_t$ is now modeled by the ARMA(p,q) system,

$$A(z)\tilde{w}_t = B(z)\varepsilon_t. \tag{6.55}$$

From (6.54), $\tilde{w}_t = (1-z)^d x_t - \mu$, and substituting into (6.55) gives

$$A(z)(1-z)^d x_t = \mu_0 + B(z)\varepsilon_t,$$

where $\mu_0 = A(z)\mu = (1 + a_1 z + \cdots + a_p z^p)\mu$ is a constant [RS79, pp.314-315].

The ARIMA(p,d,q) model is adequate for modeling processes that are non-stationary in both the mean and variance. For processes that exhibit non-stationarity only in the variance, power transformations on the time series may generate a transformed series with constant variance. For example, the logarithmic transformation of a series, $\ln x_t$ (the base is irrelevant), will give a series with constant variance [Wei90, p.83]. Other transformations such as the square root transformation $\sqrt{x_t}$ are possible, and in general the class of such transformations are referred to as variance-stabilizing transformations.

In summary, the goal behind time series analysis is the derivation of a suitable (often parsimonious) model which adequately explains the stochastic nature of the time series observations. The primary application of such a model is forecasting, or prediction, of future behavior of the system under investigation. To this end the task of time series analysis attempts to model the process which generates the entire series over some interval in time.

6.6 Interrupted Time Series Experiments

Apart from prediction, another important application of time series analysis is within the framework of Interrupted Time Series Experiments (ITSE). Interrupted time series experiments involve measuring some quantity over time (hence generating a time series). At some point in time an *intervention* is introduced. The goal of the analysis is to determine whether the observed measurements differ before and after the intervention. If the pre- and post-intervention time series differ significantly, then it may be possible to conclude that the intervention had some *effect*. An example of this approach is often taken when testing the effectiveness of new drugs. Patients are given a placebo for some length of time, after which the drug is introduced. With the drug acting as the intervention, the pre and post-intervention time series are compared. If the post-intervention time series differs significantly, it may be concluded that the drug had some significant effect.

Interventions essentially appear as edges of various forms in a time series. Ten different forms of interventions are shown in Figure 33, where the dotted vertical line represents the onset of the intervention [Got81, p.50]. Interventions can in general be modeled by linear transfer functions. It is common to chose an appropriate transfer function based on an a priori expectation of the hypothesized effect [CC79, p.265]. Pre-

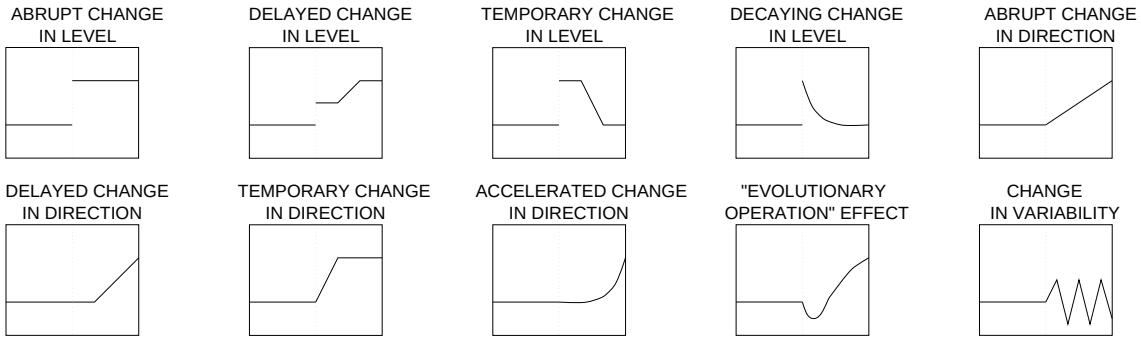


Fig. 33. Models of interrupted time series experiments interventions. Adopted from [Got81, p.50 (Fig. 6.6)].

and post-intervention time series are generally modeled as ARIMA systems allowing representation of non-stationary processes. In the following section the notion of a piecewise-ARIMA process is introduced and an automatic intervention detection method is presented based on wavelet multiscale edge detection. Interventions are assumed to be abrupt, time-limited changes in the mean of piecewise-ARIMA time series.

6.7 Piecewise Autoregressive Integrated Moving Average Time Series

Stochastic processes can in general be represented by autoregressive integrated moving average (ARIMA) models. In the context of interrupted time series experiments (ITSE), empirical observations are modeled by

juxtaposed ARIMA models separated by an expected intervention. The intervention is typically modeled by a transfer function based on a priori expectations. In this section, a model is developed for stochastic processes composed of multiple ARIMA processes demarked by compactly supported interventions. Each ARIMA segment is modeled by a bounded ARIMA process, while the interventions are modeled by an abrupt, time-limited change in the mean. The overall time series is thus composed of ARIMA "pieces" delimited by sharp interventions and is referred to as a piecewise-ARIMA time series, or PARIMA. The motivation behind PARIMA modeling is the automatic (computational) detection of the bandlimited interventions. In contrast to general ARIMA and ITSE modeling, the PARIMA model is characterized by the expected interventions and not by the ARIMA segments themselves. That is, if the interventions can be located, it is assumed that the time series segments between interventions are ARIMA sequences.

6.7.1 The Time-Bounded PARIMA Process

In this section the nature of the stochastic PARIMA model segments expected between interventions is presented. These segments are modeled by ARIMA systems within a given time interval. Formally,

$$\begin{aligned}
(1-z)^{d_l} x_{t_l} &= \mu_{t_l} + \sum_{k=0}^{\infty} h_k \varepsilon_{t_l - k} \\
&= \mu_{t_l} + (h_0 + h_1 z + h_2 z^2 + \cdots)\varepsilon_{t_l} \\
&= \mu_{t_l} + H(z)\varepsilon_{t_l} \\
&= \mu_{t_l} + \frac{B(z)}{A(z)}\varepsilon_{t_l} \\
&= \mu_{t_l} + \left(\frac{b_0 + b_1 z + b_2 z^2 + \cdots + b_{q_l} z^{q_l}}{a_0 + a_1 z + a_2 z^2 + \cdots + a_{p_l} z^{p_l}}\right)\varepsilon_{t_l},
\end{aligned}$$

for some time interval $t_l \in [a,b]$ with $d_l \geq 0$. That is, each PARIMA segment, within the l^{th} interval, is modeled by the finite-parameter ARIMA(p_l, d_l, q_l) model. Each segment is characterized by the familiar statistical properties of ARIMA processes, namely with interval mean,

$$E[x_{t_l}] = \overline{x_l} = \mu_{t_l} = \mu_l, \quad a \leq t_l \leq b,$$

variance,

$$E[(x_{t_l} - \overline{x_l})^2] = \sigma_l^2 \sum_{k=0}^{\infty} h_k^2, \quad a \leq t_l \leq b,$$

autocovariance,

$$\gamma_k(x_{t_l}) = \sigma_l^2 \sum_{j=0}^{\infty} h_j h_{j+k}, \quad \forall k, \quad a \leq t_l \leq b,$$

and autocorrelation,

$$\zeta_k(x_{t_l}) = \frac{\gamma_k(x_{t_l})}{\gamma_0(x_{t_l})} \quad \forall k, \quad a \leq t_l \leq b,$$

where the linear filter h_t is (possibly) unique within each interval. If the mean μ_l and variance σ_l^2 of the l^{th} PARIMA segment are constant, then the segment can be represented by the ARIMA$(p_l, 0, q_l)$, or ARMA(p_l, q_l), model of stationary processes.[3] The PARIMA(p_l, d_l, q_l) process is schematically depicted in Figure 34.

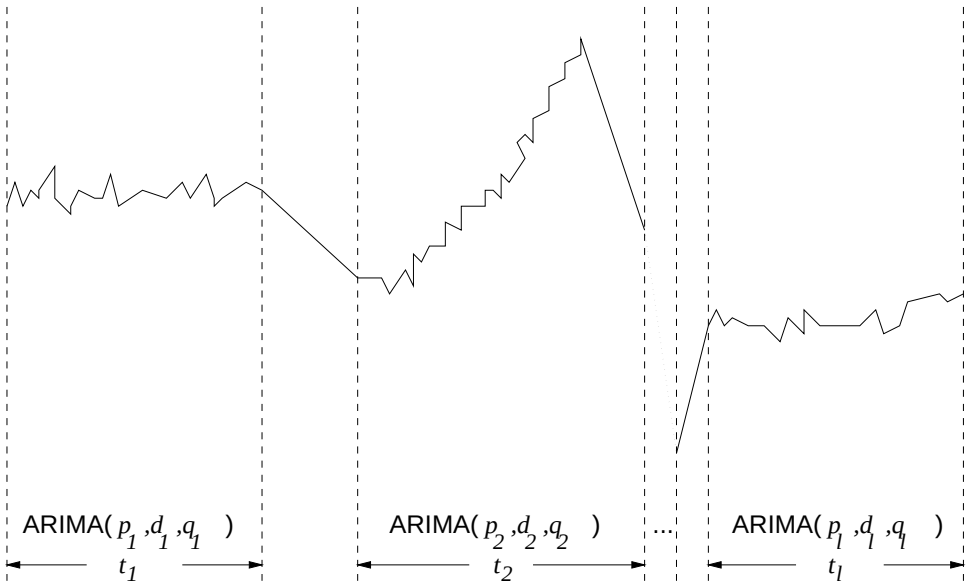


Fig. 34. Schematic depiction of PARIMA(p_l, d_l, q_l) model.

6.7.2 Automatic Intervention Detection

Interventions demarking ARIMA segments in the PARIMA model are assumed to occur within a bounded time interval. If the limited duration of interventions is known a priori, then the interventions can be modeled as multiscale edges and hence readily detectable by the wavelet transform. The specification of the interventions and their detection is described here.

Let S_l denote the ARIMA(p_l, d_l, q_l) sequence in the interval $a \leq t_l \leq b$, and let S_m denote the ARIMA(p_m, d_m, q_m) sequence in the interval $c \leq t_m \leq d$ where $a < b < c < d$. Let $\mathrm{I}_{l,m}(T)$ denote the intervention occurring in the interval $b < t_{l,m} < c$, so that $\mathrm{I}_{l,m}(T)$ demarks the sequences S_l, S_m where S_l precedes sequence S_m in time, as depicted in Figure 35. The duration of the intervention is denoted by T and is assumed to be bounded, i.e., $T_{min} \leq T \leq T_{max}$.

[3] To be consistent with accepted modeling practice, it is assumed that each PARIMA time series segment contains at least 50 observations.

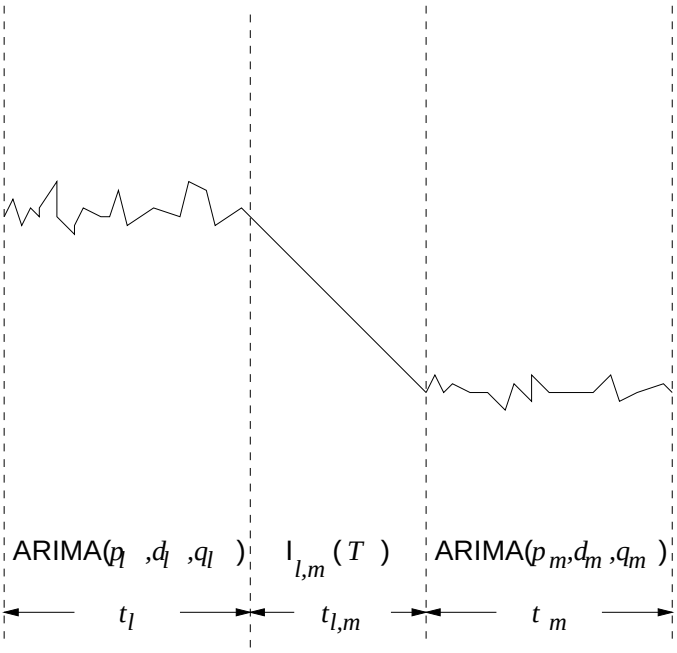


Fig. 35. Schematic depiction of PARIMA intervention model.

Theorem 2 *Assuming the observed PARIMA sequence is measured with a uniform sampling period s_p, and given length-2 scaling and wavelet functions ϕ, ψ, where ψ approximates the first derivative of a smoothing function θ, interventions of bounded duration T can be found in the wavelet transform at $j > log_2(\frac{s_p}{T_{min}}) + 1$ levels of the dyadic wavelet decomposition.*

Proof: The first part of Theorem 2 was proven by Mallat where it was shown that singularities of a function can be detected from the wavelet transform modulus maxima [MH92]. Mallat's result is based on the fundamental definition of the derivative, i.e.,

$$\frac{d}{dx}x_t = \lim_{k\to 0}\frac{x_{t+k} - x_t}{k},$$

where given the appropriate wavelet, the local wavelet modulus maxima corresponds to regions of highest slope, i.e., edges. By detecting local derivative maxima, in the limit, the most pronounced maxima correspond to singularities which are regions of infinite slope, i.e., step edges. The second part of the proof is based on sampling rate requirements. In order to sample a signal without aliasing artifacts, the sample rate must be at least twice the signal frequency. This is the well known Nyquist sampling frequency. For interventions of period T, this implies

$$s_r \quad > \quad 2\frac{1}{T}, \tag{6.56}$$

where $s_r = 1/s_p$ denotes the sampling rate. Rewriting (6.56) in terms of the sampling period,

$$s_p \quad < \quad \frac{1}{2}T. \tag{6.57}$$

Multiplying both sides of (6.57) by 2 gives

$$T > 2s_p,$$

indicating that the intervention period must be greater than twice the sample period. Under the dyadic wavelet transform, the signal is subsampled by the averaging scaling function ϕ at successive scales proportional to 2^j, i.e.,

$$2^j T > 2s_p. \tag{6.58}$$

Equation (6.58) represents the relationship between the scaled signal frequency and the sampling rate, showing resolution levels at which the signal is subsampled enough to eliminate aliasing artifacts. Solving for j,

$$\begin{aligned} 2^{j-1}T &> s_p \\ j &> \log_2(\frac{s_p}{T}) + 1 \end{aligned}$$

gives the decomposition levels where the signal is subsampled at frequencies greater than the Nyquist frequency. Since $T_{min} < T_{max}$, in order to detect interventions of period T, the signal must be decomposed to at least $j > \log_2(\frac{s_p}{T_{min}}) + 1$ level by the dyadic wavelet transform with sampling rate $s_r = 1/s_p$. □

As an example of the practical implication of Theorem 2, consider an intervention duration that is twice as long as the sampling period, i.e., $T = 2s_p$. In this case $j > \log_2(\frac{1}{2}) + 1 = 0$ which states that the first level of decomposition is sufficient for aliasing-free signal analysis. Since $T = 2s_p$, the sampling frequency is twice the signal frequency, $s_r = 2(1/T)$, i.e., the intervention is sampled at the Nyquist frequency so subsampling beyond the first level is not required. If, on the other hand, $T = \frac{1}{2}s_p$, then $j > \log_2(2) + 1 = 2$, which shows that more than two decomposition levels are required in order to eliminate aliasing. Referring to frequency, in this case $\frac{1}{2}(\frac{1}{T}) = s_r$ which states that the sampling frequency is half the signal frequency and must be increased by a factor of 4 to eliminate aliasing. In the case when $T = s_p$, to eliminate aliasing, the sampling frequency must be increased by a factor of 2, or equivalently, the signal must be subsampled by a factor of 2. In the context of PARIMA modeling, implications of Theorem 2 are best realized in the case where all interventions can be assumed to be of the same duration. In this case the wavelet transform provides a uniform partitioning method of the PARIMA sequence into multiple ARIMA sequences, with Theorem 2 giving the minimum number of decompositions required for intervention detection. If the expected intervention duration is short, only a few wavelet decomposition levels will be required easing the computational burden. Note that Theorem 2 does not guarantee that all interventions will be detected.

As an example of a PARIMA process, consider the sequence generated by the feedback ARMA sequence

$$x_{t_l} = \mu_{t_l} + x_{t_l-1} + \varepsilon_{t_l}, \quad a \le t_l \le b,$$

where $\varepsilon_{t_l} \sim N(0,(5/4)^2)$ was chosen arbitrarily with the intention of generating small perturbations about the mean for a given interval. Samples were generated uniformly every 18ms, with each constant-mean interval lasting an average of 375ms with a Poisson distribution. That is, interventions were distributed with an average inter-arrival time of 375ms. An intervention was induced by the instantaneous change of mean μ_{t_l}, distributed by the feedback relation

$$\mu_{t_l} = \mu_{t_{l-1}} + \varepsilon_\mu,$$

where $\varepsilon_\mu \sim N(0,128^2)$ was chosen arbitrarily with the intention of generating large perturbations. Thus for a given interval $a \leq t_l \leq b$, the sequence x_{t_l} is ARMA$(1,0)$ with constant mean μ_{t_l} and constant variance ε_{t_l}. Interventions are modeled by instantaneous changes in the mean for a new intervention arrival. The duration of the intervention is one sample period long, i.e., $T = 18$ms. The original time series is plotted in Figure 36 with overlayed detected interventions (dashed lines) at two decomposition levels (after thresholding—see below). Note that due to the inherent spatial decimation of the wavelet transform the modulus maxima plot has been scaled in the abscissa by a factor of 2^j. Notice also that not all interventions were detected at this scale ($j = 2$). This is due to aliasing since intervention and sampling periods coincide.

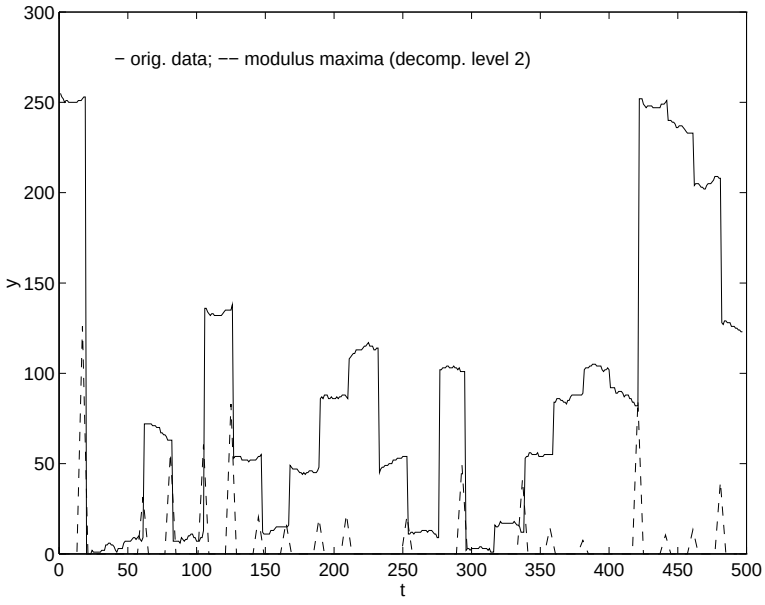


Fig. 36. PARIMA-modeled time series with modula maxima.

6.7.3 PARIMA Sequence Partitioning

Once the modulus maxima points corresponding to interventions have been detected in the wavelet transform, the ARIMA segments can be isolated by wavelet coefficient thresholding prior to reconstruction. The

intent is to substitute data points originally corresponding to interventions with easily distinguishable values, e.g., zero. This is accomplished by decimating wavelet and scaling coefficients that correspond to significant maxima locations. Upon reconstruction, decimated transform coefficients will generate near-zero values with negligible noise possibly being introduced due to neighboring regions.

Wavelet coefficient thresholding methods in the context of statistical analysis were recently popularized by Donohoe and Johnstone. These authors introduced two general thresholding rules, namely

$$\begin{aligned} T_{hard}(w;t) &= wI(|w| > t) \\ T_{soft}(w;t) &= sgn(w)(|w| - t)wI(|w| > t) \end{aligned}$$

where w refers to a wavelet coefficient and $t > 0$ is the given threshold value. Hard-thresholding is a 'keep-or-kill' strategy, while soft-thresholding is considered 'shrink-or-kill' [Nas94]. A particularly popular choice of threshold value is the *universal* threshold given by

$$t_{uv} = \hat{\sigma}_j \sqrt{2\log(n)},$$

where n is the overall sample size and $\hat{\sigma}_j$ is the sample wavelet deviation at scale j [JS94b]. In most thresholding applications the design of thresholding multipliers is aimed at reducing noise (see [DJKP96, CM95]). In the context of PARIMA models, the wavelet thresholding rule is based on existence of significant modulus maxima at the corresponding spatial location. Modulus maxima values are themselves first subject to thresholding in order to eliminate noisy edge artifacts from the eventual reconstruction.

Modulus maxima values may contain insignificant maxima, i.e., modulus maxima points pertaining to small spatially local perturbations, presumably due to noise. For this reason, a maxima thresholding filter is required prior to wavelet thresholding. This is accomplished by trimming (removing) insignificant maxima values at each scale of resolution. An interesting approach is prescribed by Carmona [Car93]. Within each resolution level j, a level of significance α is chosen in the interval $(0,1)$. The $100(1-\alpha)$ percentile, denoted by $h_j(\alpha)$, is computed from the histogram $\mathcal{H}_j$ of maxima values at resolution level j. If the absolute value of a wavelet maxima, denoted by $|M\{x_t\}(j)|$, is smaller than $h_j(\alpha)$, the maxima at the current location is suspected of being the result of the noise component of the given signal and is decimated. The problem of the histogram approach is in the specification of the width of the histogram intervals. Furthermore, any statistical measure of the modula maxima values is usually dependent on an assumption of the values' distribution (with the normal distribution being a frequent candidate). In the case of the modula maxima, the distribution is not known. It is known, however, that due to the selection of maxima most of the energy in the modulus maxima signal is devoted to singularities (edges) in the original signal. Therefore, a pragmatic alternative to the histogram approach is the decimation of maxima values of small amplitude. This is readily accomplished by the hard

thresholding

$$T_{hard}[M\{x_t\}(j)] = M\{x_t\}(j)I(|M\{x_t\}(j)| > \alpha\ddot{M}\{x_t\}(j)),$$

where $\ddot{M}\{x_t\}(j)$ denotes the range of maxima values at level j. Since only low-amplitude values need to be decimated, α should be small. With $\alpha = 0.05$ for example, $100(1-\alpha)\%$ of the range of values is preserved. Denoting the threshold parameter by $t_M = \alpha\ddot{M}\{x_t\}(j)$, adds another parameter to the PARIMA model. The trimmed maxima values are used within the indicator function for wavelet coefficient thresholding prior to reconstruction.

Given trimmed modulus maxima information, wavelet coefficients are hard-thresholded so that the transform reconstruction yields zero values in place of interventions. Soft-thresholding of the coefficients (wavelet shrinkage) would result in wavelet interpolation yielding smoothed intervention regions. Since the goal is to isolate ARIMA sequences between interventions, wavelet coefficients are hard-thresholded (decimated) by the following rule

$$T_{hard}[W\{x_t\}(j)] = W\{x_t\}(j)I(|M\{x_t\}(j)| > 0),$$

where, at location t and scale j, $W\{x_t\}(j)$ and $M\{x_t\}(j)$ denote the wavelet coefficient and modulus maxima, respectively. Thresholding is also applied to the scale coefficients ensuring that thresholded regions yield near-zero values upon reconstruction. Reconstruction generates noise due to the influence of neighboring wavelet coefficients. The level of resultant noise is dependent on the length of the reconstruction kernels. That is, more noise will be generated by wavelet and scaling filters of longer length.

The reconstructed time series, from scale $j = 2$ modula maxima detection, is plotted in Figure 37. Isolated segments are presumed to be ARIMA sequences and are subject to individual ARIMA model specification as discussed in §6.5 and §6.4.

6.7.4 PARIMA Model Parameters

The proposed PARIMA model is characterized by compactly-supported interventions between ARIMA sequences. The PARIMA model is described by the following model parameters:

- T: average (expected) intervention duration; only one parameter is required for homogeneous interventions, requiring a search for all interventions at a constant number of resolutions of the wavelet transform. Inhomogeneous interventions increase the complexity of the PARIMA model drastically and must be modeled by a variable duration describing each intervention necessitating detection of interventions at various levels of the wavelet transform. The intervention duration parameter T determines the wavelet transform decomposition level j.

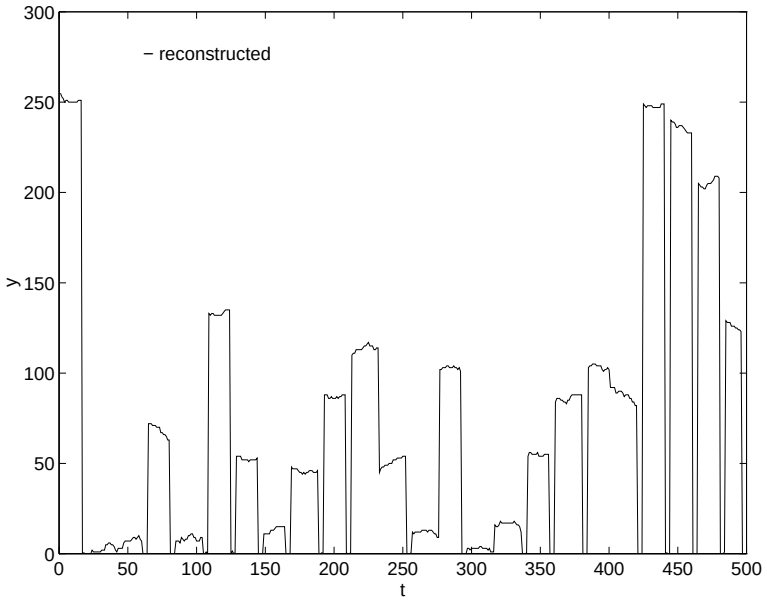


Fig. 37. Partitioned PARIMA-modeled time series.

- $\{p_l, d_l, q_l\},\ l \in [1, n]$: ARIMA parameters describing each segment between the $n+1$ detected interventions. These are implicit parameters in the PARIMA model. Each triple must be estimated by traditional ARIMA modeling techniques (e.g., Box-Jenkins). In all, $3n$ ARIMA parameters are required to model the entire observed PARIMA sequence. If certain ARIMA subsequences are assumed to be generated by a single process, the PARIMA model complexity may be reduced. For example, if it is assumed that the PARIMA system is a combination of two competing ARIMA processes interrupted by an intervening process, then the PARIMA model will require only 6 ARIMA parameters, namely the two sets $\{p_1, d_1, q_1\}$, and $\{p_2, d_2, q_2\}$.
- t_M: modula maxima thresholding parameter. The threshold value is determined by fixing the value of the significance level α. This value is required to be small.
- ϕ, ψ: scaling and wavelet functions, respectively. The choice of these functions influences the intervention detection mechanism. The wavelet ψ must approximate the first derivative of a smoothing function, and is recommended to be length-2. For these reasons the Haar wavelet is a suitable choice.

CHAPTER VII

EYE MOVEMENT MODELING

In the development of a gaze-contingent system, a model of eye movements is necessary for the exploration of vision and its underlying visual stimuli. The need here is to confidently classify eye movements within natural human viewing patterns. Assuming eye movements composed of dynamic fixations (i.e., proper fixations and smooth pursuit movements) denote overt locations of visual attention, localization of these features is crucial to a gaze-contingent analysis and synthesis of visual information.

Due to its simplicity and ease of implementation, a particularly attractive strategy for eye movement modeling involves linear time-invariant (LTI) filtering. Previous eye movement classification strategies utilizing linear filters are briefly discussed in this section. A conceptual Piecewise Auto-Regressive Integrated Moving Average (PARIMA) model of conjugate eye movements is then proposed. The PARIMA model is a piecewise-LTI representation of stochastic signals. The analytical framework of the PARIMA model features a wavelet-based strategy for eye movement segmentation. Implementational issues are discussed, and a video frame-based technique is offered for classification of eye movements into smooth pursuits, fixations, and saccades.

7.1 Linear Filtering Approach to Eye Movement Classification

Classification of eye movement data through linear filtering has been extensively studied, and is not without controversy. Simplified linear models of eye movements are attractive for eye movement data partitioning due to the availability and ease of use of linear filters. The primary objective is to identify eye movements in terms of their signal characteristics.

In contrast to the linear models of the oculomotor system described in §IV, here the goal is to specify filters describing the observed (external) signal characteristics of eye movements. The approach taken for the specification of these filter models follows time series modeling, where the goal is to specify a filter such that given an input signal, s_t, the output, x_t, resembles white noise.[1] Due to the Slutzky effect (see §VI), if such a filter can be found, then the inverse (sometimes called reverse) filter will generate the observed time series given white noise as its input. In this case the filter (or its inverse) completely describes the observed time series. Hence, the goal of eye movement signal methodology presented here is to model the observed signal

[1]Note that in §IV these symbols were reversed so that s_t denoted noise (input to the oculomotor system) and x_t symbolized the output of the system (eye movements).

by an inverse of the filters used to represent the oculomotor system. This modeling strategy is represented in Figure 38 (note the dual use of symbols s_t and x_t).

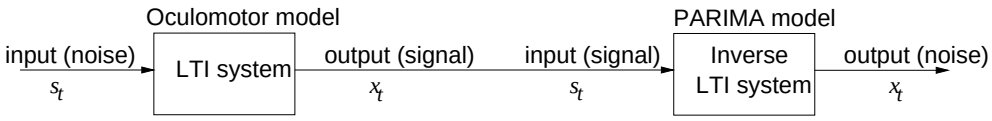


Fig. 38. Linear filter modeling strategy.

A major controversy surrounding eye movement modeling through linear filtering centers on the inadequate representation of the nonlinear nature of the oculomotor system (see §7.6.1 below). Contemporary oculomotor models attempt to explain the inherent nonlinearities of the neural substrate, although linear methods for eye movement identification are still in use today [Car96]. In spite of their simplistic representation of the the underlying processes, linear filtering strategies provide a good first approximation of the neural controller signal [WNS84].

Buizza and Avanzini reject frequency domain analysis in favor of the time domain [BA83a]. The authors are primarily interested in smooth pursuit signals and remove the saccadic component by linearly interpolating between saccade onset and termination.

Cabiati et al. present a real-time saccade detection algorithm through linear high-pass filtering [CPSZ83]. Denoting the observed eye movement signal by the input s_t, and the LTI filter output by x_t, the high pass filter is defined as

$$x_t \quad = \quad \frac{1}{n}\sum_{-k}^{k} g_k s_{t-k},$$

where $k \in [-3,3]$ with coefficients $\{-3,-2,1,0,1,2,3\}$ and $n = 2k+1$.

Karsh and Breitenbach concentrate on fixations by partitioning raw eye movement data [KB83]. The critical step in their implementation concerns the data sampling rate and the choice of number of data points (cluster size) needed for fixation classification. The onset of fixation is determined through an averaging process. Using a 60Hz sampling rate, the authors show varying scanpath interpretations with cluster sizes of 2–8 points and suggest 6 data points as suitable.

Shebilske and Fisher warn of the dangers of choosing an inappropriate cluster size in the context of reading studies [SF83]. A cluster size of 2, for example, tends to overestimate the number of words fixated and un-

derestimates fixation durations. Shebilske and Fisher make a point of differentiating between global effects (e.g., statistics over the whole page) and local effects (e.g., statistics over a single sentence) when evaluating reading performance.

A specific aim of the present investigation is the evaluation of the wavelet approach for eye movement classification. Due to the inherent linear nature of the wavelet transform, it is undoubtedly an oversimplification of the underlying process. Nevertheless, Gyaw and Ray suggest the feasibility of using wavelet transform zero-crossings as a tool for classification of biosignal patterns [GR94]. The authors use DWT zero crossings as studied by Mallat (see §5.3) to characterize Electro-Cardiogram (ECG) signals. The present wavelet-based technique for saccade localization is based on Mallat's strategy for detection of modulus maxima instead of zero crossings. Similar to the high-pass filtering saccade detection of Cabiati et al., the present model uses a linear filtering approach for saccade localization, utilizing the Haar wavelet is used at multiple scales (see below).

In contrast to Buizza and Avanzini's model, the present modeling strategy operates partially in the frequency domain through wavelet analysis. The main advantage of wavelets over traditional frequency domain methods (e.g., Fourier transform) is the spatial localization property of the compactly supported wavelet filters (see §V). Similar to Buizza and Avanzini's model, data corresponding to saccades is removed. Unlike their model, interpolation between saccade onset and termination is not performed in the analysis.[2]

The present wavelet-based PARIMA model identifies (dynamic) fixations by assuming these patterns are delineated by saccades. In a sense, fixation identification follows a deductive argument, i.e., with the saccadic component removed, the remaining signal is assumed to be composed of all other types kinds of eye movements, e.g., fixations (defined by noise-like miniature movements), smooth pursuits, or the slow phase of nystagmus. This deductive strategy alleviates the concerns raised by Shebilske and Fisher regarding cluster size allocation for fixation detection. Cluster size determination is instead relegated to saccade detection, where signal identification is dependent on the chosen temporal interval.

The goal of the proposed model is to detect dynamic fixations in eye movement data. As such, the model's purpose is one of pattern recognition. Although criteria for eye movement patterns are derived from known characteristics of the oculomotor system, the objective is not a model of the neural substrate itself. Rather, the proposed model is a (dynamic) fixation algorithm based on the detection of saccades.

[2]Interpolation is performed, however, in the Volume Of Interest visualization of eye movements (see §VIII), and in the preattentive VOI strategy for video processing (see §XIII and §13.1.1). Inter-saccade signal interpolation is not directly related to the present description of the model.

7.2 Conceptual Specification of the PARIMA Model

In the proposed PARIMA eye movement model, the three principal types of eye movements (saccades, fixations, and smooth pursuits) are identified through the detection of saccades. Saccades are modeled as time-limited mean discontinuities of the time series with intervention duration $E[T] = [10, 100]$ms. The expected saccade duration of 10–100 ms is based on reported saccade characteristics [Fin92]. Fixations and smooth pursuits are modeled as two competing ARIMA processes denoted by the parameters $\{p_f, 0, q_f\}$, $\{p_s, d_s, q_s\}$, respectively. Note that fixations are modeled as ARMA sequences. All other types of eye movements, e.g., microsaccades, tremors and shifts, are assumed to be noise contained within the three principal movement types and are not explicitly recognized by the proposed model.

Considering the feedback nature of the oculomotor system, and using the linearity assumption for observed eye movement signals, the proposed PARIMA linear filter models conjugate eye movements as shown in Figure 39. The output of the PARIMA model x_t is symbolized by a circuit switch representing the neural

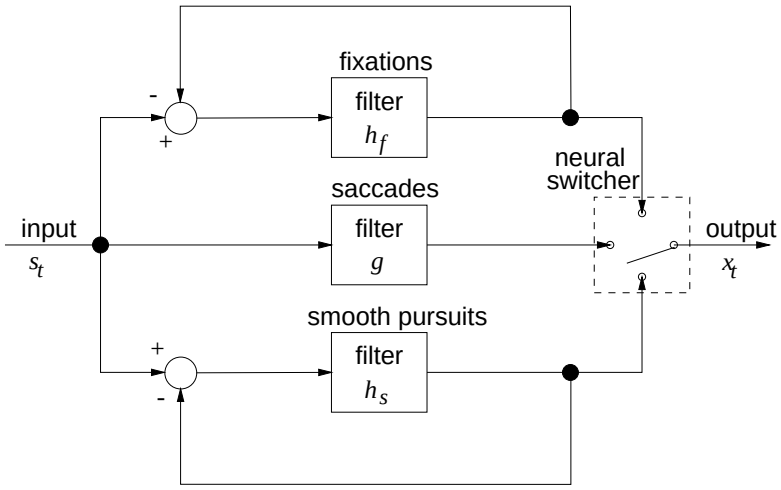


Fig. 39. Block diagram of a simple linear system modeling conjugate movements.

integrator. Modeling the oculomotor system, the *neural switch* selects the required type of eye movement resulting in a temporally integrated multi-component signal. From a signal processing perspective, at any given time, the active filter describing the current signal segment is the one generating white noise. In the oculomotor system, the input signal s_t is characterized by white noise, and the filter generating the signal x_t is driven by an internal copy of the visual environment. In the signal model, s_t represents the observed signal. Provided the appropriate filter representing the inverse of the oculomotor counterpart is active, the generated output x_t will resemble noise. Conjugate movements, such as vestibular or optokinetic nystagmus, are temporally represented by "switching" the system between saccades and smooth pursuits. This type of

saccadic bypass of the pursuit feedback coincides with proposed mechanisms of neural integration based on the simplified assumption of linear summation (see for example [Car77, p.288]).

7.2.1 Fixations

Fixations are characterized by the miniature eye movements tremor, drift, and microsaccades. In the PARIMA model representation, the stochastic signal corresponding to fixations is described by the linear time-invariant filter h_f. Miniature eye movements correspond to the variance of the system. Specifically, the output of the system x_t is assumed to be white noise. In the time domain, fixations are modeled by the following equation

$$\begin{aligned} x_t &= \sum_{k=0}^{\infty} h_{f_k}(s_{t-k} - x_{t-k}) \\ &= h_{f_0}(s_t - x_t) + h_{f_1}(s_{t-1} - x_{t-1}) + h_{f_2}(s_{t-2} - x_{t-2}) + \cdots, \end{aligned} \tag{7.1}$$

or in the z-domain,

$$X(z) = H_f(z)(S(z) - X(z)).$$

Supposing that the filter H_f represents a stable and causal system, e.g.,

$$H_f(z) = h_{f_0} + h_{f_1} z + h_{f_2} z^2 + \cdots,$$

then the filter $H_f(z)$ can be represented by a rational function in z or ARMA model of the form

$$\begin{aligned} H_f(z) &= \frac{B_f(z)}{A_f(z)} \\ &= \frac{b_{f_0} + b_{f_1} z + b_{f_2} z^2 + \cdots + b_{f_q} z^q}{a_{f_0} + a_{f_1} z + a_{f_2} z^2 + \cdots + a_{f_p} z^p}. \end{aligned} \tag{7.2}$$

The ARMA coefficients $\{a_f\}$ and $\{b_f\}$ can be obtained through Padé approximation (see [RS79, §4.11] for the derivation and subsequent algorithm). Substituting the rational approximation of Equation (7.2) in Equation (7.1) gives

$$\begin{aligned} X(z) &= H_f(z)(S(z) - X(z)) \\ &= \frac{B_f(z)}{A_f(z)}(S(z) - X(z)). \end{aligned} \tag{7.3}$$

Expanding Equation (7.3) in the time domain,

$$\begin{aligned} a_{f_0} x_t + a_{f_1} x_{t-1} + \cdots + a_{f_p} x_{t-p} = \\ b_{f_0}(s_t - x_t) + b_{f_1}(s_{t-1} - x_{t-1}) + \cdots + b_{f_q}(s_{t-q} - x_{t-q}), \end{aligned}$$

gives

$$\begin{aligned} a_{f_0} x_t + a_{f_1} x_{t-1} + \cdots + a_{f_p} x_{t-p} + b_{f_0} x_t + b_{f_1} x_{t-1} + \cdots + b_{f_q} x_{t-q} = \\ b_{f_0}(s_t) + b_{f_1}(s_{t-1}) + \cdots + b_{f_q}(s_{t-q}), \end{aligned}$$

which is succinctly written in summation form as

$$\sum_{k=0}^{p} a_{f_k} x_{t-k} + \sum_{k=0}^{q} b_{f_k} x_{t-k} = \sum_{k=0}^{q} b_{f_k} s_{t-k}, \tag{7.4}$$

resulting in the general ARMA model of fixations in terms of the autoregressive (AR) coefficients $\{a_f\}$, and the moving average (MA) coefficients $\{b_f\}$. This model is succinctly represented by the notation ARMA(p_f,q_f) where p_f and q_f denote the number of AR and MA coefficients, respectively, for each identified fixation segment. That is, each fixation is represented by a (possibly) different number of coefficients.

The rational approximation of $H(z)$ and hence the ARMA(p_f,q_f) model is a parsimonious simplification of the inverse simple linear fixation model described in §IV. Recall that fixations there are modeled by an essentially identical linear feedback system as that used for smooth pursuits, except for the implicit assumption of stationarity of fixations. Save for this constraint, the smooth pursuit feedback system is derived from the ARMA(p_f,q_f) fixation model by examining Equation (7.4) in the z-domain

$$\begin{aligned} \sum_{k=0}^{p} a_{f_k} x_{t-k} + \sum_{k=0}^{q} b_{f_k} x_{t-k} &= \sum_{k=0}^{q} b_{f_k} s_{t-k} \\ A(z)X(z) + B(z)X(z) &= B(z)S(z) \qquad (7.5) \\ X(z)(A(z)+B(z)) &= B(z)S(z) \\ \frac{X(z)}{S(z)} &= \frac{B(z)}{A(z)+B(z)}. \qquad (7.6) \end{aligned}$$

The collection of terms in Equation (7.5) is performed by appropriately padding summation terms if $p \neq q$. For example, if $p < q$ then the summation involving the $\{a_{f_k}\}$ coefficients is padded with $q-p$ zero terms. The analogous operation is used for the summation involving the $\{b_{f_k}\}$ coefficients if $q < p$. Multiplying by $1/A(z)$ both the numerator and denominator of the right hand side of Equation (7.6) gives

$$\begin{aligned} \frac{X(z)}{S(z)} &= \frac{\frac{B(z)}{A(z)}}{1+\frac{B(z)}{A(z)}} \\ &= \frac{H(z)}{1+H(z)}, \qquad (7.7) \end{aligned}$$

which represents the noise-to-signal ($X(z)/S(z)$) ratio of the system. This is the inverse filter of the linear feedback model of smooth pursuits discussed in §IV where $X(z)/S(z)$ represented signal-to-noise.

The ARMA(p_f,q_f) model tacitly assumes mean stationarity of the signal. The stationarity assumption can be denoted explicitly by extending the ARMA(p_f,q_f) model to ARIMA notation, with the numeral 0 standing in for the parameter d_f, i.e., ARIMA(p_f,0,q_f). The assumption of a stationary mean reflects the expected temporal clustering of observed measurements of true fixations about the point of regard.

7.2.2 Smooth Pursuits

The stationarity assumption invoked in the ARMA(p_f,q_f) model of fixations is relaxed in modeling smooth pursuit movements. The feedback model is derived as for the fixation model, with coefficients $\{a_s\}$ and $\{b_s\}$ distinguished by subscript s. The number of coefficients is also distinct from the fixation model denoted by parameters p_s, q_s. Since stationarity cannot be assumed for smooth pursuits, the parameter d_s is made explicit, giving the full ARIMA(p_s, d_s, q_s) model designation. The resulting simple linear feedback filter's transfer function is identical to the one given in Equation (7.7).

The only difference between the fixation and smooth pursuit models (apart from distinct model parameters) is the imposed condition of stationarity on the fixation signal. This expectation is reflected by setting the implicit d_f parameter to zero. For smooth pursuits, this parameter is necessarily non-zero since a trend in the signal mean is expected. The parameter d_s specifies the number of "differencing" operations required on the signal in order to eliminate this trend. In the Box-Jenkins approach, a signal showing a trend in the mean (an ARIMA process) is differentiated enough times so that it can be adequately described by an ARMA model (see §VI).

In applying the above model to the analysis of eye movements, the parameters $\{p_f, 0, q_f\}$ and $\{p_s, d_s, q_s\}$ for fixations and smooth pursuits, respectively, need not ever be explicitly determined. Since observed eye movement data is an expected conjugate signal composed of fixations, smooth pursuits, and saccades, the full analysis of the signal requires identification of the three signal components followed by the determination of parameters for each component. With respect to the 5 parameters $\{p_f, 0, q_f\}$ and $\{p_s, d_s, q_s\}$, the complexity of a full analysis increases r-fold with $r = m + n$ where m, n are the number of classified fixations and smooth pursuits, respectively. That is, each fixation instance requires the determination of a distinct set of coefficients. Although all fixations may be approximated by the same linear filter system, it is unlikely that the same filter coefficients can be used for each fixation segment. Thus, if in a given finite eye movement sequence m fixation instances are identified, m sets of coefficients will be required, one for each signal segment corresponding to a fixation. Similarly for smooth pursuit sequences. For this reason, the fixation and smooth pursuit parameters $\{p_f, 0, q_f\}$ and $\{p_s, d_s, q_s\}$ are never calculated. These parameter tuples are only used to conceptually distinguish fixations from smooth pursuits. In practice, smooth pursuits are simplistically regarded as dynamic fixations, i.e., non-stationary vs. stationary.

For temporal visualization of eye movements and hence visualization of overt visual attention, the goal of the analysis is the localization of the dynamic course of foveal vision. Identification of (dynamic) fixations follows a deductive argument. Since eye movements are assumed to be generated by a "switched circuit" depicted in Figure 39, it follows that elimination of one of the circuit paths will result in a signal characterized

solely by the remaining components. In essence, the goal of identification of dynamic fixations reduces to the localization of saccades.

7.2.3 Saccades

The ARMA model of saccades is a linear filter designed to detect short-term time series interventions. Specifically, the filter is designed to detect edges, or step functions, in time. The ARMA linear system is fully specified by the filter g in the time domain,

$$\begin{aligned} x_t &= \sum_{k=0}^{\infty} g_k s_{t-k} \\ &= g_0 s_t + g_1 s_{t-1} + g_2 s_{t-2} + \cdots, \end{aligned} \tag{7.8}$$

or in the z-domain,

$$X(z) = G(z)S(z),$$

with transfer function given as the noise-to-signal ratio,

$$\begin{aligned} \frac{X(z)}{S(z)} &= G(z) \\ &= g_0 + g_1 z + g_2 z^2 + \cdots. \end{aligned}$$

By choosing the Haar wavelet with coefficients $\{1/\sqrt{2}, -1/\sqrt{2}\}$, the transfer function becomes

$$\begin{aligned} \frac{X(z)}{S(z)} &= G(z) \\ &= \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}} z, \end{aligned}$$

which is a scaled inverse of the filter modeling saccades in the oculomotor plant (see §IV). In the time domain, the filter modeling the observed signal is specified by the linear moving average (MA) model,

$$\begin{aligned} x_t &= g_0 s_t + g_1 s_{t-1} \\ &= \frac{1}{\sqrt{2}} s_t - \frac{1}{\sqrt{2}} s_{t-1}, \end{aligned} \tag{7.9}$$

or an ARMA(0,0,1) sequence. In practice, Equation (7.9) is applied at a diminished temporal scale over the subsampled signal. The temporal scale is governed by the expected duration of saccades (10-100ms) and on the data sampling rate.

7.2.4 Wavelet Model of Temporal Time Series

The proposed framework for temporal analysis of eye movement time series is the Discrete Wavelet Transform (DWT), described in §V and §VI and specifically in §5.3, §6.7 and §6.7.2. The DWT is chosen for

its spatiotemporal localization property. In this section, a general wavelet-based technique is presented for temporal saccade detection within a time series signal representation of eye movements. Implementation strategies are recommended in the following section.

Recall the wavelet transform of s_t at scale j and position (time) t is the convolution product

$$\{W_\psi s_t\}(j) = s_t * \psi_j(t),$$

with wavelet ψ and implicit translation parameter k. Saccades corresponding to sharp variation points are detected by finding the local maxima of the modulus $|\{W_\psi s_t\}(j)|$, assuming the wavelet ψ approximates the first derivative of a smoothing function (see §5.3). This criterion is satisfied with the choice of the Haar wavelet. At each scale j, local modulus maxima are located by finding the points where $|\{W_\psi s_t\}(j)|$ is larger than its two closest neighbor values, and strictly larger than at least one of them [MH92]. That is, a modulus maxima $M\{s_t\}(j)$ is located at scale j and location t if:

$$|\{W_\psi s_{t-1}\}(j)| \leq |\{W_\psi s_t\}(j)| \geq |\{W_\psi s_{t+1}\}(j)|, \text{ and}$$

$$\begin{cases} |\{W_\psi s_t\}(j)| > |\{W_\psi s_{t-1}\}(j)|, & \text{or} \\ |\{W_\psi s_t\}(j)| > |\{W_\psi s_{t+1}\}(j)|. \end{cases}$$

Modulus maxima values are subject to the hard thresholding rule,

$$T_{hard}[M\{s_t\}(j)] = M\{s_t\}(j) I(|M\{s_t\}(j)| > \alpha \ddot{M}\{s_t\}(j)),$$

where $\ddot{M}\{s_t\}(j)$ denotes the range of maxima values at level j, with modulus maxima threshold parameter $\alpha = 0.05$. To yield zero values in the time series at the location of interventions upon reconstruction, i.e., to isolate the ARIMA sequences between interventions, wavelet coefficients are hard-thresholded (decimated) by the following rule

$$T_{hard}[\{W_\psi s_t\}(j)] = \{W_\psi s_t\}(j) I(|M\{s_t\}(j)| > 0),$$

where, at location t and scale j, $\{W_\psi s_t\}(j)$ and $M\{s_t\}(j)$ denote the wavelet coefficient and modulus maxima, respectively.

To complete the specification of the saccade detection model, the Haar scaling function is used for temporal decomposition, and the wavelet transform decomposition level j is derived from the eye movement sampling rate. The current experimental apparatus operates with an average eye movement sample period of $s_p = 18$ms, giving a temporal decomposition level of

$$j > \log_2\left(\frac{s_p}{T_{min}}\right) + 1$$

$$> \quad 1$$
$$\geq \quad 2.$$

7.3 Implementation Recommendations

The goal of the proposed model is to detect dynamic fixations in eye movement data. As such, the model's purpose is one of pattern recognition. Although criteria for eye movement patterns are derived from known characteristics of the oculomotor system, the objective is not a model of the neural substrate itself. Rather, the proposed model is a (dynamic) fixation algorithm based on the detection of saccades.

A number of computational strategies are available for saccade detection, each dependent on an appropriate representation of the raw eye movement data. Defining raw Point Of Regard (POR) eye movement data by tuples $p_i = (x_i, y_i, t_i)$, for $i \in [1, n]$, the set of data samples $\{p_n\}$ defines a three-dimensional eye movement trajectory in space-time. The choice of an appropriate strategy for trajectory partitioning (e.g., through saccade detection) depends on the trajectory's mathematical representation. Different methodologies may be suitable for this task. For example, the samples may be approximated by B-Splines subject to a regularization constraint, where either time is used as the parametric variable, e.g.,

$$x = s_x(t), \quad y = s_y(t),$$

or the curve is parameterized by an arbitrary variable u, e.g.,

$$x = s_x(u), \quad y = s_y(u), \quad t = s_t(u),$$

where s represents the spline. The former strategy emphasizes the dynamic form of the three-dimensional curve, where saccade onsets and terminations are represented by a cluster of control knots proportional to the degree of saccade amplitude. For example, a sharp discontinuity corresponding to a saccade will require the insertion of several knots, proportional to the spline's smoothing parameter. The latter representation emphasizes the geometric form of eye movements in x, y, t. In this case, a saccade will typically have a parameter space (i.e., u) distance between the beginning and end of the saccade proportional to its arc length. Numerous automatic methods are available for spline curve fitting, for example see [Die93].

Alternative representations of the eye movement trajectory rely on a direct mapping of the raw data tuples, i.e., without *a priori* curve approximation. First, vectors can be used to represent the sampled points, e.g., (x_t, y_t), in which case the PARIMA time series analysis applies directly in a multivariate sense. Second, the sampled data can be mapped to image sequence frames $f(x_i, y_i, t_i)$, defined as characteristic functions of the

sample points:

$$f(x_i, y_i, t_i) \quad = \quad \begin{cases} 255 & \forall\, x_i, y_i, t_i, \;\; i \in [1, n], \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the application of wavelet-based image processing strategies for spatiotemporal edge detection, discussed in §V, is straightforward.

The PARIMA model of eye movements is independent of the representation of the data. The model's conjugate description of eye movements can be used to guide eye movement trajectory segmentation in all representations, although the implementation of the saccade detection method is different. In the latter two representations (vectors and frames), the LTI wavelet-based filtering approach is directly applicable.

In all but the frame-based approach, synchronization between eye movement samples and stimulus video frames is implicit. The frame-based approach provides the flexibility to vary the temporal distribution of characteristic functions (i.e., the frames) at the cost of reduced temporal resolution. That is, data samples are temporally pooled on frames depending on the ratio $r = s_f / s_p$ where s_p is the data sampling period, and s_f is the inter-frame period (inverse of frame rate). For example, if the eye movement data is sampled at a period of 18ms, but frames are distributed at a rate of 16fps (inter-frame period 62.5ms), then $r \approx 3.5$, meaning that 3-4 data samples will be pooled per frame. The mapping of sample points on frames is then expressed as:

$$f(x_i, y_i, \left\lfloor \frac{t_i}{t_f} \right\rfloor) = \begin{cases} 255 & \forall\, x_i, y_i, t_i, \;\; i \in [1, n] \\ 0 & \text{otherwise,} \end{cases} \tag{7.10}$$

where the fraction $\lfloor t_i / t_f \rfloor$ maps the sample data points' time stamps onto frame indices denoted by t_f. For example, if $t_i = \{0, 18, 36, 54, 72, 90, \ldots\}$ represents the time stamps of the first 6 data samples, then the first four samples are mapped onto the first (zeroth) frame, and the next two samples are mapped onto the next frame, e.g., the corresponding frame indices are $t_f = \{0, 0, 0, 0, 1, 1, \ldots\}$.

In the current implementation, the frame-based approach is chosen in order to synchronize eye movement sample data with stimulus video frames. This implementational decision sacrifices temporal resolution for ease of representation of the eye movement data in stimulus video frame coordinates.

7.3.1 Frame-based Implementation of the PARIMA Model

To facilitate computational means of fixation detection through wavelet-based image sequence analysis, raw eye tracker data is composed into a 16fps video sequence where the eye tracker data is represented by white pixels on a black background. Video frame resolution matches the eye tracker resolution (currently 512×256 at 60Hz). In video format, eye tracker data is submitted to anisotropic 3D wavelet transform analysis. Due to limited computational resources, analysis is limited to 128 frames (8 seconds worth) of data. Eye movements

are considered multivariate time series where spatiotemporal eye movement samples are represented by the characteristic function defined by Equation (7.10) above. Equation (7.10) specifies the mapping of raw POR data onto initially empty video frames composed of 0-intensity pixels.

It is assumed values $f(x_0, y_0, t_0)$ are correlated with their neighbors $f(x_0 + \Delta x, y_0 + \Delta y, t_0 + \Delta t)$, over some measurably small volume $(\Delta x, \Delta y, \Delta t)$. Numerous multivariate time series analysis methods exist, including tests for normality, statistical linear vs. non-linear independence, and the specification of non-linear models. In the present context, the primary concern is the detection of dynamic fixations within the eye movement series. To this end, eye movements are tacitly assumed to be non-stationary, statistically linearly dependent signals, where fixations and smooth pursuit movements are modeled as Auto-Regressive Integrated Moving Average (ARIMA) sequences delineated by discontinuities of limited duration. In other words, eye movements are modeled as piecewise-ARIMA (PARIMA) sequences where dynamic fixations are localized between discontinuous saccades. Smooth pursuit movements and fixations are distinguished as ARIMA and ARMA sequences, respectively, i.e., fixations are characterized by a stationary mean, whereas the mean of smooth pursuit movements is assumed to be nonstationary.

7.4 Three-dimensional Considerations in the Frame-Based Implementation

Extending the general one-dimensional wavelet time series model of eye movements to the frame-based implementation in three dimensions requires a spatial decomposition step prior to temporal analysis. Spatial decomposition is required to overcome the frame-to-frame correspondence problem of single-pixel raw eye tracker locations. Temporal analysis of the DWT is carried out on a per pixel basis between video sequence frames. The goal of the temporal analysis is to locate discontinuities occurring between frames. In essence, by applying the wavelet filter between frame pixels, discontinuities are located in the transform by finding high amplitude wavelet coefficients (i.e., pixels of value over a given threshold). In general, a pixel value exceeds the threshold only if there is a significant intensity change between the pixel location in two successive frames, e.g.,

$$f(x,y,t) - f(x,y,t+1) > T.$$

The difference between successive frame pixels is expressed by the wavelet coefficients, given the appropriate choice of wavelet function (e.g., the Haar wavelet). Two successive pixels generally present the following cases:

1. $f(x,y,t)$: black, $f(x,y,t+1)$: black
2. $f(x,y,t)$: black, $f(x,y,t+1)$: white
3. $f(x,y,t)$: white, $f(x,y,t+1)$: black
4. $f(x,y,t)$: white, $f(x,y,t+1)$: white

where a white pixel represents raw eye tracker data, the so-called Point Of Regard, or POR. Case 1 represents a steady black ‘background’ where no POR was recorded, i.e., no eye movement occurred across this location. Case 4 represents a steady white ‘foreground’ suggesting a steady eye movement (at least between these two frames). Both cases 1 and 4 will not be identified as a discontinuity since there is no change between pixel values from frame to frame.

Consider for the moment the simplistic case of ‘perfect’ eye movements composed of only ‘perfect’ (e.g., noise-free) fixations and ‘perfect’ saccades, where fixations do not vary from pixel-to-pixel over time. Saccades simply change pixel locations where space-invariant fixations occur. In this case, the one-dimensional DWT, working on a per-pixel basis, would easily locate fixations by detecting saccades as fixation endpoints. Cases 2 and 3 represent a change in POR, where in the ‘perfect’ visual system, this change can be interpreted as a fixation onset (case 2) or fixation cessation (case 3).

Real eye movements, however, vary over space. In particular, the non-saccade eye movements sought by the present DWT strategy, tend to shift in space. A fixation may vary in time over a small neighborhood of pixels, i.e., pixels subtended by some small visual angle. Smooth pursuit movements, depending on velocity, will also drift over a small number of pixels between a small number of frames, depending on the temporal resolution of the eye movement video sequence. Although this variation may be small between successive frames, cases 2 and 3 above can no longer be interpreted simply as onset or cessation of fixations. In reality cases 2 and 3 may still reflect factual fixations provided that a variance of a small number of pixels is considered. This uncertainty is referred to as the correspondence problem between video frames representing eye movement data.

To overcome the correspondence problem, a realistic spatial pixel neighborhood matching natural eye movement spatial variance must be considered in the three-dimensional DWT analysis. In the above illustrative case, if the neighborhood is extended to a sufficient number of pixels, then pixels that were ‘misaligned’ are brought into overlap. As long as overlapping pixels are present in the video stream, the onset and cessation of dynamic fixation events will be correctly classified by the temporal DWT as in the simple case. Extending the local pixel neighborhood is equivalent to either spreading (copying or zooming) individual pixel values over some small region, or subsampling pixel values by the equivalent amount. The idea is to give up a certain level of resolution in trade for provision of greater spatial variance. Subsampling video frames containing eye movement information is naturally performed by the scaling function of the two-dimensional DWT.

To properly classify eye movements, a sensible number of spatial decomposition levels must be determined. The number of decomposition levels corresponds to the extent of 2D spatial scaling prior to the temporal

analysis. The criterion for the extent of spatial scaling is governed by maximal spatial eye movement deviations over inter-frame durations at the given spatial resolution of the eye tracker data. The resolution of the available eye tracker is 512 pixels horizontally and 256 pixels vertically at a sampling rate of 60 Hz. To calculate the visual angle subtended by the eye tracker, the dimensions of the monitor where the visual stimulus is displayed must be considered. Presently a $21''$ television is used. The horizontal and vertical display dimensions are obtained from the following ratios,

$$\frac{3}{5} :: \frac{\text{height}}{21}, \quad \frac{4}{5} :: \frac{\text{width}}{21},$$

gives 16.8×12.6 width $\times$ height, in inches. The effective resolution in dots per inch (dpi) is found by dividing the number of pixels by the monitor dimensions,

$$\frac{512}{16.8} = 30.47 \doteq 30\text{dpi (horizontal)}, \quad \frac{256}{12.6} = 20.32 \doteq 20\text{dpi (vertical)}.$$

With no decomposition, it is assumed that each pixel corresponds to a true point of regard as provided by the eye tracker. Using the effective horizontal resolution of 30dpi, each pixel roughly covers $1/30 = 0.03$ inches of the stimulus display. Assuming a 60cm viewing distance, each pixel subtends $2\tan^{-1}(.03/(2\times 23.622)) = 0.07°$ visual angle, where 23.622 is the viewing distance in inches. Each level of decomposition has a two-fold effect on the subtended visual angle: first, the effective eye tracker resolution is decreased by 2 (assuming dyadic scaling); second, each pixel representing eye tracker data now represents twice the number of pixels in either horizontal or vertical direction. For example, at 1 level of decomposition, the resolution of the eye tracker data is reduced to approximately 15 dpi, while each pixel is spread over a 2×2 region. That is, the width of pixels representing the point of regard is now 2, giving a width of $2/15 = 0.133$ inches. Denoting the radius of the base of the visual angle by r, calculated as half the width, the visual angle θ subtended by the POR region is given by

$$\begin{aligned} \theta &= 2\tan^{-1}\left(\frac{r}{D}\right) \\ &= 2\tan^{-1}\left(\frac{0.0667}{23.622}\right) \\ &\doteq 0.32°. \end{aligned}$$

Extending these calculations over successive dyadic decompositions produces values given in Table 9. The significance of the subtended visual angle by the POR is that the pixel region at each decomposition level contains POR data within that visual angle. At three decomposition levels, for instance, all recorded eye movements within a region of 4.84° visual angle will be present in the 8×8 pixel region. Furthermore, each single pixel at the original resolution will extend over the entire region. This is repeated over all frames in the video sequence of eye movements. In this way the matching region between frames has been extended to consider spatially varying eye movements over 4.84° visual angle. If a POR corresponding to a fixation is present at some location in one frame and varies no more than 4.84° then assuming the fixation persists into

TABLE 9
Resulting subtended visual angle of POR at dyadic spatial subsampling levels.

Decomposition level	Effective resolution				POR width		Visual angle
	width	×	height	dpi	pixels	in	degrees
1	256	×	128	15	2	0.13	.32
2	128	×	64	8	4	0.5	1.22
3	64	×	32	4	8	2.0	4.84
4	32	×	16	2	16	8.0	19.22
5	16	×	8	1	32	32.0	68.22
6	8	×	4	0.5	64	128.0	139.48

the next image frame, its subsequent POR will appear within the subsampled region overlapping the current POR location. In this case the temporal DWT will detect no significant change in the overlap between the regions. Continuing the table values further exceeds the usefulness of subsampled data for temporal processing. Eventually, if the maximum spatial decomposition is reached, decomposed frames will contain one pixel giving the erroneous interpretation of a steady fixation at the central location of the visual field. Excessive spatial compression results in a loss of positional information.

To make use of the visual angle values presented in Table 9, pursuit movement velocities should be considered over inter-frame periods to match the temporal DWT analysis. The velocity of the slow phase of smooth eye movements ranges roughly from 10° – 50° s^{-1} [Car77, §3, p.37]. Since the eye movement video sequence is composed at 16 fps, the inter-frame period is $1/16 = 62.5$ ms. The DWT temporal analysis at the previously specified two temporal decomposition levels examines pixel differences at half this resolution, i.e., at a period of 125 ms. The expected range of smooth pursuit velocities over a period of 125 ms is 1.25° – 6.25° as derived below:

$$\frac{10^\circ}{1s} = \frac{.01^\circ}{1ms} = \frac{.625^\circ}{62.5ms} = \frac{1.25^\circ}{125ms}, \quad \frac{50^\circ}{1s} = \frac{.05^\circ}{1ms} = \frac{3.125^\circ}{62.5ms} = \frac{6.25^\circ}{125ms}.$$

To match the expected spatial extent of the slow phase of smooth pursuit movements, the closest decomposition level offered by the dyadic spatial DWT is 3 providing detection of velocities not exceeding 38.7° s^{-1}. Higher decomposition levels run the risk of over-averaging POR data.

7.5 Automatic Algorithm Specification

The PARIMA eye movement model relies on the spatio-temporal detection of saccades in eye movement data. In the current implementation, automatic saccade detection is performed over data assembled in video frames as described above. That is, eye movement data is assembled into a video sequence $f(x,y,t)$. The following steps specify the automatic saccade detection algorithm performed through the application of the

anisotropic three-dimensional wavelet transform.

1. Spatial 2D (intra-frame) wavelet decomposition.
2. Temporal 1D (inter-frame) wavelet decomposition.
3. Temporal 1D (inter-frame) modulus maxima detection.
4. Temporal Modulus maxima thresholding.
5. Temporal wavelet coefficient decimation.
6. Temporal 1D (inter-frame) reconstruction.
7. Spatial 2D (intra-frame) projection.
8. Region of interest (eye movement data point) grouping.

Each step is summarized below.

7.5.1 Spatial 2D (Intra-Frame) Wavelet Decomposition

Each frame is decomposed to 3 levels. Expressing this decomposition concisely from §5.7,

$$\begin{aligned}
f^{j-1}_{\phi\phi\cdot}(x,y,t) &= \sum_{k,m}(h_k \otimes h_m) f^{j}_{\phi\phi\cdot}(2x+k,2y+m,t) \\
f^{j-1}_{\psi\phi\cdot}(x,y,t) &= \sum_{k,m}(g_k \otimes h_m) f^{j}_{\phi\phi\cdot}(2x+k,2y+m,t) \\
f^{j-1}_{\phi\psi\cdot}(x,y,t) &= \sum_{k,m}(h_k \otimes g_m) f^{j}_{\phi\phi\cdot}(2x+k,2y+m,t) \\
f^{j-1}_{\psi\psi\cdot}(x,y,t) &= \sum_{k,m}(g_k \otimes g_m) f^{j}_{\phi\phi\cdot}(2x+k,2y+m,t)
\end{aligned}$$

gives the four components of the 2D wavelet transform. Collectively, these decomposition components are denoted by the 2D wavelet transform:

$$\{Wf(x,y,t)\}_{xy}(j) = \{f^{j}_{\phi\phi\cdot}(x,y,t), f^{j}_{\psi\phi\cdot}(x,y,t), f^{j}_{\phi\psi\cdot}(x,y,t), f^{j}_{\psi\psi\cdot}(x,y,t)\}$$

This step is performed to overcome the inter-frame pixel correspondence problem by essentially averaging spatial eye movement data on a per-frame basis.

7.5.2 Temporal 1D (Inter-Frame) Wavelet Decomposition

The entire sequence, treated as a now one-dimensional signal, is decomposed temporally to 2 levels. Using the notation from §5.9,

$$\begin{aligned}
f^{j-1}_{\cdot\cdot\phi}(x,y,t) &= \sum_{k} h_k f^{j}_{\cdot\cdot\phi}(x,y,2t+k), \\
f^{j-1}_{\cdot\cdot\psi}(x,y,t) &= \sum_{k} g_k f^{j}_{\cdot\cdot\phi}(x,y,2t+k),
\end{aligned}$$

giving the 1D temporal DWT:

$$\{Wf(x,y,t)\}_t(j) = \\ f^j_{\cdot\cdot\phi}(x,y,1), f^j_{\cdot\cdot\psi}(x,y,2), \ldots, f^j_{\cdot\cdot\phi}(x,y,n-1), f^j_{\cdot\cdot\psi}(x,y,n).$$

High- and low-pass filtered frames are rearranged forming two $n/2$ sequences.

7.5.3 Temporal 1D (Inter-Frame) Modulus Maxima Detection

To detect temporal discontinuities, only the 2D spatially subsampled frame quadrants are used. That is, a modulus maxima $M\{f^j(x,y,t)\}$ is located at scale j and location t if:

$$|f^j_{\phi\phi\psi}(x,y,t-1)| \leq |f^j_{\phi\phi\psi}(x,y,t)| \geq |f^j_{\phi\phi\psi}(x,y,t+1)|, \text{ and}$$

$$\begin{cases} |f^j_{\phi\phi\psi}(x,y,t)| > |f^j_{\phi\phi\psi}(x,y,t-1)|, & \text{or} \\ |f^j_{\phi\phi\psi}(x,y,t)| > |f^j_{\phi\phi\psi}(x,y,t+1)|. \end{cases}$$

Temporal modulus maxima values correspond to step edges in time, or saccades. Eye movement data corresponding at these temporal locations are deleted.

7.5.4 Temporal Modulus Maxima Thresholding

Modulus maxima values are subject to the hard thresholding rule,

$$T_{hard}[M\{f^j(x,y,t)\}] = M\{f^j(x,y,t)\}I(|M\{f^j(x,y,t)\}| > \alpha\ddot{M}\{f^j(x,y,t)\}),$$

where $\ddot{M}\{f^j(x,y,t)\}$ denotes the range of maxima values at level j, with modulus maxima threshold parameter $\alpha = 0.05$.

7.5.5 Temporal Wavelet Coefficient Decimation

Wavelet coefficients are hard-thresholded (decimated) by the following rule

$$T_{hard}[\{W_{\phi\phi\psi}f(x,y,t)\}(j)] = \{W_{\phi\phi\psi}f(x,y,t)\}(j)I(|M\{f^j(x,y,t)\}| > 0),$$

where, at location t and scale j, $\{W_{\phi\phi\psi}f(x,y,t)\}(j)$ and $M\{f^j(x,y,t)\}$ denote the wavelet coefficients and modulus maxima, respectively.

7.5.6 Temporal 1D (Inter-Frame) Reconstruction

The entire sequence is treated as a one-dimensional signal and the 1D inverse DWT is applied on a per-pixel basis taking care to properly interleave whole image frames as required (see Equation (5.64)). Using the interleave operator $\bowtie$, image frames are arranged for reconstruction at level j by:

$$f^{j-1}_{\cdot\cdot\phi\bowtie\psi}(x,y,2t+p) \quad = \quad (1-p)\,f^{j-1}(x,y,t) + (p)\,f^{j-1}(x,y,t),$$

for $p \in \{0,1\}$. Reconstruction is then written as:

$$f^{j}_{\cdot\cdot\phi}(x,y,2t+p) = (1-p)\sum_k \widetilde{h}_k f^{j-1}_{\cdot\cdot\phi\bowtie\psi}(x,y,t-k) + (p)\sum_k \widetilde{g}_k f^{j-1}_{\cdot\cdot\phi\bowtie\psi}(x,y,t-k),$$

giving the spatially decomposed function $f^j(x,y,t) = \{W f(x,y,t)\}_{xy}(j)$.

7.5.7 Spatial 2D (Intra-Frame) Projection

Instead of reconstructing the 2D spatially subsampled frames, the scaled frame quadrants are projected to the original frame dimensions. This is done since the coarse-scale 2D wavelet quadrants do not provide any useful information. Projecting subsampled values results in spatial zooming where neighboring pixels are folded into a larger region (a large pixel essentially), i.e.,

$$\begin{matrix} f^j(2x+k,2y+m,t) \\ f^j(2x+k,2y-m,t) \\ f^j(2x-k,2y+m,t) \\ f^j(2x-k,2y-m,t) \end{matrix} = \begin{cases} f^{j-1}_{\phi\phi\cdot}(x,y,t) & \text{if } f^{j-1}_{\phi\phi\cdot}(x,y,t) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

for $k,m \in [0,1]$.

7.5.8 Region Of Interest (Eye Movement Data Point) Grouping

The final step of the algorithm considers any non-zero pixel to be a valid eye movement data point (point of regard, or POR). The above processing steps effectively remove any samples identified as saccades. This step merges all non-zero pixel regions and finds the 2D centroid of the merged region. Any non-zero pixels outside the search area will create separate regions. The search region used for iterative pixel grouping is the number of pixels subtended by the foveal visual angle (see §IX).

7.6 Limitations of the Frame-Based PARIMA Implementation

It is important to reemphasize that although the PARIMA model is based to a large extent on the functional characteristics of the oculomotor system, it is not a model of the system itself. Instead, it is a stochastic model of the observed signal. That is, no implication is made that the oculomotor neural substrate behaves in the

proposed manner, only that measured signals may be characterized in terms of patterns associated with the three major types of eye movements. In this sense, the PARIMA model is a three-tiered pattern recognition algorithm delineating fixations and smooth pursuits through the detection of saccades. As such, the PARIMA model is subject to the following limitations.

7.6.1 Linear Assumption

The linear assumption of the PARIMA model does not properly express the underlying nonlinearity of the oculomotor system. Eye movement interactions should reflect these inherent nonlinearities [WNS84]. Winters et al. argue that the "linear summation" or the "additivity hypothesis" is inadequate in explaining conjugate eye movements. Linear summation refers to the simplified view of the neural integrator which expects a conjugate eye movement resulting from the summation of independent eye movement trajectories. For this superposition to apply, they argue, either the system must be linear or any nonlinearities must cancel. Through power spectra analysis of smooth pursuit movements, Wong showed that 2nd, 3rd, and 4th order nonlinearities are present in the human ocular system, manifest as the modulation between the stimulating frequencies [Won90]. Wong suggests that these nonlinearities can be detected by the sum-of-sinusoids method. Continuing this work, Sa proposed a model using Wiener filters to estimate the magnitude and power spectra of smooth pursuits [Sa95].

In contrast, the wavelet-based linear filter model presented here is clearly an inadequate model of the inverse of the oculomotor system. As a first approximation to the observed signal, however, the PARIMA strategy is an attractively simple yet parsimonious description of a complex process. From a pattern recognition standpoint, since the multiscale saccade detection performed by the wavelet transform is equivalent to the Canny edge detector, the technique is numerically optimal [Can86]. The wavelet-based edge detection offers flexibility in the choice of wavelet and scaling functions as well as non-dyadic resolution scales.

7.6.2 Wavelet Filter Length

A problem associated with the short length Haar wavelet and the dyadic multiresolution scale is the inability to detect all step edges. At present the length-2 Haar wavelet is used. Short duration edges at dyadic sample boundaries will not be detected at fine resolution scales. This is due to an edge occurring between dyadic sample points. A numerical example of a missed step edge is shown in Table 10. The four-element signal contains a step edge between the second and third elements. The one-level Haar wavelet decomposition of the signal ($\mathbf{d}^1$) fails to locate the edge (the edge is detected at the next decomposition level).

TABLE 10
Numerical 1D DWT example of missed edge.

Decomposition				
$\mathbf{f}^2 = \mathbf{c}^2$:	0	0	4	4
$\mathbf{d}^1$:		0		0
$\mathbf{c}^1$:		0		$\frac{8}{\sqrt{2}}$

7.6.3 Frame-Based Implementation

The choice of frame-based representation of eye movement data for analysis trades temporal resolution for synchronism with video frame stimuli. That is, the presented solution is time-driven since raw data is sub-sampled to fit the display rate of the video stimuli. Alternative implementation recommendations were made in §7.3. Both spline-based and the direct time series analysis methods are data-driven in the sense that no imposition is made on the data sampling rate beyond the limitation of the eye tracker instrument. The time-based method is empirically evaluated in §XI. Further work is required to test and compare all four methods.

7.6.4 Misclassification of Manifold Eye Movements

The premise of conjugate eye movements being delineated by saccades is an oversimplification resulting in an underestimation of manifold eye movements such as optokinetic nystagmus. Theoretically, smooth pursuits are identified by the PARIMA model, provided the observed signal is indeed smooth. Nystagmus movements, however, are characterized by a combination of smooth pursuits and saccades, referred to as the slow and fast phases of nystagmus, respectively. The PARIMA model, as presently configured, will not identify nystagmus as such, instead nystagmus movements will be packetized into consecutive smooth pursuit segments. The PARIMA model effectively “chops up” nystagmus into its slow phase components.

7.6.5 Off-line Implementation of PARIMA Model

The computational complexity of the PARIMA model currently prevents real-time implementation. Conceivably, a real-time version of the algorithm is possible by limiting the extent of the temporal analysis. In theory, only two consecutive eye movement samples (at temporal sampling rate of 60Hz) are needed for a local (albeit noisy) estimate of saccades. The present implementation design sacrifices efficiency for flexibility of the multidimensional signal processing technique. At present it is possible to test various wavelet and scaling functions at several decomposition levels.

7.6.6 Context-Free Analysis

Since the PARIMA model is based on stochastic time series modeling techniques, it is appropriate to consider the the model's forecasting power. However, the model is inappropriate for eye movement forecasting since it evaluates the signal outside the context of the visual environment. For long term prediction of future locations of the point of regard, the visual context must be considered. Since the PARIMA model does not consider visual information, by itself it is not suited for prediction of visual scanpaths. On the other hand, the short term forecasting of eye movements may be useful for real-time applications. That is, given the examination of the trend of a non-saccadic eye movement (e.g., a pursuit as classified by the PARIMA model), it may be possible to predict the general direction of the movement. If calculated quickly, this capability may be suitable for minimizing the latency of gaze-contingent systems. For example, assuming a small enough delivery delay of raw eye position by the eye tracker, and the availability of previous eye movement history (e.g., sufficiently large system "short-term" memory), the system may be able to anticipate the direction of pursuit movements. Although this prediction would be limited in temporal extent, it may provide enough predictive power to combat eye tracker latency.

7.7 Summary

In this section a model of eye movements is presented from a signal processing perspective. The model assumes eye movements to be linearly dependent time series composed of three types of signals: a stationary component (fixations), a non-stationary component (smooth pursuits), and discontinuities (saccades). Fixations and smooth pursuits are modeled as Auto-Regressive Integrated Moving Average (ARIMA) stochastic linear systems, while saccades are modeled as short-term Moving Average (MA) step discontinuities.

The algorithmic implementation of eye movement classification is carried out by the anisotropic discrete wavelet transform (ADWT). Eye movements are represented as video data where sampled points of regard are represented by white pixels over black video frames constituting sparse matrix representations. The ADWT is utilized to spatially average intra-frame data as well as for inter-frame signal step detection.

The wavelet-based algorithm is a flexible multidimensional signal analysis framework suitable for limited eye movement classification. The multiscale saccade (edge) detection is numerically optimal since it is equivalent to the Canny edge detection technique. Detection of temporal step edges (saccades) delineates the signal into ARIMA segments resulting in a piecewise-ARIMA (PARIMA) model of conjugate eye movements upon reconstruction.

CHAPTER VIII

VOLUMES OF INTEREST

> "How does one *represent* a fixation in a two-dimensional, analog plot? Does one use a single point, a number, a disk of a certain size...?"
>
> – Peter R. Coles [Col83, p.4]).

The questions raised by Coles reflect the visualization problem studied in this section. Two-dimensional representations of fixations suffers from three inadequacies:

1. inability to quantify the duration of a fixation,
2. inability to represent the order of fixation points (without explicit labeling), and
3. inability to characterize the nature of the fixation change.

The latter point deals with the possibility that alternative eye movements may be present in a fixation change, e.g., smooth pursuit, or saccade. The former two points are a consequence of the lack of temporal dimension in a two-dimensional plot of eye movements. Temporal information is lost due to the projection of three-dimensional data onto a two-dimensional plane. In this section, a three-dimensional visualization technique is introduced which explicitly represents the temporal dimension.

The three-dimensional eye movement visualization technique relies on the identification of dynamic fixations. In this implementation, dynamic fixations are identified by the wavelet-based Piecewise Auto-Regressive Integrated Moving Average (PARIMA) model of three-dimensional stochastic processes introduced in §VII.[1] Fixations, represented by discrete pixel regions in a video sequence, match foveal intra-frame loci of attention, referred to as Regions Of Interest (ROIs). Sampled at 18ms and mapped onto a 16fps video sequence, roughly 3-4 fixation points may occupy each video frame. Points within a small locus are effectively averaged by the PARIMA analysis to define intra-frame ROIs. Inter-frame ROIs are merged to visualize foveal *Volumes Of Interest* (VOIs), providing a depiction of dynamic foveal vision.

In general, the VOI model of eye movements amalgamates distinct (i.e., multiple viewers') scanpaths into a consolidated spatio-temporal description. Figure 40 shows the abstract conceptualization of the aggregate VOI model. Vertical lines represent time slices, e.g., uniformly sampled video frames, as a temporal reference. Figure 40 highlights a hypothetical individual scanpath included in the VOI collection. This scanpath may correspond to a past observation of actual view patterns, or may present a candidate future scanpath.

[1] Note that the visualization is independent of the analysis method.

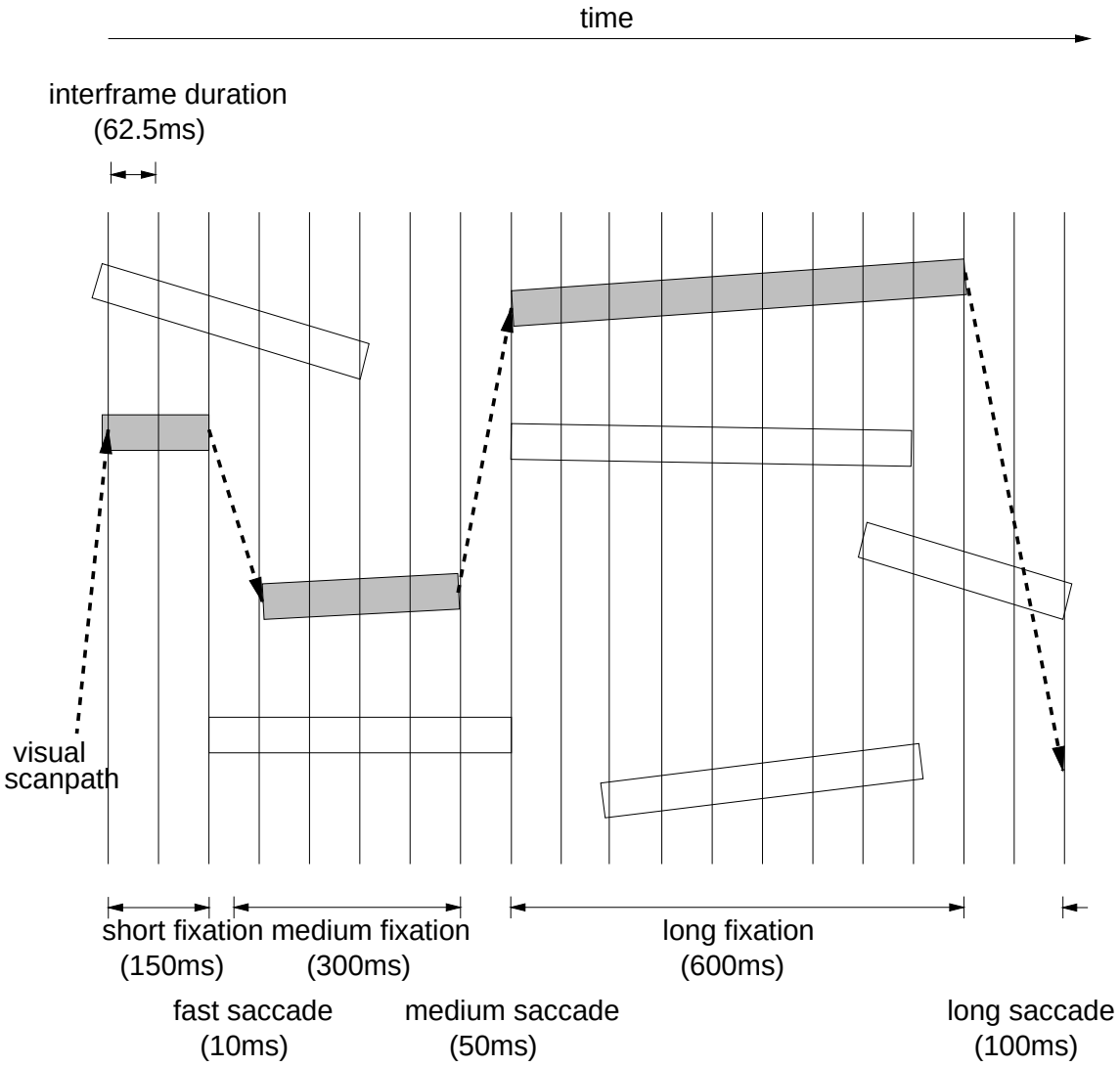


Fig. 40. Graphical Volume Of Interest model.

That is, observed eye movement patterns, represented by Volumes Of Interest, are concatenated to form a history of multiple viewers' scan patterns over a particular sequence. Since VOIs identify observed loci of attention, they serve as indicators of potential future visual attractors.

Overlapping VOI cross-sections of the highlighted scanpath represent fade-in and fade-out temporal ramps (see §9.1). VOI diameters normally match the dimension of the subtended foveal region. As gaze shifts (denoted by dashed arrows in Figure 40), the VOI from where gaze departed is attenuated. Prior to gaze localization, the VOI is gradually amplified in anticipation of gaze shift. Peripheral VOIs (unshaded VOIs in Figure 40) are also modulated in a similar fashion. Essentially, the general VOI model considers historical VOIs as future predictors of gaze patterns.

8.1 Synthesis of Volumes Of Interest

The Volume Of Interest (VOI) assembly consists of a concatenation of individual Regions Of Interest (ROIs) contained in discrete video frames. Let $\{ROI_k^j\}$ denote a list of ROIs where j represents the index of the ROI on video frame k. Each video frame may contain numerous disjoint ROIs. There is no restriction on the number of ROIs on any given frame, although it is assumed there are no overlapping ROIs. The total number of video frames typically corresponds to the duration of the gaze-contingent stimulus. Currently 128 frames are used, but this does not pose any restrictions on the VOI assembly algorithm. Let $\{VOI_m\}$ denote a list of VOIs. Each VOI is defined as a list of ROIs where the ROIs correspond to intersections between the VOI and consecutive video frames. There is no restriction on the number of VOIs in the space-time volume defined by the video frame dimensions and the number of frames, although it is assumed no VOIs overlap. Initially the VOI list is empty. Header structures $ROIh$, $VOIh$ point to the ROI and VOI lists, respectively. The VOI assembly algorithm iteratively processes each ROI associated with video frame f. The VOI list is then searched for a VOI that extends to the previous video frame in the sequence, i.e., frame $f-1$. If the euclidian distance between the current ROI and the intersection of such a VOI with frame $f-1$ is below a threshold r, then ROI_f^j is conjoined with the VOI. This operation extends the VOI to frame f. If no such VOI is found, ROI_f^j is made to be the start of a new Volume Of Interest.

In the current visualization, the threshold r is held constant corresponding to the foveal 5° visual angle at the prescribed viewing distance (see §IX). Choosing this parameter for the VOI dimension results in a visualization of dynamic foveal vision in space-time. A possible alternative choice of r is the variance of gaze position with respect to a given target. Experimental gaze error in the visual tracking paradigm is discussed in §12.5.3. A one-dimensional representation of intra-frame error is shown in Figure 74. VOI visualization of this type of variance metric may be suitable for the representation of a scalable attentional window, such

as proposed by Kosslyn (see 2.1.8).

8.2 Graphical VOI Construction

Volumes Of Interest are created in 3-space as wireframe tubes connecting VOI segments. Each VOI is defined as a discrete list of VOI and video frame intersections, or ROIs. Each inter-frame segment of the VOI is constructed by linearly interpolating between ROI_f^v and ROI_{f+1}^v where the ROI superscript v now refers to the VOI of which the ROI is a member. The inter-frame, or inter-ROI segment is constructed by creating 9 rectangular polygons per quadrant of the circular volume as shown in Figure 41. This set of polygons is forthwith referred to as a surface patch. Each patch quadrant is defined by angle $\Omega \in \{0°, 90°, 180°, 270°\}$

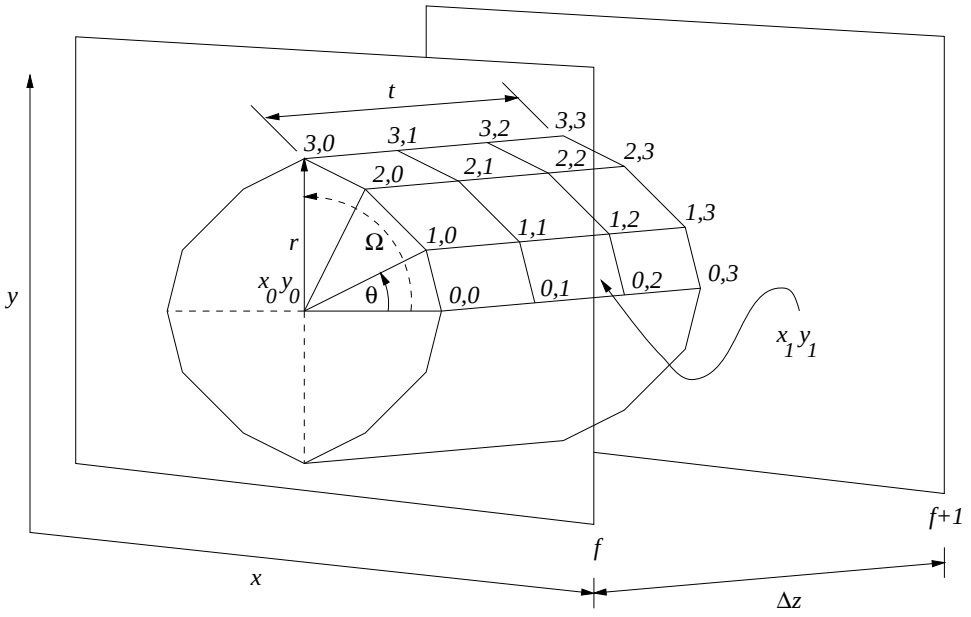


Fig. 41. Graphical VOI scaffolding.

and is symmetrically replicated to form the circular volume. The coordinates of the control points used to define the patches are dependent on the interpolant $t \in \{0, 1/3, 2/3, 1\}$. Surface patches are created as a function of angle $\theta \in \{0°, 30°, 60°, 90°\}$. Boundary control point coordinates are calculated by

$$x' = x_0 + r\cos\theta, \quad y' = y_0 + r\sin\theta, \quad z' = f,$$
$$x'' = x_1 + r\cos\theta, \quad y'' = y_1 + r\sin\theta, \quad z'' = f + 1,$$

where r is the radius of the volume matching the dimensions of the subtended foveal region, f and $f+1$ are the frame z-coordinates appropriately scaled to sufficiently spread out the graphical environment (e.g., scaled

by a constant of 1000 pixels), and (x_0, y_0) and (x_1, y_1) are ROI centers on frames f and $f+1$, respectively. Interior control point coordinates are interpolated on parameter t:

$$x_{i,j} = (1-t)x' + (t)x'', \quad y_{i,j} = (1-t)y' + (t)y'', \quad z = (1-t)z' + (t)z''.$$

The coordinates of the control points are calculated in a double loop, the order of which is indicated by the (i, j) vertex indices in Figure 41. Control points are maintained in a polygonal list which contains data structures for face vertices and normals. Vertex normals are maintained in a separate vertex list.

Along with the VOIs, rectangles symbolizing video frames are generated to provide visual cues for temporal ordering and duration. The distance between frames (Δz in Figure 41) is constant representing inter-frame duration. In the current environment, $\Delta z = 62.5$ms, the inter-frame period corresponding to the 16fps display rate. Every 16th frame is texture mapped with the corresponding video sequence image.[2] All constructs form one graphical object are subject to rendering and standard three-dimensional operations (rotate, scale, translate).

8.2.1 Rendering and User Interaction Considerations

Rendering of the Volumes Of Interest includes hidden surface removal and smooth shading. Since the entire graphical environment is composed of a single object (the collection of VOIs), it is a good candidate for hidden surface removal by the Binary Space Partition (BSP) algorithm. Vertex normals are calculated as averages of adjacent face normals. These are used to Gouraud shade the facets of the VOIs. The BSP and Gouraud shading algorithms are well-known and can be found in most computer graphics textbooks (see for example [FvFH90, pp.675-680 and §16.2.4, pp.736-738, respectively]).

Standard three-dimensional operations are available, including object rotations and translations, as well as viewpoint (camera) roll, pan, tilt, and truck. The truck operation allows camera translation along the z-axis providing a "fly-through" of the video volume. The fly-through provides a frame-by-frame dynamic visualization of scanpaths and allows close inspection of VOI-frame intersections. A three-dimensional scanpath is rendered within the VOIs representing the pixel-to-pixel gaze locations. The pixel-wide scanpath gives the user an idea of eye tracker accuracy since it shows the degree of pixel error between fixation locations and stimulus target.

[2] The number of texture mapped frames is constrained by the system memory capacity. Each texture mapped frame actually contains several images each with a different transparency (alpha-channel) value. With enough memory every frame could conceivably be texture mapped to give the full visual effect of the video content.

8.3 Comparison of Two- and Three-dimensional Eye Movement Visualizations

The two-dimensional representation of eye movements has not significantly changed since Noton and Stark introduced the "scanpath" [NS71a, NS71b]. The scanpath depiction of eye movements shows the route taken by the subject's point of regard (POR). Unfortunately, the scanpath is a two-dimensional projection of a three-dimensional phenomenon. Figure 42 shows a simple scanpath over 4 seconds of a CNN video clip (shown at 16fps, greyscale—experimental conditions are described in §X). Due to the loss of temporal information, it

Fig. 42. Traditional 2D eye movement visualization.

is difficult to ascertain

1. the direction of the scanpath,
2. the duration of potential fixations, and
3. the type of eye movements evoked to change fixation locations.

For example Figure 42, it is difficult to tell whether the scanpath originated at the television anchor's eye, or the "timebox" in the lower right corner of the image. Assuming the two clusters at these two points are fixations, it is difficult to tell how much time was spent looking at these locations. The movement from one location to the other may or may not be a saccade. Because the original stimulus was not a still image it is possible that something in the field of view appeared evoking a smooth pursuit (the anchor scratching his right eyebrow with his left hand for example). This ambiguous two-dimensional representation of motion-related eye movements is especially problematic in studies where stimulus motion is a factor.

Several visualization techniques are available which alleviate these difficulties to some extent. Arrows or time line plots indicating eye movement direction address the first point above. The second difficulty is alleviated by popular “raindrop” displays [Iscan94]. Raindrop displays illustrate fixations with circles increasing in size with fixation duration. Although the relative sizes of raindrops helps identify longer dwell times, it is difficult to quantifiably judge the time spent in fixations. The third difficulty is often ignored when two-dimensional static images are used as the visual stimulus. Assuming the head is stabilized, the lack of stimulus motion precludes the consideration of motion-related eye movements e.g., smooth pursuits. As a result, saccades are usually considered the sole mechanism responsible for instigating fixation changes. In the case of moving stimuli, e.g., video, the two-dimensional visualization of eye movements may ambiguously represent either saccades and smooth pursuits.

Three-dimensional visualization provides explicit representation of the temporal component of eye movements. Figure 43 shows the same 4-second scanpath as shown in Figure 42 in three dimensions. The scanpath,

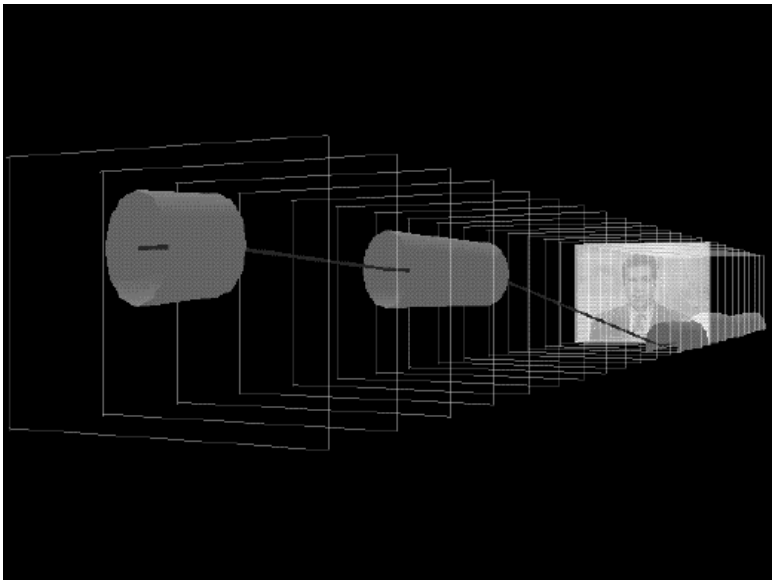


Fig. 43. Eye movement visualization with Volumes Of Interest.

shown as a thin three-dimensional line within Volumes Of Interest, unambiguously depicts eye movements originating at the central location of the frame (the anchor’s eyebrow), jutting slightly downward (to the eye), and then continuing to the lower right corner of the frame (the timebox). The duration of eye movements is referenced by the presence of the video frame outlines. The first fixation roughly extends over two or three frames, suggesting a dwell time of about 125 milliseconds (the video rate is 16fps). The length of VOIs provides relative dwell time information just as the raindrops do, but with the video frames as reference, it

is easier to quantify these durations. Finally, resembling old fashioned slinky toys, the flexibility of VOIs disambiguates smooth pursuit eye movements from saccades. That is, VOIs represent identified (dynamic) fixations. In Figure 43, any part of the scanpath enclosed by a VOI is considered a fixation. In the current example, fixations have been identified by the PARIMA model discussed in §VII, although any suitable fixation algorithm can be used for this purpose.

8.4 Aggregate Volumes Of Interest

Volume Of Interest visualization naturally extends to the representation of aggregate eye movements gathered from multiple subjects. The study of multiple subjects' viewing patterns gives insight into the nature of visual stimuli that attracts gaze and therefore presumably visual attention. Characterization of stimulus in terms of *visual attractors* is essential for predicting the dynamic time course of human vision. This is a fundamental open problem in vision research. Notable examples of work in this area include the developments of algorithmic strategies that attempt to explain human visual search strategies through machine-based simulation (see §15.4).

In their early work on eye movements, Noton and Stark investigated viewing patterns of multiple subjects [NS71a, NS71b]. Although recorded scanpaths exhibited significant variability in how different individuals view a scene (inter-subject variability), or for that matter how an individual's scanpaths differ from session to session (intra-subject variability), the authors identified "informative details" as common fixation points. To show this graphically, several images were used to illustrate scanpath variability and the common location of interesting features. Integrating multiple scanpaths over a single two-dimensional image makes the illustration of common fixated details difficult. Figure 44 shows scanpaths recorded from 7 subjects over the latter 4 seconds of the anchor man sequence (for experimental conditions, see §XII). The visualization of multiple scanpaths exasperates the problems associated with the two-dimensional representation.

The Volume Of Interest visualization extends the representation of individual scanpaths in space-time to the display of aggregate eye movement trajectories. Figure 45 depicts the same scanpaths shown in Figure 44 in three dimensions. The interactive nature of the visualization permits closer inspection of VOI-frame intersections yielding two-dimensional Regions Of Interest at particular points in time, as shown in Figure 46. The importance of relative dwell times is clearly seen in the VOI representation. The aggregate representation shows the convergence of multiple scanpaths over the anchorman's face through most of the sequence. Peripheral regions such as the timebox are fixated sporadically propounding the region as one of diminished relevance.

Fig. 44. Aggregate scanpaths.

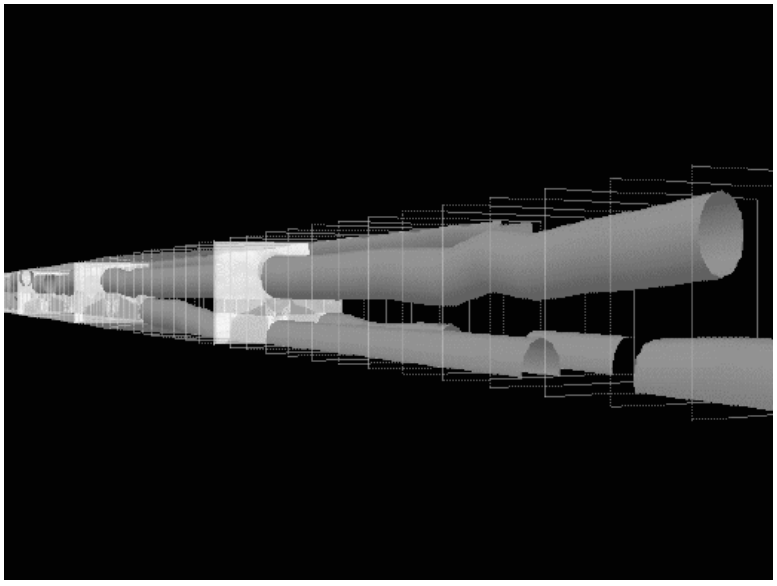


Fig. 45. Aggregate Volumes Of Interest.

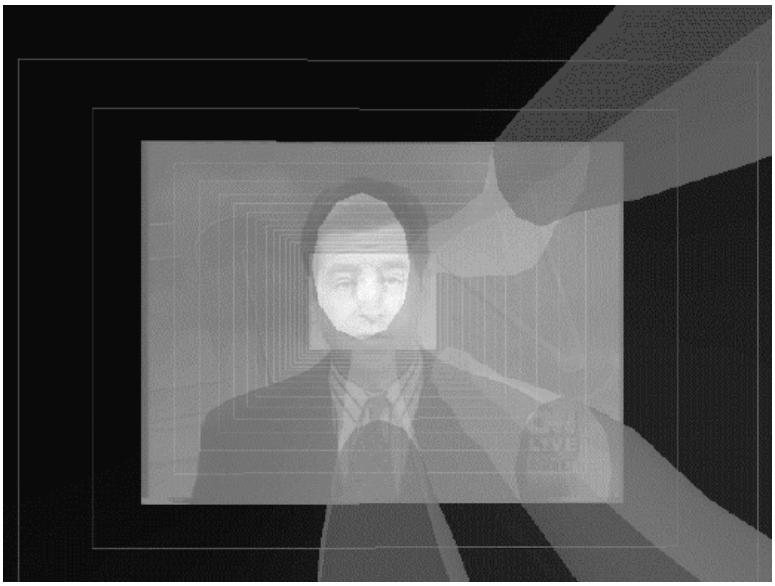


Fig. 46. Inspection of aggregate VOI-frame intersection.

Volumes Of Interest representing spatio-temporal segments of the stimulus constitute potential attractors of visual attention. After its consolidation, the individual scanpath loses its significance in the aggregate model. At any point in time the intersections of multiple VOIs and a video frame constitute potential spatio-temporal candidates for attentive inspection as evidenced by their historical selection by previous viewers. In this sense, aggregate VOIs are similar to Noton and Stark's informative details identified over still images except VOIs depict these details in space-time.

Aggregate VOI visualization presents an interesting speculation of the "what" and "where" duality of visual attention. Representing aggregate VOIs by opaque volumes, in Figure 47(a), provides an illustration of the exclusionary "what" of visual attention. Opaque VOIs depict the restrictive aspect of foveal vision. At the same time, attention seems to have a simultaneous pre-attentive component responsible for selecting the next focus of attention. Pre-attention, in this sense, is the "where" of visual attention. Representing aggregate VOIs by transparent (or rather translucent) volumes, in Figure 47(b), provides an illustration of the apparently simultaneous covert attentional mechanism. On the one hand, VOIs offer a representation of the dynamic high-fidelity attentional channel (overt foveal vision), limited in its spatial extent. On the other hand, VOI transparency symbolizes the peripheral influence of pre-attention.

(a) Opaque representation. (b) Transparent representation.

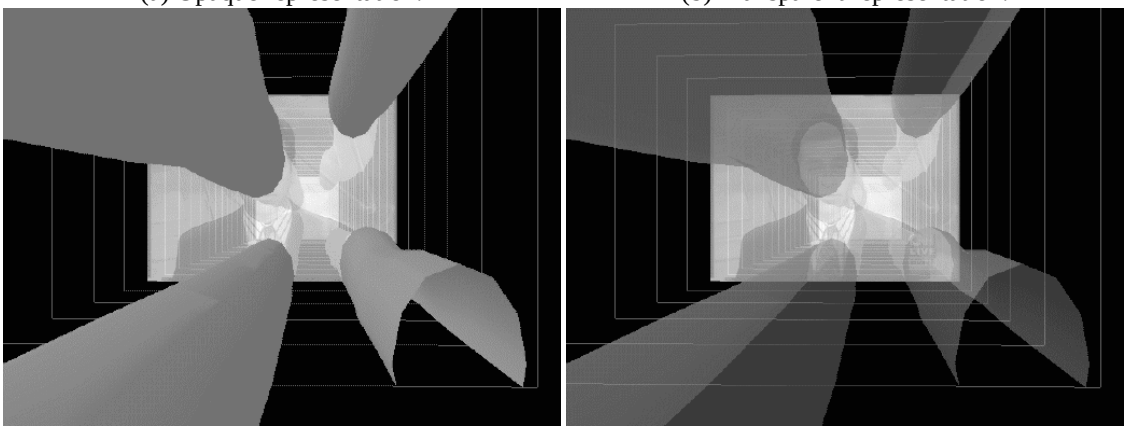


Fig. 47. Transmissivity of aggregate VOIs.

CHAPTER IX

GAZE-CONTINGENT VISUAL COMMUNICATION

In order to match Human Visual System (HVS) resolution capabilities, a multiresolution method is developed for preserving multiple Regions Of Interest (ROIs) in images.[1] Regions Of Interest are maintained at high (original) resolution while peripheral areas are degraded. The multiresolution method follows the variable resolution reconstruction with MIP-wavelets technique developed in §5.10.3. Most ROI-based gaze-contingent schemes concentrate on preserving a single foveal region, usually attempting to match the visual acuity of the HVS. The multiple ROI method described here offers three variants of peripheral degradation including linear, nonlinear, and HVS acuity-matching resolution mapping. Pixel degradation is carried out relative to each ROI. The Voronoi diagram, a well-known planar partitioning construction in computational geometry, is used to partition the image relative to the multiple ROIs centers.

Region Of Interest (ROI) image processing aims at presenting a single high resolution area to the fovea. The goal of limiting high resolution to a foveal "spotlight of attention" is to minimize bandwidth requirements, while matching early vision capabilities of the HVS in the periphery. Typically, peripheral imagery is degraded in order to minimize information content prior to encoding or transmission. Approaches range from luminance attenuation, smoothing, and locally adapted transform coding techniques (i.e., local clamping of DCT coefficients). On the other hand, feature detection algorithms that locate visually interesting information (multiple ROIs, essentially) tend to treat the whole image as the periphery and do not typically provide a foveal ROI. The objective of the method presented here is to provide a foveal ROI and also represent peripheral ROIs (pROIs) as potential future foci of gaze (but not necessarily attention). Moreover, the periphery around each ROI is degraded matching the HVS acuity function. For gaze-contingent displays, this type of processing may be more suitable than single-ROI methods since its aim is to preserve preview benefit.

9.1 Background

In gaze-contingent (GC) applications (such as flight simulators), emphasis has usually been placed on representing the foveal ROI, while homogeneously degrading the periphery [Koc87, LTFW+89]. In the Super Cockpit Visual World Subsystem, Kocian considered visual factors including contrast, resolution and color in the design of a head-tracked GC display. In their Simulator Complexity Testbed (SCTB), Longridge et al. included an eye-slaved ROI as a major component of the Helmet Mounted Fiber Optic Display (HMFOD).

[1] Regions Of Interest are also known as Areas Of Interest (AOIs).

This ROI provided a high resolution inset in a low resolution (presumably homogeneous) field which followed the user's point of regard. The precise method of peripheral degradation was not described apart from the criteria of low resolution. However, the authors did point out that a smooth transition between the ROI and background was necessary in order to circumvent the possibility of a perceptually disruptive edge artifact. Recently, more sophisticated approaches have been proposed for ROI-based video coding [NLO94, ST94]. While these schemes concentrate on the representation of the foveal ROI, the periphery is processed by conventional means such as smoothing or quantization of transform coefficients. While the transition from the high-resolution ROI to the periphery may be smooth, it does not necessarily match the HVS acuity function.

Various multiple-ROI image processing schemes have been proposed for feature-detection tasks where features are either preserved or enhanced, while the rest of the imagery is decimated in some way. Of particular relevance are multiresolution, pyramidal schemes [PW92, San90]. Sandon developed a connectionist network model of guided visual attention, while Pölzleitner and Wechsler used distributed associative memories (DAMs) in preattentive mode to find relevant segments in the field of view. In both cases a Gaussian pyramid was used to subsample the scene into lower resolution levels. The purpose of these schemes is to locate peripheral ROIs in order to simulate guided visual attention. In contrast to foveal ROI coding schemes, the periphery seems to be of greater importance in these approaches, although again the transition from each ROI to its surround does not necessarily match the HVS acuity function.

The multiresolution method for ROI representation given here follows the variable resolution reconstruction with MIP-wavelets technique developed in §5.10.3. Selective scaling of wavelet coefficients is not a new approach. A similar method to the one presented here was shown by Nguyen et al. [NLO94]. In order to enhance relative reconstruction quality, *a priori* weighting factors were introduced defining a region-based weighted l^2 metric. The weighting factors were considered as quantitative decimating factors in the relative distortion contributions in each region. In their paper, only region-based spatial weighting was considered. The work focused on video encoding, where each frame was synthesized from a fixed subband representation (multiresolution structure). ROIs were selected according to a motion criterion, where ROIs were obtained from a segmentation map which isolated moving objects from the background. To preserve the hierarchy of relevant spatial information in the decimation process, the ROIs were projected onto the subband domain. Simple uniform threshold quantization was used on wavelet coefficients obtained using Daubechies-4 filters. The effect of the projection and subsequent uniform thresholding resulted in enhanced (or rather preserved) fidelity within the ROI, with the background regions blurred. Each ROI boundary was clearly defined due to the authors' assumption of independent encoding of the prediction error (PE) signal within an ROI. In essence, the scheme identified a hierarchy of ROIs within a frame (based on region segmentation) and wavelet coefficients within each ROI were subjected to uniform thresholding. Reportedly, the processed image sequence

induced the observer to naturally focus on the (presumably) most interesting ROI. The main aspect in which this method differs from the present method is the uniform decimation of the ROI coefficients. The present goal is to match the abrupt but smooth gradient of the HVS spatial acuity function. Projection of an ROI onto the subband domain results in abrupt resolution modulation at the reconstructed ROI boundary. This effect can be demonstrated by rounding up the scaling factor p to 1 within the ROI region and decimating wavelet coefficients corresponding to background regions. The resultant images possess smaller Mean Squared Error (MSE) in the ROI-projected images compared to images processed by scaling coefficients. This is due to the fact that rounding up the scaling factor to 1 within the ROIs preserves a proportionately larger high resolution region than when the coefficients are linearly interpolated within ROIs. The boundaries between levels of resolution in the image reconstructed without coefficient scaling are clearly visible, however. This may be suitable for the purposes of compression, but it does not match the spatial sensitivity of the HVS.

Another interesting approach was described by Abdel-Malek and Bloomer [AB90]. The authors presented multiresolution images synthesized from a Laplacian pyramid representation. The Laplacian pyramid is an instance of the DWT where the smoothing filter is a Gaussian-like averaging function, and bandpass (Laplacian) images are obtained by differencing adjacent Gaussian images. The HVS acuity function is approximated by rectangular regions concentric to the point of regard. As in the work of Nguyen et al., no effort is made to smooth boundaries between reconstructed resolution levels. In fact, the authors state that no extra reconstruction filter is required to hide transition zones. Transition zones are not evident in the presented resultant images, but this may be due to the fact that high resolution is restricted to 2×2 pixel neighborhoods around zero-crossings detected at the bottom level of the pyramid.

The current wavelet approach follows the classical pyramid representation providing comparable functionality and coding efficiency. The filtering approach presented here provides the means to represent ROIs of any shape, while simultaneously ensuring smooth transition between reconstructed resolution regions. In the current scheme, circular ROIs were chosen to better match the circular foveal region of the HVS.

The intent of the present method is to demonstrate a scheme to juxtapose the representation of pROIs, as produced by feature-detection tasks, with a foveal ROI as found in gaze-contingent display modalities. Utilizing a multiresolution spatial pyramid, the objective here is not to find visual attractors, but to offer a method of image representation suitable for attentive and pre-attentive viewing. The work described here centers on the reconstruction of the image from prefiltered (texture) maps of the original image. The novel aspect of the approach is the ability to maintain several ROIs within the image while gradually degrading the periphery around each ROI. ROI shape is circular (although it need not be) and the peripheral degradation function can be chosen from several variants with respect to spatial distance from the center of the ROI. In this imple-

mentation, the periphery can be degraded using a linear, nonlinear, or HVS acuity-matching function. The resolution of an arbitrary peripheral pixel depends on its distance to the closest ROI.

Peripheral ROIs pose an apparent paradox: why should peripheral regions be represented at high resolution when the peripheral visual system lacks the ability to resolve peripheral detail? In the context of replicating the HVS by a computational strategy (e.g., building a synthetic retina) only one foveal region makes sense. However, from the perspective of building a gaze-contingent system, peripheral ROIs address the system's temporal inability to faithfully track the human's gaze in real-time. There is an inherent delay in the system, due to eye tracking latency, which is inevitably manifested by a temporal lag between the viewer's change of gaze (e.g., via a saccade) and the system's translation of the foveal ROI. Gaze-contingent systems are, in essence, *reactionary* with respect to the subject's gaze, and are unavoidably late in updating the foveal ROI, causing foveal vision to fall on a region of low resolution. The inherent delay in updating the foveal region to high resolution may cause impaired perception through a lack of *preview benefit*. The reason for pROIs is not to provide high resolution to match peripheral acuity, but to anticipate saccades, thereby addressing the inherent latency present in the gaze-contingent system. In this sense, a gaze-contingent system provided with peripheral ROIs is *anticipatory* with respect to gaze.

To balance bandwidth minimization requirements with the dynamic representation of multiple ROIs (foveal and peripheral), the system should gradually attenuate pROIs when the observer maintains detectable fixations. Similarly, a fully anticipatory system should predict fixation changes and gradually amplify pROIs in anticipation of fixation changes. Gradual modulation of pROIs is required so that sudden onsets and withdrawals of pROIs do not distract attention from its natural course. A temporal ramp may be used for this purpose (see [ST94]).

9.2 Resolution Mapping

As alluded to in §5.10.3, ROI-based reconstruction of the image from its wavelet transformation relies on the choice of a mapping function. A mapping function, denoted by l, maps resolution from the multiresolution pyramid to *image space*. The choice of a mapping function is a precursor to reconstruction of the image and is crucial to the the final representation of the image. It is important to note that resolution information in the pyramid is distributed nonlinearly (by decreasing powers of 2 if the multiresolution pyramid is dyadic in nature). Since reconstruction is carried out in image space (dependent on the pixel location (x,y) in the final image), the resultant percent resolution distribution is obtained by taking the inverse of the constant 2

raised to the mapping function, i.e., % resolution$/100 = 1/2^l$.[2] In the current implementation, three mapping functions are developed: linear, nonlinear, and empirical HVS acuity-matching. The linear and nonlinear mapping functions were chosen as approximate lower and upper bounds, respectively, to the HVS matching function, in terms of percent resolution. Each mapping function segments the image into concentric resolution regions, or bands. In all three implementations, resolution within the central 5° of each ROI is consistent and equal. Although this is not a restriction imposed by the image reconstruction, the size of each ROI is maintained consistently across mapping functions (by the choice of R) so that different peripheral degradation methods could be readily compared. The three mapping functions are given below:

1. linear,
$$l = \frac{d}{R};$$
2. nonlinear,
$$l = A(1 - e^{-\lambda \frac{d}{R}});$$
3. HVS acuity-matching,
$$l = -\frac{ln(\text{empirical \% resolution at pixel distance}/100)}{\ln(2)}.$$

The parameter d is the pixel distance from the ROI center, and R is the radius of the highest resolution region (foveal region).[3] The derivation of R is based on an empirical HVS acuity function (see §9.2.1). For the nonlinear mapping function, A is the asymptote approximated at the image boundary (here A = 2.35). To consistently preserve resolution within the radius of the highest resolution region, λ is chosen so that $l = 1$ at pixel distance R. That is,

$$1 = A(1 - e^{-\lambda}),$$

so that

$$\lambda = \ln(\frac{A}{A-1}).$$

The HVS acuity mapping was originally obtained in terms of percent resolution, that is, it was specified as a function in resolution space and thus the above mapping function (in image space) is the inverse of that empirical function. All three functions are plotted in Figure 48 showing the mapping functions in image space, and the corresponding relative resolution.[4] The concentric resolution bands in image space are shown in Figure 49 with 2 ROIs. Lighter areas are reconstructed at higher resolution, black rings are level boundaries. For comparison between mapping functions, in terms of statistical image quality, and image reproduction examples, see [DM95].

---

[2]Percent resolution refers to relative resolution in the reconstructed image assuming 100% resolution in the original.

[3]The current system implementation uses $R = 52$ slightly overestimating the NTSC television display resolution at 50dpi (see text).

[4]To exaggerate the spatial distribution effect, Figures 48 and 49 use $R = 105$.

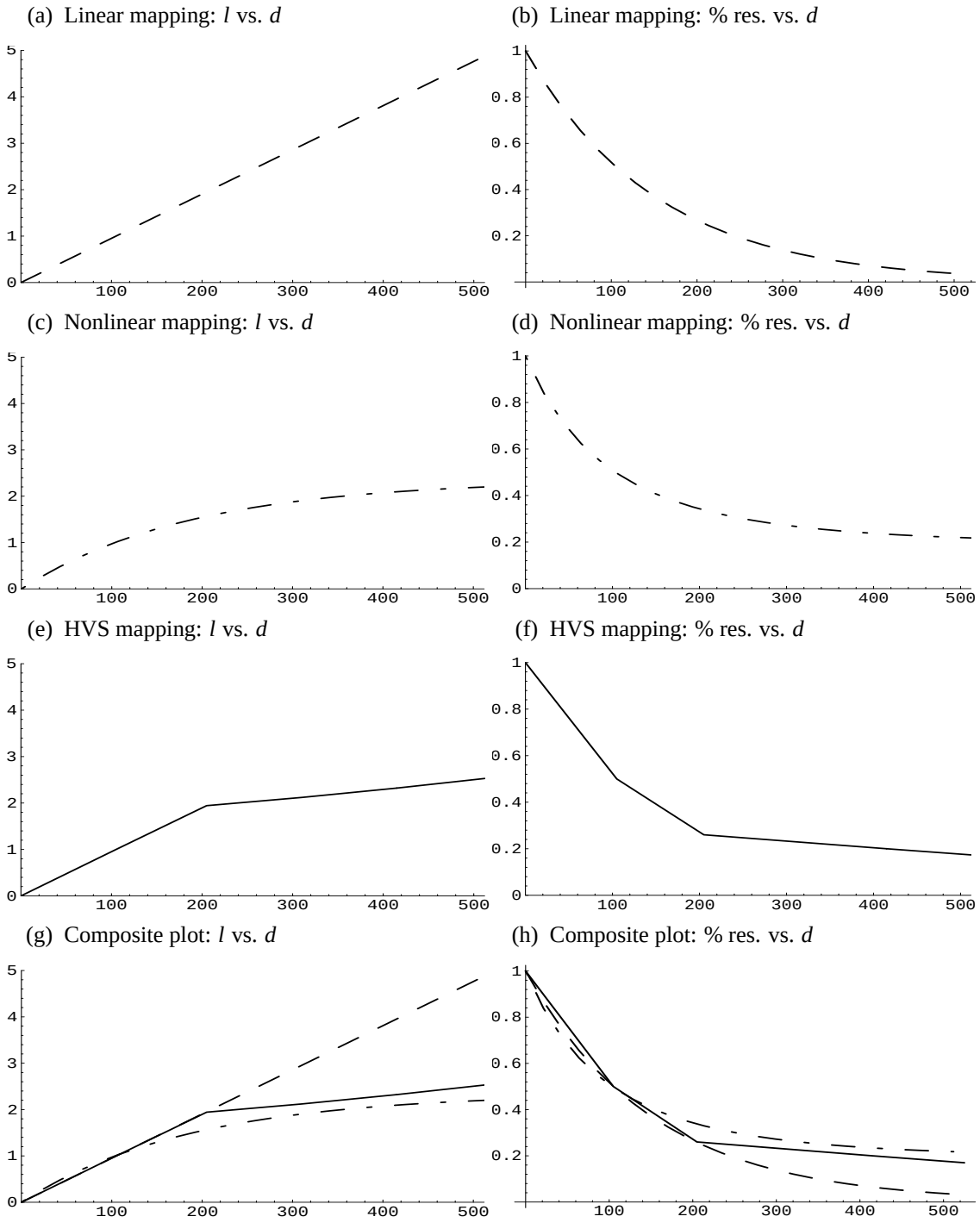


Fig. 48. Resolution mapping functions (assuming 100dpi screen resolution).

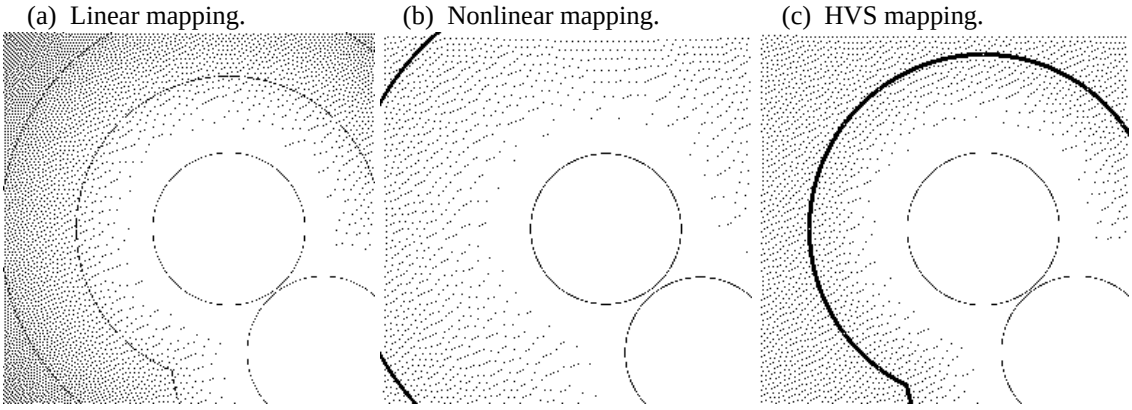


Fig. 49. Resolution bands in image space (assuming 100dpi screen resolution).

9.2.1 HVS Acuity-Matching Mapping

The HVS acuity-matching mapping was derived from empirical MAR (minimum angle of resolution) data [FGT89]. Assuming a 60cm viewing distance, acuity as a function of eccentricity is calculated and shown in Table 11. Letting θ represent MAR, minimum separability D is calculated by

TABLE 11
Resolution as function of eccentricity at 60cm viewing distance.

Eccentricity (deg.)	MAR (deg. at 75cm)	Min. Separability (inches, at 60cm)	Max. Resolution (dots per inch)
5	.14	.06	33
10	.28	.12	17
15	.32	.13	15
20	.38	.16	13
25	.45	.19	11

$$D = \frac{2r\tan(\theta/2)}{2.54},$$

where r is the viewing distance (60cm), 2.54 is the scaling factor used to convert the result to inches. Maximum resolution is calculated by assuming a resolution of 2 dots per minimum separability distance and converting to dots per inch, e.g., 33 = 2/0.06dpi.

Visual acuity at 5° eccentricity is roughly 50% of acuity at the fovea [Irw92]. Percent resolution is approximated by treating maximum resolution from Table 11 relative to 100% resolution at the fovea, i.e., 33dpi maximum resolution ≈ 50%, 17 maximum resolution ≈ 26%, etc. This form of linear approximation of visual acuity underestimates foveal resolvability at 66dpi. True foveal acuity is better described by the Modulation Transfer Function (MTF) which considers resolvable spatial frequencies of the retinal photoreceptors. The

MTF theoretically specifies the eye's resolvability limit—frequencies beyond the MTF cannot be resolved. The inter-cone spacing in the fovea is roughly 2.2μm. Within 1° visual angle about 136 foveal cones are contained in an area of $\sim$300μm. By the sampling theorem, this suggests a resolvable spatial Nyquist frequency of 68 c/deg. Empirical tests suggest an effective resolution of 60 c/deg [DD88, p.46]. Treating pixels as cyclic stimulus, 60 c/deg implies a resolvable resolution of 60/0.41 dots per inch, or 145.53dpi at 60cm viewing distance.

Using percent resolution and assumed screen display resolutions of 100, 50, 38, and 30 dots per inch, resolution level distances were calculated and are shown in Table 12. Resolution level distances are used to

TABLE 12
Resolution levels (in pixels).

Eccentricity	0-5°	5°	10°	15°	20°	25°
Resolution	100%	50%	26%	23%	20%	17%
Diameter subtended						
(cm)	–	5.2	10.4	15.8	21.2	26.7
(in)	–	2.1	4.1	6.2	8.3	10.5
(pixels @ 100dpi)	–	210	410	620	830	1050
(pixels @ 50dpi)	–	105	205	310	415	525
(pixels @ 38dpi)	–	80	156	236	315	399
(pixels @ 30dpi)	–	63	123	186	249	315

determine diameters of decreasing resolution areas. At 100dpi, the highest resolution region within each ROI, for example, is a circular region with a diameter of 210 pixels. Since resolution is distributed by decreasing powers of two, desired relative resolution at eccentricities are mapped into resolution space using the function for the band level, $l = -\ln(\%\ resolution)/\ln 2$. Percent resolution between bands is linearly interpolated prior to the log mapping.

A slight overlap is provided by the representation of foveal ROIs in order to cover the dynamic spatial variability of fixations. This is accomplished by slightly overestimating the resolution of a standard NTSC television display. The television's *x*- and *y*-dimensions are derived from the Pythagorean Theorem using the television's known diagonal measurement and the screen's aspect ratio. For example, a 21$''$ television measures roughly 16.8in $\times$ 12.6in. The CCIR square pixel format specifies a 640 $\times$ 480 pixel array, giving a resolution of 640/16.8 = 480/12.6 $\approx$ 38dpi. The current implementation of the reconstruction algorithm slightly overestimates the screen resolution at 50dpi. The fovea subtends 5°, covering $(60/2.54)\tan(5) \approx 2.06$in of the screen at 60cm, which at 38dpi is about 80 pixels, consistent with the value derived in Table 12. Using 105 pixels to represent foveal regions, instead of 80 pixels, gives 105in/38dpi = 2.76in corresponding to $\tan^{-1}(2.76/(60/2.54)) \approx 6.6°$ visual angle. The spatial distribution of fixations typically does not exceed to

0.4° full visual angle (± 0.2°) [Car77, p.105]. The 105 pixel-wide foveal ROI representation provides about 1.6° full angle spatial overlap, or ± 0.8° giving sufficient coverage in either (horizontal) direction.

9.3 Multiple ROI Image Segmentation

To include multiple ROIs within the reconstructed image, the image is partitioned into multiple regions. Image filtering is performed on a per-pixel basis, where the desired resolution at each pixel location is determined by the mapping function, relative to the center of only one of multiple ROIs. To select the appropriate ROI, each pixel is subjected to a membership test. This test involves measuring the distance from the pixel location to each ROI center. Using the Euclidian distance metric, the resolution level of the pixel is determined by the mapping function with respect to the closest ROI center.

Formally, the set $S = \{p_1, \ldots, p_n\}$ of n points in the plane, defined by ROI centers, defines a partition of the plane into n regions $V_1, \ldots, V_n$ such that any pixel in the region V_i is closer to the point p_i than to any other $p_j \in S$. This definition of the planar partitioning specifies the *Voronoi diagram* where each V_i is a convex polygonal region called the *Voronoi polygon* of the point p_i in S (for an alternate definition and construction of the Voronoi diagram, see [PS85, §5.5]). Voronoi diagrams have found diverse purposes in a number of disciplines. For example, in archaeology, Voronoi polygons are used to map the spread of the use of tools in ancient cultures, and in ecology, the Voronoi diagram of various territorial animals is used to investigate the effects of overcrowding (see [PS85, §5.2.2] for references). In image processing, the Voronoi diagram has recently been employed to order codewords in a principal component analysis codebook search algorithm for a vector quantization based coding system [LT96]. An example of the Voronoi diagram is shown in Figure 50(a).

A graphic representation of wavelet coefficient scaling (as discussed in §5.10.3) of an arbitrary image at two resolution levels is shown in Figure 50(b). White regions represents coefficients scaled by constant 1, black regions represent coefficient decimation (scaling by 0), and intermediate regions are scaled by linearly interpolated values in the interval $(0, 1)$. Note that the boundaries between linearly interpolated regions, i.e., boundaries between ROIs, are (by construction) Voronoi edges. In other words, Regions Of Interest, where wavelet coefficients are scaled producing spatially degrading resolution, constitute Voronoi polygons about the ROI centers. Although Voronoi partitioning may have little to do with the Human Visual System, it is nevertheless an elegant and natural way to partition the image plane.

(a) Voronoi planar partitioning. (b) Two-level wavelet coefficient scaling.

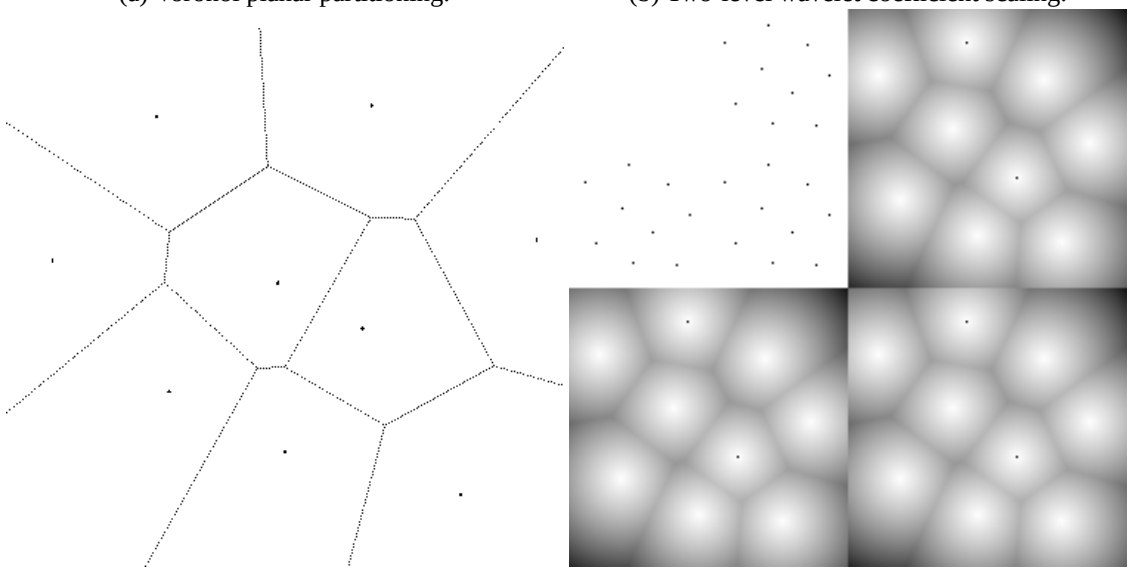


Fig. 50. Example of Voronoi partitioning.

9.4 Multiple ROI Image Reconstruction Examples

Examples of the variable resolution wavelet reconstruction technique are shown in Figures 51 and 52. The *cnn* sequence image was processed with two artificially placed ROIs, over the anchor's right eye and the "timebox" found in the bottom right corner of the image, respectively. Figure 51 shows, in the left column, the original image and its copy with circles imprinted over the ROIs.

The right column of Figure 51 depicts the extent of wavelet coefficient scaling in frequency space. The upper left quadrants of the wavelet space images show the wavelet scaling within the lower frequency bands of the wavelet transform representation of the image. Notice the distribution of the concentric resolution bands and the extent of white pixel regions. Pixel luminance symbolizes the degree of coefficient decimation, i.e., white pixels represent scaling by 1, black pixels represent scaling by 0.

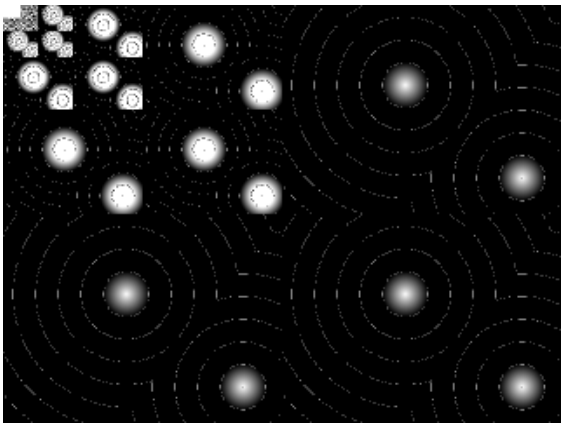
In the linear mapping, resolution bands are brought together to generate sharp degradation with respect to ROI centers. Coefficients are decimated at levels 1 (the bottom of the pyramid, or highest frequency bands), 2, 3, and 4. Nonlinear mapping spreads out resolution bands resulting in gradual degradation. Fewer coefficients are decimated at level 3 suggesting that more information is preserved (compare also the extent of white pixel regions in the nonlinear mapping to the HVS mapping, especially at decomposition level 2).

Reconstructed images are shown in Figure 52. To see the degradation effects, note the texture of the anchor man's tie, and the resolution of the anchor's right shoulder. Due to the length of the Daubechies-6 wavelets,

more information is contained within wavelet coefficients generating a smoother resolution reconstruction. Conversely, blocking artifacts are easily detected in the images processes with the length-2 Haar wavelets. The Haar images are provided as a reference when comparing resolution degradation in the images processed with Daubechies-6 wavelets.

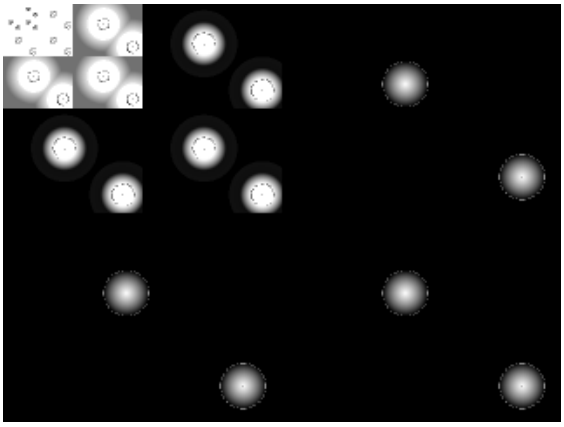
(a) Original image.

(b) Linear mapping.



(c) Imprinted ROIs.

(d) HVS mapping.



(e) Nonlinear mapping.

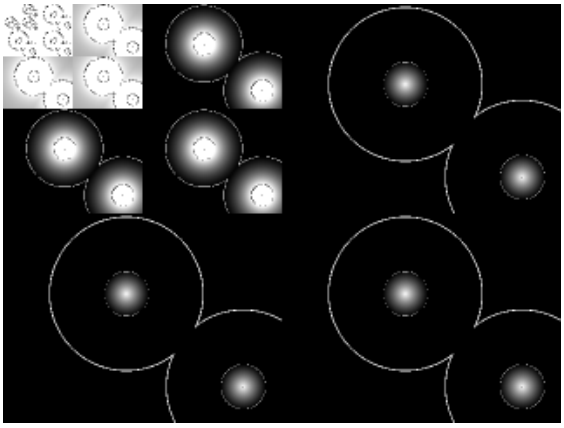


Fig. 51. Wavelet coefficient resolution mapping (assuming 50dpi screen resolution).

(a) Haar linear mapping.

(b) Daub-6 linear mapping.

(c) Haar HVS mapping.

(d) Daub-6 HVS mapping.

(e) Haar nonlinear mapping.

(f) Daub-6 nonlinear mapping.

Fig. 52. Image reconstruction (assuming 50dpi screen resolution).

CHAPTER X

EXPERIMENTAL METHOD AND APPARATUS

Three experiments were performed to test the adequacy of the wavelet-based eye movement signal processing technique and variable resolution video representation. Experimental objectives are outlined below.

1. Experiment 1 (Eye movement modeling): The purpose of this experiment is to evaluate the wavelet-based model of eye movements. Specifically, recorded eye movements are analyzed to test for correspondence of predicted and observed saccade locations.
2. Experiment 2 (Gaze-contingent VOI detection): The goals of this experiment are: (1) to visualize successive scan patterns of individuals over video sequences, and (2) to collect individual and aggregate Volumes Of Interest over video sequences from multiple subjects.
3. Experiment 3 (Gaze-contingent visual representation): The aim of this experiment is to test whether peripheral regions of the image can be degraded imperceptibly. This objective is significantly distinct from testing sensory-guided human performance (see XIII).

The objectives of the three experiments are related in the sense that (1) Experiment 1 tests the adequacy of the PARIMA eye movement model, (2) Experiment 2 applies the model to characterize eye movement patterns in terms of VOIs, and (3) Experiment 3 utilizes the VOIs to degrade digital imagery in a gaze-contingent manner. All gaze-contingent experiments were carried out in the Virtual Environments Laboratory, Department of Computer Science, Texas A&M University.

This section describes the gaze-contingent video display system developed for the purpose of recording eye movement data during simultaneous real-time video viewing. The system architecture, problems and lessons learned are discussed. The hardware and software configurations are described in §10.1 and §10.2, respectively. Experimental designs and results are described in §XI, §XII, and §XIII.

10.1 Hardware

The main hardware components of the eye movement recording system include an ISCAN eye tracker, an SGI 2-processor Onyx® RealityEngine2™ host computer equipped with a Sirius Video™ broadcast-quality video capture/display board, and a 21in standard NTSC television.[1] The front end of the eye tracker is shown in Figure 53. The subject setup includes a head/chin rest which provides limited head stability while main-

[1] Silicon Graphics, Onyx, RealityEngine2, are registered trademarks of Silicon Graphics, Inc. Sirius Video is a trademark of Silicon Graphics, Inc. As of this writing, see URL: `http://www.sgi.com/Misc/external.list.html` for a complete list of trademarks.

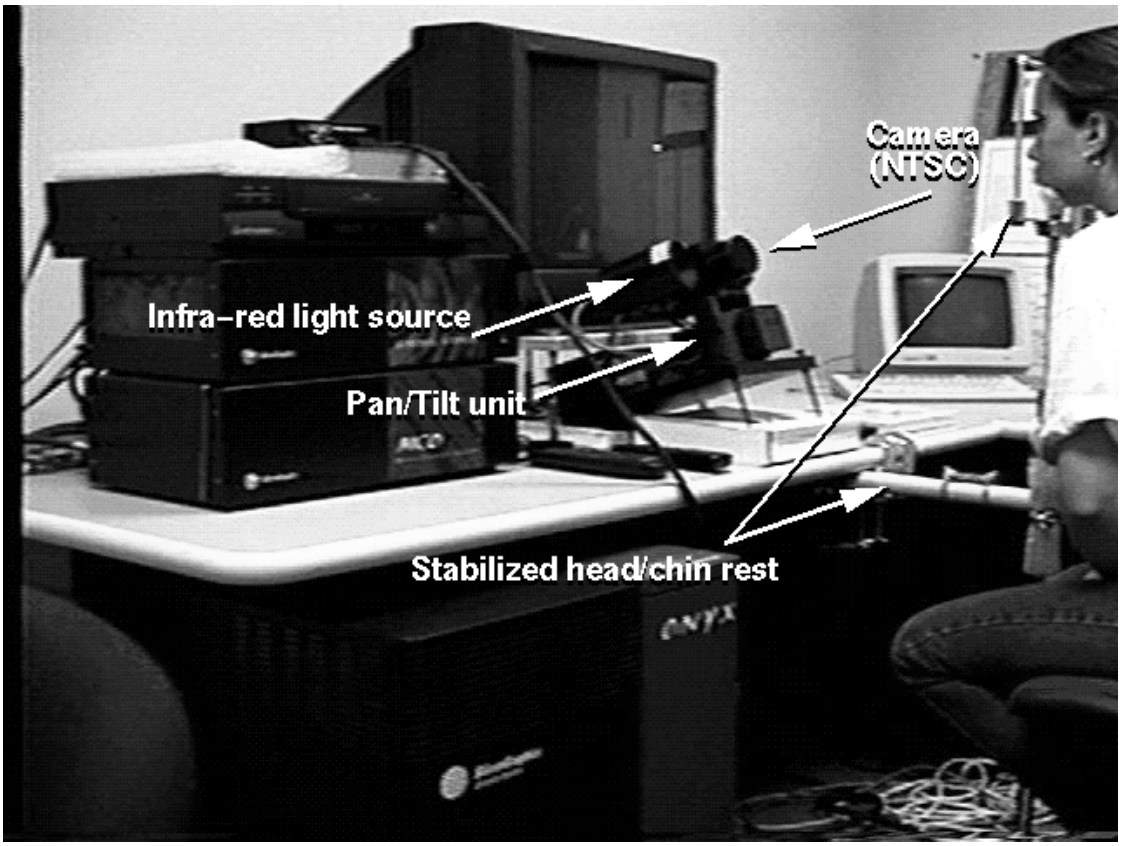


Fig. 53. Virtual Environments Laboratory: eye-tracking apparatus.

taining constant viewing distance for all subjects. Because the head-rest sits atop a monopod (not a tripod) base, subjects tended to tilt the head-rest while leaning on it. A simple clamp mechanism was installed to prevents this. The clamp mechanism connects the vertical head-rest monopod to the table on which the remote eye tracker rests. A long round wooden cylinder (broom handle) is used to position the head-rest relative to the television so that the eye-screen distance is maintained at about 60cm. This cylinder is fastened to the vertical head-rest pole at one end, and to a flat block of wood at the other. Butterfly (wing) nuts are used to allow positioning of the horizontal cylinder. Once measured, the assembly is tightened to stabilize the head-rest.

10.1.1 Eye Tracker

The infrared video eye tracker is composed of a desk-top digital camera and infrared light source connected to a dedicated personal computer (PC). Using proprietary software and hardware, the PC calculates the subject's real-time (60Hz) fixation position from the video image of the subject's corneal reflection of the infrared light source (first Purkinje image). In the present experimental setup, the eye tracker is treated as a black box delivering real-time fixation (x,y) coordinates over a 19.2 Kbaud RS-232 serial connection.

One of the inherent problems in any gaze-contingent system is the latency of the eye tracker. In the case of the ISCAN tracker, although its sampling frequency is 60Hz, the calculations required to obtain fixation position (e.g., calculations needed to disambiguate eye movements from head movements) incur an additional latency. The vendor guaranteed, however, that an updated fixation data word would be available on the serial line within a period not exceeding 18ms.

10.1.2 Video Format

The Sirius Video™ subsystem provides broadcast-quality video capture and playback. Standard NTSC video is captured from standard VHS video running on an off-the-shelf VCR. The NTSC video format (NTSC/component 525, CCIR square-pixel) provides 525 lines of resolution allowing 640×480 video frames [SGI95, p.240,p.268]. The SGI software libraries provide capabilities for the transfer of video frames encoded in RGBA format, with each channel represented by 8 bits for a total of 32 bits per pixel (other formats are also available but are not currently used). An in-house program was developed to to collect video frames in memory. Video frames are composed by interlacing NTSC video fields. Video playback is also facilitated through the Sirius board and an in-house program. Video frames are bisected into NTSC video fields before memory-to-video transfer.

Video display rate is constrained by the availability of processors and the number of concurrently competing processes. In order to meet the NTSC display rates the original target rate was set to 30fps. Testing has shown

that 30fps display rates are possible provided no other directly competing processes exist simultaneously on the machine. If, however, concurrent processes are competing for CPU resources, the display rate drops dramatically. In the case of the eye tracking system, a process concurrent with the video display routine is required to sample the serial port. Due to the competitive nature of these dual processes, both objectives of target sample and display rates could not be met simultaneously. Instead, a balance was sought where motion in the video sequence could still be perceived while eye tracker data was obtained at the fastest rates possible.

To afford a fast eye movement sampling rate, the target display was decreased to 15fps to match minimum motion perception requirements of the human visual system. Frame rates as low as 5fps have been suggested as a critical minimum rate for acceptable subjective quality (in the context of audio enhanced video conferencing) [PH95]. The critical interstimulus interval (i.s.i.) range required for the hybrid perception of stroboscopic and continuous motion is approximately 32-64ms (30-15fps) [Mor80]. This range corresponds to the two qualities of human vision responsible for seamless perception of television imagery: the *critical fusion frequency* of about 30Hz which allows flicker-free perception of strobe-like 30fps NTSC video (60 interlaced fields per second), and the perception of *apparent motion* of spatially displaced objects with i.s.i of roughly 64ms (15fps).

Testing has shown that a maximum sustained display rate of 16fps is achievable in conjunction with eye tracker data collection at (approximately) 60Hz. The multithreaded system developed for this task is described in §10.2 below.

The length of video sequences for recording and playback (to and from memory) is limited by the relatively small amount of RAM. The Onyx computer contains two MIPS® R4400™ processors with a 320 megabyte base memory.[2] With each pixel represented by 4 bytes, one video frame requires approximately 1.23 megabytes of storage. Theoretically, approximately 260 frames of uncompressed video may be stored in RAM. Compared to RAM, disk access is extremely slow and must be avoided in order to ensure fast video display rates. Since much of the available memory is consumed by the operating system and the user programs required to control video display, experience has shown that a maximum of 128 video frames may be stored in RAM before disk swaps are required.

The system limitations described above dictate an empirically optimum video sequence duration and display rate for experimentation. Video sequences composed of 128 NTSC (640 $\times$480) frames were shown at 16fps providing 8 seconds of stimulus duration. Since the primary goal of the experiments was the testing of the gaze-contingent variable resolution display strategy, monochrome (greyscale) video was used.

[2]MIPS is a registered trademarks, and R4400 is a trademark of MIPS Technologies, Inc.

10.2 Software

Experiments were carried out in the Virtual Environments Laboratory. The laboratory environment, with an operator and subject present, is shown in Figure 54. The gaze-contingent video (gcv) display system was

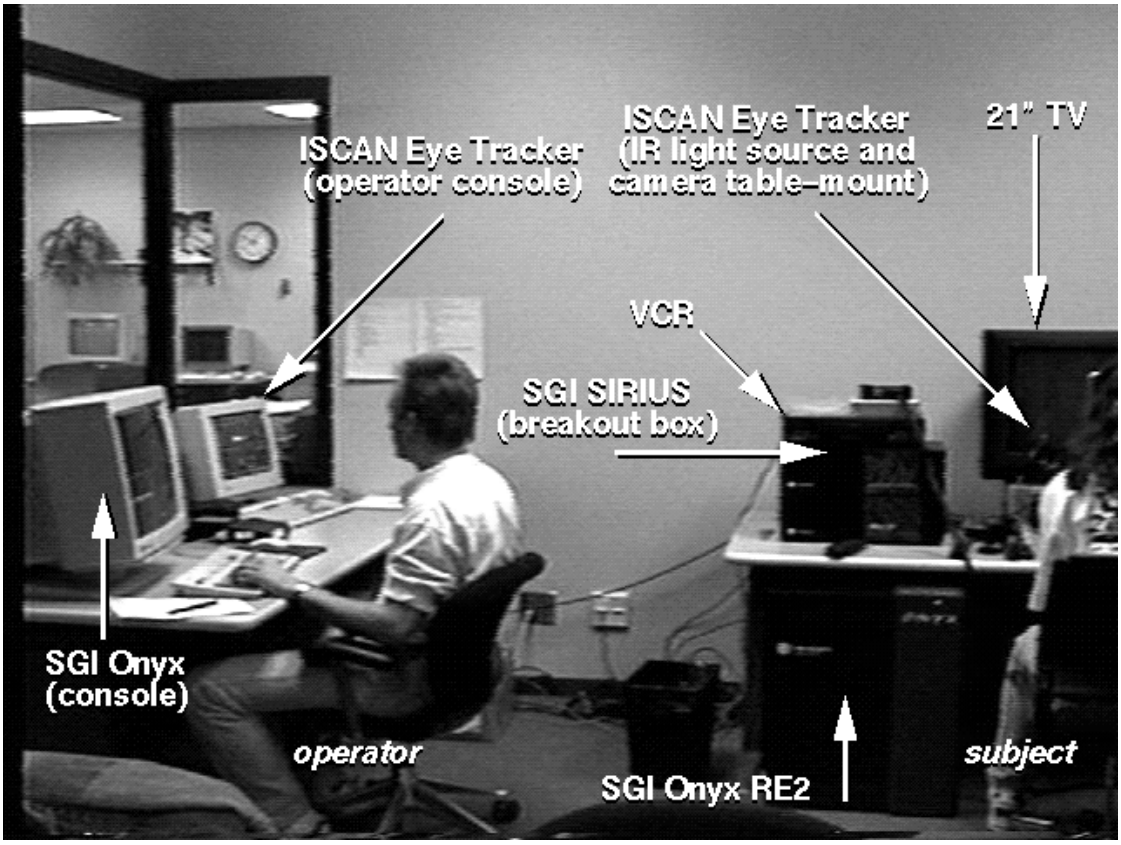


Fig. 54. Virtual Environments Laboratory: laboratory setup.

developed to simultaneously display video and obtain raw eye movement data. Due to hardware limitations (see §10.1 above), the system evolved over several revisions. Although initial attempts were developed independently, the final version in many ways parallels the virtual environment system described by Jacoby et al. [JAE96]. The implementation of the multi-process, shared memory gcv system evolved for very similar reasons (common hardware and software platforms), and in the case of shared memory, with the help of one of the authors. The minimization of end-to-end latency and effective update rate was the common driving force behind gcv's and Jacoby et al.'s designs (see [JAE96] for details).

Eye movement sample and video display rates of the gcv system are measured by taking time-stamp differ-

ences at specific points in the source code during each cycle of the time critical software loops.[3] Cycle times (periods) are inverted (i.e., 1/period = rate) yielding update rates for that software loop. In the case of the video display, a data structure was created which maintains the index of the currently displayed frame. This frame counter advances as soon as the video update rate exceeds the desired frame rate (16fps). The averaged sampled cycle time over the duration of the displayed sequence is reported giving an effective display rate. Note that (especially) on a UNIX system, sampled cycle times need to be averaged to account for the underlying stochastic variation of the observed context-switchable process [JAE96]. The sampling rate of the eye tracker data collection component is obtained similarly. In this case the next eye movement tuple (x,y) is not collected from the serial port buffer until the desired period has elapsed (18ms).

The `gcv` system organization was broken into seven main sub-processes (threads, or light-weight processes) to take advantage of the operating system's (IRIX™ version 5.3) real-time library extensions (via arguments to the system calls `sysmp` and `schedctl`). During development, it was found that the bottleneck of the system was the transfer of a video frame from memory to the Sirius board. This transfer is accomplished through a sequence of calls to the Video Library (`vlGetNextFree, vlGetActiveRegion`), the memory transfer (`memcpy`), and a final Video Library call (`vlPutValid`). Although the system manual pages offer `memcpy` as the fastest available memory transfer method, due to the voluminous nature of video data, this point in the video transfer loop was identified as the source of congestion. In early stages of development, when the `gcv` system was run as a single process, this video transfer delay prevented timely collection of eye movement data from the serial port buffer. Separating the competing software modules into concurrent process threads allowed more frequent context switching, thereby allocating system resources to the threads more fairly. A further gain in update rate performance was achieved by locking the video transfer process onto its own processor. Since the eye tracker data collection process requires a much smaller amount of data transfer (on the order of bytes), it was locked together with the remaining bookkeeping threads on the other available processor. Finally, all processes were given the same non-degrading priority to ensure fair treatment. In all, the following six sub-processes were spawned by the parent process via the `sproc` system call:

`MAIN` The main (parent) process: spawns 6 child processes and waits for the quit signal.

`VID` The video bookkeeping thread: in charge of maintaining video status, e.g., current frame number being

[3] Preliminary versions of the system used the C callable function `gettimeofday` which, according to the system manual pages, has a resolution of 1ms. Examination of various SGI users' comments and critiques of this utility on various SGI-related discussion newsgroups (e.g., `comp.sys.sgi.bugs`, `comp.sys.sgi.graphics`, `comp.sys.sgi.hardware`, `comp.sys.sgi.misc`) revealed that this mechanism may be unreliable and that its resolution may in fact be as low as 10ms. Finding this inadequate, an alternative time-stamp `dmGetUST`, part of the Digital Media Library™, was used instead as suggested by various SGI users. This time-stamp is maintained in SGI hardware and has a reported nanosecond resolution (it is a 64-bit number representing nanoseconds since the last system reboot). It has been found more than adequate for millisecond timing purposes although it is non-portable.

displayed, video play, video stop, etc. Signals process DRW.

TRK The tracking thread: in charge of maintaining tracking status (number of samples, sample rate, etc.) and the recording of eye movement data, known as the current Point Of Regard (POR). Signals process DSC.

DRW The drawing thread: solely in charge of signaling processes DVD and DSC.

DVD The draw-to-video thread: initiates memory to video transfer.

DSC The draw-to-screen thread: initiates screen update. This thread displays the current video frame on the Onyx monitor with an overlay of the current POR. Although this feature is visually informative for the experimenter, due to the video-to-framebuffer transfer, it is also a point of congestion. It is mostly used as a debugging aid and is turned off during experimentation.

GUI The graphical user interface thread: maintains vigil over user inputs.

Thread synchronization is achieved through the use of semaphores, provided by UNIX intrinsic system function calls (semget, semctl, semop). The gcv software system organization is shown in Figure 55, where semaphores are identified by a symbol representing a small flag. The raised flag symbolizes a semaphore signal operation, while the lowered flag depicts a semaphore wait. Semaphores are named after the process that they control, except for processes GUI, VID and TRK, which are controlled by the semaphore syn.

From the experiences described in [JAE96], it was decided to dissociate the RS-232 serial port driver process from the gcv system. Whether this programming strategy provides a benefit to the run-time performance of the gcv system is not known (evidence in [JAE96] suggests that it does), it certainly facilitates the development of a serial port driver. The result is the svr program which is solely responsible for accessing the serial port, reading the raw eye movement data, and updating a segment of shared memory from which the gcv system obtains the raw POR data. The gcv system obtains a pointer to the shared memory segment by first issuing the system call shmget with an agreed key to obtain the proper identification, and then by calling shmat to attach the pointer. Access to the memory is controlled by a semaphore previously created by svr. In effect, svr acts as a server/writer process always writing to the shared buffer, and gcv acts as the client/reader process always reading from the shared buffer. The svr program is responsible for destroying the semaphore and shared memory segments. This protocol requires that svr must be running prior to starting the gcv system.

Besides ease of development, the independence of the serial port interface provides two other important benefits. First, since the gcv system only reads from shared memory and not directly from the serial port, a phantom server may be substituted in place of the svr. In fact, a random eye movement simulator was created in this manner. The program sim was created to generate eye movements adhering to expected

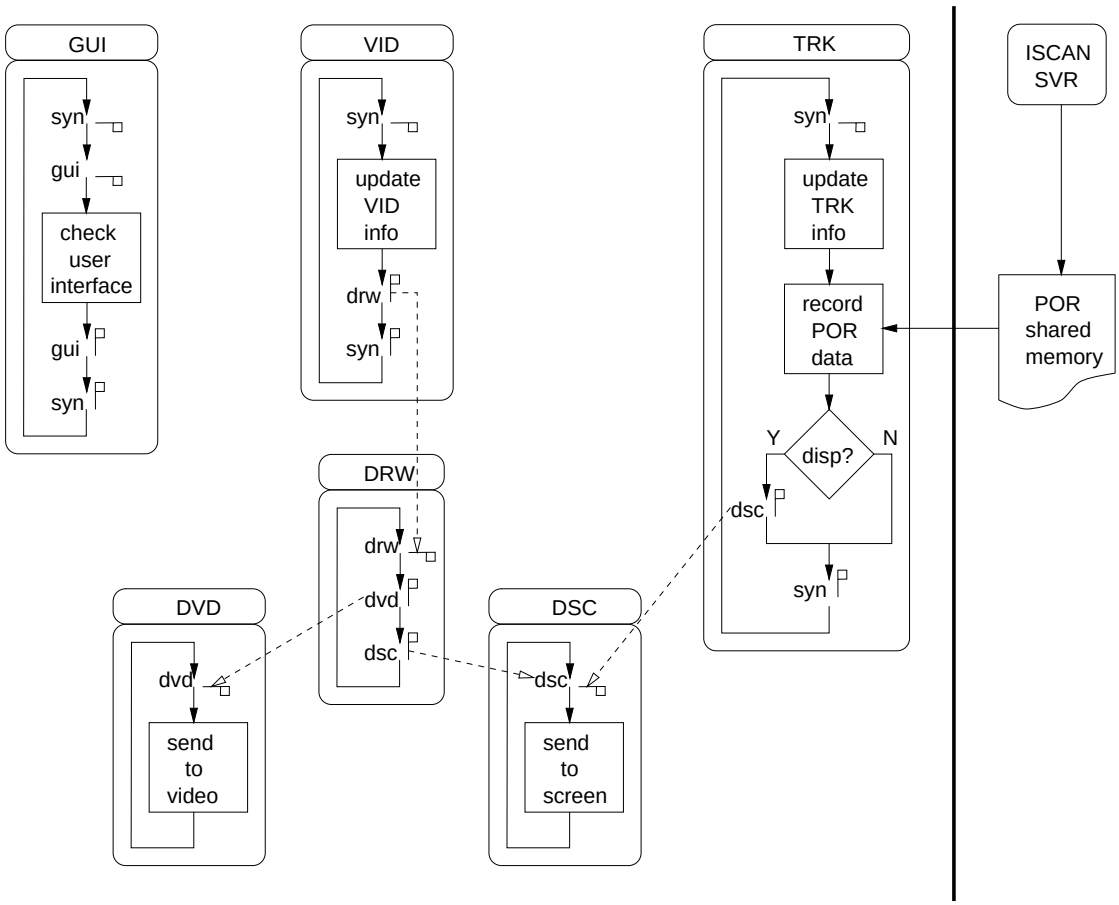


Fig. 55. Eye tracking software system organization.

stochastic characteristics of natural eye movements, but with a random spatial distribution. In a sense, `sim` may be either thought of as a blind individual with perfectly (in the statistical sense) functioning eyes, or simply as a context-free eye movement data generator. Second, since the real POR server (`svr`) only writes to shared memory, eye movements may be recorded over any type of imagery displayable on the television set. An example of an alternative eye movement analysis program is a gaze-contingent image display system. Instead of displaying video, single images may be shown for varying durations. The `gci` system was created to provide this functionality. Instead of accepting a desired video display rate, `gci` accepts a desired image display duration. Since there are no practical memory constraints associated with the single image (unlike the 128 recommended maximum frames for `gcv`), prolonged single image display durations are possible.

10.3 Calibration Procedures

Each experimental trial included three calibration steps. Calibration was divided into two procedures: *external* and *internal*. External calibration pertains to the proprietary eye tracker instrument calibration procedure specified by the manufacturer. Internal calibration pertains to the procedure developed within the `gcv` system in order to measure eye tracker accuracy. The eye tracker was externally calibrated using the vendor's proprietary 9-point calibration procedure. `Gcv` internal calibration was developed over 30 points. External calibration points were positioned to match internal outliers. The layout of the internal calibration points denoted by the symbol + is shown in Figure 56(a). The point arrays are framed for clarity, horizontal and vertical lines do not appear during calibration. Figure 56(b) shows the position of external calibration points (depicted by circles) overlayed on top of internal calibration points. The location of calibration points is

(a) Internal calibration points.

(b) External (eye tracker) calibration points overlayed over internal calibration points.

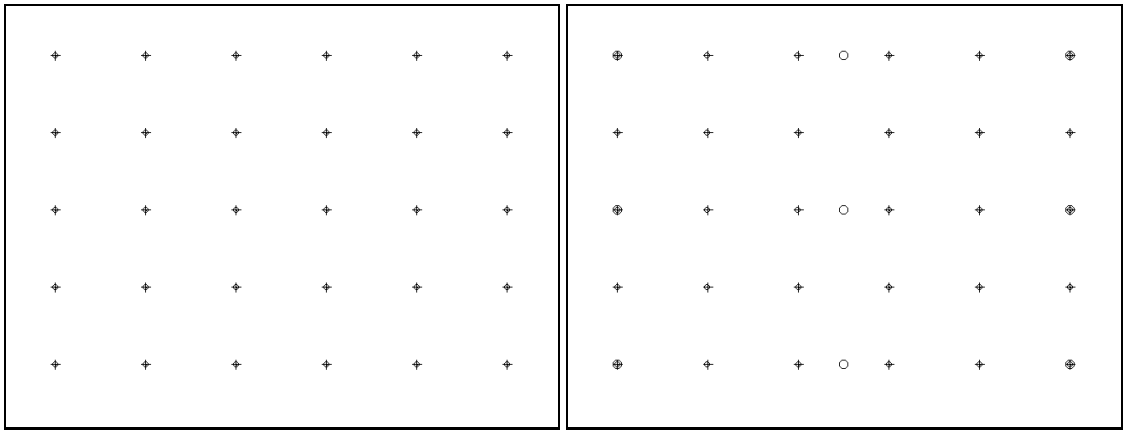


Fig. 56. Calibration stimulus.

described in two coordinate systems: the internal (`gcv`) image coordinate space, and the television coordinate space. Computation of a mapping transformation from (external) eye tracker coordinates to (internal) image coordinates is described in §10.4. The following description of the layout of both internal and external calibration points is in image coordinate space.

The description of the calibration point layout is based on a viewing distance of 60cm (23.622in). Viewing distance is maintained by a stabilized head/chin rest described above. Using television resolution specifications (NTSC standard), the stimulus dimensions are obtained from the following equations,

$$\frac{x}{y} = \frac{640}{480} = \frac{4}{3}$$
$$x^2 + y^2 = 21^2$$

where solving for the unknowns gives $y = 12.6$in., and $x = 16.8$in. Denoting the length of the base of the visual angle by r, the horizontal visual angle θ_x subtended by the subject is given by

$$\begin{aligned} \theta_x &= 2\tan^{-1}\left(\frac{r}{2D}\right) \\ &= 2\tan^{-1}\left(\frac{16.800}{47.244}\right) \\ &\doteq 39.2°, \end{aligned}$$

where $D = 23.622$ is the viewing distance in inches. The vertical visual angle θ_y subtended by the subject is obtained similarly,

$$\begin{aligned} \theta_y &= 2\tan^{-1}\left(\frac{r}{2D}\right) \\ &= 2\tan^{-1}\left(\frac{12.600}{47.244}\right) \\ &\doteq 29.9°. \end{aligned}$$

The effective resolution in dots per inch (dpi) of the television is found by dividing the number of pixels by the monitor dimensions,

$$\frac{640}{16.8} = \frac{480}{12.6} \doteq 38\text{dpi}.$$

The internal calibration point distribution is based on a 108 horizontal and 92 vertical pixel displacement starting at the top-left pixel location (50,50). The bottom-right calibration point is located at (590,418). At the above resolution, the horizontal inter-point distance is

$$r_x = \frac{108 \text{ pixels}}{38 \text{ dpi}} = 2.842\text{in},$$

resulting in the horizontal inter-point subtended visual angle of about 7° as derived below:

$$\theta_x = 2\tan^{-1}\left(\frac{r}{2D}\right)$$

$$
\begin{aligned}
&= 2\tan^{-1}\left(\frac{2.842}{47.244}\right) \\
&\doteq 6.88° \approx 7°.
\end{aligned}
$$

Similarly, the vertical calibration point displacement of 92 pixels results in a subtended visual angle of about 6°.

Internal calibration is performed twice, immediately after external calibration (before stimulus display), and immediately after the stimulus display. Hence internal calibration is used to check the accuracy of the eye tracker before and after the stimulus display, to check for instrument "slippage". The 30 internal calibration points are shown in random order in a semi-interactive manner similar to the external calibration procedure. As each point is drawn on the screen (the subject is presented with an $\times$ to minimize aliasing and flicker effects of the analog display), the `gcv` system waits for a input key before sampling eye movement data for a period of 800ms. The input key delay allows the operator to observe eye stability on the eye tracker's eye monitor. The eye is judged to be stable once the eye tracker has repositioned the eye in the center of the camera frame. Recorded point of regard data is mapped to image coordinates in in real-time. Calibration data is stored in a flat text file for later evaluation.

10.4 Eye Tracker-Image Coordinate Space Mapping Transformation

Two different coordinate spaces represent stimulus imagery and gaze position, respectively. Dimensions of stimulus imagery are based on the dimensions of the video display (640 $\times$ 480 pixels) while gaze position is dependent on the resolution of the eye tracker (512 $\times$ 256 pixels). Since gaze position information is sought in image coordinates, a transformation is sought mapping eye tracker coordinates to image coordinates. The mapping is graphically depicted in Figure 57, where image and eye tracker dimensions are shown approximately to scale. In this section, a linear mapping is derived between eye tracker and image coordinates,

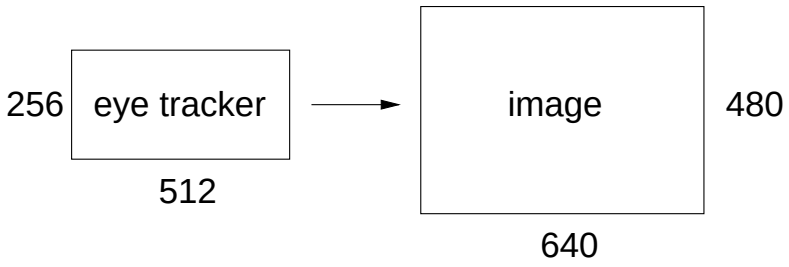


Fig. 57. Eye tracker-image coordinate space mapping transformation.

following a consideration of nonlinear display distortions presented below.

10.4.1 Stimulus Display Distortions

Video data presented on the television display is subject to a geometric transformation due to several possible optical distortions present in the display [RR93, p.58]. Robinett and Roland consider spherical aberration (SA), coma, astigmatism (AST), field curvature (FC), distortion and chromatic aberrations in the context of designing a stereoscopic head-mounted display. The authors describe nonlinear field distortion as the effect of straight lines appearing curved on the display. Although this type of severe aberration was not observed in the current television display, internal calibration points appeared slightly displaced from expected locations due to the curvature of the picture tube. This distortion is known as the pin-cushion effect [FvD82, p.105]. Pixel data is effectively spread out concentrically from the center of the screen, most noticeably in the corners of the display. The global effect is hardly noticeable, especially in viewing imagery, but local pixel perturbations are significant. To illustrate, the upper left calibration point at location (50,50) in image coordinates is displayed at a location near (21,21) in television coordinates.[4]

One possible compensation method for the optical nonlinearity of the display is a predistortion of the stimulus image. For example, to correctly display a straight line in the internal image, a curved line needs to be drawn on the external device, balancing out the optical distortion. This method is computationally expensive since it needs to be performed over the entire image. In the present eye tracking application, it is more important to obtain correct POR measurement instead of ensuring undistorted display of the stimulus. That is, instead of a predistortion transformation of the television input, a suitable method is sought for real-time, point-by-point correction of the eye tracker output.

The pin-cushion effect can be compensated for automatically, if an appropriate transformation can be derived from the eye tracker's measurements of displayed pixel locations. That is, if eye tracker coordinates can be passed through the same display distortion as the image coordinates, e.g., pixel information from both sources appearing on the same display device, the distortion effects will effectively cancel. This is the motivation behind two eye tracker data transformation methods described below, which are based on two different measurement techniques. In these sections, the terms "transformation" and "correction" are used interchangeably.

10.4.2 Eye Tracker Coordinate Transformation

The first attempt at eye tracker data correction, termed "manual", was performed by measuring the pixel locations of the internal calibration points as they appeared on the television screen. Pixel coordinates were obtained in inches and converted to pixel values. The second attempt, termed "automatic", used the eye

[4]The TV measurement is an approximate count of the television's RGB pixel triads.

tracker's fine-grained cursor position readout to estimate the coordinates of the displayed calibration points. Since the video signal is sent through the eye tracker, the calibration points are displayed on the eye tracker stimulus monitor. The eye tracker cursor can then be positioned over the calibration points by using the keyboard arrow keys. A status readout on the monitor shows the cursor's *x*- and *y*-coordinates in eye tracker coordinate space. Both manual and automatic approaches can be used to transform eye tracker coordinates. Manually and automatically measured points are shown in the left and right columns of Figure 58, respectively. The point arrays are framed for clarity, horizontal and vertical lines do not appear during calibration. Measured points are represented by circles joined with a line to the corresponding calibration point rep-

(a) Manually measured. (b) Automatically measured.

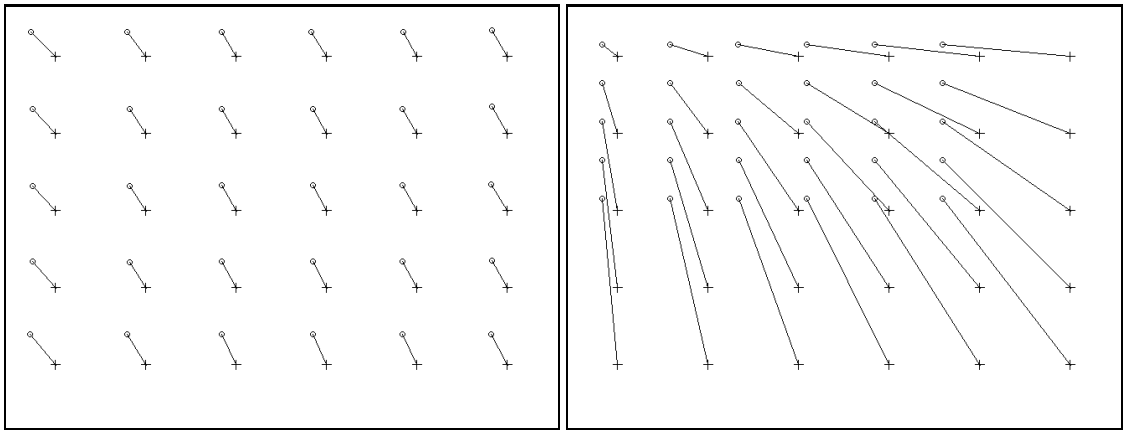


Fig. 58. Eye tracker-image coordinate transformation measurements.

resented by a $+$. The top row shows raw measurements of both manual and automatic approaches. The apparently larger discrepancy seen in the automatic approach is a result of the eye tracker's limited vertical resolution (approximately half the image height). The bottom row shows measured points corrected by two transformation methods described below.

10.4.3 Manual Transformation of Eye Tracker Data

The manual compensation method relies on bilinear interpolation among measured error values at four corner points closest to a received POR data point. The interpolation resolution is limited to the number of calibration points (30) stored in the "error lattice". Given an arbitrary POR (x,y) in the eye tracker coordinate space, the point is first linearly mapped from the eye tracker coordinate range to the image coordinate space, i.e., $(x,y) \mapsto (s,t)$, where s,t are image coordinates. The closest error points in image space are found by using truncated integer values of the (s,t) coordinates as indices into the error lattice, with (x_{00},y_{00}) denoting the

closest upper-left error point, i.e.,

$$\begin{aligned}(x_{00},y_{00}) &= (x_{\lfloor s\rfloor,\lfloor t\rfloor},y_{\lfloor s\rfloor,\lfloor t\rfloor}),\\ (x_{01},y_{01}) &= (x_{\lfloor s\rfloor,\lfloor t\rfloor+1},y_{\lfloor s\rfloor,\lfloor t\rfloor+1}),\\ (x_{10},y_{10}) &= (x_{\lfloor s\rfloor+1,\lfloor t\rfloor},y_{\lfloor s\rfloor+1,\lfloor t\rfloor}),\\ (x_{11},y_{11}) &= (x_{\lfloor s\rfloor+1,\lfloor t\rfloor+1},y_{\lfloor s\rfloor+1,\lfloor t\rfloor+1}).\end{aligned}$$

The values (s,t) are then converted to interpolation parameters by the following rule:

$$(s,t) = (s-\lfloor s\rfloor, t-\lfloor t\rfloor),$$

defining s as the x-coordinate interpolant and t as the y-coordinate interpolant. Finally, new image coordinates (x',y') are obtained through bilinear interpolation:

$$\begin{aligned}x' &= (1-t)\left[(1-s)x_{00}+(s)x_{01}\right]+(t)\left[(1-s)x_{10}+(s)x_{11}\right]\\ y' &= (1-t)\left[(1-s)y_{00}+(s)y_{01}\right]+(t)\left[(1-s)y_{10}+(s)y_{11}\right]\end{aligned}$$

Corrected calibration points are shown in Figure 58(e).

10.4.4 Automatic Transformation of Eye Tracker Data

The automatic measurement method abandoned bilinear interpolation in favor of Lagrange's method of least squares [LS86, §2.5]. Lagrange's method was chosen in order to estimate the matrix **B** expressing the two-dimensional transformation required to bring the observed points into alignment with the calibration points. Minimizing the error of the transformation parameters over the thirty sample points gives a one-time calculation applicable to all raw data points as they are obtained. Thus instead of a lookup table, as is required by the bilinear interpolation method, the transformation compensation required as a result of Lagrange's minimization is a 3×2 matrix. The form of this matrix and the estimation of its coefficients are described below.

Switching notation for calibration points from §10.4.3, denote the sampled (observed) lattice points by $\{x_{i1},x_{i2}\}$ and the internal calibration points by $\{y_{i1},y_{i2}\}$ for $i\in[1,30]$. Assuming matrix **B** can be found, the transformation required to bring the observed points into alignment is, in general for $i\in[1,n]$, expressed by Equations (10.1) and (10.2):

$$\begin{bmatrix} y_{11} & y_{12}\\ y_{21} & y_{22}\\ \vdots & \vdots\\ y_{n1} & y_{n2}\end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12}\\ 1 & x_{21} & x_{22}\\ \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2}\end{bmatrix}\begin{bmatrix} \alpha_1 & \alpha_2\\ \beta_{11} & \beta_{12}\\ \beta_{21} & \beta_{22}\end{bmatrix} \tag{10.1}$$

$$\begin{bmatrix} y_{i1} & y_{i2}\end{bmatrix} = \begin{bmatrix} 1 & x_{i1} & x_{i2}\end{bmatrix}\begin{bmatrix} \alpha_1 & \alpha_2\\ \beta_{11} & \beta_{12}\\ \beta_{21} & \beta_{22}\end{bmatrix}, \tag{10.2}$$

or in matrix notation,

$$\mathbf{Y} = \mathbf{XB}.$$

Writing out Equation (10.2) gives

$$\begin{aligned} y_{i1} &= \alpha_1 + \beta_{11} x_{i1} + \beta_{21} x_{i2} \\ y_{i2} &= \alpha_2 + \beta_{12} x_{i1} + \beta_{22} x_{i2} \end{aligned}$$

with α_1 and α_2 denoting the translation parameters and β_{ij} denoting scale/rotation factors. The sought matrix **B** is a two-dimensional homogeneous coordinate transformation matrix.

Matrix **B** is estimated by Lagrange's method of least squares, or the multivariate multiple regression model [Fin74, §4.3–§4.5]. The general linear model describing $i \in [1,n]$ observations of p random variables $y_k,\ k \in [1,p]$, from q predictors $x_j,\ j \in [1,q]$ is specified by the following p separate univariate equations:

$$\begin{aligned} \begin{bmatrix} y_{i1} & y_{i2} & \cdots & y_{ip} \end{bmatrix} &= \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p \end{bmatrix} \\ &+ x_{i1} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \end{bmatrix} \\ &+ x_{i2} \begin{bmatrix} \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \end{bmatrix} \\ &\ \ \vdots \\ &+ x_{iq} \begin{bmatrix} \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix} \\ &+ \begin{bmatrix} \varepsilon_{i1} & \varepsilon_{i2} & \cdots & \varepsilon_{ip} \end{bmatrix}, \end{aligned}$$

where α_k and β_{jk} are the linear regression parameters relating x_{iq} to y_{ip} with random error ε_{ip}. The sample estimate is expressed in matrix form by Equations (10.3) and (10.4):

$$\begin{bmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1q} \\ 1 & x_{21} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nq} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 & \cdots & \hat{\alpha}_p \\ \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1p} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{q1} & \hat{\beta}_{q2} & \cdots & \hat{\beta}_{qp} \end{bmatrix} + \begin{bmatrix} \hat{\varepsilon}_{11} & \hat{\varepsilon}_{12} & \cdots & \hat{\varepsilon}_{1p} \\ \hat{\varepsilon}_{21} & \hat{\varepsilon}_{22} & \cdots & \hat{\varepsilon}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\varepsilon}_{n1} & \hat{\varepsilon}_{n2} & \cdots & \hat{\varepsilon}_{np} \end{bmatrix} \tag{10.3}$$

$$\begin{aligned} \begin{bmatrix} y_{i1} & \cdots & y_{ip} \end{bmatrix} &= \begin{bmatrix} 1 & x_{i1} & \cdots & x_{iq} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 & \cdots & \hat{\alpha}_p \\ \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1p} \\ \hat{\beta}_{21} & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{q1} & \hat{\beta}_{q2} & \cdots & \hat{\beta}_{qp} \end{bmatrix} + \\ &\begin{bmatrix} \hat{\varepsilon}_{i1} & \hat{\varepsilon}_{i2} & \cdots & \hat{\varepsilon}_{ip} \end{bmatrix}, \end{aligned} \tag{10.4}$$

or in matrix notation,

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}} + \hat{\mathbf{E}},$$

where matrix $\mathbf{X}$ is composed of rows of values of the predictor variables for n observations, $\hat{\mathbf{E}}$ is the $n \times p$ matrix of sample residuals or errors, and matrix $\hat{\mathbf{B}}$ is the $(q+1) \times p$ matrix of partial regression coefficients for predicting each outcome measure from the p independent variables.

In order to estimate the matrix $\mathbf{B}$ by the method of least squares, the squared sample residuals of n sampled observations are minimized. The sum of squared residuals for one outcome measure is one diagonal element of $\hat{\mathbf{E}}^T\hat{\mathbf{E}}$, and their sum is the trace of $\hat{\mathbf{E}}^T\hat{\mathbf{E}}$ [Fin74, p.110].[5] That is, the sum of squared residuals is given by

$$tr(\hat{\mathbf{E}}^T\hat{\mathbf{E}}) = tr\left[(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})\right],$$

which is minimized by setting the partial derivatives with respect to the elements of $\hat{\mathbf{B}}$ to zero and solving. The resulting normal equations are

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{B}} = \mathbf{X}^T\mathbf{Y}.$$

The system is left-multiplied by $(\mathbf{X}^T\mathbf{X})^{-1}$ to obtain the estimate of $\hat{\mathbf{B}}$:

$$\begin{aligned} \hat{\mathbf{B}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{G}^{-1}\mathbf{X}^T\mathbf{Y}, \end{aligned}$$

where $\mathbf{G}^{-1}$ is known as the pseudo-inverse of $\mathbf{X}$. For details on the implementation of this method, the estimability criterion, and the invertibility of $\mathbf{G}$, see [Fin74, §4.4].

In the particular case of the internal calibration points, setting the parameters $p = 2$ and $q = 2$ results in a 3×2 estimate matrix $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix}$$

where the matrix $\mathbf{G} = \mathbf{X}^T\mathbf{X}$ is the 3×3 symmetric covariance matrix

$$\mathbf{G} = \mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix}.$$

For the 30 automatically measured points, the solution of matrix $\hat{\mathbf{B}}$ is estimated by the matrix

$$\hat{\mathbf{B}} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} 7.58 & -22.00 \\ 1.33 & 0.00 \\ 0.00 & 2.00 \end{bmatrix}.$$

[5] The trace of a square matrix $\mathbf{M}$, denoted by $tr(\mathbf{M})$, is defined as a sum of its diagonal elements.

Reverting to the calibration point notation of §10.4.3, corrected POR data (x',y') is obtained from raw POR data (x,y) through the transformation,

$$\begin{aligned} x' &= 7.58 + 1.33x \\ y' &= -22.00 + 2.00y. \end{aligned}$$

As expected, the translation and scale factors correspond to the dimensions and relative spatial location of eye tracker and image coordinates. In practice, a restricted subset of the eye tracker coordinate space was used to estimate corresponding locations of calibration points in image space. The scale value of 1.33 corresponds to the scale factor between the eye tracker and image *x*-coordinates, $[32,438]$ and $[50,590]$, respectively. The dimension ratio $(590-50)/(438-32)$ gives 1.33. The *y*-coordinate scale value of 2.00 is obtained given eye tracker and image *y*-coordinate ranges, $[36,220]$ and $[50,418]$, respectively. Translation factors may be similarly derived from the relative location of the upper-left corners of the coordinate spaces, shown overlayed, approximately to scale in Figure 59.

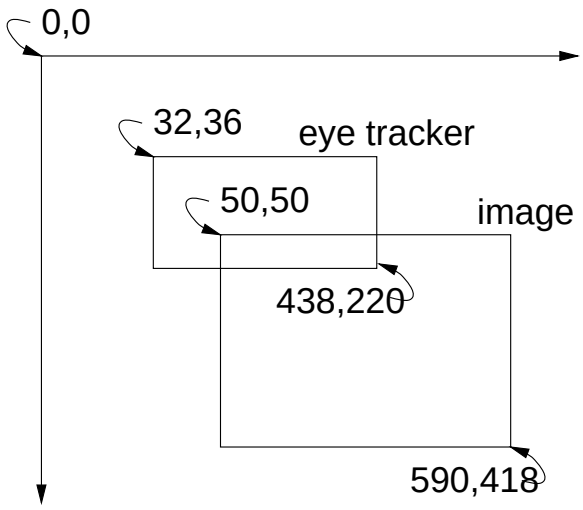


Fig. 59. Eye tracker-image coordinate space overlay.

10.4.5 Comparison of the Transformation Methods

The measurement of observed calibration point locations (manual or automatic) is independent from the correction method used. In sections §10.4.3 and §10.4.4 above, the manual measurement method was paired with the bilinear interpolation correction and the automatic measurement was paired with the method of least squares. Corrected calibration points are shown in Figure 60. Either correction method can be used with each form of measurement. However, the automatic method directly circumvents display effects since eye tracker

(a) Bilinear interpolation. (b) Least squares.

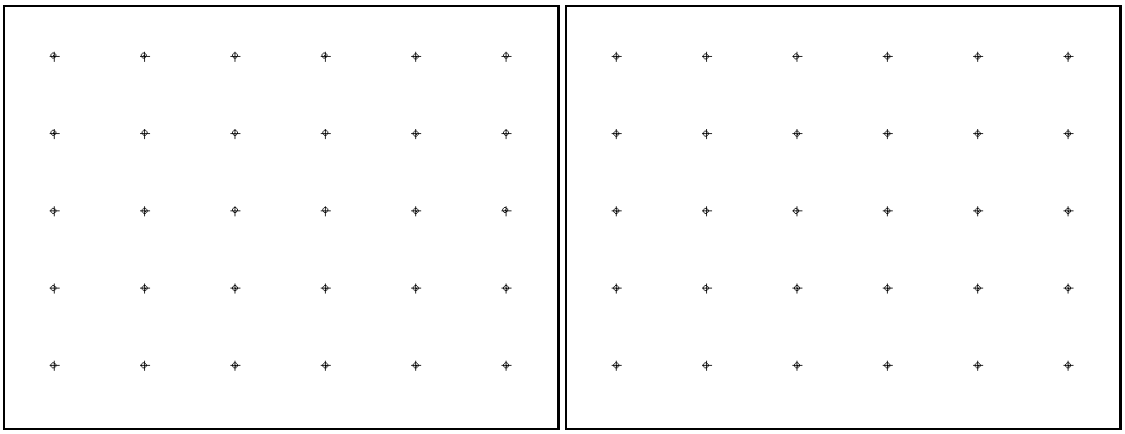


Fig. 60. Eye tracker-image coordinate transformation results.

measurements of image calibration points are carried out on the stimulus display, and hence are subject to the same distortion. Furthermore, the eye tracker provides a more precise measurement technique for estimating calibration point locations thereby eliminating the introduction of another source of measurement error. For these reasons the automatic method is incorporated into the `gcv` system.

In general, the method of least squares is capable of evaluating potential rotational errors. In the present case only the linear regression model was considered in §10.4.4, since nonlinear display distortions effectively cancel out. Conceivably, higher order terms in the regression model could provide more robust correction parameters should this need arise.

CHAPTER XI

EXPERIMENT 1: EYE MOVEMENT MODELING

The goal of this experiment is to evaluate the wavelet-based model of eye movements. Specifically, recorded eye movements are analyzed to test for correspondence of predicted and observed saccade locations.

11.1 Video Sequences

Three 8-second, 16fps video sequences are used as stimulus. Each sequence is composed of a single white dot (10-pixel diameter) subtending roughly 1° visual angle. The dot randomly simulates fixations, saccades and smooth pursuit eye movements. That is, fixation durations, saccadic repositioning and smooth pursuit movements are drawn from random distributions of a range of values corresponding to known eye movement characteristics. Three sequences are used for all subjects, denoted by *sim1*, *sim2*, and *sim3* and referred to as "random-dot" sequences.

Fixations are modeled by ARMA feedback sequences,

$$\begin{aligned} x_{t_x} &= \mu_{t_x} + x_{t_x-1} + \varepsilon_{t_x}, \quad a \le t_x \le b, \\ y_{t_y} &= \mu_{t_y} + y_{t_y-1} + \varepsilon_{t_y}, \quad a \le t_y \le b, \end{aligned}$$

where $\varepsilon_{t_x} = \varepsilon_{t_y} \sim N(0,(5/4)^2)$ were chosen to simulate small perturbations about the mean for a given interval. Variations roughly correspond to a 3×3 square pixel region which, measuring $3\sqrt{2}$ pixels along the diagonal, subtends 1/4° visual angle. The typical spatial distribution of fixation dwell times roughly extends to 1/4° (full angle), 75% of the time [Car77, p.105]. Samples were generated uniformly every 18ms, with each constant-mean interval lasting an average of 375ms with a Poisson distribution. That is, interventions (saccades) are distributed with an average inter-arrival time of 375ms. An intervention is induced by the instantaneous change of means μ_{t_x} and μ_{t_y} (stimulus dot location), distributed by the feedback relation

$$\begin{aligned} \mu_{t_x} &= \mu_{t_{x-1}} + \varepsilon_{\mu_{t_x}}, \\ \mu_{t_y} &= \mu_{t_{y-1}} + \varepsilon_{\mu_{t_y}}, \end{aligned}$$

where $\varepsilon_{\mu_{t_x}} \sim N(0,32^2)$ and $\varepsilon_{\mu_{t_y}} \sim N(0,16^2)$ is chosen to simulate spatial saccade amplitudes. The μ_{t_x} and μ_{t_y} distributions roughly correspond to a quarter of the size of the eye tracker's rectangular resolution window (twice the 32×16 standard deviations giving a 128×64 pixel rectangle). Smooth pursuits are randomly initiated whenever the stimulus dot happens to fall in marginal regions of the image defined to be 100 pixels

wide. Pursuits are simulated by the ARMA feedback sequences,

$$\begin{aligned} x_{t_x} &= \mu_{t_x} + x_{t_x-1} + \varepsilon_{t_x}, \quad a \leq t_x \leq b, \\ y_{t_y} &= \mu_{t_y}, \end{aligned}$$

where $\varepsilon_{t_x} =\sim N(11,(3)^2)$ was chosen as a constant increment in the x-direction during the pursuit duration. The increment falls within a range of $[1,21]$ pixels, which at 27dpi subtends 0.06-1.34° visual angle. The position of the stimulus dot is updated every 62.5ms (for 16fps frame rate) giving a velocity range of roughly $[1,20]$°/s. The increment value is made negative if the stimulus dot happens to start at the right image margin. The resultant dot movements occur at random y-coordinates and traverse the screen from left-to-right or right-to-left at a random constant velocity.

11.2 Experimental Trials

Each experimental session consists of multiple subjects tested individually in experimental trials. Each trial consists of single presentations of three random-dot video sequences. Trials are limited to a total of three video presentations since loading of the video into memory requires 7 minutes and the experiment is designed to process each individual in roughly 30 minutes. Each trial consists of the following steps:

1. Brief introduction. When a subject enters, s/he is asked to sit in front of the eye tracker head rest. The equipment is briefly described with emphasis on the eye tracker infra-red (IR) light source and camera. It is pointed out that the IR assembly contains a standard overhead projector bulb as the light source. This is done to alleviate any preconceived fears regarding the apparatus (some subjects were under the impression that eyelid movement would be restricted). Subjects are assured that the experiment is physically unobtrusive.
2. No training is given to subjects.
3. Video presentation. Each video presentation consists of the following substeps:
   (a) External calibration. Once the subject has settled into the head rest, the eye tracker is calibrated, as discussed in §10.3.
   (b) Internal calibration. Immediately following external calibration, the internal calibration procedure is performed to record the initial accuracy of the eye tracker. The calibration results are stored in a text file for later analysis (the file name convention adopted is (`<seq_name>.1.clb`).
   (c) Stimulus display. The video is presented twice in succession each time eye movements are recorded and stored (the file name conventions adopted are (`<seq_name>.por`).
   (d) Internal calibration. Immediately following stimulus presentation, the internal calibration procedure is performed once again to record the final accuracy of the eye tracker. The calibration results

are stored in a text file for later analysis (the file name convention adopted is (`<seq_name>.-2.clb`).

11.3 Subjects

A total of 7 subjects (4 female, 3 male) participated in Experiment 1. The age distribution was mean 21, minimum 19, and maximum 23. The subjects were recruited from an introductory course (CPSC 203) for non-engineering majors offered by the Department of Computer Science. Subjects' majors ranged from Biology (BIOL) to Political Science (POLS). The undergraduate level distribution was 0 freshmen, 2 sophomore, 2 juniors and 3 senior. All subjects had good vision with 1 subject wearing glasses, 2 wearing contact lenses. There were no subjects from the Departments of Computer Science or Electrical Engineering.

11.4 Experimental Design

The predominant problem which hampers eye movement experiments is the invariability of human scanpath patterns over complex scenery. Scanpath variability emanates from two sources. First, natural scenery is viewed differently by different individuals. Although certain elements in the scene may be fixated consistently between individuals, the order of fixations may vary greatly. Second, an individual's scanpaths may differ greatly on successive views of the same scene [NS71a, NS71b].

To limit scanpath variability, subjects may be coerced to concentrate on certain visual features by the introduction of a performance criterion. Simply asking subjects to follow a prominent object in a video sequence greatly reduces inter- and intra-subject variability. This *visual tracking* experimental paradigm is useful in testing peripheral sensitivity. Note that this paradigm differs from the visual search task since the object to be fixated is generally conspicuous and easily found. The visual tracking paradigm also differs from typical tasks used in peripheral sensitivity experiments since the point of fixation is not restricted to a given screen location, e.g., the screen center. Instead, the directed paradigm resembles smooth pursuit tasks since the subject is instructed to maintain gaze over a given object, whether it is still or in motion.

Random-dot video sequences were presented in serial order. Since there was no control factor, no prescribed presentation order was necessary. It is not expected that subjects viewed sequences differently, but this point is not explicitly tested in the analysis.

11.5 Results

The objective of Experiment 1 is to evaluate the PARIMA model of eye movements. Specifically, since the PARIMA model is based on the detection of saccades, the analysis centers on comparison of expected and detected saccade locations. Descriptive statistics are presented regarding "hit rate" (percentage detection of total saccades) and "correctness rate" (percentage detection of correct saccades). These statistics are then used to make a rough qualitative assessment of the PARIMA model and its susceptibility to Type I and II errors (false positives and false negatives).

Usefulness of PARIMA saccade detections in the video sequences are contingent on two factors of the eye tracking experiment: (1) accuracy of the eye tracker, and (2) accuracy of the eye tracker during the viewing task. Gaze position is not verified in this experiment since gaze position does not bear significance on the evaluation of the signal processing strategy. Gaze position is examined in Experiment 3 where peripheral image manipulation depends on viewers' collocation of gaze with the intended regions of interest. Prior to the saccade detection analysis, eye tracker accuracy is tested in the following two sections.

11.5.1 Verification of Eye Tracker Accuracy

The measurement of gaze depends on the accuracy of the eye tracking instrument. Eye tracker accuracy was measured by internal calibration procedures described in §10.3. Measurements were taken before and after stimulus viewing trials, as discussed in §11.2. These procedures provide the basis for two statistical measures: (1) the overall accuracy of the eye tracker, and (2) the amount of instrument slippage during stimulus viewing. The latter measurement gives an indication of the instrument accuracy during the viewing task, i.e., by recording loss of accuracy between the before- and after-viewing calibration procedures.

Eye tracker readings were obtained over 30 internal calibration points as described in§10.3. Each calibration point measurement consists of eye tracker samples about the calibration point over an 800ms period (approximately 44 individual data points). Raw sample points falling in the exterior 10-pixel wide borders are ignored. This is due to the eye tracker's property of generating (0,0) values during blinks (confirmed by the vendor). For this reason, any time a raw sample point is close enough to the location (0,0) (within 10 pixels), it is removed from further consideration. An average of valid data points (centroid) is obtained and the error between the centroid and calibration point is calculated. Each two-dimensional euclidian distance measurement is converted to the the full visual angle dependent on the viewing distance and calculated resolution of the television screen. Thus each calibration run contains 30 average measured deviations at each calibration point in terms of visual angle. A graphical example of this measurement is shown in Figure 61. The internal calibration locations are represented by $+$, sample measurements are represented by individual pixel dots,

(a) Calibration before stimulus viewing (avg. error: 1.40°).

(b) Calibration after stimulus viewing (avg. error: 1.77°).

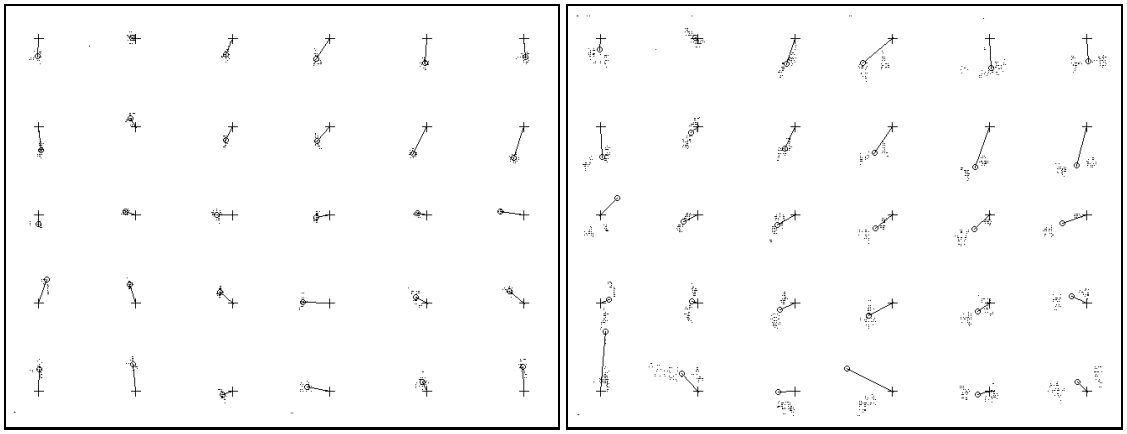


Fig. 61. Typical per-trial calibration data (subject # 21).

and centroid gaze positions are represented by circles, joined with the corresponding calibration point by a line. The length of the line is the average error deviation in pixels. This distance, r is converted to visual angle θ by the calculation

$$\theta = 2\tan^{-1}\frac{r}{2D},$$

where D is the viewing distance. The error distance r is measured in the same units as the viewing distance D, dependent on the resolution of the display.

To quantify the overall eye tracker accuracy succinctly, the average calibration error is obtained from each set of calibration points in order to calculate an overall average statistic of the eye tracker. The resulting average instrument error is an average statistic over all calibration runs performed in Experiment 1. The calculated mean value is 1.87°. This is not a particularly informative statistic since the data does not appear to fit a normal distribution. The histogram of average errors is shown in Figure 62. Since the average error data appears skewed, a more meaningful statistic is the median value, which ignores the influence of outliers. Its value is 1.41°. Using similar reasoning for reporting a dispersion statistic, the interquartile range (iqr) is utilized instead of the standard deviation for its robust response to outliers. The iqr value is 0.97°. These findings indicate an overall acceptable performance, not far off from the vendor's claimed accuracy (roughly 1°visual angle).

11.5.2 Verification of Eye Tracker Slippage

Quantification of the before- and after-viewing eye tracker error provides a measure of instrument accuracy during the viewing task. A graphical example of this measure is shown in Figure 63 which is a composite

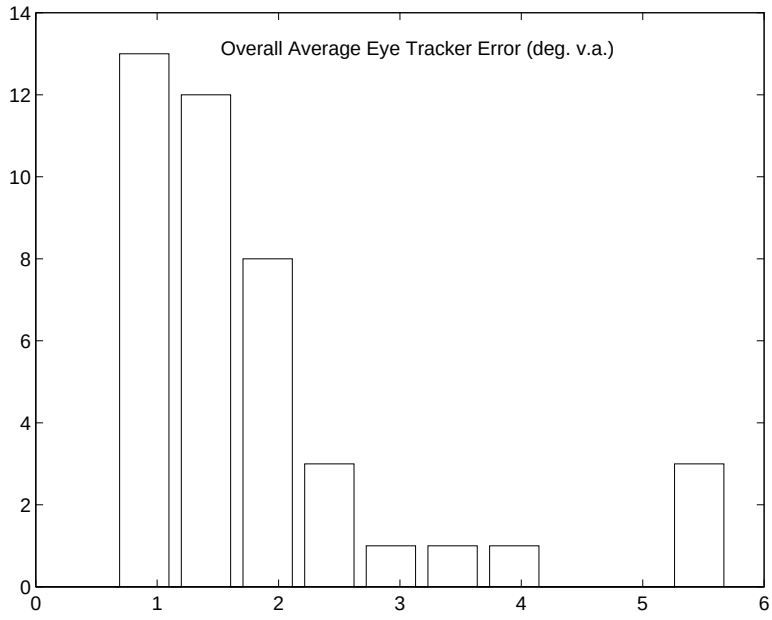


Fig. 62. Overall eye tracker error histogram.

plot of Figures 61 (a) and (b). Notice that measured eye positions in relation to calibration points coincide well overall. To quantify this correspondence, a one-way ANOVA was performed on the means of the before- and after-viewing average error measures. Table 18 lists the ANOVA measures in §C. Error mean boxplots are shown in Figure 64. On average, no significant slippage is detected by this statistic. Note that ANOVA in this case is not very informative since it does not consider eye tracker slippage on a per-trial basis. That is, the ANOVA only reports significant correspondence of the mean measurements.

To examine eye tracker slippage on a per-trial basis, differences of average errors were calculated between the before- and after-viewing calibration runs on a per-run basis. Difference measurements fit a skewed distribution, as shown in Figure 65. Due to the apparent skewed distribution, statistical measures robust to outliers are used. The median error is -0.25°, and interquartile range is 0.40°. These values quantify the close correspondence of example pre- and post-viewing calibration measurements shown graphically in Figure 63. Overall, the tracker accuracy varies roughly a quarter of a degree visual angle over the 8-second viewing task. Over some calibration points, accuracy improves while over others it degrades. Since the peak frequency is close to 0, generally the accuracy before and after viewing the stimulus remains stable overall.

11.5.3 Evaluation of PARIMA Model of Eye Movements

The PARIMA model of eye movements characterizes fixations and smooth pursuits by detecting saccades modeled by mean discontinuities. The resultant eye movement analysis is a reconstructed version of the

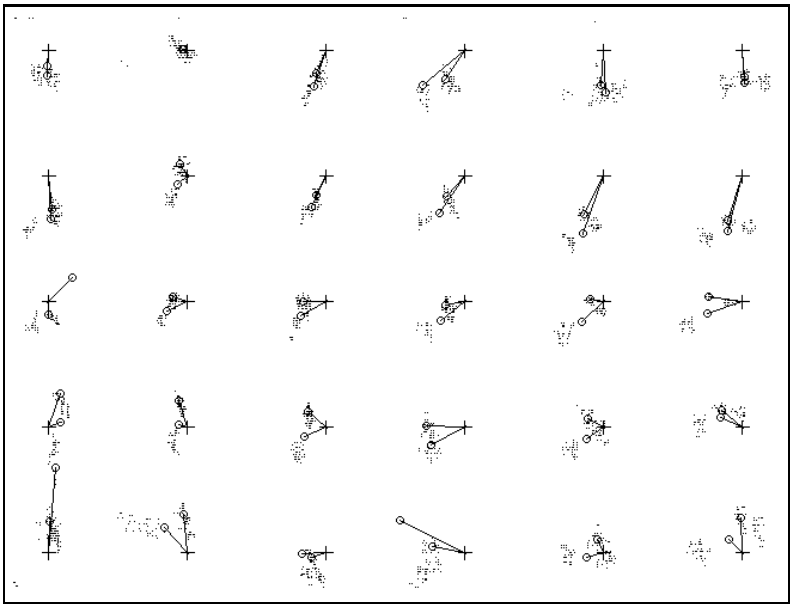


Fig. 63. Composite calibration data showing eye tracker slippage (subject # 21).

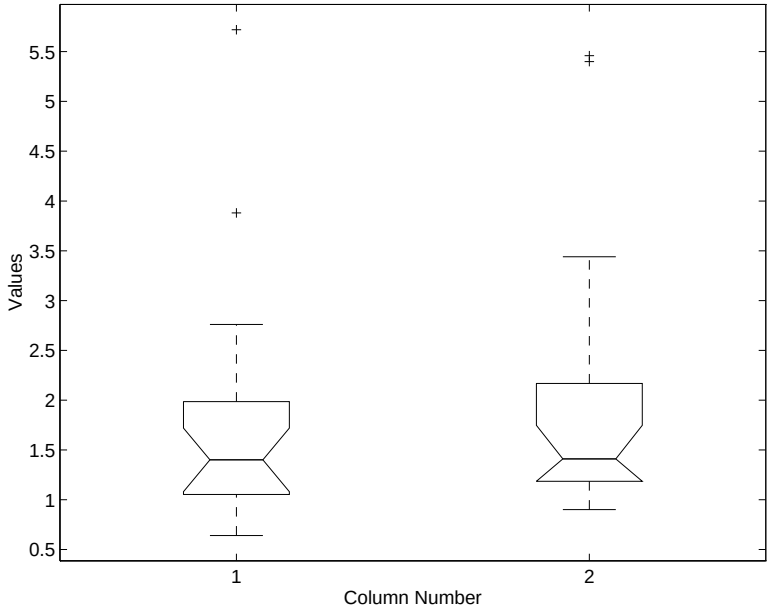


Fig. 64. Pre- vs. post-stimulus viewing average calibration error boxplots.

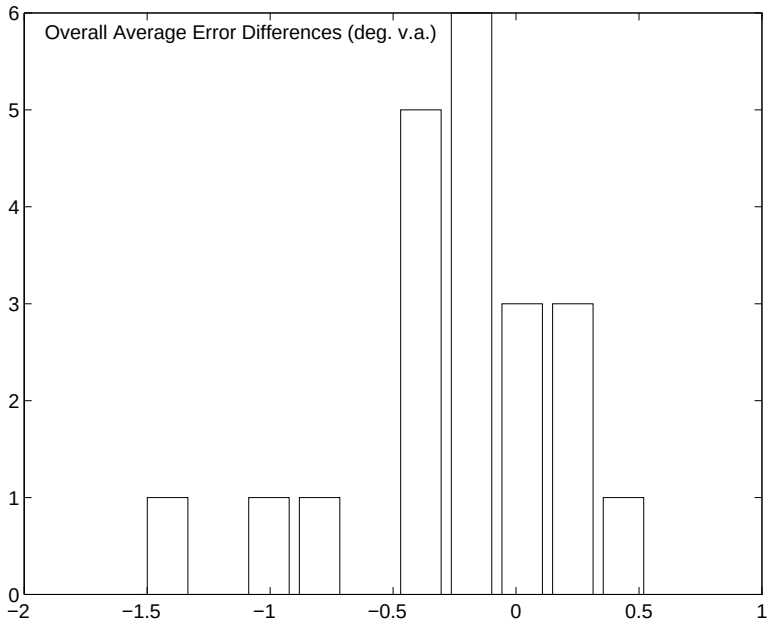


Fig. 65. Overall difference error histogram.

original eye movement signal with removed saccadic events. The model is evaluated through comparison of detected and expected saccade locations. Mean “hit” and “correctness” rates are discussed. Data for calculation of these statistics is given in Tables 19–21.

The mean percent “hit rate” of the model is estimated at 45%. This statistic is obtained by calculating the ratio of correctly identified saccades versus the total number expected over each experimental trial. For example, the first random dot sequence (*sim1*) contained 14 expected saccades of varying spatial amplitude. Analysis of subject 21’s scan patterns identified 9 of these saccades correctly (64% hit rate). Hit rates for all subjects were averages to produce the mean hit rate of 45%.

The mean “correctness rate” of the model is estimated at 29%. This estimate is obtained by taking the ratio of correctly identified saccades versus the total number of saccades detected in the (individual’s) eye movement signal. For example, 27 saccades were detected in subject 21’s eye movement patterns over sequence *sim1* (see Table 19). Of these, 9 saccades matched the expected saccade location giving a 27% correctness rate of the model for this scan pattern.

Saccade detection rate was qualitatively analyzed to test for saccade spatial amplitude bias. That is, at first glance it appeared that saccade spatial amplitude may influence detection, e.g., large saccades may be more easily detected. To test this, expected saccade spatial amplitudes were calculated (far-right columns in in

Tables 19–21). Saccade detections across individuals are plotted against saccade amplitude in Figure 66. No

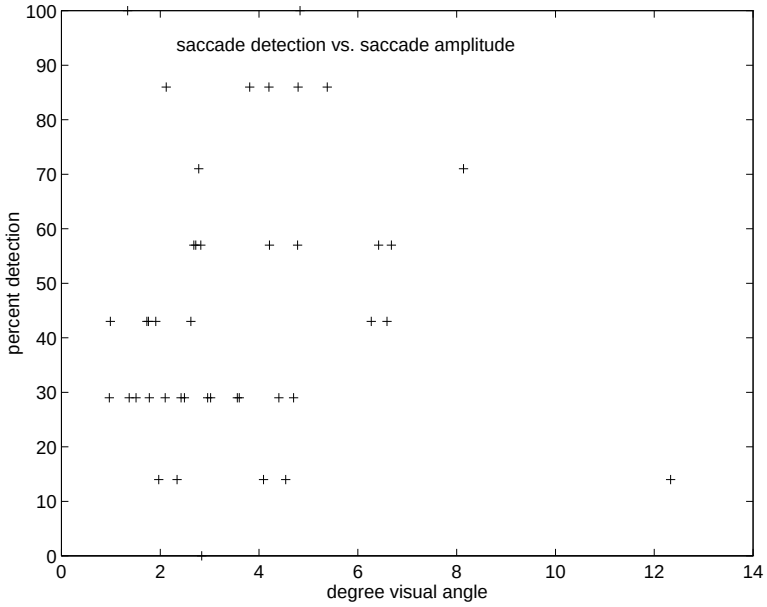


Fig. 66. Percent saccade detection vs. saccade spatial amplitude.

obvious trend is visible in the plot of Figure 66. That is, data points appear to be roughly uniform about the vertical line $x = 3.63$, which is the mean of saccade amplitudes of the visual stimulus. In particular, both saccades with spatial amplitudes of 1.34° and 4.83° visual angle were always detected (100 percent detection) suggesting that when present, the PARIMA model had no difficulty locating either of these two saccade signatures. This is not surprising since once the spatial decomposition threshold is set, the three-dimensional wavelet analysis should perform discontinuity detection equally well no matter what the spatial saccade amplitude (so long as it is over the spatial threshold).

11.6 Discussion

The hit and correctness rates identified above indicate the PARIMA model's possible susceptibility to Type I and Type II errors. The hit rate suggests a Type II (false negative) error, i.e., failure to identify a true saccade. The correctness rate suggests a Type I (false positive) error, i.e., identification of false saccades. Low estimated values of both measures imply that the PARIMA model classifies too many non-saccades and suffers from a fairly large miss rate. In essence, it seems the PARIMA model overestimates saccades. At least four possible problems may be responsible for the somewhat poor initial evaluation of the PARIMA model: (1) stringent classification of saccade matches, (2) experimental bias toward high velocity saccades, (3) incorrect

initial estimation of model parameters, and (4) confounding saccade-like events.

The criteria used to judge saccade matches (generation of positive or null saccade detection values in Tables 19–21) was very strict. Saccade matches were identified if observed saccades matched expected locations exactly. That is, if a saccade between video frames 001-002 was expected, an observed saccade was deemed a match if frames 001-002 were not part of any VOIs, i.e., if frames 001-002 were part of a VOI hole. No error tolerance was given, i.e., if frame 002 was part of a VOI, the observed saccade was deemed a mismatch. This may be too stringent a criterion since it does not allow any delay between the stimulus saccade and the observed saccade. More liberal estimates may be used which provide a larger temporal interval for saccade matches. Continuing the current example, providing a one-frame tolerance level would produce a saccade match for the expected saccade between frames 001-002 if an observed saccade occurred between frames 002-003, or between frames 001-002 where frame 002 is the first frame of a VOI. Further relaxation schemes may improve the hit rate estimate. The point is, the exact match criteria used in the current analysis probably underestimates the power of the PARIMA model.

Related to the stringent saccade match criteria is the distribution of the stimulus saccades. Fast stimulus saccades may evoke delayed saccade responses in the eye movement signal. That is, a fast stimulus saccade may incur a longer response time from the subject. At present, practically all stimulus saccades completed within one video frame duration (62.5ms). The range of stimulus saccades used is $[.97, 12.33]°$ visual angle, generating saccade velocities of $[15.52, 197.28]°/s$. Slower saccades may generate different response times which may give more robust estimates of the PARIMA model's capability.

PARIMA analysis of eye movements in all cases used constant values for wavelet spatial and temporal decomposition thresholds (3 spatial and 2 temporal decomposition levels were used). Criteria for initial choices of these parameters are given in the model description, in §VII. Since the above analysis suggests a fairly high Type II error rate, the model should be evaluated using larger spatial decomposition levels. A larger spatial decomposition effectively smoothes the spatial variability of the eye movement signal, limiting saccade detection to large spatial discontinuities.

Finally, only a weak attempt was made to distinguish possible confounding events from saccades. In particular, blinks may be falsely interpreted as saccades. The eye tracker generates a data value of $(0, 0)$ during full eyelid closure (loss of pupil in the eye tracker optics). Prior to full lid closure, however, the eye tracker signal corresponding to initial eyelid movement resembles signals typically associated with saccadic activity. That is, eyelid movement causes an apparent saccade towards the upper left corner of the image. At present, care is taken to remove values at the point $(0, 0)$ and at a close distances to the image boundaries. However, sam-

pled data remain within central image regions corresponding to apparent saccades generated by blink onsets just prior to full eyelid closure. A graphical example of this problem is shown in Figure 67. Note the dot

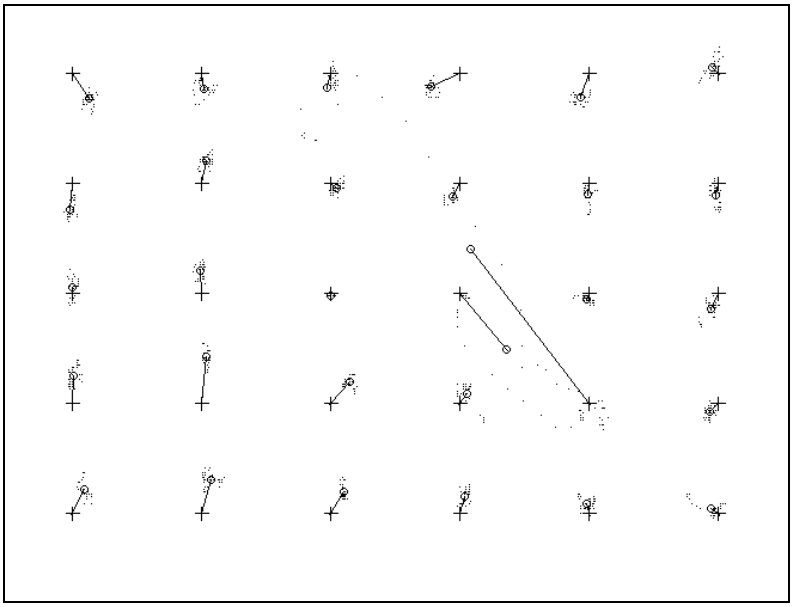


Fig. 67. Calibration data showing partial eye blink (subject # 28).

patterns representing sampled POR data at the two calibration points with largest error. These dot patterns, along with the vector defined by the internal calibration point and the calculated centroid of the POR data, clearly indicate a path towards the upper-left regions of the screen. These eye tracker samples were recorded during partial eyelid closures (just prior to blinks) and are not removed from analysis (the average error of the calibration shown in Figure 67 is 1.41°, with range $[0.13, 10.33]$°). A more sophisticated blink detection algorithm (based on pupil measures for example) may improve the PARIMA model.

In summary, the initial evaluation of the PARIMA model performed here probably underestimates the power of the model. Presently it appears as if the PARIMA model overestimates saccades. Suggestions were given for more robust analysis of the data, as well as for the variation of both model parameters and stimulus characteristics in future experiments.

CHAPTER XII

EXPERIMENT 2: GAZE-CONTINGENT VOI DETECTION

The goals of this experiment are: (1) to visualize successive scan patterns of individuals over video sequences, and (2) to collect individual and aggregate Volumes Of Interest over video sequences from multiple subjects. Recorded individual and aggregate scan patterns are obtained for comparison against ideal observer patterns (in Experiment 3).

12.1 Video Sequences

Two 8-second, 16fps video sequences are used as stimulus. Each sequence is presented in its original state as digitally captured from an analog video source (a VCR). First and last frames of the *cnn* sequence are shown in Figure 68. Not shown in Figure 68 are the first and last frames of the *flight* sequence. In this sequence, a

(a) *cnn* (frame 001). (b) *cnn* (frame 128).

Fig. 68. Unprocessed video sequences.

Navy fighter chase jet, in the lower right portion of the image, is following a smaller target plane which starts at the top right portion of the first frame. The target plane performs a half-roll evasion maneuver arcing from the top-right to the top-left screen regions through the bottom of the screen. The horizon rotates as the chase plane pursues. The chase plane remains steady relative to the frame since it bears the camera. Lighting on the chase plane changes as it emerges from shadow to sunlight.

Individuals' successive scan patterns are recorded over the *cnn* sequence while the *flight* sequence is used to

obtain an individual's representative scan pattern for comparison versus the ideal observer.

12.2 Experimental Trials

The experimental session consists of multiple subjects tested individually in experimental trials. Each trial consists of three presentations of the *cnn* sequence and two presentations of the *flight* sequence. Each trial consists of the following steps:

1. Brief introduction. When a subject enters, s/he is asked to sit in front of the eye tracker head rest. The equipment is briefly described with emphasis on the eye tracker infra-red (IR) light source and camera. It is pointed out that the IR assembly contains a standard overhead projector bulb as the light source. This is done to alleviate any preconceived fears regarding the apparatus (some subjects were under the impression that eyelid movement would be restricted). Subjects are assured that the experiment is physically unobtrusive.
2. Training. Prior to the start of the trial, the *flight* sequence is shown three times to the subject from video tape. Training is done so that subjects can familiarize themselves with the video content and can identify the target object for the visual tracking task. No training is performed over the *cnn* sequence and subjects are not told of its content in advance.
3. Video presentation. Each video presentation follows the following substeps:
   (a) External calibration. Once the subject has settled into the head rest, the eye tracker is calibrated, as discussed in §10.3.
   (b) Internal calibration. Immediately following external calibration, the internal calibration procedure is performed to record the initial accuracy of the eye tracker. The calibration results are stored in a text file for later analysis (the file name convention adopted is (`<seq_name>.1.clb`).
   (c) Stimulus display. The video is presented in succession with eye movements recorded and stored each time (the file name conventions adopted are (`<seq_name>.1.por`, `<seq_name>.2.por`, and `<seq_name>.3.por`, with only two data sets recorded for the *flight* sequence).
   (d) Internal calibration. Immediately following stimulus presentation, the internal calibration procedure is performed once again to record the final accuracy of the eye tracker. The calibration results are stored in a text file for later analysis (the file name convention adopted is (`<seq_name>.2.clb`).

12.3 Subjects

A total of 18 subjects (8 female, 10 male) participated in Experiment 2. The age distribution was mean 19.61, minimum 18, and maximum 23. The subjects were recruited from an introductory course (CPSC 203) for non-engineering majors offered by the Department of Computer Science. Subjects' majors ranged from Business (BUSN) to Psychology (PSYC). The undergraduate level distribution was 6 freshmen, 6 sophomore, 3 juniors and 3 seniors. All subjects had good vision with 1 subject wearing glasses, 5 wearing contact lenses. There were two subjects from the Departments of Computer Science or Computer Engineering.

12.4 Experimental Design

Video sequences were presented in serial order. Since there was no control factor, no presentation order was used. Subjects viewed three presentations of the *cnn* sequence and two presentations of the *flight* sequence. The visual tracking paradigm was used over the *flight* sequence, and no viewing instructions were issued over the *cnn* sequence (free viewing paradigm). It is not expected that subjects viewed the *flight* sequence differently, but this point is not tested in the analysis.

12.5 Results

The objectives of Experiment 2 are the collection of aggregate VOIs over the *cnn* sequence and an individual's scan patterns over the *flight* sequence.

The choice of subjects' scan patterns for the construction of individual and aggregate VOIs is contingent on two factors of the eye tracking experiment: (1) accuracy of the eye tracker, and (2) accuracy of the eye tracker during the viewing task. In the case of the individual VOI collection, the choice of the subject's scan patterns is also contingent on the accuracy of gaze position over the intended (ideal observer) Regions Of Interest. Gaze position is not verified for subjects viewing the *cnn* sequence as there is no predetermined scanpath. Prior to the visualization of the individual and aggregate VOIs, eye tracker and gaze position accuracy is given in the following two sections.

12.5.1 Verification of Eye Tracker Accuracy

The measurement of gaze depends on the accuracy of the eye tracking instrument. Eye tracker accuracy was measured by internal calibration procedures described in §10.3. Measurements were taken before and after stimulus viewing trials, as discussed in §12.2. These procedures provide the basis for two statistical measures: (1) the overall accuracy of the eye tracker, and (2) the amount of instrument slippage during stimulus

viewing. The latter measurement gives an indication of the instrument accuracy during the viewing task, i.e., by recording loss of accuracy between the before- and after-viewing calibration procedures.

Eye tracker readings were obtained over 30 internal calibration points as described in§10.3. Each calibration point measurement consists of eye tracker samples about the calibration point over an 800ms period (approximately 44 individual data points). Raw sample points falling in the exterior 10-pixel wide borders are ignored. This is due to the eye tracker's property of generating (0,0) values during blinks (confirmed by the vendor). For this reason, any time a raw sample point is close enough to the location (0,0) (within 10 pixels), it is removed from further consideration. An average of valid data points (centroid) is obtained and the error between the centroid and calibration point is calculated. Each two-dimensional euclidian distance measurement is converted to the the full visual angle, dependent on the viewing distance and calculated resolution of the television screen. Thus each calibration run contains 30 average deviations at each calibration points measured in terms of visual angle. A graphical example of this measurement is shown in Figure 69. The

(a) Calibration before stimulus viewing (avg. error: 0.96°).

(b) Calibration after stimulus viewing (avg. error: 1.10°).

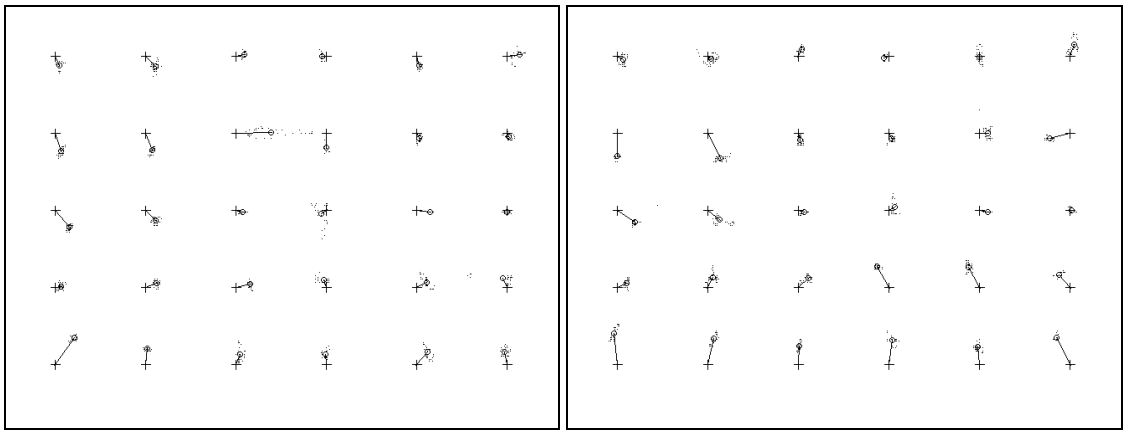


Fig. 69. Typical per-trial calibration data (subject # 29).

internal calibration locations are represented by +, sample measurements are represented by individual pixel dots, and centroid gaze positions are represented by circles, joined with the corresponding calibration point by a line. The length of the line is the average error deviation in pixels. This distance, r is converted to visual angle θ by the calculation

$$\theta = 2\tan^{-1}\frac{r}{2D},$$

where D is the viewing distance. The error distance r is measured in the same units as the viewing distance D, dependent on the resolution of the display.

For aggregate VOI creation, seven subjects were chosen whose average calibration error was subjectively low (about 1-2° visual angle). To quantify the overall eye tracker accuracy for these individuals succinctly, the average calibration error is obtained from each set of calibration points in order to calculate an overall average statistic of the eye tracker. The resulting average instrument error is an average statistic over these subjects' calibration runs performed in Experiment 2. The calculated mean value is 1.18°. This is not a particularly informative statistics since the data does not appear to fit a normal distribution. The histogram of average errors is shown in Figure 70. Since the average error data appears skewed, a more meaningful statistic is

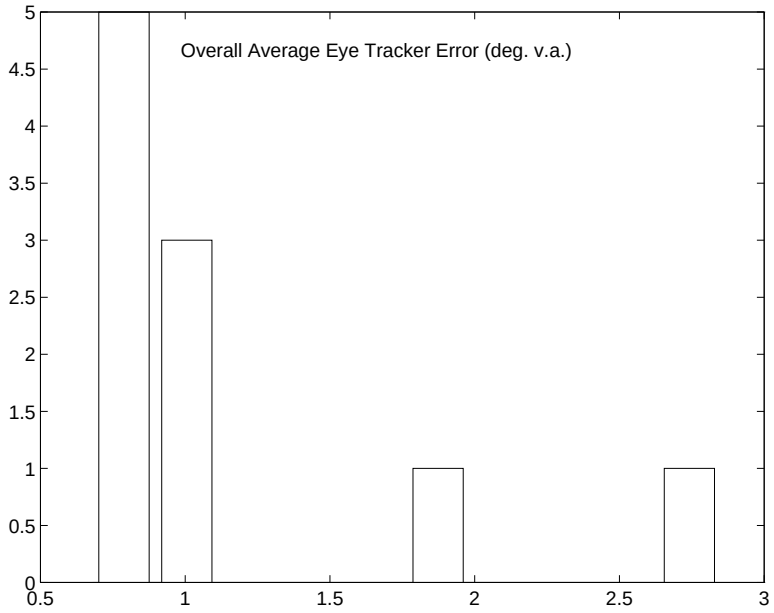


Fig. 70. Overall eye tracker error histogram.

the median value, which ignores the influence of outliers. Its value is 0.90°. Using similar reasoning for reporting a dispersion statistic, the interquartile range (iqr) is utilized instead of the standard deviation for its robust response to outliers. The iqr value is 0.30°. These findings indicate an overall acceptable performance, not far off from the vendor's claimed accuracy (roughly 1°visual angle). The average error for the individual VOI collection was 0.83° before stimulus viewing and 0.86°after. Eye movement patterns provided by this individual are considered slightly better than average. When asked whether the subject had any previous visual tracking training, the subject answered in the negative but noted hunting as an enjoyable hobby. The subject is referred to as "hunter".

12.5.2 Verification of Eye Tracker Slippage

Quantification of the before- and after-viewing eye tracker error provides a measure of instrument accuracy during the viewing task. A graphical example of this measure is shown in Figure 71 which is a composite plot of Figures 69 (a) and (b). Notice that measured eye positions in relation to calibration points coincide well

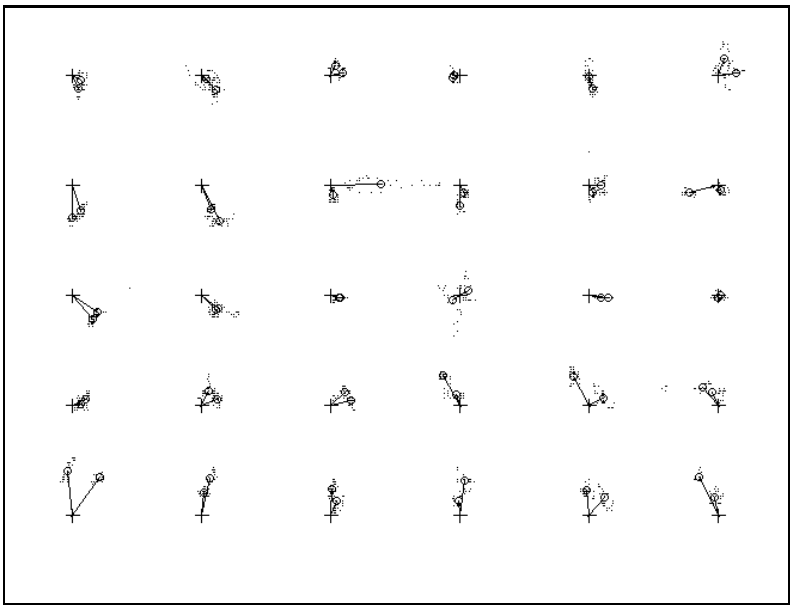


Fig. 71. Composite calibration data showing eye tracker slippage (subject # 29).

overall. To quantify this correspondence, a one-way ANOVA was performed on the means of the before- and after-viewing average error measures. Table 22 lists the ANOVA measures in §D. The error mean boxplot is shown in Figure 72. On average, no significant slippage is detected by this statistic. Note that ANOVA in this case is not very informative since it does not consider eye tracker slippage on a per-trial basis. That is, the ANOVA only reports significant correspondence of the mean measurements.

To examine eye tracker slippage on a per-trial basis, differences of average errors were calculated between the before- and after-viewing calibration runs on a pre-run basis. Difference measurements fit a skewed distribution, as shown in Figure 73. Statistical measures robust to outliers are used. The median error is -0.14°, and interquartile range is 0.38°. These values quantify the close correspondence of example pre- and post-viewing calibration measurements shown graphically in Figure 71. For subject “hunter”, the average error between the pre- and post-viewing calibration measurements is 0.03°. Overall, the tracker accuracy before and after viewing the stimulus remains fairly stable.

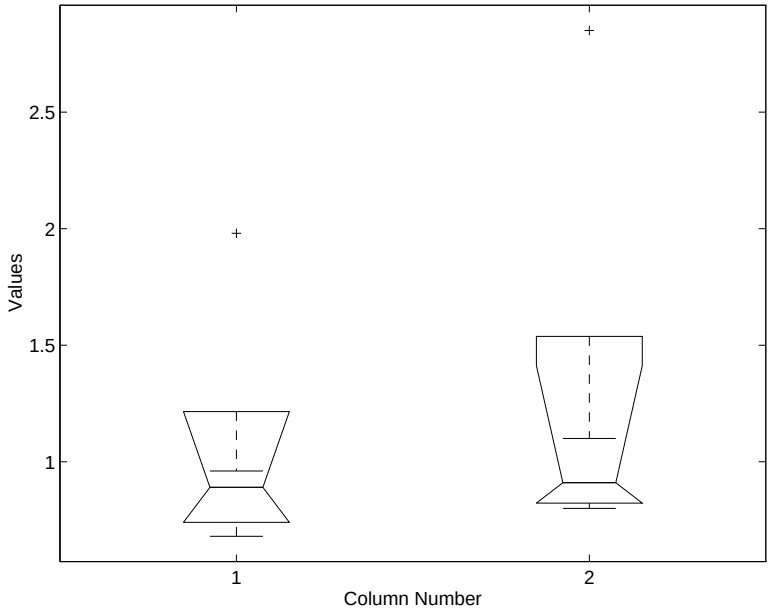


Fig. 72. Pre- vs. post-stimulus viewing average calibration error boxplots.

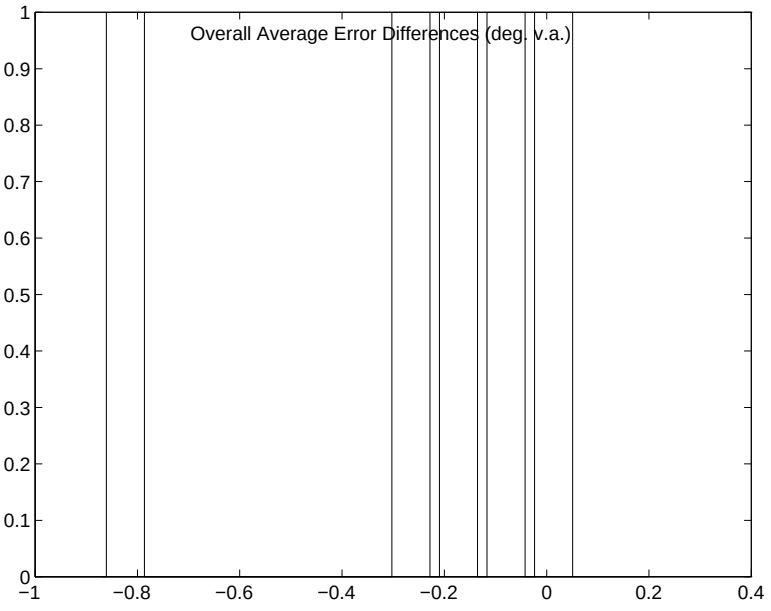


Fig. 73. Overall difference error histogram.

12.5.3 Verification of Gaze Position

To verify the individual's gaze position with respect to expected Regions Of Interest (ROIs) of an ideal observer, eye gaze position error was calculated on a per-frame basis over all participating subjects' data. Estimated Volumes Of Interest (VOIs) from the raw Point Of Regard (POR) data identified the subject's fixation locations. These locations were then compared to expected (ideal) VOI locations. Reference (ideal) VOIs were compared to the subject's observed VOIs by dismantling the VOIs into VOI-frame intersections (ROIs). The error measure was calculated as the distance between ideal and observed ROIs for each frame and converted to degrees visual angle. A median error value of the subject's VOI data represents a measure of the gaze error over the entire sequence. Outlier measured error values corresponding to suspected blinks were dropped from the median calculation. The overall median gaze position error for subject "hunter" over the *flight* sequence is 1.07° visual angle, with iqr 1.03°.

The per-frame error of "hunter"'s scan patterns with respect to the ideal observer is shown in Figure 74. Note

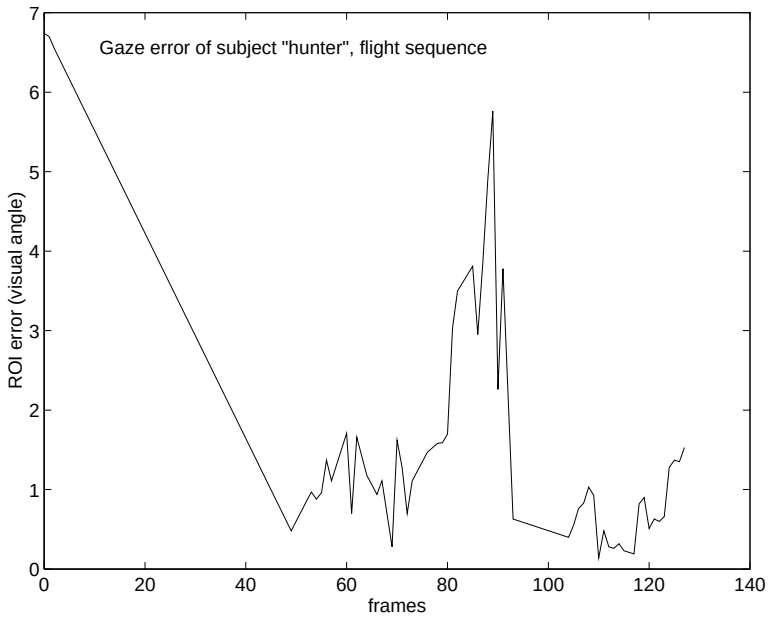


Fig. 74. Per-frame gaze errors.

missing frame data is linearly interpolated in the graph (e.g., over frames 3-49). Missing frame data corresponds to the lack of Regions Of Interest at saccadic locations in the VOI record (i.e., saccades identified by the PARIMA model). During onset of the video sequence, according to the PARIMA model, the subject performed six saccades over frames 7-9, 9-12, 20-22, 29-30, 31-32, and 35-39 (some of these saccades may be blinks). The subject was observed acquiring the object during these first few seconds (roughly 2.5s, 40

frames at 16fps). The large errors at frames 81-91 correspond to the subject's observed lag behind the target object and subsequent saccade performed to catch up to the object. The PARIMA model identified saccades over frames 82-85, 89-90, and 91-93. Figure 75 (a) shows the subject's VOIs just after acquisition of the target object. The target plane is seen in the texture-mapped frame between first two VOIs. Figure 75 (b) shows the subject's VOIs just prior to the saccadic activity over frame 82-93. Re-established fixations over

(a) View just after target acquisition. (b) View just prior to saccadic activity.

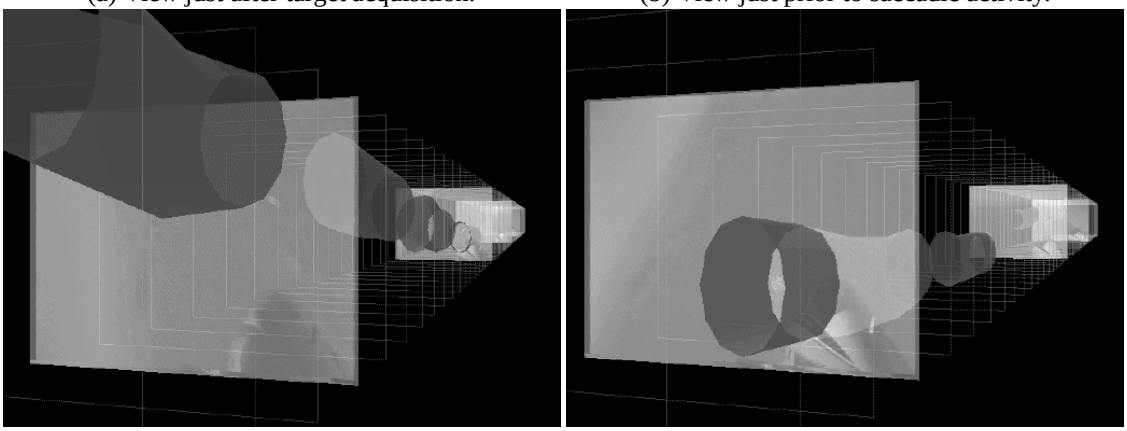


Fig. 75. Individual subject's ("hunter") VOIs over the *flight* sequence.

the target object can be seen behind the second texture-mapped image frame. The three-dimensional scan-path is not shown in Figure 75, but is visible in Figure 76. The view in Figure 76 is slightly off-centered to

(a) View over early portion of the sequence. (b) View over later portion of the sequence.

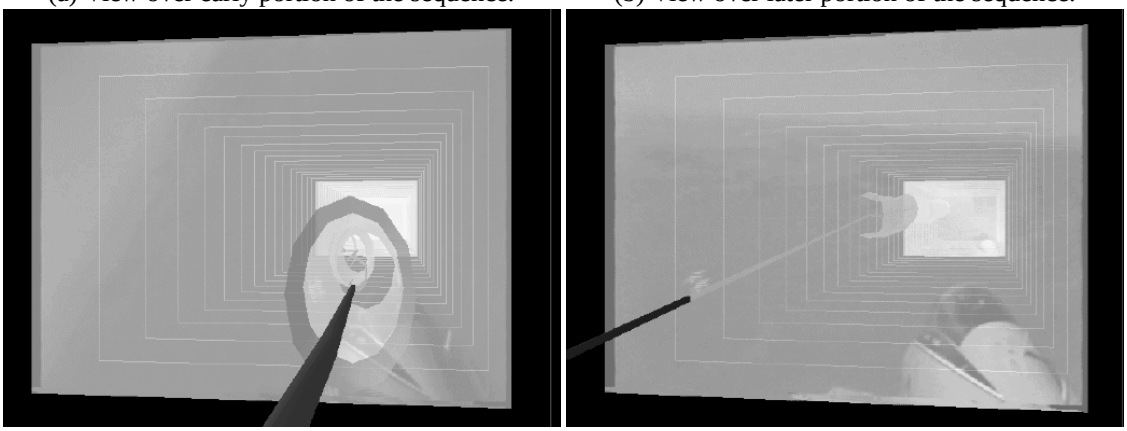


Fig. 76. Individual subject's ("hunter") VOIs and scanpath over the *flight* sequence.

expose the scanpath and frame intersections. In Figure 76 (a), the target plane is seen just to the left of the

scanpath-frame intersection. Figure 76 (b) shows the target plane slightly above the scanpath-frame intersection. The scanpath in the latter image, drawn by linearly-interpolating between VOI segments, approximates the saccades detected over frames 89-93.

12.6 Discussion

Aggregate VOIs over the seven qualifying subjects was presented and discussed in §VIII. During creation of the aggregate representation, an interesting observation was made regarding an individual's successive scan patterns over the *cnn* sequence. Recall that each of the seven subjects viewed the *cnn* sequence three times in succession. Subject 7 appears to have made saccades over several regions of the video frame in the first two trials, but maintained an almost steady fixation over the central region in the last trial. (Subject 7's pre- and post-viewing average eye tracker error was 0.89° and 0.83°, respectively.) Figures 77 shows the scanpath and VOIs of the first trial. Figure 77 (b) shows diverse eye movements from the central region of the screen, to

(a) 2D scanpath. (b) 3D VOIs.

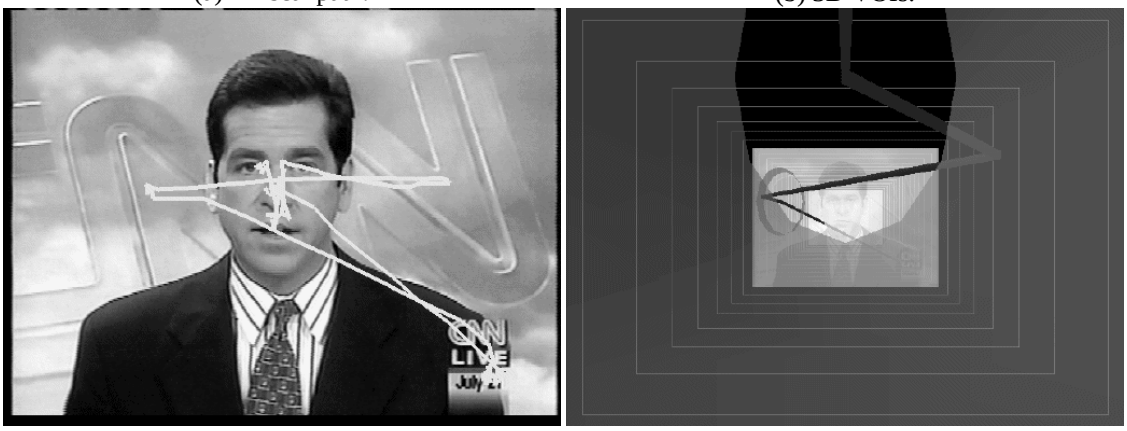


Fig. 77. Individual subject's (subject # 7) first scanpath and VOIs over the *cnn* sequence.

the right portion, and then to the left side of the screen. A very short fixation is made at the right side of the image. Figure 78 shows the scanpath and VOIs of the second trial. Figure 78 (b) indicates a saccade from the left portion of the screen to the region just above the timebox, followed by a saccade to timebox. Although not clearly seen in Figure 78 (b), interaction in the 3D visualization environment showed the scanpath terminating over the anchorman's tie. Figure 79 shows the scanpath and VOIs of the third trial. The subject's eye movements are restricted to the facial area of the video sequence. Figure 79 (b) shows the VOIs over the sequence. Note that the current view is within a VOI (the 3D scanpath line is removed for clarity).

Note the marked difference in eye movement variability across three viewing trials over the same video

(a) 2D scanpath. (b) 3D VOIs.

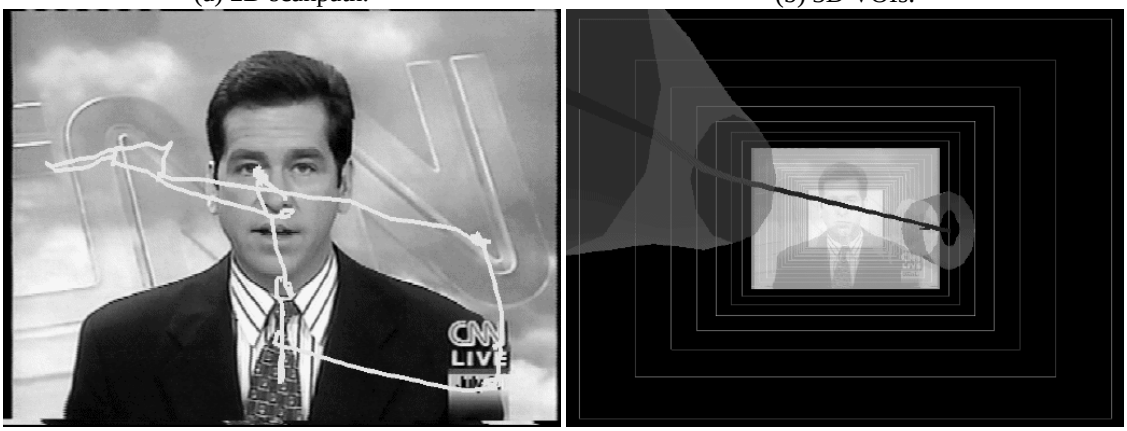


Fig. 78. Individual subject's (subject # 7) second scanpath and VOIs over the *cnn* sequence.

(a) 2D scanpath. (b) 3D VOIs.

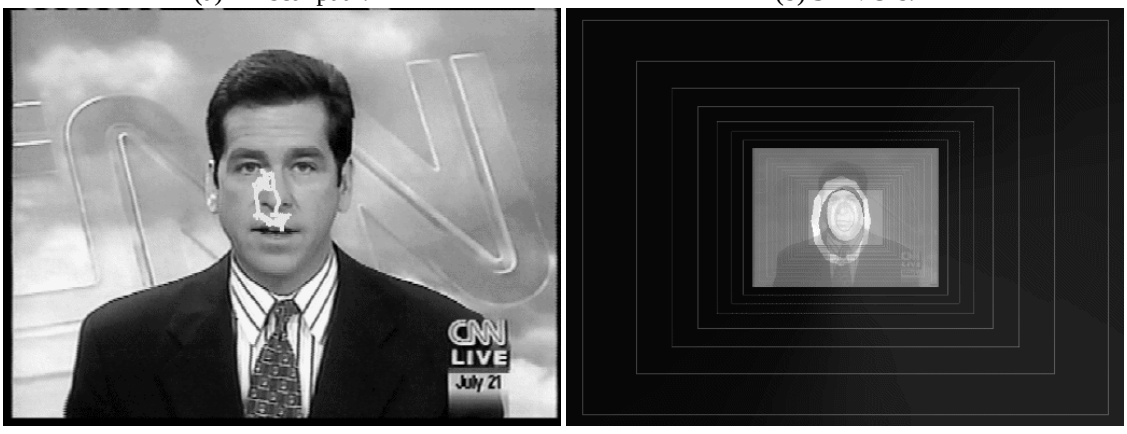


Fig. 79. Individual subject's (subject # 7) third scanpath and VOIs over the *cnn* sequence.

sequence. An interesting question arises regarding the reason for the diminished spatial variation in the subject's eye movement patterns. One explanation is that attention is drawn to motion events (in this instance the movement of the anchor's lips). Another reason is the viewer's formation of a mental map of the image contents. The contents of the *cnn* sequence did not change from trial to trial. To speculate, it may be that that having viewed the sequence twice previously, the subject decided to focus attention on one portion of the image, being already familiar with the peripheral content.

To further investigate perception of the periphery, Experiment 3 degrades peripheral regions of three different video sequences. The aggregate VOIs obtained over the *cnn* sequence in this experiment are used as spatiotemporal Regions Of Interest in an effort to predict locations of viewers' foci of attention. VOI visualization of subject "hunter"'s scan patterns (Figures 75 and 76) suggests fairly good correspondence of this subject's eye movements to the ideal observer. The analysis of gaze position error over corresponding VOIs substantiates this observation. Missing VOI segments (VOI "holes") represent the PARIMA eye movement model's prediction of the subject's saccades. The adequacy of this representation for gaze-contingent video representation is tested in Experiment 3.

CHAPTER XIII

EXPERIMENT 3: GAZE-CONTINGENT VISUAL REPRESENTATION

The goal of this experiment is to test whether peripheral regions of the image can be degraded imperceptibly. This objective is significantly distinct from testing sensory-guided human performance. Human perceptual characteristics must be distinguished from those of performance. An example of this distinction, pointed out by Shebilske, arises in the context of reading performance [She93]. Perceptually noticeable degradation may be introduced without impeding performance (e.g., comprehension of text). In the present case, the focus of the experiment is the imperceptible impairment of the stimulus (video).

The specific aim of this experiment is to test the variable resolution mapping provided by the wavelet based gaze-contingent video representation. Two resolution mapping functions are tested against an unprocessed sequence. The resolution mappings are denoted by LIN (for linear), HVS (for human visual system), and ORG (for original, or unprocessed). Both LIN and HVS mappings introduce peripheral degradation of digital imagery. HVS degradation follows a resolution function derived from empirical quantification of human visual acuity. LIN is a lower bound function (in terms of relative resolution) generating more rapid degradation of the periphery. The multiple-ROI degradation processing strategy is described in §IX.

13.1 Video Sequences

Three 8-second, 16fps video sequences are used as stimulus. Each sequence is degraded by the LIN, HVS, and ORG mappings with respect to designated intra-frame ROIs. Three different intra-frame ROI localization strategies are used in each of four experimental sessions. Frames from the unprocessed sequences are shown in Figure 80. Not shown in Figure 80 are the first and last frames of the *flight* sequence. In this sequence, a Navy fighter chase jet, in the lower right portion of the image, is following a smaller target plane which starts at the top right portion of the first frame. The target plane performs a half-roll evasion maneuver arcing from the top-right to the top-left screen regions through the bottom of the screen. The horizon rotates as the chase plane pursues. The chase plane remains steady relative to the frame since it bears the camera. Lighting on the chase plane changes as it emerges from shadow to sunlight. Sequence *flight* is used in two sessions, subject to a different ROI placement strategy in each session.

The *brain2* sequence simulates a virtual environment, *Exploring the Brain Forest*, being developed at the Scientific Visualization Laboratory, Department of Computer Science, Texas A&M University [MBD96]. *Exploring the Brain Forest* presents hierarchical views of the brain at several levels of scale from a global

(a) *brain2* (frame 001). (b) *cnn* (frame 001).

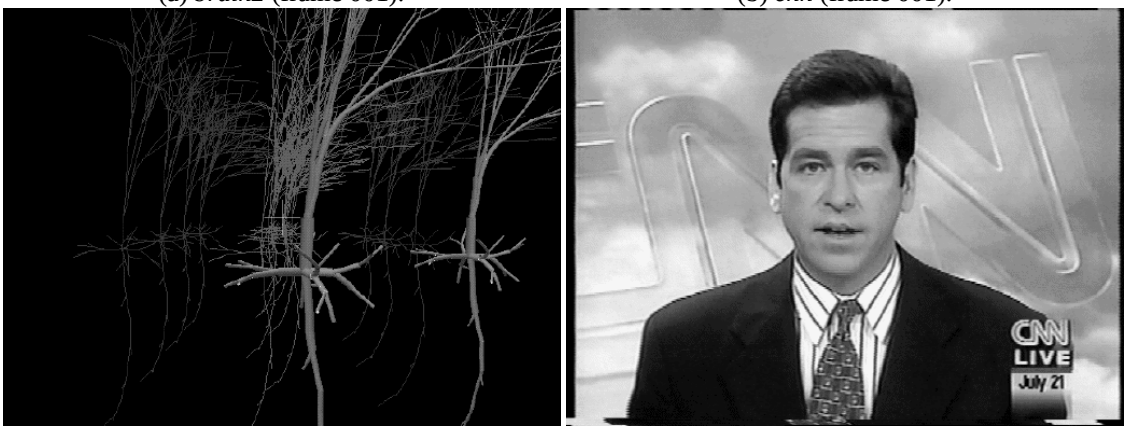


Fig. 80. Unprocessed video sequences.

overview to immersion within its forest of neurons and glial cells. The *brain2* video sequence simulates a "fly-through" of such a neural forest. To simulate the proposed attentive graphical display of this complex virtual environment, two attentive techniques are used. First, a scale-dependent strategy is used to represent model neurons at two levels of geometric detail. Neurons are rendered as cylindrical models near the expected central (foveal) regions and as wireframe models at greater distances. Second, the ROI-based wavelet reconstruction described herein forms the gaze-contingent strategy. Video frames are processed at the LIN, HVS and ORG degradation levels contingent upon expected locations of intra-frame ROIs.

The *cnn* sequence is chosen for its content of a human face. The sequence contains relatively little motion, but is representative of potential imagery found in telephony. The sequence is also well suited for visual attention and eye movement studies since it contains several distinguishable features which may serve as attentional attractors.

13.1.1 VOI Strategies

Three Volume Of Interest strategies are used to vary the ROI localization within video sequences.

1. The *ideal* session utilizes a VOI based on an "ideal observer". The location of the visual target (ROI) was manually measured for each video frame of the sequence. The resulting VOI is assembled from the list of ROIs.
2. The *preat* (for preattentive) session utilizes a VOI obtained from one subject's trial run in Experiment 2. The subject provided eye movement patterns considered slightly better than average. When asked whether the subject had any previous visual tracking training, the subject answered in the negative but noted hunting as an enjoyable hobby. The subject is referred to as "hunter" and provides model eye movement patterns for the preattentive session.

This strategy is preattentive since placement of the ROI anticipates the viewer's future point of regard. To clarify, a human subject's eye movement patterns typically include "breaks" of "holes" in the resulting VOI due to saccades, or blinks. In contrast, the VOI of an "ideal observer" is a continuous dynamic fixation. Intra-frame ROIs are defined as intersections of VOIs and video frames. In the case of discontinuous VOIs derived from human subjects, certain frames occur between continuous VOI segments resulting in a null VOI-frame intersection. The lack of an intra-frame ROI results in overall degradation of the frame; inclusion of such a frame in the video sequence results in a sudden, brief loss of resolution. Since human vision is practically blind during saccades and blinks, this resolution loss theoretically does not pose a problem provided it corresponds to a blink or saccade. The prediction of a saccade of blink, however, is not currently possible. For this reason, since it is believed that a sudden loss of resolution would impede perception, VOI "holes" need to be filled in by extending past or future VOI segments in the temporal VOI stream.

Extension of future VOI segments to the past results in an anticipatory video stream, in terms of foveal ROI placement. Consider frame f which happens to occur within a VOI "hole", as depicted in Figure 81. Assume also that the last VOI (VOI_{i-1}) terminates at frame $f-k$, and that the next

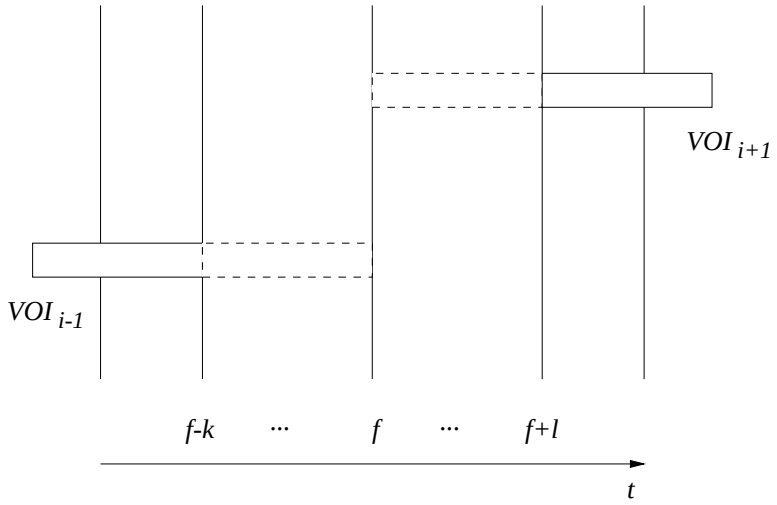


Fig. 81. Schematic VOI extension.

VOI (VOI_{i+1}) commences at frame $f+l$. Since by the PARIMA model of eye movement the frames $f \in [f-k, f+l]$ constitute a discontinuity (e.g., due to a saccade), the spatial locations of VOI_{i-1} and VOI_{i+1} need not correspond. To ensure that perception is not disrupted, an artificial Region Of Interest must be inserted into frame f. Determination of the location of the ROI at frame f is the source of the

current problem. To re-emphasize, ideally the frame should not contain an ROI and should be subject to maximal degradation only if it can be reasonably expected that a saccade or blink will occur between frames $f-k$ and $f+l$. This is not a practical expectation, however, due to the unpredictable nature of saccades and blinks (blinks more so than saccades).

A pragmatic solution to the VOI "hole" problem is the extension of either VOI_{i-1} or VOI_{i+1} over frame f. Extension of VOI_{i-1} into the future is a reactive solution since it assumes that the fixation will persist until frame f. Extension of VOI_{i+1} into the past is an anticipatory strategy since a region of high resolution is introduced into the video stream prior to the expected arrival of a fixation. In essence, this strategy is based on the assumption of attention preceding a fixation change and is adopted in the *preat* VOI strategy.

Note that both the *ideal* and *preat* strategies are limited to the choice of viewing paradigm where it can be reasonably expected that viewers will follow a similar viewing pattern. That is, the viewing paradigm must be based on a suitable viewing task, e.g., visual tracking.

3. The *agg* (for aggregate) strategy utilizes an aggregate VOI constructed from eye movement patterns of several viewers in Experiment 2 (see §XII). The rationale for this strategy is the evaluation of the aggregate VOI construction as a gaze prediction method. The expectation here is that multiple viewers will identify all the potential spatio-temporal segments of the video stream that need to be presented at full resolution for complete perception of the sequence. Following this argument, the hypothesis is that perception will not be impeded in only if these segments are displayed at high resolution. Unlike *ideal* and *preat*, however, the *agg* strategy does not explicitly rely on a common visual scanpath.

13.1.2 Experiment Sessions

Four experiment sessions were conducted based on the following combinations of video sequence and VOI strategy: (1) *flight* (ideal), (2) *flight* (preat), (3) *brain2* (ideal), and (4) *cnn* (agg). Ideal observer VOIs were created for both the *flight* and *brain2* sequences. For the *flight* sequence, the VOI follows the target plane, as illustrated in Figure 82. The *flight* sequence is also used in combination with the preattentive VOI strategy. Both degradation strategies use the *flight* sequence for its easily identifiable target suitable for the visual tracking paradigm. Figure 83 shows the VOI of subject "hunter" over the *flight* sequence. Missing VOI sections are "holes" corresponding to discontinuities identified by the PARIMA model.

In the *brain2* sequence, the ideal observer's VOI follows an artificially overlayed crosshair in the center of the video frame, as shown in Figure 84. The crosshair serves as a visual cue easily located by subjects.

(a) Left perspective. (b) Right perspective with zoom.

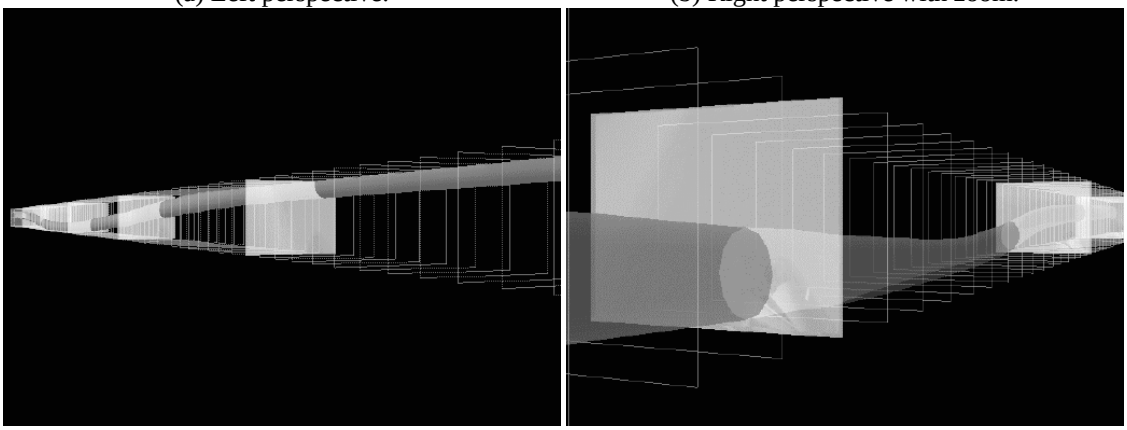


Fig. 82. Expected VOIs: *flight* (ideal observer).

(a) Left perspective. (b) Right perspective with zoom.

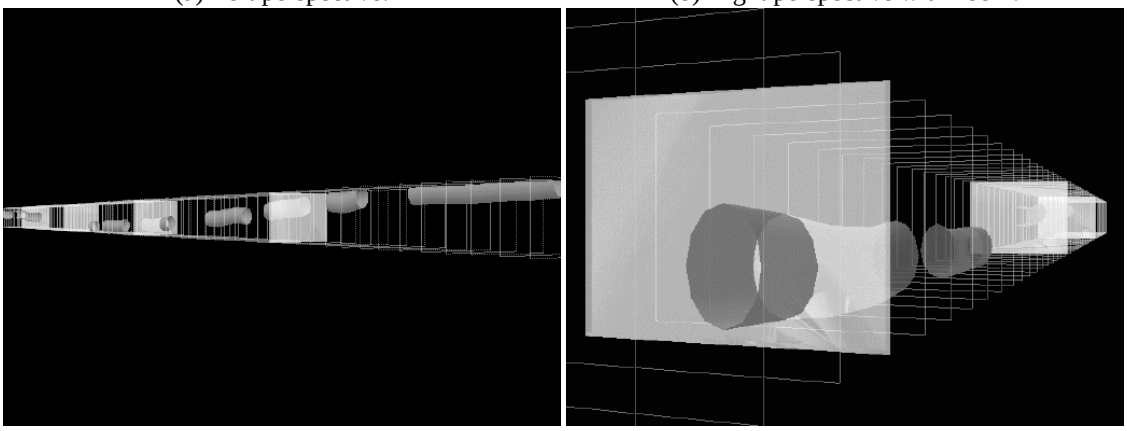


Fig. 83. Expected VOIs: *flight* (preattentive).

(a) Left perspective. (b) Left perspective with zoom.

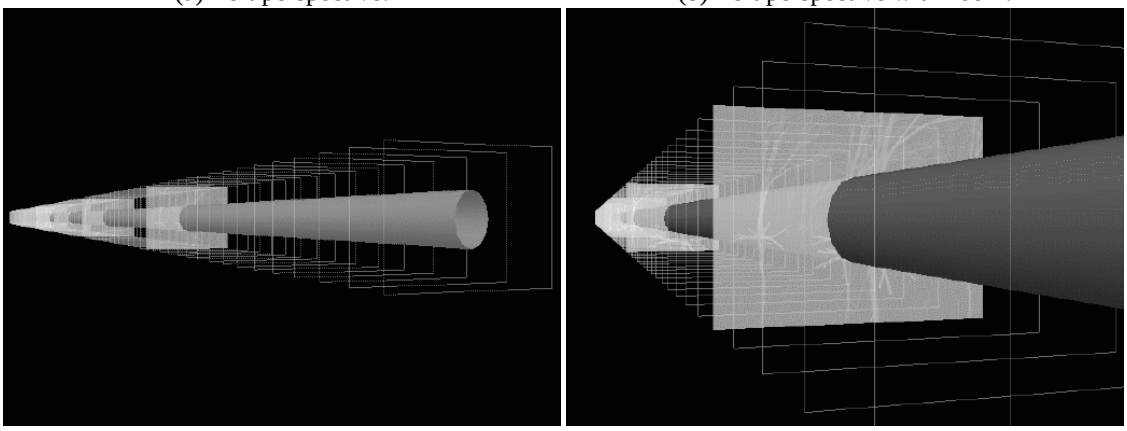


Fig. 84. Expected VOIs: *brain2* (ideal observer).

Figure 85 shows the aggregate VOI formed over the *cnn* sequence by subjects in Experiment 2. Although

(a) Left perspective. (b) Off-center perspective with zoom.

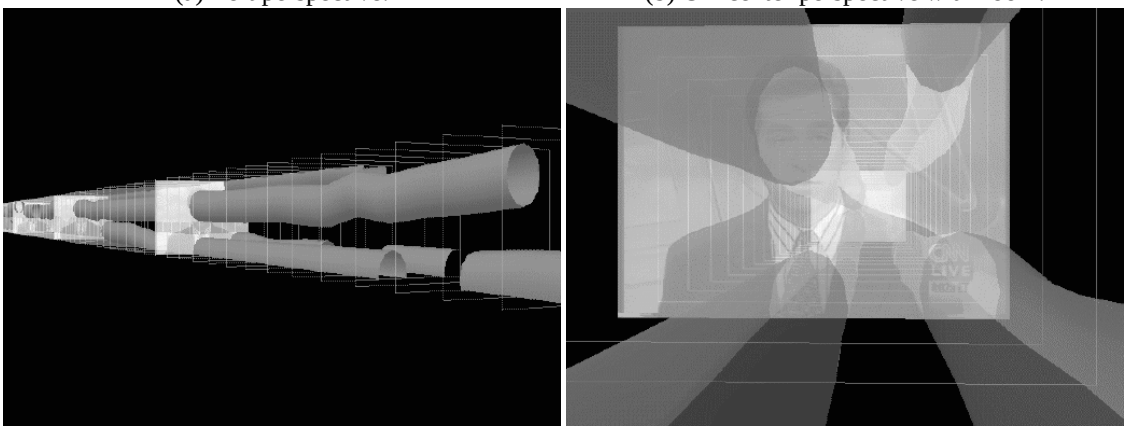


Fig. 85. Expected VOIs: *cnn* (aggregate).

expected interesting spatial regions correspond to VOI location, the aggregate model does not extend VOIs as is done in the *preat* model. Hence, high resolution regions are "turned off" when no VOI is present.

13.2 Experimental Trials

Each experimental session consists of multiple subjects tested individually in experimental trials. Each trial consists of three presentations of one video sequence processed in the LIN, HVS, and ORG manner. Trials are limited to three video presentations since loading of the video into memory requires 7 minutes and the experiment is designed to process each individual in roughly 30 minutes. Each trial consists of the following steps:

1. Brief introduction. When a subject enters, s/he is asked to sit in front of the eye tracker head rest. The equipment is briefly described with emphasis on the eye tracker infra-red (IR) light source and camera. It is pointed out that the IR assembly contains a standard overhead projector bulb as the light source. This is done to alleviate any preconceived fears regarding the apparatus (some subjects were under the impression that eyelid movement would be restricted). Subjects are assured that the experiment is physically unobtrusive.
2. Training. Prior to the display of LIN, HVS, ORG sequences, the subject is shown the ORG sequence from video tape. The sequence is shown three times with an explanation of digital artifacts (blockiness, blurry regions, jagged edges and the like). It is mentioned that the ORG sequence contains no artifacts and that subsequent sequences (to be shown in no specific order) should be judged relative to the ORG

sequence.

3. Video presentation. Each video presentation consists of the following substeps:
   (a) External calibration. Once the subject has settled into the head rest, the eye tracker is calibrated, as discussed in §10.3.
   (b) Internal calibration. Immediately following external calibration, the internal calibration procedure is performed to record the initial accuracy of the eye tracker. The calibration results are stored in a text file for later analysis (the file name convention adopted is (`<seq_name>.<lin|hvs|org>.1.clb`).
   (c) Stimulus display. The video is presented twice in succession; each time eye movements are recorded and stored (the file name conventions adopted are (`<seq_name>.<lin|hvs|org>.1.por` and `<seq_name>.<lin|hvs|org>.2.por`).
   (d) Internal calibration. Immediately following stimulus presentation, the internal calibration procedure is performed once again to record the final accuracy of the eye tracker. The calibration results are stored in a text file for later analysis (the file name convention adopted is (`<seq_name>.<lin|hvs|org>.2.clb`).
   (e) Stimulus judgment. Following the second internal calibration step the subject is told to sit back and relax and to mark the perceived level of impairment on a 5-point impairment scale shown in Table 13. The impairment table follows the CCIR Recommendation 500 rating scale provided in [ST94]. Values of 1 and 5 are assigned to the IMPERCEPTIBLE and VERY ANNOYING judgment, respectively.

13.3 Subjects

A total of 16 subjects (7 female, 9 male) participated in Experiment 3. The age distribution was mean 19.4, minimum 18, and maximum 22. The subjects were recruited from the Department of Psychology undergraduate pool. Subject majors ranged from Agricultural Biology (AGBL) to Psychology (PSYC). The undergraduate level distribution was 8 freshmen, 3 sophomore, 4 juniors and 1 senior. All subjects had good vision with 3 subjects wearing glasses, 2 wearing contact lenses. There were no subjects from the Departments of Computer Science or Electrical Engineering, and all were considered naive in terms of their ability to judge effects of image processing methods.

Subjects were randomly divided into equal groups of 4 for each of 4 experimental sessions. Each session pertained to a particular video sequence (*flight* (ideal), *flight* (preat), *brain2* (ideal), *cnn* (agg)) following a randomized block experimental design.

TABLE 13
5-point impairment scale.

1. **Sequence 1**

   - ☐ IMPERCEPTIBLE
   - ☐ PERCEPTIBLE, BUT NOT ANNOYING
   - ☐ SLIGHTLY ANNOYING
   - ☐ ANNOYING
   - ☐ VERY ANNOYING

2. **Sequence 2**

   - ☐ IMPERCEPTIBLE
   - ☐ PERCEPTIBLE, BUT NOT ANNOYING
   - ☐ SLIGHTLY ANNOYING
   - ☐ ANNOYING
   - ☐ VERY ANNOYING

3. **Sequence 3**

   - ☐ IMPERCEPTIBLE
   - ☐ PERCEPTIBLE, BUT NOT ANNOYING
   - ☐ SLIGHTLY ANNOYING
   - ☐ ANNOYING
   - ☐ VERY ANNOYING

13.4 Experimental Design

Processed video sequences (LIN, HVS, and ORG) were presented in random order (see Table 35 in § E). Note that in Table 35 some subjects from Experiment 4 were included. Experiment 4 was a supplemental experiment run as an extension of Experiment 3.

The stimulus video was shown twice to each subject after viewing the training video. This constitutes a modification of the CCIR Recommendation 500 double stimulus impairment method (sequence ABAB where A is the unprocessed sequence). The resulting sequence order used is ABB.

13.5 Results

The objective of Experiment 3 is to evaluate whether there is any perceptible difference in the different resolution mappings under tracking and free viewing conditions. A specific hypothesis being tested is whether the spatial resolution degradation of the HVS mapping is imperceptible under either or both conditions. The resolution mapping hypothesis is examined through the ANalysis Of VAriance (ANOVA) of subjective quality testing. Subjects rated the differently processed video sequences under different viewing conditions. An unprocessed (ORG) sequence was used as the control stimulus factor. If there is no statistically significant difference between mean ratings of the HVS and ORG sequence, then it is reasoned that there is no perceptible effect of the HVS mapping. Results of this analysis are reported here.

Usefulness of subjective ratings of the video sequences are contingent on three factors of the eye tracking experiment: (1) accuracy of the eye tracker, (2) accuracy of the eye tracker during the viewing task, and (3) accuracy of gaze position over the intended Regions Of Interest. The impairment analysis implicitly assumes that subjects aimed their gaze at the intended high-resolution ROI targets. Prior to the impairment analysis, this assumption is tested in the following two sections which evaluate eye tracker and gaze position accuracy of the experiment.

13.5.1 Verification of Eye Tracker Accuracy

The measurement of gaze depends on the accuracy of the eye tracking instrument. Eye tracker accuracy was measured by internal calibration procedures described in §10.3. Measurements were taken before and after stimulus viewing trials, as discussed in §13.2. These procedures provide the basis for two statistical measures: (1) the overall accuracy of the eye tracker, and (2) the amount of instrument slippage during stimulus viewing. The latter measurement gives an indication of the instrument accuracy during the viewing task, i.e., by recording loss of accuracy between the before- and after-viewing calibration procedures.

Eye tracker readings were obtained over 30 internal calibration points as described in §10.3. Each calibration point measurement consists of eye tracker samples about the calibration point over an 800ms period (approximately 44 individual data points). Raw sample points falling in the exterior 10-pixel wide borders are ignored. This is due to the eye tracker's property of generating (0,0) values during blinks (confirmed by the vendor). For this reason, any time a raw sample point is close enough to the location (0,0) (within 10 pixels), it is removed from further consideration. An average of valid data points (centroid) is obtained and the error between the centroid and calibration point is calculated. Each two-dimensional euclidian distance measurement is converted to the the full visual angle dependent on the viewing distance and calculated resolution of the television screen. Thus each calibration run contains 30 average deviations at each calibration point measured in terms of visual angle. A graphical example of this measurement is shown in Figure 86. The internal calibration locations are represented by $+$, sample measurements are represented by individual

(a) Calibration before stimulus viewing (avg. error: 0.86°).

(b) Calibration after stimulus viewing (avg. error: 1.46°).

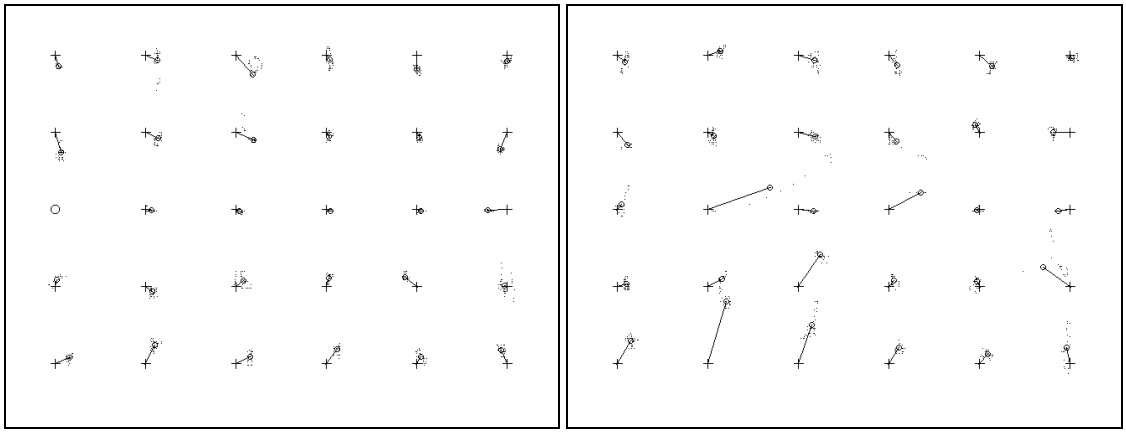


Fig. 86. Typical per-trial calibration data (subject # 11).

pixel dots, and centroid gaze positions are represented by circles, joined with the corresponding calibration point by a line. The length of the line is the average error deviation in pixels. This distance, r is converted to visual angle θ by the calculation

$$\theta = 2\tan^{-1}\frac{r}{2D},$$

where D is the viewing distance. The error distance r is measured in the same units as the viewing distance D, dependent on the resolution of the display.

To quantify the overall eye tracker accuracy succinctly, the average calibration error is obtained from each

set of calibration points in order to calculate an overall average statistic of the eye tracker. The resulting instrument error measure is an average statistic over all calibration runs performed in Experiment 3. The calculated mean value is 2.48°. This is not a particularly informative statistics since the data does not appear to fit a normal distribution. The histogram of average errors is shown in Figure 87. Since the average error

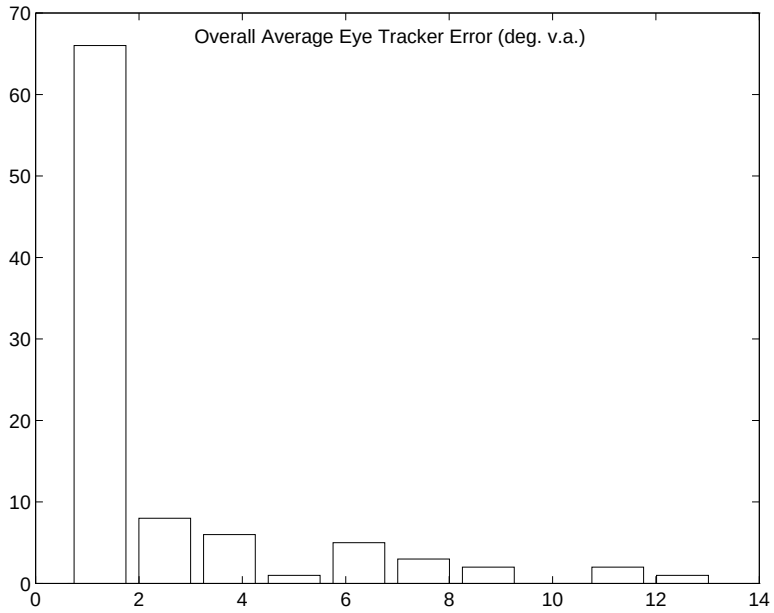


Fig. 87. Overall eye tracker error histogram.

data appears skewed, a more meaningful statistic is the median value, which ignores the influence of outliers. Its value is 1.42°. Using similar reasoning for reporting a dispersion statistic, the interquartile range (iqr) is utilized instead of the standard deviation for its robust response to outliers. The iqr value is 1.10°. These findings indicate an overall acceptable performance, not far off from the vendor's claimed accuracy (roughly 1° visual angle).

13.5.2 Verification of Eye Tracker Slippage

Quantification of the before- and after-viewing eye tracker error provides a measure of instrument accuracy during the viewing task. A graphical example of this measure is shown in Figure 88 which is a composite plot of Figures 86 (a) and (b). Notice that measured eye positions in relation to calibration points coincide well overall. To quantify this correspondence, a one-way ANOVA was performed on the means of the before- and after-viewing average error measures. Table 23 lists the ANOVA measures in §E. Error mean boxplots are shown in Figure 89. On average, no significant slippage is detected by this statistic. Note that ANOVA in this case is not very informative since it does not consider eye tracker slippage on a per-trial basis. That is,

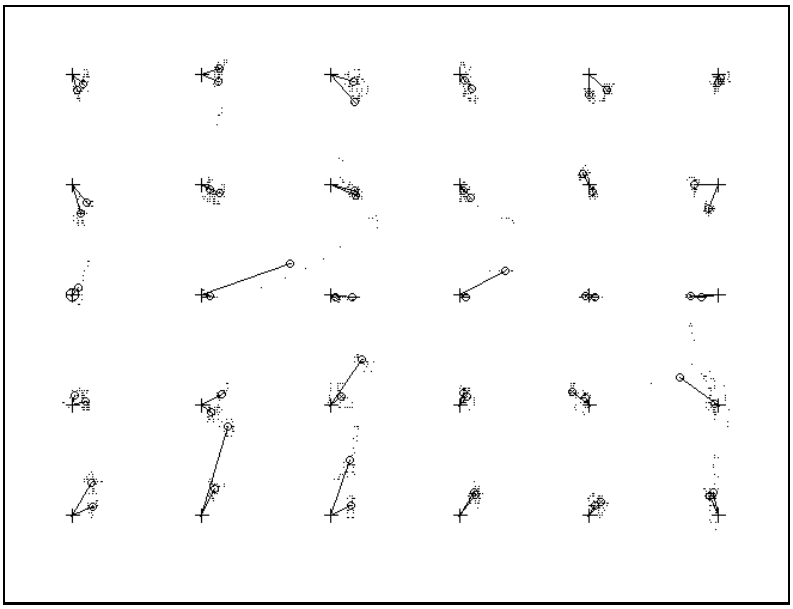


Fig. 88. Composite calibration data showing eye tracker slippage (subject # 11).

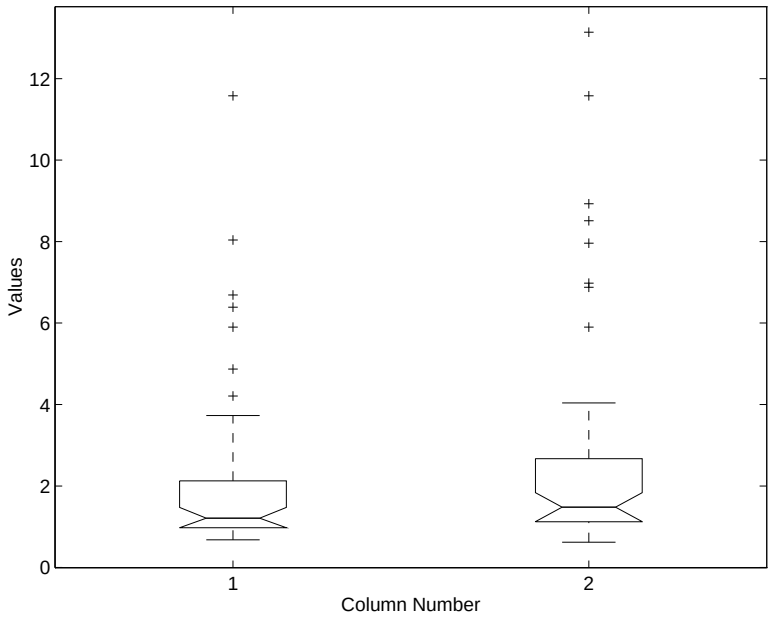


Fig. 89. Pre- vs. post-stimulus viewing average calibration error boxplots.

the ANOVA only reports significant correspondence of the mean measurements.

To examine eye tracker slippage on a per-trial basis, differences of average errors were calculated between the before- and after-viewing calibration runs on a pre-run basis. Difference measurements fit a skewed distribution, as shown in Figure 90. Due to the apparent skewed distribution, statistical measures robust to

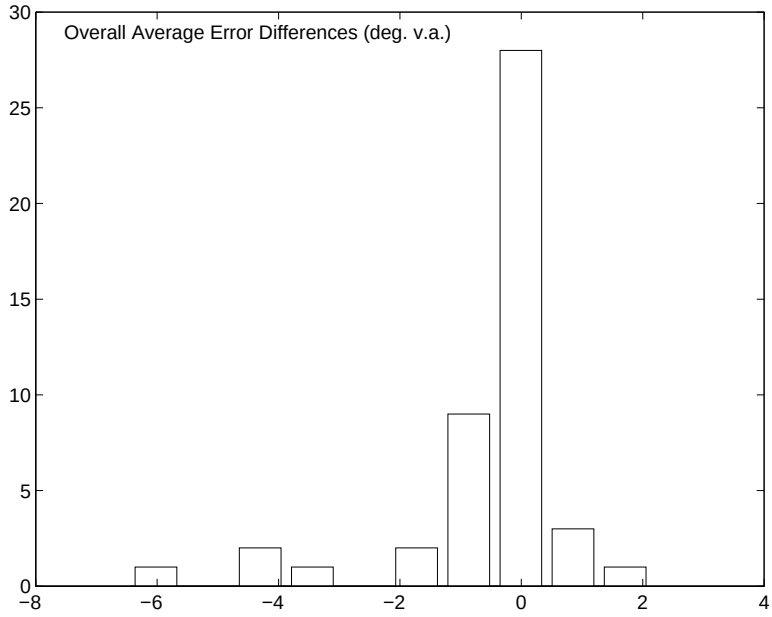


Fig. 90. Overall difference error histogram.

outliers are used. The median error is -0.18°, and interquartile range is 0.54°. These values quantify the close correspondence of example pre- and post-viewing calibration measurements shown graphically in Figure 88. Overall, the tracker accuracy varies roughly a fifth of a degree visual angle over the 8-second viewing task. Over some calibration points, accuracy improves while over others it degrades. Since the peak frequency is close to 0, generally the accuracy before and after viewing the stimulus remains stable overall.

13.5.3 Verification of Gaze Position

To verify subjects' gaze position with respect to expected Regions Of Interest (ROIs), eye gaze position error was calculated on a per-frame basis over all participating subjects' data. Estimated Volumes Of Interest (VOIs) from the raw Point Of Regard (POR) data identified subjects' fixation locations. These locations were then compared to expected (ideal) VOI locations. Reference (ideal) VOIs were compared to each subject's observed VOIs by dismantling the VOIs into VOI-frame intersections (ROIs). The error measure was calculated as the distance between ideal and observed ROIs for each frame and converted to degrees visual angle.

In the case of the aggregate VOI stimulus, subjects' intra-frame fixation locations are compared to the closest expected intra-frame ROI (based on a Euclidian measure). A median error value from each subject's VOI data is used as a representative measure of the gaze error over the entire sequence. Outlier measured error values corresponding to suspected blinks were dropped from the median calculation.

Overall gaze position error is shown over each sequence in Table 24. Not surprisingly, the *brain2* (ideal) sequence incurred the least amount of gaze positional error. This was expected since subjects were instructed to maintain steady gaze position at a central location of the display. Subjects did not appear to have significant difficulty in performing this task. Two-way Analysis of Variance of the medians indicates no significant difference in mean values (means of medians, $p = 0.7178$). However, the difference between experimental VOI paradigms is significant ($p < 0.0000$). Statistical data is given in Table 25. One-way ANOVA of the median of means between resolution mappings supports the similarity of degree error. Data for pairwise mean of median comparisons is given in Tables 26, 27, and 28.

The two-way ANOVA of means between viewing conditions suggests that the VOI presentation strategy (ideal, preat, agg) is a factor in observed gaze position error. To investigate further, one-way analysis of variance was performed between the mean of median gaze errors within each experimental VOI paradigm. In some cases data had to be truncated to perform the analysis resulting in slightly different mean values for individual conditions between comparisons.

1. Sequence *flight* (ideal) vs. *flight* (preat): Mean gaze error boxplots are shown in Figure 91. The difference between means of medians (5.0419° vs. 1.9571°, respectively) is somewhat significant ($p < 0.05$) suggesting a smaller gaze error during the *flight* (preat) sequence than during *flight* (ideal).
2. Sequence *flight* (ideal) vs. *brain2* (ideal): Mean gaze error boxplots are shown in Figure 92. The difference between means of medians (3.7050° vs. 0.6278°, respectively) is significant ($p < 0.01$), suggesting a smaller gaze error during the *brain2* (ideal) sequence than during *flight* (ideal).
3. Sequence *flight* (ideal) vs. *cnn* (agg): Mean gaze error boxplots are shown in Figure 93. The difference between means of medians (5.0419° vs. 3.4210°, respectively) is not significant suggesting comparable gaze error between the two sequences. This is somewhat surprising since the *flight* (ideal) sequence contained only one VOI while the *cnn* (agg) sequence contained several. Multiple VOIs were expected to distract viewers generating a significantly large gaze position error.
4. Sequence *flight* (preat) vs. *brain2* (ideal): Mean gaze error boxplots are shown in Figure 94. The difference between means of medians (2.0006° vs. 0.6278°, respectively) is significant ($p < 0.01$), suggesting a smaller gaze error during the *brain2* (ideal) sequence than during *flight* (preat).
5. Sequence *flight* (preat) vs. *cnn* (agg): Mean gaze error boxplots are shown in Figure 95. The difference between means of medians (1.9571° vs. 3.4210°, respectively) is somewhat significant ($p < 0.05$),

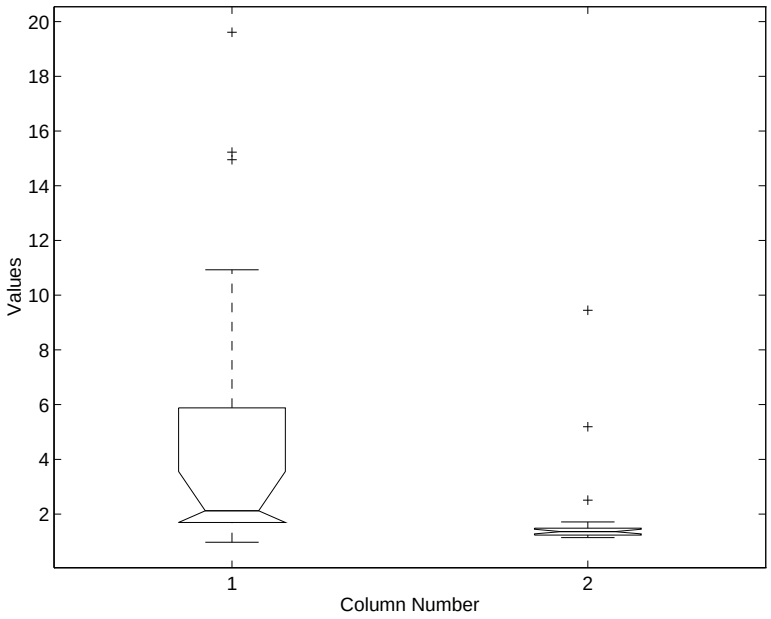


Fig. 91. Mean gaze errors, *flight* (ideal) vs. *flight* (preat).

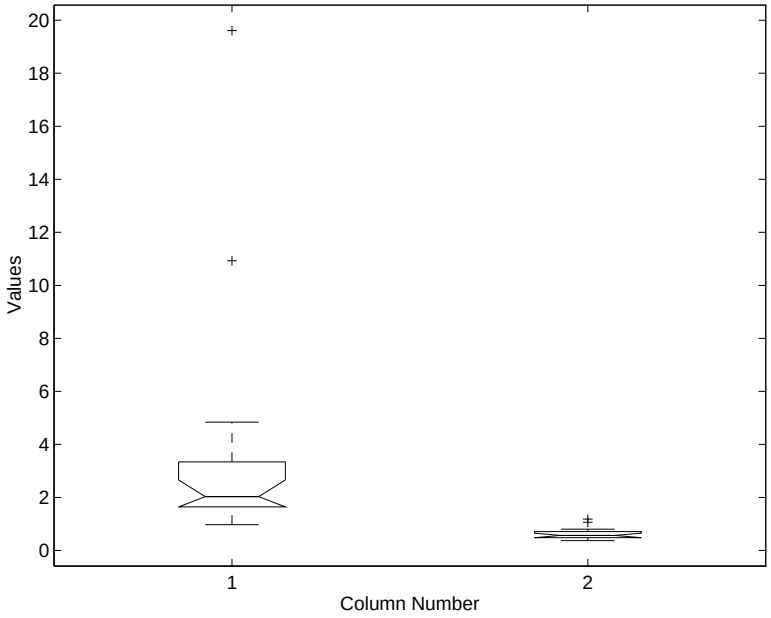


Fig. 92. Mean gaze errors, *flight* (ideal) vs. *brain2* (ideal).

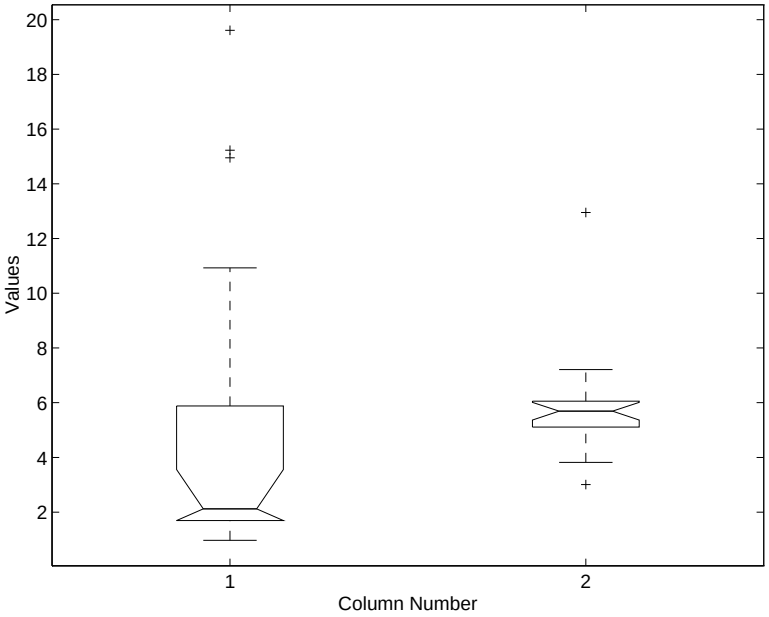


Fig. 93. Mean gaze errors, *flight* (ideal) vs. *cnn* (agg).

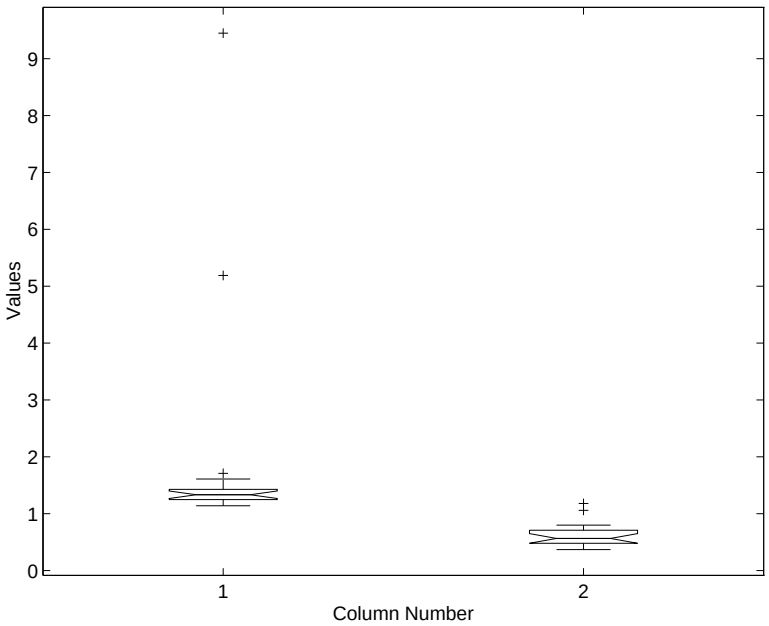


Fig. 94. Mean gaze errors, *flight* (preat) vs. *brain2* (ideal).

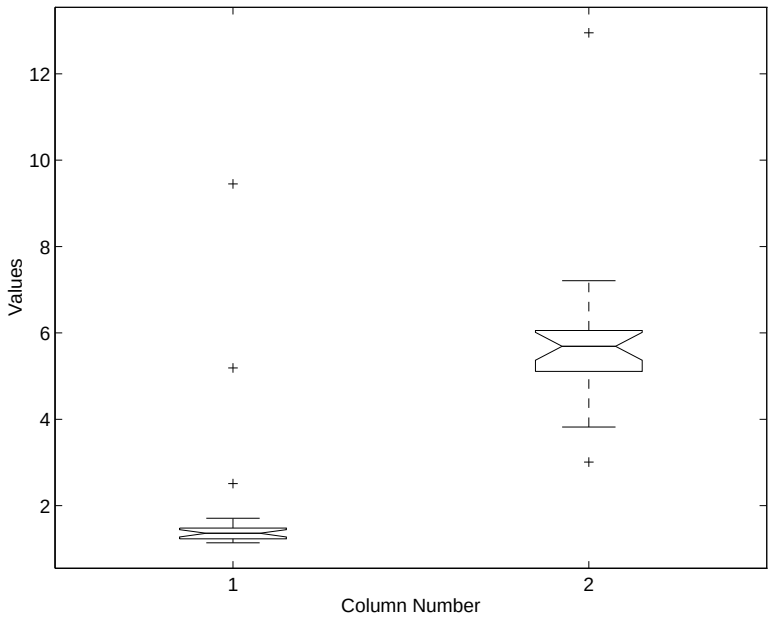


Fig. 95. Mean gaze errors, *flight* (preat) vs. *cnn* (agg).

suggesting a smaller gaze error during the *flight* (preat) sequence than during *cnn* (agg).

6. Sequence *brain2* (ideal) vs. *cnn* (agg): Mean gaze error boxplots are shown in Figure 96. The difference between means of medians (0.6278° vs. 3.5539°, respectively) is strongly significant ($p < 0.01$), suggesting a much smaller gaze error during the *brain2* (ideal) sequence than during *cnn* (agg).

The one-way ANOVA tables for the above comparisons are given in Tables 29, 30, 31, 32, 33, and 34.

Of all viewing conditions, the aggregate VOI strategy, utilizing multiple intra-frame regions of interest, appears to present the most difficult visual tracking task. On the one hand, qualitative observations of this task suggest that multiple regions, when easy to differentiate from background imagery, tend to disrupt normal viewing patterns. Subjects exhibited some confusion as to which ROI to fixate, and in general, found these sequences annoying. Subjective quality ratings support this observation and are presented in the next section. On the other hand, quantitative gaze error measurements suggests that, on average, viewers were capable of fixating a particular intra-frame ROI with surprising accuracy (overall median gaze error 2.98° visual angle). Interestingly, this median of medians error measurement is not significantly different from the ideal VOI condition, and is only somewhat worse (in term of statistical significance) from the preattentive condition.

Overall, the analysis of gaze error suggests two important points: (1) viewers have little difficulty in matching gaze to expected locations within fixation (*brain2* (ideal) sequence) and tracking (both *flight* sequences) paradigms, and (2) the degree of error indicates close proximity to the intended region of interest. On average, recorded gaze positions did not vary more than roughly the foveal dimension as projected on the stimulus

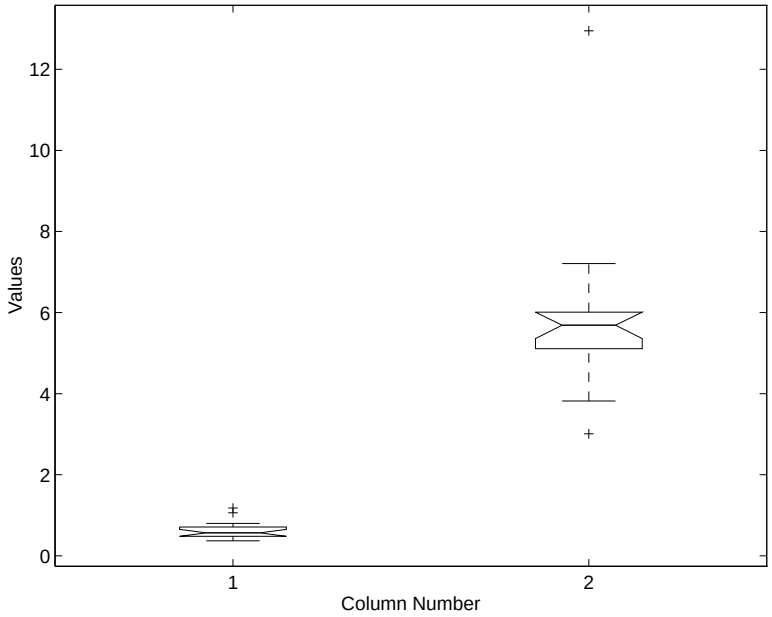


Fig. 96. Mean gaze errors, *brain2* (ideal) vs. *cnn* (agg).

display. Interestingly, performance is somewhat better over the preattentive sequence than over the ideal observer sequence. This is somewhat surprising since viewers were expected to foveally match the ideal target. Qualitatively, gaze of the subject ("hunter"), whose VOI data was used to generate the (preat) sequence, tends to slightly fall behind the moving target. It is interesting to note that in this experiment, "hunter"'s viewing patterns provide a better prediction of scan patterns than the ideal observer.

13.5.4 Impairment Perception Analysis Over All Conditions

Mean results for the 16 subjects were analyzed by two-way ANOVA and are shown in Table 36 in §E. Columns in the ANOVA table refer to the different mappings (LIN, HVS, ORG), rows refer to the effect of different sequences (*flight* (ideal), *flight* (preat), *brain2* (ideal), *cnn* (agg)). The important statistical consideration in this analysis is the test for variability. The analysis finds no evidence of variability across resolution mappings (columns), but there is strong indication ($p < 0.0000$) of variability across different viewing conditions (ideal, preat, agg). Furthermore, there appears to be evidence of interaction between row-column pairs ($p < 0.03$).

At first glance these findings suggest no perceived difference in mappings. However, due to the strong evidence of variability across viewing conditions, interactions between each condition must be further analyzed to gain insight into the results. The goal of the analysis is to ascertain whether there is any perceived difference between the LIN, HVS, and ORG mappings. Overall analysis in this case is meaningful only in the

sense that it indicates a statistical difference between sequence viewing conditions.

13.5.5 Impairment Perception Analysis Between Conditions

The overall analysis was broken down to identify statistically significant interactions between viewing conditions (pairs of rows) in the experiment. Two-way ANOVA statistics were examined to test for significant differences between viewing conditions. The ANOVA tables are shown for selected pairs of conditions in §E. Results are interpreted as follows:

1. *flight* (ideal) vs. *flight* (preat): no significant difference was found between resolution mappings, nor between viewing conditions. No significant interactions between factors were found. This finding suggests that there was no statistically significant difference in perceived quality among the resolution mappings or between the viewing conditions.

   The significant observation here is that there is no apparent difference between the ideal observer and preattentive method of resolution degradation. Differences between resolution mappings are examined within each viewing condition below.

   Since there is no statistically significant difference between the ideal and preat processing strategies between the *flight* viewing conditions, only the ideal condition is used in comparison between the remaining three conditions.
2. *flight* (ideal) vs. *brain2* (ideal): no significant difference was found between resolution mappings, nor between viewing conditions. No significant interactions between factors were found.
3. *flight* (ideal) vs. *cnn* (agg): a significant difference was found between resolution mappings ($p < 0.02$), and between viewing conditions ($p < 0.0000$). Strong interaction between the factors also appeared ($p < 0.01$).
4. *brain2* (ideal) vs. *cnn* (agg): a significant difference was found between resolution mappings ($p < 0.01$) and between viewing conditions ($p < 0.01$). The interaction between factors was notably less significant ($p < .1$).

The analysis between conditions suggests no significant difference between the preattentive and ideal viewing conditions, and no significant difference between sequence types over the ideal (and hence preattentive) viewing condition. By deductive reasoning, the viewing condition that appears to stand out is the aggregate condition. Analysis within conditions is carried out below in an attempt to narrow down the effect that appears to set the aggregate experiment apart from the other conditions. Individual analysis within conditions are also performed to draw conclusions regarding resolution mappings. From the above it appears there is no significant difference between mappings except in the case when the apparently confounding aggregate

factor is considered.

13.5.6 Impairment Perception Analysis Within Conditions

The primary interest in performing the within-condition analysis is to test for significance between impairment ratings of the three resolution mapping factors (LIN, HVS, ORG). Mean results for each viewing condition are shown in Figure 97. One-way ANOVA was performed over results from each viewing condition:

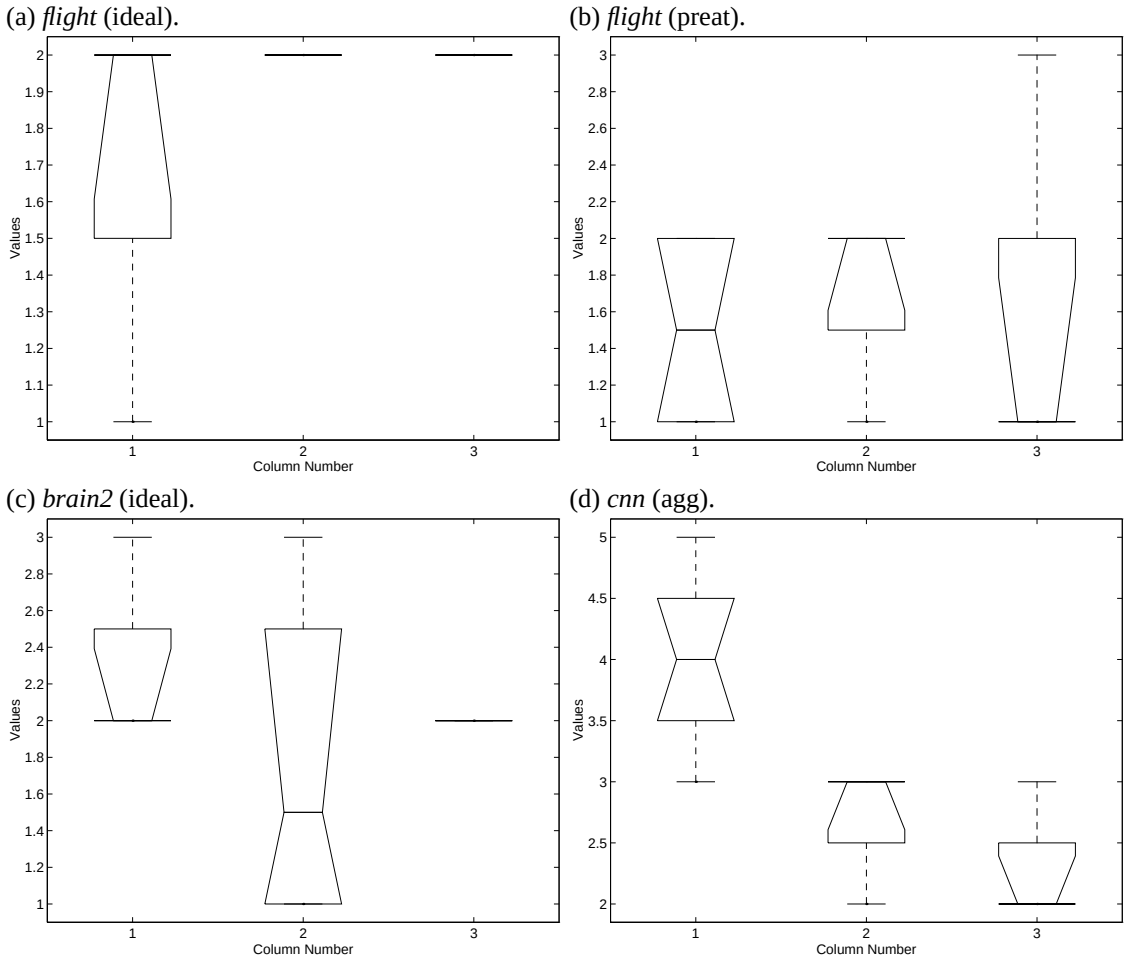


Fig. 97. Mean boxplots between resolution mapping ratings within viewing conditions (columns 1, 2, and 3 correspond to LIN, HVS, and ORG mappings, respectively.)

1. *flight* (ideal): no significant difference between the means was found preventing the rejection of the null hypothesis $\mathcal{H}_o$.
2. *flight* (preat): no significant difference between the means was found preventing the rejection of the null hypothesis $\mathcal{H}_o$.

3. *brain2* (ideal): no significant difference between the means was found preventing the rejection of the null hypothesis $\mathcal{H}_o$.
4. *cnn* (agg): a significant difference ($p < 0.01$) between the means was found strongly suggesting the rejection of the null hypothesis $\mathcal{H}_o$. This indicates a significant difference between resolution mapping under this viewing condition. The boxplot suggests that the LIN condition is significantly different from the HVS and ORG factors. To confirm this, each pair of columns (resolution mapping factors) was analyzed. One-way ANOVA tables for these pairs are given in §E and are summarized here:
   (a) LIN vs. HVS: statistically significant difference ($p < 0.05$);
   (b) LIN vs. ORG: statistically significant difference ($p < 0.05$);
   (c) HVS vs. ORG: difference is not statistically significant.

   It appears that the linear mapping generated perceptually significant effects from both the HVS and ORG mappings.

13.6 Discussion

The boxplots in Figures 97 (a)–(d) reflect the somewhat curiously low rating of the unprocessed (ORG) sequence. The expected rating of the ORG sequence was imperceptible (mean of 1). There are two possible reasons for this result: (1) the close viewing distance for all conditions, and (2) the naiveté of the subjects. The former criterion was imposed to simulate a large field of view ($> 30°$ visual angle), necessary for providing a large enough spatial region to test peripheral degradation. At the required viewing distance (60cm), however, imperfections of a standard NTSC television become evident. Aliasing of fine lines, for example, becomes quite visible. In one case, a subject commented on the "slightly annoying" quality of the calibration symbols ($\times$). The subject had to be instructed to refrain from judging the calibration symbols and concentrate on the stimulus video. Clearly, high-resolution displays are needed to prevent the confounding effects of low-resolution monitors with the effects of image processing techniques under investigation.

The latter explanation of the the low rating of the ORG sequence may be a result of the subjects' inexperience in the judgment of digitally processed imagery. Expected artifacts were described in vague terms such as fuzziness, jagginess, and blurriness in order to avoid technical jargon such as ringing effects, or block artifacts. Nevertheless, subjects may have had difficulty in recognizing impairments.

The important quality of the analysis is not so much the absolute impairment rating, but rather the relative score. The assumption used in the following interpretation of results is that a statistically insignificant difference between the unprocessed sequences and a processed one suggests an insignificant factor, i.e., imperceptible degradation.

Interpretation of the statistical analysis suggests the HVS resolution mapping produced imperceptible degradation of the video stream under ideal and preattentive visual tracking conditions. This result was expected since the degradation strategy was designed to match human visual acuity. Two subjects reported not to have seen any difference between the HVS and ORG sequences (one following the *brain2* (ideal) sequence, the other following *flight* (ideal) presentation). Surprisingly, the linear mapping (LIN) also generated imperceptible degradation results (under ideal and preattentive conditions). The linear mapping was expected to generate perceptible degradation.

Results between the ideal and preattentive conditions suggest that VOI-based prediction of eye movements is effective for visual tracking tasks. Before this contention was tested, an aggregate VOI was compiled from several subjects' eye movement patterns during a visual tracking task. Since the aggregate VOI qualitatively resembled the ideal observer's, the goal of this experiment was to test the minimal condition of using only one subject's patterns as a predictor for resolution degradation in the video stream (*flight*). Results suggest that one eye movement pattern is sufficient for this purpose.

VOI-based prediction of eye movements may not be effective for natural viewing tasks. This hypothesis was tested through the degradation of the *cnn* (agg) sequence. Results show that under natural viewing conditions (no imposed viewing task), aggregate VOIs impair perception (or perhaps more fairly natural eye movements). It is suspected that the problem resides in the discontinuous nature of the VOIs. In this experiment, VOIs were not extended as in the preattentive case. The resulting video sequence possessed a bubbling quality of high resolution regions. That is, in a low resolution peripheral region, a high resolution inset suddenly appears.

The bubbling effect of aggregate VOIs reflects the poor predictive power of VOIs in free viewing situations. The discontinuities within VOIs correspond to observed saccades made by viewers over the course of the video sequence. This is a historical record. Replaying a sequence processed through aggregate VOIs may depict locations where viewers looked in the past but there is no guarantee that viewers will repeat these patterns without prior instruction. In this sense, the VOI model of eye movements may detect saccades in an eye movement record, but this record does not serve well as predictor of future eye movement patterns. A computational method of eye movement prediction over arbitrary (natural) scenery is an open problem in vision research.

The bubbling effect of aggregate VOIs generates a type of sudden onset stimulus suspected of drawing gaze. It was observed that viewers tended to follow the high resolution insets as they appeared. Eye movements

were not as autonomous as observed in Experiment 2, rather they had a forced appearance. It is interesting to note that the HVS processed sequence produced statistically imperceptible results even though forced eye movements were observed. The distinction between perception and performance mentioned earlier may offer a partial explanation of the observed results. At the milder degradation level (HVS), subjects may have had their eye movements impaired by the sudden onset ROIs, but still managed to perceptually form a coherent representation of the sequence.

The latter speculative comments bring up an intriguing question: is it possible to guide gaze while imperceptibly degrading peripheral areas? Gaze is guided to a certain degree in the directed aspect of visual media. Television commercials or movies are directed so that the audience follows some informational criteria such as action in a movie, or a marketed product. This type of direction may also be possible in visually intensive applications such as virtual reality. The aggregate VOI experiment suggests that VOIs may be effective for this purpose. It may be possible to forcefully guide gaze in a virtual environment so long as performance, not necessarily perception, is not impeded.

CHAPTER XIV

CONCLUSION

This dissertation addresses the design of a gaze-contingent system aimed at matching the requirements of human visual attention. Neurophysiological and psychophysical literature is surveyed to gain insight into the functionality and limitations of the human visual system. Tacitly assuming the point of gaze coincides with the focus of attention, the objective is three-fold: (1) the study of the dynamic aspect of eye movements, (2) the visualization of attention in space-time, and (3) the quantification of perceptual limitations of human foveal and peripheral vision. The dissertation targets three specific aims:

1. Development of the Piecewise Auto-Regressive Integrated Moving Average (PARIMA) time series model of eye movements. This model is used in the analysis of recorded eye movement patterns for spatiotemporal localization of saccades, fixations, and smooth pursuit movements.
2. Visualization of spatiotemporal eye movements through *Volumes Of Interest*, providing an aggregate graphical representation of visual attention.
3. Implementation of a wavelet-based video resolution degradation method matching human visual acuity.

The first and third objectives are realized using the discrete wavelet transform due to its properties of multiresolution frequency localization and multiscale subsampling. A real-time eye tracker is used for eye movement data acquisition and model evaluation through subjective quality experiments.

14.1 Time Series PARIMA Model of Eye Movements

The wavelet-based time series model of eye movements offers a simple yet robust technique for spatiotemporal localization of saccades. Through saccade detection, the model identifies fixations and smooth pursuit movements. The PARIMA model approximates nonlinear eye movements through the simplified assumption of linear summation. For this reason it is not an adequate model of the underlying neural oculomotor processes. Due to its spatiotemporal localization property, the pyramidal multiresolution wavelet transform offers flexible detection of multiscale interventions (i.e., edges or saccades). Computationally, this edge detection technique is known to be optimal. The PARIMA representation provides a parsimonious expression of piecewise signal components corresponding to dynamic fixations. Overall, the PARIMA model is a powerful mathematic-theoretical framework for eye movement analysis.

Empirical evidence suggests the PARIMA model's adequacy for eye movement modeling. Although the aggressive statistical analysis exposed the model's susceptibility to Type I and Type II errors, the model's utility is demonstrated through eye movement visualization and its application in gaze-contingent image represen-

tation.

14.2 Three-dimensional Volume Of Interest Eye Movement Visualization

Three-dimensional eye movement visualization offers three advantages over two-dimensional approaches:

- identification of scanpath onset and termination and movement direction,
- quantification of dwell times, and
- unambiguous representation of transitional eye movements.

The first two benefits are provided by the explicit representation of the temporal dimension. The third depends on the underlying framework used for eye movement classification (i.e., the PARIMA model). Utilizing a suitable analytical substrate, Volumes Of Interest explicitly denote dynamic fixations delineated by saccades. VOIs generalize the two-dimensional representation of eye movements by providing a visual temporal reference frame.

Aggregate Volumes Of Interest extend the identification of interesting two-dimensional spatial features to three-dimensional spatiotemporal components of the visual stimulus. This is particularly useful for the study of motion stimulus. Localization of dynamic fixations through aggregate VOIs also gives insight into the mechanism of visual attention. Aggregate VOIs furnish an adequate prediction of eye movements under restricted conditions, e.g., visual tracking, but not free-viewing.

14.3 Gaze-Contingent Resolution Degradation

Evidence suggests that under certain conditions, a significant amount of information may be withheld in gaze-contingent visual displays with little perceptible effect. Surprisingly, experimental results suggest that under the visual tracking paradigm, effects of linear and acuity-matching resolution mapping are imperceptible. Dyadic linear mapping results in roughly 50% resolution degradation over 97% of the image (50% resolution beyond the 105×105 foveal ROI over a 640×480 image frame). This information reduction has significant implications for gaze-contingent image and video compression. Further research is required to test HVS-matching multidimensional digital compression of imagery, e.g., degradation of color, contrast, and motion.

The visual tracking paradigm utilized in human subject experiments limits the generalizability of results. Visual tracking assumes *a priori* positional gaze information. In free-viewing, the effects of the linear mapping degradation are clearly visible. In fact, using this degradation, the sudden onset of displayed high resolution areas tends to distract natural viewing patterns, as demonstrated by the aggregate VOI condition. Further

research is needed to determine whether (a) a milder form of degradation can offer imperceptible results with significant savings in terms of compression, and/or (b) linear (or stronger) degradation effects can be used to induce eye movement patterns affecting perception without impeding performance. Current results provide positive support for both avenues of research.

Subjective impairment ratings suggest imperceptible effects of resolution degradation milder than the linear mapping function. Temporal ramps used to gradually modulate Regions Of Interest may further enhance the perceptive quality of processed video.

Observations made during linear mapping experiments suggest the possibility of directing induced eye movements to specific locations through the use of noticeable degradation acting as a cue for visual attention. This form of *directed* viewing paradigm may be adequate for preservation of visual performance at the cost of perceived quality impairment.

14.4 Summary

In conclusion, the PARIMA model of eye movements and the wavelet-based methods presented in this dissertation offer a suitable framework for the development of gaze-contingent visual displays. The PARIMA model is an adequate linear approximation of the nonlinear signal, delineating fixation and smooth pursuit movements by saccade discontinuities. The utility of the wavelet-based techniques is demonstrated through the generation of both (1) meaningful three-dimensional visualization of eye movements through Volumes Of Interest, and (2) imperceptible spatiotemporal image resolution degradation.

CHAPTER XV

FUTURE DIRECTIONS

Utilizing the framework for gaze-contingent visual communication adopted in this dissertation, recommendations for future research are made within three application contexts, namely Gaze-Contingent Virtual Reality, Multi-Component Visual Representation, and Computational Modeling of Visual Attention. Open problems within general areas as well as specific shortcomings of present methods are identified, and suggestions for improvements are proposed.

15.1 Gaze-Contingent Virtual Reality

Virtual environments today lack realism. Real-time display of visually rich scenery is encumbered by the demand to render excessive numbers of polygons. This problem is especially severe in virtual reality. To minimize refresh latency, image quality is often sacrificed for speed. The gaze-contingent representation strategy developed in this dissertation is a potential solution to this problem. Specifically, the wavelet-based image degradation methodology is directly applicable to graphics rendering [Fou95]. The multiresolution image reconstruction seamlessly extends to the representation of 3D graphical objects. The use of an appropriate HVS-matching resolution mapping should generate graphical models matching human visual resolvability. This type of rendering is applicable to graphical environments where eye trackers are used as an ocular interface, e.g., gaze-contingent virtual reality and graphical simulators.

A prototype for gaze-contingent multiresolution representation is the virtual environment, *Exploring the Brain Forest*, being developed at the Scientific Visualization Laboratory, Department of Computer Science, Texas A&M University [MBD96]. Two attentive graphical techniques for display-time compression are currently being investigated. The first scale-dependent strategy is contingent upon displaying model neurons at four levels of geometric detail. The second proposed strategy tracks the viewer's eye and matches imagery to the gaze-contingent perceptual limitations of the human visual system.

15.1.1 Scale-dependent Geometric Modeling

Exploring the Brain Forest presents hierarchical views of the brain at several levels of scale from a global overview to immersion within its forest of neurons and glial cells. To simulate the attentive graphical display of this complex virtual environment, the scale-dependent strategy represents model neurons at four levels of geometric detail. Different geometric models are utilized at different ranges (scales) from the viewer under

this *discrete model-switching* display strategy. The virtual environment is modeled on a three-dimensional gaze-contingent segmented stage. The presentation of different geometric models is evoked at different locations of the stage, as shown schematically in Figure 98. Near the expected central (foveal) region, neurons

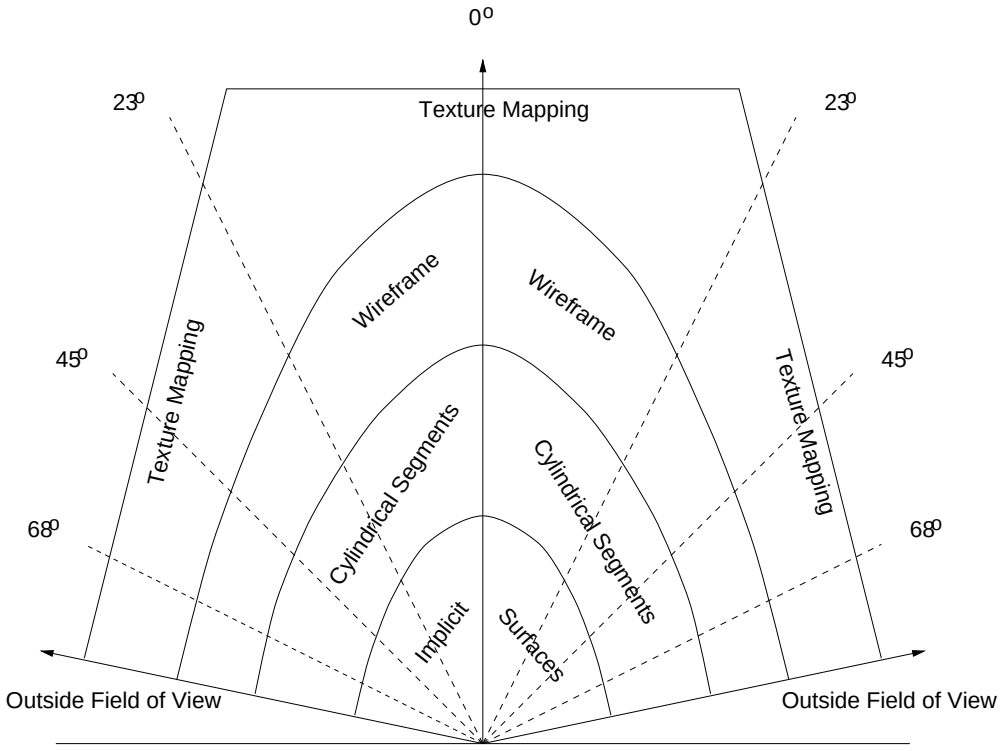


Fig. 98. Gaze-contingent segmented stage.

are represented by implicit surfaces. Further away, in terms of depth and eccentricity, the model is switched to a cylindrical segment representation of neurons, where 7-9 polygons are used per cylinder circumference. Further still, the model is changed to wire-frame curves of variable thickness in 3-space. At the boundaries of the stage, dense neuronal forests are texture mapped onto the stage's polygonal backdrops.

15.1.2 Gaze-contingent Geometric Modeling

The wavelet-based multiresolution strategy developed here for image degradation can be used for multiresolution geometric modeling of figures, curves, and surfaces. [CG95, DLR95]. Cohen and Gortler show the utility of wavelet decompositions for B-spline control point and least squares editing and discuss the implementation of a multiresolution curve and surface modeler. DeRose, Lounsbery and Reissell use wavelets to represent multiresolution surfaces of arbitrary topology. In general, objects are subdivided by refining the control point lattice by projecting to subspaces of different scales (e.g., application of an appropriate scaling

function at a given decomposition level). The control point lattice denoting figures, curves, and surfaces is analogous to an *n*-dimensional image. Therefore, the multiresolution image representation strategy presented in this dissertation is applicable to graphical objects. Application of this strategy to neuronal modeling extends the discrete model-switching strategy to a *seamless multiresolution model* strategy. The gaze-contingent stage segments serve as indicators of resolution level, cueing successive representations of the model (e.g., dyadic approximations).

15.1.3 Outlook

The need for Gaze-Contingent Virtual Reality (GCVR) systems has been recognized and prototypical systems utilizing dynamic foveal regions of interest are currently under development [Jac97, McC97]. The work in this dissertation should be transferred to virtual reality environments using miniature eye tracking optics installed in high-resolution head-mounted displays (HMDs).[1] Following successful testing, wavelet-based attentive graphical rendering should be evaluated in the GCVR. It is the opinion of the author that with sufficiently fast graphics hardware, gaze-contingent rendering may prove more manageable than digital imagery due to the potentially smaller amount of control point data needed to represent graphical objects.

15.2 Multi-Component Attentive Visual Representation

Most display techniques assume an isotropic, or homogeneous, representation of visual information, although the human visual system (HVS) does not process this information *in toto*. JPEG and MPEG are prominent examples of codecs which homogeneously compress imagery through isotropic quantization. Selective quantization schemes are being developed, however, in anticipation of automatic control algorithms [ISO97]. As shown in this dissertation, resolution is one component along which anisotropic image degradation is capable of matching the limitations of the HVS. A multi-component degradation strategy for visuotopic scene representation is proposed below, followed by recommendations for real-time attentive display of imagery.

15.2.1 Visuotopic Scene Representation

Research presented in this dissertation deals with peripheral degradation of spatial resolution. The human visual system is peripherally limited along more than one visual dimension. A multi-component, visuotopic image degradation strategy along six dimensions is proposed in §3.3, and is repeated here for convenience:

[1] Currently HMDs support active matrix LCDs with 640 × 480 resolution providing a 60° (diagonal) field of view (e.g., the V6 from Virtual Research, Santa Clara, CA). Leading manufacturers of video-based eye trackers (e.g., ISCAN and ASL) offer installation of miniature optics into these HMDs.

1. **Spatial Resolution** should remain high within the foveal region and smoothly degrade within the peripheral, matching human visual acuity.
2. **Temporal Resolution** must be available in the periphery. Sudden onset events are potential attentional attractors.
3. **Luminance** should be coded for high visibility in the peripheral areas since the periphery is sensitive to dim objects.
4. **Chrominance** should be coded for high exposure almost exclusively in the foveal region, with chromaticity decreasing sharply into the periphery. This requirement is a direct consequence of the high density of cones and parvocellular ganglion cells in the fovea.
5. **Contrast** sensitivity should be high in the periphery, corresponding to the sensitivity of the magnocellular ganglion cells found mainly outside the fovea.

Special consideration should be given to sudden onset, luminous, high frequency objects (i.e., suddenly appearing bright edges).

The above multivariate quantization of peripheral vision requires a processing strategy capable of analyzing and synthesizing local image regions along all of the above dimensions. It is not known at this time whether the wavelet strategy can satisfy this requirement. Luminance and contrast may be better analyzed in image space instead of the wavelet domain, although the wavelet transform has been used for contrast enhancement [LH94]. The wavelet transform appears suitable for chrominance analysis, synthesis, and degradation. Degradation of the color component may require a form of wavelet coefficient decimation in color space. The wavelet transform is particularly well suited to the analysis of spatial and temporal frequencies (e.g., spatiotemporal edges), as shown in this dissertation. An interesting research direction is a unified representation (encoding) of images (and video) suitable for processing each and all of the above components.

15.2.2 Real-Time Attentive Display Recommendations

Real-time wavelet-based visuotopic display of imagery relies on fast anisotropic synthesis of image frames. At present, image representation involves wavelet decomposition, coefficient filtering, and reconstruction. For an image of appropriate size, e.g., 640×480, this set of operations prohibits real-time execution. Two possible solutions, which may alleviate processing requirements, are discussed below.

First, the wavelet coefficient filtering strategy generates results equivalent to MIP-mapping, however, the recursive nature of the wavelet reconstruction is more time intensive than the MIP-mapping synthesis. In terms of image size, n, the computational complexity of wavelet reconstruction is $O(n\log(n))$, compared to the $O(n)$ complexity of MIP-mapping. MIP-mapping absorbs the initial $O(n\log(n))$ cost during decomposition.

There is a corresponding memory-performance tradeoff between the two approaches, however. Wavelet decomposition requires no more memory than the original image, while MIP-map subimages require roughly 30% more storage per frame [DM95]. If the extra memory requirements are not prohibitive, MIP-mapping reconstruction provides faster display since most of the processing cost is absorbed during decomposition.

A second possible solution involves pre-filtering of a bank of images ready for real-time, gaze-contingent display. *A priori* identification of potential regions of interest may be used to pre-filter images, resulting in a set of images for each frame of video. During each saccade, the image containing ROIs at appropriate locations may be substituted for the upcoming frame. The current diameter of the foveal ROI is 105 pixels. For a 640×480 image, roughly 24 foveal ROIs are needed to cover the entire display. A brute force pre-filtering approach requires a set of 24 images per video frame. For gaze-contingent video, this approach is clearly prohibitive due its intensive memory requirements. The requirement of holding 24 frames in memory is not unreasonable if gaze-contingent static images are used as stimulus. Thus, it is recommended that real-time attentive display strategies be initially evaluated over static scenery. The program `gci` has been developed for gaze-contingent viewing of a single frame and is suitable for this purpose.

Both solutions presented here for real-time image reconstruction are dependent on an appropriately fast gaze prediction strategy. This is discussed in §15.3.3.

15.3 Eye Movement Analysis

The development of a gaze-contingent system depends on a real-time eye movement classification strategy. The PARIMA model, presented in this dissertation, provides an adequate linear filtering strategy capable of segmenting nonlinear eye movement signals into dynamic fixations through the localization of saccades. However, experimental results exposed the current implementation's susceptibility to Type I and II errors. Strategies for refinement of the model are given in §15.3.1. Extension of the model to real-time eye movement analysis is suggested in §15.3.2, and real-time prediction of eye movements is discussed in §15.3.3.

15.3.1 PARIMA Model Parameterization

The currently chosen PARIMA parameters overestimate saccades. Although subjective quality ratings of identified VOIs support the model's overall acceptable performance (e.g., pre-attentive viewing condition of *flight* sequence), it is expected that refinement of PARIMA parameters will yield better accuracy. Strategies for parameter refinement rely on iterative testing of parameters and model evaluation.

It is suspected that a higher spatial decomposition level will result in better detection of inter-frame fixations. Higher spatial decomposition results in coarser subsampling providing a more robust solution of the correspondence problem as discussed in §7.4.

As suggested in §XI, empirical evaluation of the model's performance should relax the criteria of saccade matches. Because of the inherent eye tracking delay between stimulus presentation and point of regard data collection, eye movements cannot be expected to match perfectly with temporal stimulus (e.g., onset). Typical saccade duration and eye tracker latency may be used as delay criteria. For example, a delay of one to two video frames (62-125ms) may provide a more robust analysis of the model.

15.3.2 Real-Time Wavelet-based Eye Movement Analysis

Currently, analysis of eye movements is performed off-line. Each set of point of regard data is converted to a video stream and analyzed *en masse* by the three-dimensional wavelet transform. Because only two levels of temporal decomposition are used, the analysis can be performed in real-time on a small number of video frames, referred to as a "mini-sequence". The delay of this analysis depends on the time needed to (1) compose the eye movement samples into video frames, (2) spatially decompose the frames, (3) temporally decompose the mini-sequence, (4) perform the ROI detection, (5) temporally reconstruct the mini-sequence, (6) spatially project the subsampled frames, and (7) perform the inter-frame linking of ROIs into VOIs. Processing time depends on the size and number of image frames. Frame size is dependent on eye tracker resolution (currently 512×256), although a smaller frame size may be used if on-the-fly spatial subsampling may be substituted for spatial decomposition (step (2)). The number of frames depends on the required temporal decomposition level of the wavelet transform. At present, two temporal decomposition levels are needed, requiring 4 video frames. The 4-frame requirement dictates a minimum delay of $4 \times s_p$, where s_p is the eye tracker sample period. Currently, eye trackers typically provide samples every 18ms resulting in a 72ms delay. Under this configuration, 72ms represents the lower bound for two-level real-time analysis of eye movements. A one-level temporal decomposition (at least one level is required) decreases the minimum delay to 36ms but is more susceptible to noise.

Real-time analysis of eye movements may indicate the current type of eye movement a subject is executing (e.g., fixation, saccade) but it is doubtful that a VOI history could be built in real-time. VOIs are currently assembled by matching each frame's ROI with the VOI currently stored in memory. VOIs are extended to the current frame through a search of existing VOIs (see §8.1). Search time is directly proportional to the number of VOIs in the record.

To make the real-time eye movement analysis worthwhile, at least the current VOI needs to be stored in memory. That is, as soon as a saccade termination is detected, POR data should be assembled into the currently stored VOI for the purposes of short-time eye movement prediction (discussed next).

15.3.3 Real-Time Eye Movement Prediction Through Forecasting

Anticipatory gaze-contingent systems rely on some mechanism of eye movement prediction. Prediction of potential fixation points over an entire image or image sequence is an open problem in computer vision requiring computational image understanding. Identification of visual attractors through analysis of historical data, as attempted in this dissertation, does not constitute a prediction of eye movements. Prediction of eye movements based solely on eye movement data (i.e., out of the visual context) does not seem to make sense. However, the desiderata addressed here reflects the ambitious goal of long-time prediction of eye movements. That is, the goal as stated is concerned with the identification of image regions which one and all subjects will fixate at some point. Indeed, if such an algorithm were available, many computer vision problems would undoubtedly be solved. For real-time eye movement prediction, a different mind set is required viewing the problem in the short term.

The PARIMA model of eye movements is conceptually based on time series analysis (TSA). A subset of TSA methods not utilized in the PARIMA model involves forecasting of observed trends in serially correlated signals. For example, forecasting methods are used in the analysis of stock market data. Although reliable long-term forecasting strategies do not exist, fairly good approximations are possible in the short term. In the context of a real-time anticipatory gaze-contingent system, this type of approach may be sufficient to predict the direction of gaze over a small number of video frames. Consider the real-time eye movement analysis approach proposed in §15.3.2, where it is suggested that a short record of the current VOI be maintained in memory. Since each VOI is modeled by an ARIMA time series, the VOI may generate reliable short-time forecasts provided it is sufficiently long.

Time series forecasting is generally concerned with predicting trends in the data. With respect to fixations and smooth pursuits, short-time forecasts may provide prediction of the direction of these movements. This may be useful for tracking the viewer's gaze during smooth pursuit movements but may not provide useful information regarding fixations. In either case, saccades may intervene defeating the utility of this approach. Saccades are problematic since (1) they disrupt the relatively continuous fixation and pursuit signal, and (2) their short duration prevents meaningful analysis. It is unlikely that time series forecasting could be applied quickly enough to offer a prediction of saccade direction.

Real-time prediction of saccades poses a significant problem, yet to effectively overcome the eye tracker's inherent delay, this problem must be addressed. Real-time detection of saccades seems manageable since it requires identification of high-amplitude discontinuities in the signal. However, the short duration of saccades requires real-time prediction of "ballistic" saccade direction from relatively few data points. A possible fast solution to this problem may be a gradient-based one. That is, once a saccade is detected, an estimate of the saccade's direction may be obtained by calculating the gradient direction of the difference of two (or more) successive POR samples.

15.4 Computational Modeling of Visual Attention: A Survey

A central problem underlying the computational approach to gaze-contingent visual communication systems is the prediction of eye movements. A computational model of visual attention is actively being sought by computer vision researchers. At present, a fully functional model of visual attention has not yet been found. In this section, several prominent attempts at the specification of such a model are summarized, followed by recommendations for possible future research directions with emphasis on the wavelet-based approach.

Visual attention is generally described by the seamless incorporation of the attentional "what" and "where" functionality. Visual performance can be distributed between these concepts in the following manner:

- The "where" appears to function in parallel over the peripheral visual scene, acting as a selective (voluntary), or reflexive or reactive (involuntary) component.
- The "what" coincides with serial, foveal vision.

Most models of visual attention either address one of the "what" and "where" concepts, or attempt to incorporate them both. Selected models of visual attention are briefly reviewed in the following sections.

15.4.1 Koch and Ullman's Model

Koch and Ullman model selective visual attention by three different stages: (1) Salient features from a set of elementary features, computed in parallel across the visual field and represented by a set of topographic maps, are combined into a *saliency map*. (2) A Winner-Take-All (WTA) mechanism, operating on the saliency map, identifies the most conspicuous location. A pyramidal structure, such as the framework proposed for image processing and analysis by Rosenfeld (see [JR94]), is used to implement the WTA mechanism. (3) Properties of the selected location are sent to a central representation. The WTA mechanism automatically shifts to the next most conspicuous location. The shift can be biased by proximity and similarity preferences [KU85]. The elementary features encoded by the topographic maps include color, orientation, direction of movement, and (binocular) disparity. A postulated "switch" maps the properties of the *selected* (most conspicuous) location

from feature space to the more central non-topographic representation and is held as the principal expression of early selective visual attention. The topographic feature maps and the saliency map correspond to the "where" of visual attention, while the selective mapping function corresponds to the "what".

15.4.2 Ahmad's Model

Ahmad proposes a biologically-motivated efficient computational model of visual attention called VISIT (a loose acronym for a network that performs VIsual Search ITeratively) [Ahm91]. The global network structure is divided into four distinct subnetworks: (1) the *gating network* (the "what"), responsible for suppressing all activity except at a given location, (2) the *priority network* (the "where"), which determines the locations of interest, (3) the *control network*, which is responsible for sequencing and for mediating the information flow between the gating and priority networks, and (4) the *working memory*, which temporarily stores relevant information. A priority map used by the priority network is similar to Koch and Ullman's saliency map. Unlike Koch and Ullman's WTA strategy, the control network shifts attention by using both top-down and bottom-up information (similar in spirit to Wolfe's *Guided Search*—see §15.4.4 below) fed to a built-in max function.

Ahmad relates aspects of VISIT to various neurological modules involved in attentional function. Specifically, the gating system relates to the Pulvinar; the bottom-up priority map to the superior colliculus and frontal eye fields; the control network to the posterior parietal areas, and the working memory to the prefrontal cortex [Ahm91, §6.3].

15.4.3 Sandon's Model

Sandon proposes a hierarchical representation of spatial location as the basis for a connectionist network simulating attention [San90]. Specifically, a Gaussian pyramid, computed by repeatedly smoothing and reducing the input intensity image, provides multiresolution data paths to the attentional connectionist network layers allowing competition among scales for higher level processing. Since static, monochromatic images are used as input, the lowest level features used by the model are oriented edges. A central-excitatory, peripheral-inhibitory interaction among features of a given type is applied to each of the feature arrays used as input to the attentional network. A WTA mechanism within the attentional network inhibits all but the strongest activation within the image array. Similar to Ahmad's gating network, the effect of the attentional activity is to gate (select) features from specific image regions (the "where") up to higher network layers where object recognition occurs (the "what").

It is important to note the use of the Gaussian pyramid. This pyramidal structure is similar to the hierarchy of images used in MIP-mapping and to the subsampled images generated by the wavelet scaling function. In

fact, with appropriate wavelets, the wavelet representation provides an identical but superset representation of the structure used by Sandon. The wavelet representation in a sense contains twice as much information since it also contains the multiscale derivative (or difference) images generated by the wavelet function. Furthermore, multiscale edge information, corresponding to oriented edge features, is readily available in the wavelet domain.

15.4.4 Wolfe's Model

Following Treisman's notion of a feature map, Wolfe proposes a computational model of visual search called *Guided Search* [Wol93, Wol94].[2] Wolfe's feature-specific maps are used by parallel processes to analyze a scene in order to guide attention. The information from feature maps is combined in an activation map. The activation map is a weighted sum of the activations in the parallel feature processors. Attention is deployed from location to location in order of decreasing activation. Wolfe's *Guided Search* occurs when the signal from the parallel processes is often (but not always) larger than the background noise.

Because Wolfe's strategy is goal-directed it is considered by the present author to be a partial model of visual attention (e.g., covert or voluntary attention). Visual attention can also be involuntary or reflexive. Wolfe's model is mentioned here because it adheres to general attentional principles, i.e., the model divides the search task into a preliminary parallel stage (the "where") followed by a sequential search process (the "what").

15.4.5 Tsotsos' Model

Recently, Tsotsos has proposed a model of visual attention motivated by complexity analysis of the visual system [Tso95]. Tsotsos argues that vision (computational and human) is intractable if performed *in toto* over the visual scene [Tso90]. Specifically, it is suggested that unbounded visual search (where a target is either explicitly unknown in advance or is not used in the execution of the search) is NP-complete. The proof relies on a direct reduction of visual search to the Knapsack problem. Bounded search, on the other hand, is claimed to have linear time complexity in the size of the image. Although the result of Tsotsos' complexity analysis may be questionable (see commentaries of Heathcote and Mewhort, Krueger and Tsav, and Kube in [Tso90]), it leads to the generally accepted conclusion that selective attention is essential for vision.

A description of Tsotsos' attentional model appears in various places [CT92a, CT92b, Tso90, Tso95]. The model is similar to Koch and Ullman's algorithm with two distinctions: (1) the model does not rely on a

[2] Visual search can be considered a goal-directed subtask of visual attention, e.g., the covert or voluntary component. For more information on this topic, see [BKB93, BM93, DMS93, Dol93, GS93, JCB93, Liu93, TK93, VOD93].

saliency map, and (2) the WTA strategy is modified by "bias units" allowing multiple winners (see [CT92a]). The selection mechanism is similar to both Ahmad's and Wolfe's models in that both bottom-up and top-down pyramidal searches are employed (see [Tso95]). The image pyramid resembles Sandon's Gaussian structure since it is composed of progressively spatially averaged images at higher levels (see [CT92b]).

The pyramidal representation is computed bottom-up, modified by biases if available, and the most salient item is detected and localized in a top-down fashion, pruning parts of the pyramid that do not contribute to the most salient item. The top-down "attentional beam" inhibits regions outside an internal pass zone not unlike a moving center-surround receptive field.

Tsotsos classifies his model as an instance of the *selective tuning* hypothesis which claims attention is used to tune the visual processing architecture in order to overcome pyramidal computation problems and to allow task-directed processing. This class of attentional models is contrasted by the author with two other major computational hypotheses, namely the *selective routing* and *temporal tagging* hypotheses. The former contains models similar to Koch and Ullman's and Anderson, Olshausen and Van Essen's (see §15.4.6 below). The latter category contains models based on single-cell performance predictions (e.g., firing rates and frequency modulations). The main distinctions between the selective routing and selective tuning hypotheses are the inclusion of inhibitory processes used in spatial selection and a multiple-winner modification of the WTA mechanism.

15.4.6 Anderson, Olshausen, and Van Essen's Model

Falling under Tsotsos' *selective routing* class of models of visual attention, Anderson, Olshausen, and Van Essen propose a *routing circuit* model of visual attention [AOV95]. The authors address the issues of shift and scale invariance by incorporating a switching mechanism for shifting and rescaling sensory input data. The general purpose of a (low-level) switching mechanism is to rescale, reposition, and otherwise reformat visual information into a standard representation for later (higher-level) recognition processes. Visual attention plays a key role in this process since attending to an object theoretically places it into a canonical, or object-based, reference frame.

The quintessential prediction of the routing circuit model is that cortical receptive fields should be dynamic; shifting and rescaling with attention. The model provides a view of preattentive vision which, in general, provides input to a saliency map including color and texture gradients.

15.4.7 Survey Summary

The brief survey of computational models of visual attention presented here is by no means exhaustive. Four highlighted computational strategies, however, appear common to most approaches. First, visual input is popularly represented by a coarse-to-fine pyramidal structure. Second, some form of feature or saliency map is derived from the pyramidal representation (e.g., luminance, edge, or motion). Third, a (possibly multi-layer) gating network is arranged to select the most salient component(s) from the saliency map. The selection most often appears to be handled by a variant of the Winner-Take-All strategy. Fourth, a control mechanism (such as selective routing or tuning) is proposed as a means for feeding information to further processing areas and for dynamically relocating the focus of attention.

Components of computational models of visual attention are generally fashioned after their hypothesized neurological counterparts. In broad terms, the four major computational modules identified here correspond to four cortical areas implicated in visual attention: (1) The pyramidal representation corresponds to retino-geniculo-cortical receptive fields. (2) Saliency maps encode visual information as represented by feature-sensitive cells found in the striate cortex (area V1). (3) The WTA gating strategy represents neurons in either the superior colliculus, frontal eye fields, parietal cortex, or Pulvinar. (4) The control mechanism models neurons thought to reside in the Pulvinar and deep (possibly posterior parietal) layers of the cortex. For a recent, although brief, survey of computational models of visual attention see [OK95].

15.5 Computational Modeling of Visual Attention: A Proposed Framework

Computational models of visual attention rely on a pyramidal framework for representation of the visual scene as the basis for further attentional processing. In most cases, the feature or saliency map built from this coarse-to-fine representation includes either edge or luminance information. The inclusion of both features typically requires the construction of several saliency maps. Here, a wavelet-based model of visual attention is proposed which utilizes the wavelet transform's ability to represent both luminance and edge information simultaneously. In the case of multi-component imagery (e.g., color video), the wavelet transform provides a unified framework for the representation of luminance, chrominance, and spatiotemporal edge information.

The following outline presents details the application of the spatiotemporal wavelet framework to the design of a wavelet-based model of visual attention. Recommendations for model design are presented in three stages, describing (1) the wavelet *saliency pyramid* in §15.5.1, (2) the gating network (the "where") in §15.5.2, and (3) the control mechanism (the "what") in §15.5.3. Preliminary implementation directions are suggested in §15.5.4, and a proposed model evaluation methodology is discussed in §15.5.5.

15.5.1 The Saliency Pyramid (the encoding of visual input)

The bottom layer of the proposed attentional model represents the substrate used for parallel identification of visual features. Following the reasoning used in previous models, a coarse-to-fine pyramidal encoding is suggested. The wavelet transform is recommended for this purpose since it readily provides a hierarchical, pyramidal multiscale representation of the visual input. With appropriate wavelet functions, e.g., Gabor wavelets modeling retino-geniculo-cortical orientation-selective cells, the wavelet transform combines traditional first and second model elements. That is, both pyramidal average and saliency map representations of the input are contained in the wavelet structure. To emphasize this duality, the wavelet representation is termed the *saliency pyramid*. With the exception of chrominance and contrast information, a single saliency pyramid encodes multiple visual components, including resolution, luminance, and spatio-temporal frequency.

15.5.2 The Gating Network (the "where")

The next layer of the model represents the gating functionality of the superior colliculus, the frontal eye fields, the parietal cortex, and the Pulvinar. It is suggested that the gating network is represented by three subnetworks modeling the three cortical centers implicated in eye movement control: (1) the superior colliculus (involuntary attention), (2) the middle temporal (MT) area (voluntary, or covert attention), and cortical centers associated with vestibulo-ocular eye movements (reflexive attention). Gating subnetworks should be incorporated via a switching circuit alternating among voluntary, involuntary, and reflexive attentional modes. The gating network's overall function should select the most salient component from the saliency pyramid, depending on the current attentional mode. For example, if the current mode is voluntary, involved in task-directed selection, the gating network should simulate visual search. Some form of WTA strategy may be implemented as the selective mechanism, e.g., a neural network incorporating central-excitatory, peripheral-inhibitory (center-surround) interaction among features.

15.5.3 The Control Mechanism (the "what")

The control mechanism should be involved in the selection of the next focus of attention. An inhibitory feature interaction layer may be provided in the gating network which attenuates previously inspected regions. The control mechanism should model the functionality of areas within the Pulvinar and the posterior parietal cortex.

15.5.4 Preliminary Directions

The central problem identified in the context of video processing is the prediction of gaze. Since it is suspected that motion is the dominant attentional stimulus to which the visual system responds, the above model of visual attention should initially be implemented with only one temporal saliency pyramid representing motion features. The spatiotemporal pyramid should ideally be constructed in real-time, following recommendations for real-time eye movement analysis presented in §15.3.2. Other features such as color may be incorporated at a later date.

15.5.5 Proposed Model Evaluation

The purpose of the computational model of visual attention is to predict eye movement patterns over a video sequence of arbitrary length. Metrics based on Volumes Of Interest are proposed as methods for preliminary evaluation of the model. Predicted eye movement patterns may be represented by VOIs from which *objective* measures should be derived for the given stimulus. Variability measures may also be obtained to test the model's performance on successive runs over the same video sequence. Next, VOIs obtained from experimental subject trials should be used to derive *subjective* measures. An estimate of the model's performance should be drawn from a comparison of VOI-based measures. Suggestions for the development of subjective and objective metrics are given below in §15.5.5.1 and §15.5.5.2, respectively.

15.5.5.1 Subjective VOI Measure of Eye Movement Variability

VOIs obtained from experimental subject trials may be used to derive *subjective* measures of human eye movement variability over visual stimuli. Qualitatively, the VOI model depicts the variability of subjects' eye movement patterns over the course of the visual stimulus. As seen in the experimental results, a single VOI over the course of a video sequence suggests little variability in eye movements over the scene (recall an individual's three viewing trials of the *cnn* sequence discussed in §12.6). This observation may be interpreted as either due to the subject's familiarity with the content or the uninteresting quality of the content itself. Quantitatively, the number of VOIs detected over a sequence gives a number of visual features on which the subject fixated. The number of aggregate VOIs gives a number of features fixated by multiple subjects.

In essence, the VOI model may be used to derive a measure of eye movement variability over visual stimuli. Furthermore, this measure may also be used as a qualitative *degree of disorientation* of subjects presented with particular types of stimuli. For example, consider three types of video sequences: (a) a movie or commercial, (b) a sporting event such as a football game, and (c) a street surveillance scene. In (a), the directed nature of the composition suggests an expected small number of similar scanpaths. In (b), virtually one VOI

is expected centered on the player or object of interest, e.g., the football or the quarterback. In (c), the diverse amount of information (e.g., crowds, cars, etc.) suggests an expected large number of diverse (sparsely situated) VOIs.

The video sequences used in the present work are rather extreme examples of video sequences: the *flight* and *brain2* sequences contain a single (possibly) moving object suitable for visual tracking, while the *cnn* sequence is composed of relatively motionless content. In the sequences used in the visual tracking paradigm, generally a single VOI was observed, as expected. The free-viewing paradigm evoked somewhat diverse responses with identifiable fixation points through the aggregate VOI composition. At this stage only qualitative observations can be made regarding the VOIs generated over these video sequences. The variability found in the VOIs should be expressed quantitatively. This measure may be dependent on both the number of VOIs and their spatiotemporal dispersion. The resultant metric would provide a *subjective* measure of video content.

15.5.5.2 Objective VOI Quantification of Diverse Visual Stimulus

Analogous to the subjective measures of video content derived from human eye movement patterns proposed in §15.5.5.1, *objective* measures can be obtained from eye movement patterns predicted by the computational model of visual attention. This metric dichotomy is similar to subjective quality ratings of imagery compared to statistical objective measures (e.g., the mean-squared error (MSE)). In the present context, an objective measure of video content is sought based on computationally generated VOIs.

The three-dimensional wavelet analysis of eye movements is directly applicable to video processing. An objective measure of video content estimates the variability of potential visual attractors within the video sequence. This is a more difficult problem than the subjective measure since it relies on the detection of salient image features. Correct detection of visual attractors is a central issue in the development of a computational model of visual attention, discussed in §15.4. Nevertheless, certain image features, such as motion, may be readily detected computationally. Since video motion is essentially manifested by temporal edges, the wavelet-based saccade detection algorithm may serve as a suitable starting point for the development of an objective motion metric.[3] Congruence between developed objective and subjective metrics would offer encouraging progress towards a computational model of visual attention. For this reason alone, the development of objective measures of diverse stimuli should be pursued.

[3]As an informal test, the saccade detection algorithm was run over the *tennis* sequence. Surprisingly, the "informative details" detected by the algorithm corresponded to the tennis player's arm, paddle, and tennis ball, as expected. Background features were eliminated (or "inhibited") upon reconstruction.

15.6 Epilogue

Research uncovered in this investigation identifies open problems in the interdisciplinary study of human vision and visual attention. Results of the work performed in this dissertation offer a spatiotemporal framework suitable for the development and evaluation of computational models of visual attention. More work lies ahead, promising exciting discoveries in computer graphics and scientific visualization, signal processing and eye movement modeling, and gaze-contingent displays.

REFERENCES

[AB90] A. ABDEL-MALEK AND J. BLOOMER, *Visually Optimized Image Reconstruction*, in Human Vision and Electronic Imaging: Models, Methods, and Applications (vol. 1249), SPIE, 1990, pp. 330–335.

[Ahm91] S. AHMAD, *VISIT: An Efficient Computational Model of Human Visual Attention*, Ph.D. dissertation, International Computer Science Institute (ICSI), Berkeley, CA, September 1991. TR-91-049.

[Akl89] S. G. AKL, *The Design and Analysis of Parallel Algorithms*, Prentice-Hall International, Inc., Englewood Cliffs, NJ, 1989.

[And89] R. A. ANDERSEN, *Visual and Eye Movement Functions of the Posterior Parietal Cortex*, Annual Review of Neuroscience, 12 (1989), pp. 377–403.

[AOV95] C. H. ANDERSON, B. A. OLSHAUSEN, AND D. VAN ESSEN, *Routing Networks in Visual Cortex*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 823–826.

[ABMD92] M. ANTONINI, M. BARLAUD, P. MATHIEU, AND I. DAUBECHIES, *Image Coding Using Wavelet Transform*, IEEE Transactions on Image Processing, 1 (1992), pp. 205–220.

[Bar94] H. J. BARNARD, *Image and Video Coding Using a Wavelet Decomposition*, Ph.D. dissertation, Technische Universiteit Delft, Delft, The Netherlands, May 1994.

[BKB93] D. P. BIRKMIRE, R. KARSH, AND B. D. BARNETTE, *Eye Movements in Search and Target Acquisition*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 1305–1309.

[BL88a] F. E. BLOOM AND A. LAZERSON, *Brain, Mind, and Behavior*, W. H. Freeman and Company, New York, NY, 2nd ed., 1988.

[BM93] M. BÖCKER AND L. MÜHLBACH, *Communicative Presence in Videocommunications*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 249–253.

[BL88b] K. R. BOFF AND J. E. LINCOLN, eds., *Engineering Data Compendium: Human Perception and Performance*, USAF Harry G. Armstrong Aerospace Medical Research Laboratory (AAMRL), Wright-Patterson AFB, OH, 1988.

[BJ76] G. E. P. BOX AND G. M. JENKINS, *Time Series Analysis: Forecasting and Control*, Holden-Day, Inc., Oakland, CA, 1976.

[Bri95] B. BRIDGEMAN, *Dissociations Between Visual Processing Modes*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 318–320.

[Bro58] D. E. BROADBENT, *Perception and Communication*, Pergamon Press, Oxford, 1958.

[BA83a] A. BUIZZA AND P. AVANZINI, *Computer Analysis of Smooth Pursuit Eye Movements*, in Eye Movements and Psychological Functions: International Views, R. Groner, C. Menz, D. F. Fisher, and

R. A. Monty, eds., Lawrence Erlbaum Associates, Hillsdale, NJ, 1983, pp. 7–17.

[Bur81] P. J. BURT, *Fast Filter Transforms for Image Processing*, Computer Graphics and Image Processing, 16 (1981), pp. 20–51.

[BA83b] P. J. BURT AND E. H. ADELSON, *The Laplacian Pyramid as a Compact Image Code*, IEEE Transactions on Communications, 31 (1983), pp. 532–540.

[CPSZ83] C. CABIATI, M. PATORMERLO, R. SCHMID, AND D. ZAMBARBIERI, *Computer Analysis of Saccadic Eye Movements*, in Eye Movements and Psychological Functions: International Views, R. Groner, C. Menz, D. F. Fisher, and R. A. Monty, eds., Lawrence Erlbaum Associates, Hillsdale, NJ, 1983, pp. 19–29.

[Can86] J. CANNY, *A Computational Approach to Edge Detection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8 (1986), pp. 679–698.

[Car96] D. CARMODY, *Bundled Replies to Fixation Algorithm Query*. EYEMOV-L Listserv Message-ID: <01I1GUZTRQ5E8X7GY4@spcvxa.spc.edu>, February 21 1996. Available at URL: <http://listserv.spc.edu/> (last referenced June 1997).

[Car93] R. CARMONA, *Wavelet Identification of Transients in Noisy Time Series*. Preprint of talk given at the Meeting of the Interface'93 (private collection A. T. Duchowski), 1993.

[Car77] R. H. S. CARPENTER, *Movements of the Eyes*, Pion Limited, London, 1977.

[Cas96] K. R. CASTLEMAN, *Digital Image Processing*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1996.

[CM95] C. CHIANN AND P. A. MORETTIN, *A Wavelet Analysis for Time Series*. Preprint (private collection A. T. Duchowski), 1995.

[Chu92] C. K. CHUI, *An Introduction to Wavelets*, Academic Press, Inc., San Diego, CA, 1992.

[CG95] M. F. COHEN AND S. J. GORTLER, *Variational Geometric Modeling with Wavelets*, in Course Notes: Wavelets and Their Applications in Computer Graphics (SIGGRAPH 1995, 22nd International Conference on Computer Graphics and Interactive Techniques), A. Fournier, ed., ACM, New York, NY, 1995, pp. 190–199.

[CDL77] C. COHEN-TANNOUDJI, B. DIU, AND F. LALOË, *Quantum Mechanics*, John Wiley & Sons, New York, NY, 1977.

[Col83] P. R. COLES, *Introduction*, in Eye Movements and Psychological Functions: International Views, R. Groner, C. Menz, D. F. Fisher, and R. A. Monty, eds., Lawrence Erlbaum Associates, Hillsdale, NJ, 1983, pp. 1–5.

[CC79] T. D. COOK AND D. T. CAMPBELL, *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Houghton Mifflin Company, Boston, MA, 1979.

[Cra94] H. D. CRANE, *The Purkinje Image Eyetracker, Image Stabilization, and Related Forms of Stimulus Manipulation*, in Visual Science and Engineering: Models and Applications, D. H. Kelly, ed., Marcel Dekker, Inc., New York, NY, 1994, pp. 13–89.

[CS85] H. D. CRANE AND C. M. STEELE, *Generation-V Dual-Purkinje-Image Eyetracker*, Applied Optics, 24 (1985), pp. 527–537.

[CHLT94] J. B. CROMWELL, M. J. HANNAN, W. C. LABYS, AND M. TERRAZA, *Multivariate Tests for Time Series Models*, Series: Quantitative Applications in the Social Sciences, Sage Publications, Thousand Oaks, CA, 1994. Series/Number: 07-100.

[CLT94] J. B. CROMWELL, W. C. LABYS, AND M. TERRAZA, *Univariate Tests for Time Series Models*, Series: Quantitative Applications in the Social Sciences, Sage Publications, Thousand Oaks, CA, 1994. Series/Number: 07-099.

[CT92a] S. M. CULHANE AND J. K. TSOTSOS, *A Prototype for Data-Driven Visual Attention*, in Pattern Recognition, 11th Int'l Conference, vol. I: Conference A, IEEE, 1992, pp. 36–40.

[CT92b] ——, *An Attentional Prototype for Early Vision*, in Computer Vision–ECCV'92: Second European Conference on Computer Vision (Santa Margherita Ligure, Italy), New York, NY, May 1992, Springer-Verlag, pp. 551–560.

[Dau88] I. DAUBECHIES, *Orthonormal Bases of Compactly Supported Wavelets*, Communications on Pure and Applied Mathematics, XLI (1988), pp. 909–996.

[Dau92] ——, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[Dav80] H. DAVSON, *Physiology of the Eye*, Academic Press, New York, NY, 4th ed., 1980.

[DeG92] P. DE GRAEF, *Scene-Context Effects and Models of Real-World Perception*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 243–259. Springer Series in Neuropsychology.

[DD88] R. L. DE VALOIS AND K. K. DE VALOIS, *Spatial Vision*, Oxford University Press, New York, NY, 1988.

[DLR95] T. D. DEROSE, M. LOUNSBERY, AND L.-M. REISSELL, *Curves and Surfaces*, in Course Notes: Wavelets and Their Applications in Computer Graphics (SIGGRAPH 1995, 22nd International Conference on Computer Graphics and Interactive Techniques), A. Fournier, ed., ACM, New York, NY, 1995, ch. V, pp. 123–153.

[DD63] J. A. DEUTSCH AND D. DEUTSCH, *Attention: Some Theoretical Considerations*, Psychological Review, 70 (1963), pp. 80–90.

[Die93] P. DIERCKX, *Curve and Surface Fitting with Splines*, Monographs on Numerical Analysis, Oxford University Press, Inc., New York, NY, 1993.

[Dol93] T. J. DOLL, *Preattentive Processing in Visual Search*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 1291–1249.

[DMS93] T. J. DOLL, S. W. WHORTER, AND D. E. SCHMIEDER, *Simulation of Human Visual Search in Cluttered Backgrounds*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual

Meeting, Santa Monica, CA, October 1993, pp. 1310–1314.

[DJKP96] D. L. DONOHO, I. M. JOHNSTONE, G. KERKYACHARIAN, AND D. PICARD, *Density Estimation by Wavelet Thresholding*, The Annals of Statistics, 24 (1996), pp. 508–539.

[DM95] A. T. DUCHOWSKI AND B. H. MCCORMICK, *Simple Multiresolution Approach for Representing Multiple Regions of Interest (ROIs)*, in Visual Communications and Image Processing, Taipei, Taiwan, May 1995, SPIE, pp. 175–186.

[Fin92] J. M. FINDLAY, *Programming of Stimulus-Elicited Saccadic Eye Movements*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 8–30. Springer Series in Neuropsychology.

[Fin89] R. A. FINKE, *Principles of Mental Imagery*, The MIT Press, Cambridge, MA, 1989.

[Fin74] J. D. FINN, *A General Model for Multivariate Analysis*, Holt, Rinehart and Winston, Inc., New York, NY, 1974.

[FvD82] J. D. FOLEY AND A. VAN DAM, *Fundamentals of Interactive Computer Graphics*, Addison-Wesley, Reading, MA, 1982.

[FvFH90] J. D. FOLEY, A. VAN DAM, S. K. FEINER, AND J. F. HUGHES, *Computer Graphics: Principles and Practice*, Addison-Wesley, Reading, MA, 2nd ed., 1990.

[FGT89] D. H. FOSTER, S. GRAVANO, AND A. TOMOSZEK, *Acuity for Fine-Grain Motion and For Two-Dot Spacing as a Function of Retinal Eccentricity: Differences in Specialization of the Central and Peripheral Retina*, Vision Research, 29 (1989), pp. 1017–1031.

[Fou95] A. FOURNIER, ed., *Course Notes: Wavelets and Their Applications in Computer Graphics*, New York, NY, 1995, SIGGRAPH 1995, 22nd International Conference on Computer Graphics and Interactive Techniques, ACM.

[FM92] J. FROMENT AND S. MALLAT, *Second Generation Compact Image Coding with Wavelets*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, London, 1992, pp. 655–678.

[FKS85] A. F. FUCHS, C. R. S. KANEKO, AND C. A. SCUDDER, *Brainstem Control of Saccadic Eye Movements*, Annual Review of Neuroscience, 8 (1985), pp. 307–337.

[GW94] N. A. GERSHENFELD AND A. S. WEIGEND, *The Future of Time Series: Learning and Understanding*, in Time Series Prediction: Forecasting the Future and Understanding the Past, A. S. Weigend and N. A. Gershenfeld, eds., Addison-Wesley Publishing Company, Reading, MA, 1994, ch. 1, pp. 1–70. Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis held in Santa Fe, NM, May 14-17, 1992.

[GS93] J. H. GOLDBERG AND J. C. SCHRYVER, *Eye-Gaze Control of the Computer Interface: Discrimination of Zoom Intent*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 1370–1374.

[GW87] R. C. GONZALEZ AND P. WINTZ, *Digital Image Processing*, Addison-Wesley Publishing Company, Reading, MA, 2nd ed., 1987.

[GB92] R. A. GOPINATH AND C. S. BURRUS, *Wavelet Transforms and Filter Banks*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, London, 1992, pp. 603–654.

[Got81] J. M. GOTTMAN, *Time-Series Analysis: A Comprehensive Introduction for Social Scientists*, Cambridge University Press, Cambridge, 1981.

[Gre90] R. L. GREGORY, *Eye and Brain: The Psychology of Seeing*, Princeton University Press, Princeton, NJ, 1990.

[GN95] N. M. GRZYWACZ AND A. M. NORCIA, *Directional Selectivity in the Cortex*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 309–311.

[GSA95] N. M. GRZYWACZ, E. SERNAGOR, AND F. R. AMTHOR, *Directional Selectivity in the Retina*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 312–314.

[GR94] T. A. GYAW AND S. R. RAY, *The Wavelet Transform as a Tool for Recognition of Biosignals*, in Biomedical Science Instrumentation (v.30), Kansas State University, Manhattan, KS, April 22-23 1994, Instrument Society of America, pp. 63–68. Proceedings of the 31st Annual Rocky Mountain Bioengineering Symposium & 31st International ISA Biomedical Sciences Instrumentation Symposium.

[HH73] R. N. HABER AND M. HERSHENSON, *The Psychology of Visual Perception*, Holt, Rinehart, and Winston, Inc., New York, NY, 1973.

[HLM92] M. M. HAYHOE, J. LACHTER, AND P. MOELLER, *Spatial Memory and Integration Across Saccadic Eye Movements*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 130–145. Springer Series in Neuropsychology.

[HK89] P. HE AND E. KOWLER, *The Role of Location Probability in the Programming of Saccades: Implications for "Center-of-Gravity" Tendencies*, Vision Research, 29 (1989), pp. 1165–1181.

[Heg92] M. HEGARTY, *The Mechanics of Comprehension and Comprehension of Mechanics*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 428–443. Springer Series in Neuropsychology.

[Hen92] J. M. HENDERSON, *Visual Attention and Eye Movement Control During Reading and Picture Viewing*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 260–283. Springer Series in Neuropsychology.

[Hol90] S. S. HOLLAND, JR., *Applied Analysis by the Hilbert Space Method: An Introduction with Applications to the Wave, Heat, and Schrödinger Equations*, vol. 137 of Monographs and Textbooks in Pure

and Applied Mathematics, Marcel Dekker, Inc., New York, NY, 1990.

[Hub88] D. H. HUBEL, *Eye, Brain, and Vision*, Scientific American Library, New York, NY, 1988.

[Irw92] D. E. IRWIN, *Visual Memory Within and Across Fixations*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 146–165. Springer Series in Neuropsychology.

[Iscan94] ISCAN INC., *ISCAN Point-Of-Regard Data Acquisition & Fixation Analysis Software (Operating Instructions)*, 125 Cambridgepark Drive, P.O. Box 382076, Cambridge, MA 02238, PC Card Version 2.06 ed., August 29 1994.

[ISO97] ISO TECHNICAL COMMITTEE ON CODING OF AUDIO, PICTURE, MULTIMEDIA, AND HYPERMEDIA INFORMATION (ID: JTC 1/SC 29), *Digital Compression and Coding of Continuous-Tone Still Images: Extensions (Draft ID: ISO/IEC 10918-3:1997)*, International Standards Organization (ISO), 1, rue de Varembé, Case postale 56, CH-1211 Genève 20, Switzerland, 1st (monolingual) ed., 1997. Drafts may be purchased through the Internet at URL: <http://www.iso.ch/> (last referenced June 1997).

[Jac97] L. JACOBSON, *How to Exploit Human Perception (And Other Tales of Planetary Exploration)*, IRIS Universe (Number 39, Spring), (1997), pp. 71–72.

[JAE96] R. H. JACOBY, B. D. ADELSTEIN, AND S. R. ELLIS, *Improved Temporal Response in Virtual Environments Through System Hardware and Software Reorganization*, in Stereoscopic Displays and Virtual Reality Systems III, M. T. Bolas and S. S. Fisher, eds., San Jose, CA, January 30-February 2 1996, SPIE.

[Jai89] A. K. JAIN, *Fundamentals of Digital Image Processing*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1989.

[JCB93] M. R. JAMES, J. N. COLDWELL, AND A. J. BELYAVIN, *The Effect of Peripheral Visual Cues on the Control of Self-Stabilization in Roll*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 157–161.

[Jam81] W. JAMES, *The Principles of Psychology*, vol. I of The Works of William James, Harvard University Press, Cambridge, MA, 1981.

[JS94a] B. JAWERTH AND W. SWELDENS, *An Overview of Wavelet Based Multiresolution Analyses*, SIAM Review, 36 (1994), pp. 377–412.

[JS94b] I. M. JOHNSTONE AND B. W. SILVERMAN, *Wavelet Threshold Estimators for Data With Correlated Noise*. Preprint (private collection A. T. Duchowski), 1994.

[JR94] J.-M. JOLION AND A. ROSENFELD, *A Pyramid Framework for Early Vision: Multiresolutional Computer Vision*, Kluwer Academic Publishers, Norwell, MA, 1994.

[Jos95] A. W. JOSHI, *Matrices and Tensors in Physics*, John Wiley & Sons, New York, NY, 3rd ed., 1995.

[Kaa89] J. H. KAAS, *Why Does the Brain Have So Many Visual Areas?*, Journal of Cognitive Neuroscience,

1 (1989), pp. 121–135.

[Kap91] E. KAPLAN, *The Receptive Field Structure of Retinal Ganglion Cells in Cat and Monkey*, in The Neural Basis of Visual Function, A. G. Leventhal and J. R. Cronly-Dillon, eds., CRC Press, Boca Raton, FL, 1991, ch. 2, pp. 10–40. Vision and Visual Dysfunction Series, vol. 4.

[Kap84] W. KAPLAN, *Advanced Calculus*, Addison-Wesley Publishing Company, Reading, MA, 3rd ed., 1984.

[KB83] R. KARSH AND F. W. BREITENBACH, *Looking at Looking: The Amorphous Fixation Measure*, in Eye Movements and Psychological Functions: International Views, R. Groner, C. Menz, D. F. Fisher, and R. A. Monty, eds., Lawrence Erlbaum Associates, Hillsdale, NJ, 1983, pp. 53–64.

[Ken92] A. KENNEDY, *The Spatial Coding Hypothesis*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 379–396. Springer Series in Neuropsychology.

[KKP92] R. KLEIN, A. KINGSTONE, AND A. PONTERFACT, *Orienting of Visual Attention*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 46–65. Springer Series in Neuropsychology.

[KU85] C. KOCH AND S. ULLMAN, *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*, Human Neurobiology, 4 (1985), pp. 219–227.

[Koc87] D. KOCIAN, *Visual World Subsystem*, in Super Cockpit Industry Days: Super Cockpit/Virtual Crew Systems, Air Force Museum, Wright-Patterson AFB, OH, 31 March–1 April 1987, Air Force Systems Command/Human Systems Division/Armstrong Aerospace Medical Research Laboratory.

[KDG85] J. J. KOENDERINK, A. J. VAN DOORN, AND W. A. VAN DE GRIND, *Spatial and Temporal Parameters of Motion Detection in the Peripheral Visual Field*, J. Opt. Soc. Am., 2 (1985), pp. 252–259.

[Koo88] D. B. KOONS, *A Model for the Representation and Extraction of Visual Knowledge from Illustrated Texts*, M. S. thesis, Texas A&M University, College Station, TX, August 1988. Techreport TAMU-88-010.

[Kos94] S. M. KOSSLYN, *Image and Brain*, The MIT Press, Cambridge, MA, 1994.

[LS86] P. LANCASTER AND K. ŠALKAUSKAS, *Curve and Surface Fitting: An Introduction*, Academic Press, San Diego, CA, 1986.

[LR86] V. P. LAURUTIS AND D. A. ROBINSON, *The Vestibulo-ocular Reflex During Human Saccadic Eye Movements*, Journal of Physiology, 373 (1986), pp. 209–233.

[LZ91] R. J. LEIGH AND D. S. ZEE, *The Neurology of Eye Movements*, Contemporary Neurology Series, F. A. Davis Company, Philadelphia, PA, 2nd ed., 1991.

[LT96] Y.-C. LIN AND S.-C. TAI, *Dynamic Windowed Codebook Search Algorithm in Vector Quantization*, Optical Engineering, 35 (1996), pp. 2921–2929.

[Liu93] Y. LIU, *Visual Scanning, Memory Scanning, and Computational Human Performance Modeling*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 142–146.

[LH88] M. LIVINGSTONE AND D. HUBEL, *Segregation of Form, Color, Movement, and Depth: Anatomy, Physiology, and Perception*, Science, 240 (1988), pp. 740–749.

[LTFW+89] T. LONGRIDGE, M. THOMAS, A. FERNIE, T. WILLIAMS, AND P. WETZEL, *Design of an Eye Slaved Area of Interest System for the Simulator Complexity Testbed*, in Area of Interest/Field-Of-View Research Using ASPT (Interservice/Industry Training Systems Conference), T. Longridge, ed., Brooks Air Force Base, TX, 1989, National Security Industrial Association, Air Force Human Resources Laboratory, Air Force Systems Command, pp. 275–283.

[LH94] J. LU AND D. M. HEALY, JR., *Contrast Enhancement via Multiscale Gradient Transformations*, in Intl. Conference on Image Processing (ICIP), Austin, TX, November 13-16 1994, IEEE, pp. 482–486 (vol. II).

[Luc93] S. J. LUCK, *The Role of Selective Attention in the Perception of Multiple-Element Visual Arrays: Cognitive and Neural Mechanisms*, Ph.D. dissertation, University of California, San Diego, CA, 1993.

[LWL95] J. S. LUND, Q. WU, AND J. B. LEVITT, *Visual Cortex Cell Types and Connections*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 1016–1021.

[Mal91] S. MALLAT, *Zero-Crossings of a Wavelet Transform*, IEEE Transactions on Information Theory, 37 (1991), pp. 1019–1033.

[MH91] S. MALLAT AND W. L. HWANG, *Singularity Detection and Processing with Wavelets*, tech. report, Courant Institute of Mathematical Sciences, New York University, New York, NY, March 1991.

[MH92] ——, *Singularity Detection and Processing with Wavelets*, IEEE Transactions on Information Theory, 38 (1992), pp. 617–643.

[MZ92a] S. MALLAT AND S. ZHONG, *Characterization of Signals from Multiscale Edges*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 14 (1992), pp. 710–732.

[MZ92b] S. MALLAT AND S. ZHONG, *Wavelet Transform Maxima and Multiscale Edges*, in Wavelets and Their Applications, M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., Jones and Bartlett, Boston, MA, 1992, pp. 67–104.

[Mal89a] S. G. MALLAT, *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 11 (1989), pp. 674–693.

[Mal89b] ——, *Multifrequency Channel Decompositions of Images and Wavelet Methods*, IEEE Transactions on Acoustics, Speech and Signal Processing, 37 (1989), pp. 2091–2110.

[Mar80] D. MARR, *Visual Information Processing: The Structure and Creation of Visual Representations*, Phil. Trans. R. Soc. Lond. B, 290 (1980), pp. 199–218.

[Mar82] D. MARR, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman and Company, New York, NY, 1982.

[MA95] P. MAZZONI AND R. A. ANDERSEN, *Gaze Coding in the Posterior Parietal Cortex*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 423–426.

[MBD96] B. H. MCCORMICK, D. A. BATTE, AND A. T. DUCHOWSKI, *A Virtual Environment: Exploring the Brain Forest*. Presentation at the CENAC Conference, October 1996, Mexico City, Mexico (preprint available from principal author, Department of Computer Science, Texas A&M University, College Station, TX).

[McC97] D. MCCUTCHEN, *A Dodecahedral Approach to Immersive Imaging and Display*, Computer Graphics, 31 (1997), pp. 35–37.

[Mey93] Y. MEYER, *Wavelets: Algorithms & Applications*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1993. Translated and revised by Robert D. Ryan.

[Mil56] G. A. MILLER, *The Magical Number Seven, Plus or Minus Two: Some Limits On Our Capacity For Processing Information*, Psychological Review, 63 (1956), pp. 81–97.

[Mor80] M. J. MORGAN, *Analogue Models of Motion Perception*, Phil. Trans. R. Soc. Lond. B, 290 (1980), pp. 117–135.

[Mum95] D. MUMFORD, *Thalamus*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 981–984.

[Nas94] G. P. NASON, *Wavelet Regression by Cross-Validation*, tech. report, Dept. of Math., Univ. of Bristol, University Walk, Bristol, BS8 1TW, U.K., Mar. 24, 1994.

[Nel95] J. I. NELSON, *Visual Scene Perception: Neurophysiology*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 1024–1028.

[NLO94] E. NGUYEN, C. LABIT, AND J.-M. ODOBEZ, *A ROI Approach for Hybrid Image Sequence Coding*, in International Conference on Image Processing (ICIP)'94, IEEE, November 1994, pp. 245–249.

[NS71a] D. NOTON AND L. STARK, *Eye Movements and Visual Perception*, Scientific American, 224 (1971), pp. 34–43.

[NS71b] ——, *Scanpaths in Saccadic Eye Movements While Viewing and Recognizing Patterns*, Vision Research, 11 (1971), pp. 929–942.

[OK95] B. A. OLSHAUSEN AND C. KOCH, *Selective Visual Attention*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 837–840.

[Ore92] K. J. O'REGAN, *Optimal Viewing Position in Words and the Strategy-Tactics Theory of Eye Movements in Reading*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 333–355. Springer Series in Neuropsychology.

[PH95] T. N. PAPPAS AND R. O. HINDS, *On Video and Audio Data Integration for Conferencing*, in Human

Vision, Visual Processing, and Digital Display VI, San Jose, CA, February 1995, SPIE, pp. 120–127.

[PW92] W. PÖLZLEITNER AND H. WECHSLER, *Selective and Robust Perception Using Multiresolution Estimation Techniques*, in 11th International Conference on Pattern Recognition, vol. II: Conference B, IEEE, 1992, pp. 54–57.

[PSD80] M. I. POSNER, C. R. R. SNYDER, AND B. J. DAVIDSON, *Attention and the Detection of Signals*, Experimental Psychology: General, 109 (1980), pp. 160–174.

[PS95] A. POUGET AND T. J. SEJNOWSKI, *Dynamic Remapping*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., The MIT Press, Cambridge, MA, 1995, pp. 335–338.

[PS85] F. P. PREPARATA AND M. I. SHAMOS, *Computational Geometry: An Introduction*, Springer-Verlag, New York, NY, 1985.

[PTVF92] W. H. PRESS, S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P. FLANNERY, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 2nd ed., 1992.

[RR93] W. ROBINETT AND J. P. ROLLAND, *A Computational Model for the Stereoscopic Optics of a Head-Mounted Display*, in Virtual Reality, R. A. Earnshaw, M. A. Gigante, and H. Jones, eds., Academic Press, London, 1993, ch. 5, pp. 51–75.

[Rob68] D. A. ROBINSON, *The Oculomotor Control System: A Review*, Proceedings of the IEEE, 56 (1968), pp. 1032–1049.

[RS79] E. A. ROBINSON AND M. T. SILVIA, *Digital Foundations of Time Series Analysis: The Box-Jenkins Approach*, vol. 1 of Holden-Day Series in Time Series Analysis and Digital Signal Processing, Holden-Day, Inc., San Francisco, CA, 1979.

[RBC+92] M. B. RUSKAI, G. BEYLKIN, R. COIFMAN, I. DAUBECHIES, S. MALLAT, Y. MEYER, AND L. RAPHAEL, eds., *Wavelets and Their Applications*, Jones and Bartlett, Boston, MA, 1992.

[Sa95] Z. SA, *Human Smooth Pursuit Tracking Eye Movement Analysis in the Frequency Domain*, M. S. thesis, Texas A&M University, College Station, TX, May 1995.

[San90] P. A. SANDON, *Simulating Visual Attention*, Journal of Cognitive Neuroscience, 2 (1990), pp. 213–231.

[Sch89] R. J. SCHALKOFF, *Digital Image Processing and Computer Vision*, John Wiley & Sons, Inc., New York, NY, 1989.

[Ser92] A. B. SERENO, *Programming Saccades: The Role of Attention*, in Eye Movements and Visual Cognition: Scene Perception and Reading, K. Rayner, ed., Springer-Verlag, New York, NY, 1992, pp. 89–107. Springer Series in Neuropsychology.

[She93] W. L. SHEBILSKE, *Visuomotor Modularity, Ontogeny and Training High-Performance Skills with Spatial Instruments*, in Pictorial Communication in Virtual and Real Environments, S. R. Ellis, M. Kaiser, and A. J. Grunwald, eds., Taylor & Francis, Ltd., London, 1993, ch. 19, pp. 305–315.

[SF83] W. L. SHEBILSKE AND D. F. FISHER, *Understanding Extended Discourse Through the Eyes: How and Why*, in Eye Movements and Psychological Functions: International Views, R. Groner, C. Menz, D. F. Fisher, and R. A. Monty, eds., Lawrence Erlbaum Associates, Hillsdale, NJ, 1983, pp. 303–314.

[SGI95] SILICON GRAPHICS INC. (SGI), *IRIS Media Libraries™ Programming Guide*, Mountain View, CA, March 1995. Document Number 007-1799-030.

[SM90] D. L. SPARKS AND L. E. MAYS, *Signal Transformations Required for the Generation of Saccadic Eye Movements*, Annual Review of Neuroscience, 13 (1990), pp. 309–336.

[ST94] L. B. STELMACH AND W. J. TAM, *Processing Image Sequences Based on Eye Movements*, in Conference on Human Vision, Visual Processing, and Digital Display V, San Jose, CA, February 8-10 1994, SPIE, pp. 90–98.

[SA93] G. R. STONER AND T. D. ALBRIGHT, *Image Segmentation Cues in Motion Processing: Implications for Modularity in Vision*, Journal of Cognitive Neuroscience, 5 (1993), pp. 129–149.

[TK80] S. TANIMOTO AND A. KLINGER, eds., *Structured Computer Vision*, Academic Press, New York, NY, 1980.

[TK93] S. TODD AND A. F. KRAMER, *Attentional Guidance in Visual Attention*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 1378–1382.

[Tre86] A. TREISMAN, *Features and Objects in Visual Processing*, Scientific American, 255 (1986), pp. 114B–125,140.

[TG80] A. TREISMAN AND G. GELADE, *A Feature Integration Theory of Attention*, Cognitive Psychology, 12 (1980), pp. 97–136.

[Tso90] J. K. TSOTSOS, *Analyzing Vision at the Complexity Level*, Behavioral and Brain Sciences, 13 (1990), pp. 423–469.

[Tso95] ——, *Toward a Computational Model of Visual Attention*, in Early Vision and Beyond, T. V. Papathomas, C. Chubb, A. Gorea, and E. Kowler, eds., The MIT Press, Cambridge, MA, 1995, pp. 207–218.

[UL95] M. UEDA AND S. LODHA, *Wavelets: An Elementary Introduction and Examples*, tech. report, Baskin Center for Computer Engineering & Information Sciences University of California, Santa Cruz, CA, January 17 1995. UCSC-CRL 94-47.

[Vai93] P. P. VAIDYANATHAN, ed., *Multirate Systems and Filter Banks*, P T R Prentice-Hall, Inc., Englewood Cliffs, NJ, 1993.

[Van92] A. H. C. VAN DER HEIJDEN, *Selective Attention in Vision*, Routledge, London, 1992.

[VOD93] K. F. VAN ORDEN AND J. DIVITA, *Highlighting with Flicker*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 1300–1304.

[VonH25] H. VON HELMHOLTZ, *Handbuch der Physiologischen Optik (Treatise on Physiological Optics)*,

vol. III, The Optical Society of America, Rochester, NY, Translated from the Third German ed., 1925.

[Wal91] G. K. WALLACE, *The JPEG Still Picture Compression Standard*, Communications of the ACM, 34 (1991), pp. 30–45.

[WW92] A. WATT AND M. WATT, *Advanced Animation and Rendering Techniques*, Addison-Wesley, Reading, MA, 1992.

[Wei90] W. W. S. WEI, *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley Publishing Company, Inc., Redwood City, CA, 1990.

[Wil83] L. WILLIAMS, *Pyramidal Parametrics*, Computer Graphics, 17 (1983), pp. 1–11.

[WNS84] J. M. WINTERS, M. H. NAM, AND L. W. STARK, *Modeling Dynamical Interactions Between Fast and Slow Movements: Fast Saccadic Eye Movement Behavior in the Presence of the Slower VOR*, Mathematical Biosciences, 68 (1984), pp. 159–185.

[Wol93] J. M. WOLFE, *Guided Search 2.0: The Upgrade*, in Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting, Santa Monica, CA, October 1993, pp. 1295–1299.

[Wol94] ——, *Visual Search in Continuous, Naturalistic Stimuli*, Vision Research, 34 (1994), pp. 1187–1195.

[Won90] W. C. WONG, *Nonlinear Analysis of Human Optokinetic (Smooth Pursuit Tracking) Responses*, Ph.D. dissertation, Texas A&M University, College Station, TX, December 1990.

[Yar67] A. L. YARBUS, *Eye Movements and Vision*, Plenum Press, New York, NY, 1967.

[ZOCR+76] D. S. ZEE, L. M. OPTICAN, J. D. COOK, D. A. ROBINSON, AND W. K. ENGEL, *Slow Saccades in Spinocerebellar Degeneration*, Archives of Neurology, 33 (1976), pp. 243–251.

[Zek93] S. ZEKI, *A Vision of the Brain*, Blackwell Scientific Publications, Osney Mead, Oxford, 1993.

APPENDIX A

BI-ORTHOGONAL WAVELET FILTER COEFFICIENTS

Filter coefficients of the Barlaud, Burt and Adelson, Mallat, and Chui (multiplicity-2), bi-orthogonal filters are given in Tables 14, 15, 16, and 17 (rounded to fit the table width). Due to the symmetry of the sequences, only half of the Chui spline wavelet coefficients are shown in Table 17, where $h_k = h_{2-k}$, $g_k = g_{4-k}$, and similarly for the duals.

TABLE 14
Barlaud's near-orthonormal spline filters.

k	h_k	g_k	$\widetilde{h}_k$	$\widetilde{g}_k$
-4		0.0378285	0.0378285	
-3	-0.0645389	0.0238495	-0.0238495	0.0645389
-2	-0.4068942	-0.1106244	-0.1106244	-0.4068942
-1	0.4180922	-0.3774029	0.3774029	-0.4180923
0	0.7884856	0.8526987	0.8526987	0.7884856
1	0.4180923	-0.3774029	0.3774029	-0.4180923
2	-0.4068942	-0.1106244	-0.1106244	-0.4068942
3	-0.0645389	0.0238495	-0.0238495	0.0645389
4		0.0378285	0.0378285	

TABLE 15
Burt and Adelson's Laplacian pyramid filters.

k	h_k	g_k	$\widetilde{h}_k$	$\widetilde{g}_k$
-3		0.0151523	-0.0151523	
-2	-0.0707107	-0.0757614	-0.0757614	-0.0707107
-1	0.3535534	-0.3687057	0.3687057	-0.3535534
0	0.8485281	0.8586297	0.8586297	0.8485281
1	0.3535534	-0.3687057	0.3687057	-0.3535534
2	-0.0707107	-0.0757614	-0.0757614	-0.0707107
3		0.0151523	-0.0151523	

TABLE 16
Mallat's quadratic spline filters.

k	h_k	g_k	$\widetilde{h}_k$	$\widetilde{g}_k$
-3			0.0078125	0.0078125
-2			0.0546850	0.0468750
-1	0.125		0.1718750	0.1171875
0	0.375	-2.0	-0.1718750	0.6562500
1	0.375	2.0	-0.0546850	0.1171875
2	0.125		-0.0078125	0.0468750
3				0.0078125

TABLE 17
Chui's (multiplicity-2) cardinal spline filters.

k	h_k, h_{2-k}	g_{k+1}, g_{3-k}	$\widetilde{h}_k, \widetilde{h}_{2-k}$	$\widetilde{g}_{k+1}, \widetilde{g}_{3-k}$
1	0.683012701892	0.866025403784	1.000000	0.8333333
2	0.316987298108	-0.316987298108	0.500000	-0.5000000
3	-0.116025403784	-0.232050807569		0.0833333
4	-0.084936490539	0.084936490539		
5	0.031088913246	0.062177826491		
6	0.022758664048	-0.022758664047		
7	-0.008330249198	-0.016660498395		
8	-0.006098165652	0.006098165652		
9	0.002232083545	0.004464167091		
10	0.001633998562	-0.001633998561		
11	-0.000598084983	-0.001196169967		
12	-0.000437828595	0.000437828595		
13	0.000160256388	0.000320512777		
14	0.000117315818	-0.000117315818		
15	-0.000042940569	-0.000085881139		
16	-0.000031434679	0.000031434678		
17	0.000011505891	0.000023011782		
18	0.000008422897	-0.000008422897		
19	-0.000003082990	-0.000006165980		
20	-0.000002256905	0.000002256905		
21	0.000000826079	0.000001652159		

APPENDIX B

MATRIX TENSOR PRODUCTS

Given two matrices $\mathbf{A}, \mathbf{B}$ with dimensions $J \times K$ and $L \times M$, respectively, the tensor (or Kronecker, or *direct*) product of the matrices, denoted by $\mathbf{A} \otimes \mathbf{B}$, is defined as

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1K}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2K}\mathbf{B} \\ \vdots & & & \\ a_{J1}\mathbf{B} & a_{J2}\mathbf{B} & \cdots & a_{JK}\mathbf{B} \end{bmatrix}$$

where

$$a_{jk}\mathbf{B} = \begin{bmatrix} a_{jk}b_{11} & a_{jk}b_{12} & \cdots & a_{jk}b_{1M} \\ a_{jk}b_{21} & a_{jk}b_{22} & \cdots & a_{jk}b_{2M} \\ \vdots & & & \\ a_{jk}b_{L1} & a_{jk}b_{L2} & \cdots & a_{jk}b_{LM} \end{bmatrix},$$

giving $\mathbf{C}$ dimension $JL \times KM$. An element of $\mathbf{C}$ in terms of the elements of $\mathbf{A}$ and $\mathbf{B}$ is written as $\mathbf{C} \equiv [c_{jl,km}]$ where row and columns of $\mathbf{C}$ are denoted by dual symbols (jl), (km), respectively, such that $c_{jl,km} = a_{jk}b_{lm}$ [Jos95, p.167]. Relabeling $[c_{jl,km}]$ with two new indices p, q, $1 \le p \le P$, $1 \le q \le Q$, so that $\mathbf{C} \equiv [c_{pq}] = [c_{jl,km}]$, the indices p, q are given by:

$$p = (j-1)L + l, \text{ and, } q \quad = \quad (k-1)M + m.$$

For example, given

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} r & s & t \\ x & y & z \end{bmatrix},$$

$\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$ is a 6×6 matrix:

$$\mathbf{C} = \left[\begin{array}{ccc|ccc} ar & as & at & br & bs & bt \\ ax & ay & az & bx & by & bz \\ \hline cr & cs & ct & dr & ds & dt \\ cx & cy & cz & dx & dy & dz \\ \hline er & es & et & fr & fs & ft \\ ex & ey & ez & fx & fy & fz \end{array}\right].$$

The tensor product is associative,

$$(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C},$$

and distributive with respect to addition:

$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}.$$

For further details, see [Jos95].

APPENDIX C

EXPERIMENT 1 SUPPLEMENTARY MATERIAL

C.1 Experiment Approval and Consent

Experiment 1 (initially entitled “Eye Movement Modeling”) has been reviewed and approved by the Institutional Review Board (IRB)–Human Subjects in Research, Texas A&M University. The official approval form is shown in a digitized reproduction in Figure 99. All subjects signed and received an Informed Consent Form, as approved by the IRB. A blank example form is shown in a digitized reproduction in Figure 100.

C.2 Verification of Eye Tracker Slippage

Analysis of variance of eye tracker slippage is given in Table 18.

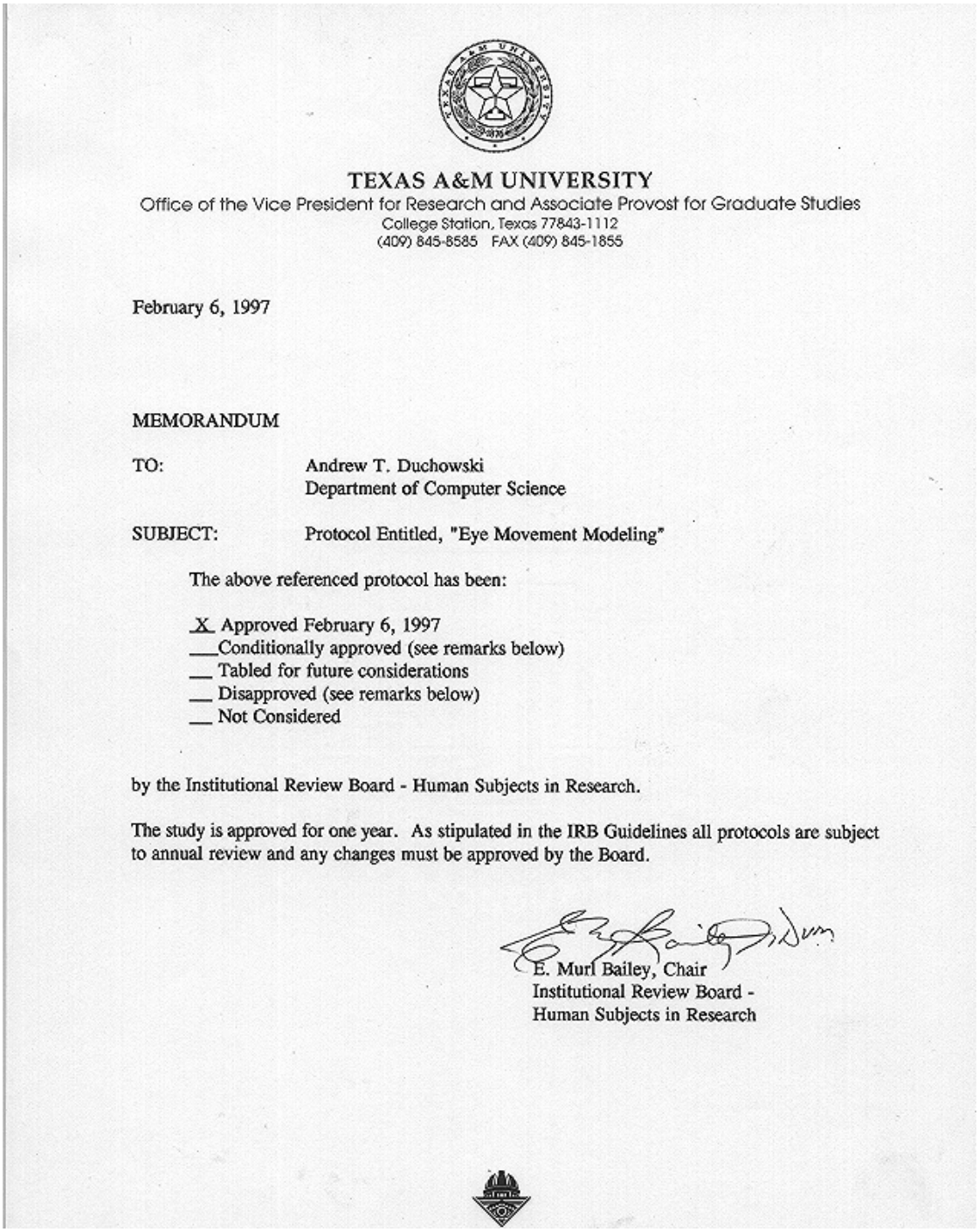
TABLE 18
Pre- vs. post-stimulus viewing average calibration error one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.5976	1	0.5976	0.3825
Error	62.5	40	1.563	
Total	63.1	41		

p = 0.5398

C.3 Evaluation of PARIMA Model of Eye Movements

Hit and correctness rates for detected saccades in observed eye movements are presented in Tables 19–21. Symbols 1 and 0 denote the positive and negative identification of a saccade, respectively.



TEXAS A&M UNIVERSITY
Office of the Vice President for Research and Associate Provost for Graduate Studies
College Station, Texas 77843-1112
(409) 845-8585 FAX (409) 845-1855

February 6, 1997

MEMORANDUM

TO: Andrew T. Duchowski
Department of Computer Science

SUBJECT: Protocol Entitled, "Eye Movement Modeling"

The above referenced protocol has been:

X Approved February 6, 1997
___Conditionally approved (see remarks below)
___ Tabled for future considerations
___ Disapproved (see remarks below)
___ Not Considered

by the Institutional Review Board - Human Subjects in Research.

The study is approved for one year. As stipulated in the IRB Guidelines all protocols are subject to annual review and any changes must be approved by the Board.

E. Murl Bailey, Chair
Institutional Review Board -
Human Subjects in Research

Fig. 99. Experiment 1 approval.

INFORMED CONSENT FORM

SUBJ. ID E 1 2 3 –
(to be filled in by experimenter)

Please read the following information pertaining to the experiment. If you choose to participate, please sign and date this document.

1. **Purpose of Study**
   The objective of the current research is modeling eye movements. That is, the experimenters are seeking a mathematical description of the way the eyes move. All experiments are conducted in the Virtual Environments Laboratory of the Computer Science Department (H. R. Bright Building, room 322).
2. **Number of Subjects**
   Approximately 30 subjects are requested to participate in this study.
3. **Experimental Procedure**
   I will be asked to view visual stimulus (3 short video clips, 8 second duration) while my eye movements are being monitored. Eye movements are recorded with the use of a video eye tracker which works by shining a low-intensity infra-red light at the eye so that the camera may be able to track the location of the pupil. The infra-red light poses no danger.

   I will not need to wear anything, ingest anything, nor provide any fluid samples. However, during calibration and stimulus viewing, **I must try to keep my head perfectly still**. This is necessary so that the eye tracker does not lose its focus on my eye. To prevent fatigue, 7-minute breaks will be given between video sequences. The entire experiment should take about half an hour.
4. **Compensation and Benefits of Participation**
   I will receive neither compsenation nor benefits from this study except for 1 course credit as designated by the Psychology Department. If I choose to withdraw from the study I will not receive the credit. I may still earn the course credit by participating in other psychology experiments, or by writing a short paper, or whatever is required by the Psychology Department (please consult the Department for further explanation).
5. **Questionnaire**
   I will be asked to fill out a short questionnaire during the study. I may refuse to answer any question that makes me uncomfortable without withdrawing from the study.
6. **Experiment Anonymity**
   If I agree to participate in this study, all recorded data will be kept anonymous. My eye movement data will be encoded by *subject number*. My name will not appear in any reports or papers resulting from this research. A separate videotape release form is provided in case the experiment is to be recorded on videotape. If I choose to withdraw from the study, no record of my participation will be kept. I may withdraw at any time during the experiment.

This research study has been reviewed and approved by the Institutional Review Board–Human Subjects in Research, Texas A&M University. For research-related problems or questions regarding subjects' rights, the Institutional Review Board may be contacted through Dr. Richard E. Miller, IRB Coordinator, Office of Vice President for Research and Associate Provost for Graduate Studies at (409) 845-1811.

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. I have been given a copy of this consent form.

Subject_____Date_____

Principal Investigator/or Authorized Representative_____Date_____
If you have any further questions concerning this research study, please contact Andrew Duchowski (H. R. Bright office 414A) at 845-9980, 775-0986, or by email at *andrewd@cs.tamu.edu*.

Fig. 100. Experiment 1 Informed Consent Form.

TABLE 19

Experiment 1 saccade detection statistics over sequence *sim1*.

Expected saccade location (inter-frame location)	Subject #							Per--subject hit rate	Saccade Amplitude (degree visual angle)
	21	22	23	26	27	28	29		
003-004	0	1	1	1	0	1	1	71	2.78
010-011	1	0	1	1	1	1	1	86	2.12
017-018	1	1	1	1	1	1	0	86	5.38
023-024	1	1	1	1	1	1	0	86	4.20
030-031	1	1	1	1	1	1	1	100	1.34
035-036	1	1	1	0	1	0	0	57	4.78
042-043	1	0	1	0	1	0	0	43	6.27
054-055	0	0	0	0	0	0	1	14	4.54
066-067	0	1	1	0	1	1	0	57	6.42
099-100	1	1	1	1	1	1	1	100	4.83
110-111	0	0	1	0	1	0	1	43	2.62
117-118	0	1	0	0	1	0	0	29	3.02
123-124	1	1	0	0	0	1	0	43	1.76
124-125	1	1	0	0	0	0	0	29	1.78
Total expected	Total found								
14	27	20	21	19	23	30	21		
	Percent correct over total expected								
	64	71	71	43	71	57	43		
	Percent correct over total found								
	33	50	48	32	43	27	29		

TABLE 20

Experiment 1 saccade detection statistics over sequence *sim2*.

Expected saccade location (inter-frame location)	Subject #							Per--subject hit rate	Saccade Amplitude (degree visual angle)
	21	22	23	26	27	28	29		
041-042	0	0	0	0	0	0	0	0	2.84
046-047	1	0	0	0	0	1	0	29	2.49
053-054	0	0	1	0	0	0	0	14	2.34
072-073	0	1	0	1	0	1	1	57	2.72
078-079	0	1	0	0	1	0	0	29	2.42
084-085	0	0	1	1	0	0	0	29	4.40
090-091	0	1	1	1	1	1	1	86	4.79
097-098	1	0	1	0	1	0	0	43	1.91
099-100	0	0	1	0	1	0	0	29	4.70
113-114	1	0	0	1	0	1	0	43	6.59
119-120	1	0	1	0	1	0	1	57	6.68
Total expected	Total found								
11	18	18	28	22	20	22	22		
	Percent correct over total expected								
	36	27	55	36	45	36	27		
	Percent correct over total found								
	22	17	21	18	25	18	14		

TABLE 21
Experiment 1 saccade detection statistics over sequence *sim3*.

Expected saccade location (inter-frame location)	Subject #							Per--subject hit rate	Saccade Amplitude (degree visual angle)
	21	22	23	26	27	28	29		
002-003	1	0	1	0	0	0	0	29	1.37
003-004	1	1	0	0	0	0	0	29	1.51
009-010	0	1	0	1	0	0	1	43	0.99
015-016	0	0	1	0	1	0	0	29	2.96
021-022	1	1	0	0	0	0	1	43	1.73
022-023	1	1	0	0	0	0	0	29	2.10
037-038	1	1	1	1	1	0	1	86	3.81
042-043	0	1	0	0	0	0	0	14	1.97
074-075	1	0	1	0	1	1	1	71	8.14
080-081	1	0	0	1	0	0	0	29	3.56
085-086	0	0	0	1	1	0	0	29	0.97
091-092	0	0	0	1	0	1	0	29	3.60
097-098	0	1	0	0	0	0	0	14	4.09
104-105	0	0	0	0	1	0	0	14	12.33
111-112	0	1	1	1	0	1	0	57	4.21
117-118	0	1	1	1	0	0	1	57	2.82
123-124	0	1	0	0	1	1	1	57	2.68
Total expected	Total found								
17	20	22	21	25	27	21	25		
	Percent correct over total expected								
	41	59	35	41	35	24	35		
	Percent correct over total found								
	32	59	29	28	22	19	24		

APPENDIX D

EXPERIMENT 2 SUPPLEMENTARY MATERIAL

D.1 Experiment Approval and Consent

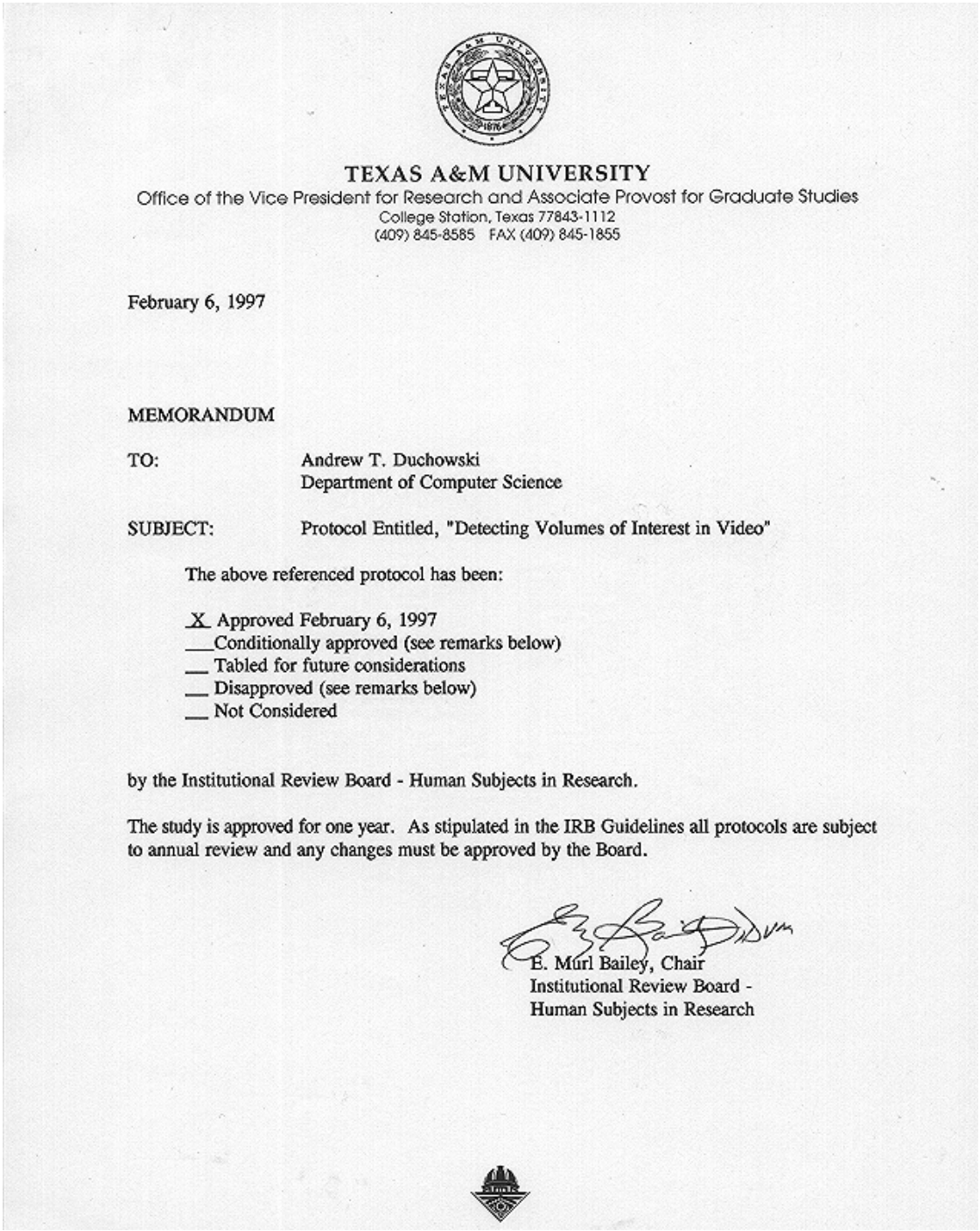
Experiment 2 (initially entitled “Detecting Volumes of Interest in Video”) has been reviewed and approved by the Institutional Review Board (IRB)–Human Subjects in Research, Texas A&M University. The official approval form is shown in a digitized reproduction in Figure 101. All subjects signed and received an Informed Consent Form, as approved by the IRB. A blank example form is shown in a digitized reproduction in Figure 102.

D.2 Verification of Eye Tracker Slippage

Analysis of variance of eye tracker slippage is given in Table 22.

TABLE 22
Pre- vs. post-stimulus viewing average calibration error one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.1488	1	0.1488	0.2845
Error	4.185	8	0.5231	
Total	4.334	9		
p = 0.6082				



TEXAS A&M UNIVERSITY
Office of the Vice President for Research and Associate Provost for Graduate Studies
College Station, Texas 77843-1112
(409) 845-8585 FAX (409) 845-1855

February 6, 1997

MEMORANDUM

TO: Andrew T. Duchowski
Department of Computer Science

SUBJECT: Protocol Entitled, "Detecting Volumes of Interest in Video"

The above referenced protocol has been:

X Approved February 6, 1997
___Conditionally approved (see remarks below)
___ Tabled for future considerations
___ Disapproved (see remarks below)
___ Not Considered

by the Institutional Review Board - Human Subjects in Research.

The study is approved for one year. As stipulated in the IRB Guidelines all protocols are subject to annual review and any changes must be approved by the Board.

E. Murl Bailey, Chair
Institutional Review Board -
Human Subjects in Research

Fig. 101. Experiment 2 approval.

INFORMED CONSENT FORM

SUBJ. ID E 1 2 3 –
(to be filled in by experimenter)

Please read the following information pertaining to the experiment. If you choose to participate, please sign and date this document.

1. **Purpose of Study**
   The objective of the current research is the detection of interesting objects in video. That is, pin-point locations of my gaze is sought when I view a short video sequence. All experiments are conducted in the Virtual Environments Laboratory of the Computer Science Department (H. R. Bright Building, room 322).
2. **Number of Subjects**
   Approximately 30 subjects are requested to participate in this study.
3. **Experimental Procedure**
   I will be asked to view visual stimulus (3 short video clips, 8 second duration) while my eye movements are being monitored. Eye movements are recorded with the use of a video eye tracker which works by shining a low-intensity infra-red light at the eye so that the camera may be able to track the location of the pupil. The infra-red light poses no danger.

   I will not need to wear anything, ingest anything, nor provide any fluid samples. However, during calibration and stimulus viewing, **I must try to keep my head perfectly still**. This is necessary so that the eye tracker does not lose its focus on my eye. To prevent fatigue, 7-minute breaks will be given between video sequences. The entire experiment should take about half an hour.
4. **Compensation and Benefits of Participation**
   I will receive neither compsenation nor benefits from this study except for 1 course credit as designated by the Psychology Department. If I choose to withdraw from the study I will not receive the credit. I may still earn the course credit by participating in other psychology experiments, or by writing a short paper, or whatever is required by the Psychology Department (please consult the Department for further explanation).
5. **Questionnaire**
   I will be asked to fill out a short questionnaire during the study. I may refuse to answer any question that makes me uncomfortable without withdrawing from the study.
6. **Experiment Anonymity**
   If I agree to participate in this study, all recorded data will be kept anonymous. My eye movement data will be encoded by *subject number*. My name will not appear in any reports or papers resulting from this research. A separate videotape release form is provided in case the experiment is to be recorded on videotape. If I choose to withdraw from the study, no record of my participation will be kept. I may withdraw at any time during the experiment.

This research study has been reviewed and approved by the Institutional Review Board–Human Subjects in Research, Texas A&M University. For research-related problems or questions regarding subjects' rights, the Institutional Review Board may be contacted through Dr. Richard E. Miller, IRB Coordinator, Office of Vice President for Research and Associate Provost for Graduate Studies at (409) 845-1811.

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. I have been given a copy of this consent form.

Subject_____Date_____

Principal Investigator/or Authorized Representative_____Date_____
If you have any further questions concerning this research study, please contact Andrew Duchowski (H. R. Bright office 414A) at 845-9980, 775-0986, or by email at *andrewd@cs.tamu.edu*.

Fig. 102. Experiment 2 Informed Consent Form.

APPENDIX E

EXPERIMENT 3 SUPPLEMENTARY MATERIAL

E.1 Experiment Approval and Consent

Experiment 3 (initially entitled “Gaze Contingent Video Processing”) has been reviewed and approved by the Institutional Review Board (IRB)–Human Subjects in Research, Texas A&M University. The official approval form is shown in a digitized reproduction in Figure 103. All subjects signed and received an Informed Consent Form, as approved by the IRB. A blank example form is shown in a digitized reproduction in Figure 104.

E.2 Verification of Eye Tracker Slippage

Analysis of variance of eye tracker slippage is given in Table 23.

TABLE 23
Pre- vs. post-stimulus viewing average calibration error one-way ANOVA.

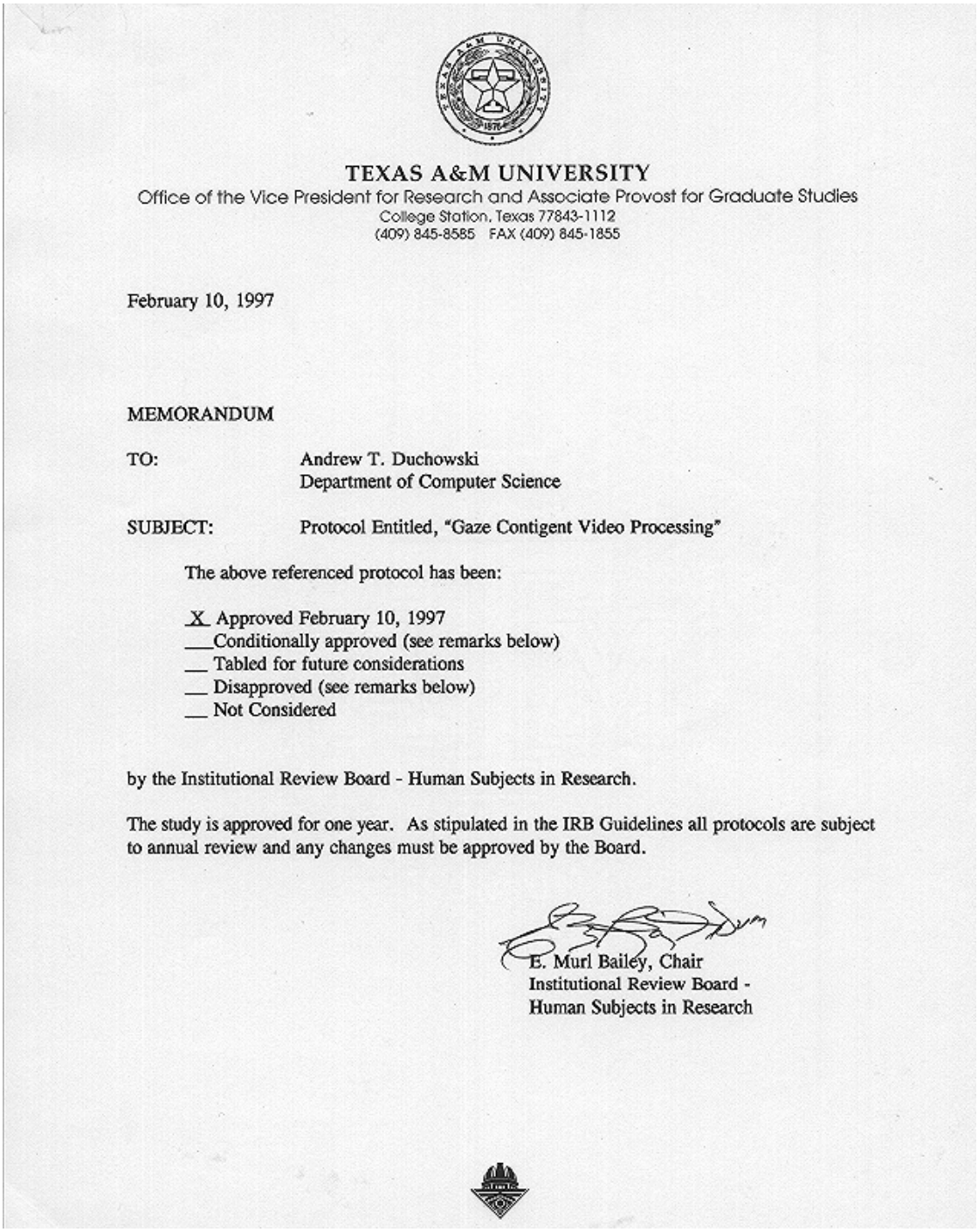
ANOVA TABLE				
Source	SS	df	MS	F
Columns	7.392	1	7.392	1.063
Error	639.9	92	6.955	
Total	647.3	93		
p = 0.3053				

E.3 Verification of Gaze Position

Overall median gaze position error for different viewing conditions is given in Table 24. Two-way ANOVA

TABLE 24
Experiment 3 gaze error.

Video sequence	Median of medians (deg. visual angle)	Median of iqr variance
flight (ideal)	2.09	1.94
flight (preat)	1.36	1.32
brain2 (ideal)	0.55	0.35
cnn (agg)	2.98	1.63



TEXAS A&M UNIVERSITY
Office of the Vice President for Research and Associate Provost for Graduate Studies
College Station, Texas 77843-1112
(409) 845-8585 FAX (409) 845-1855

February 10, 1997

MEMORANDUM

TO: Andrew T. Duchowski
Department of Computer Science

SUBJECT: Protocol Entitled, "Gaze Contigent Video Processing"

The above referenced protocol has been:

_X_ Approved February 10, 1997
___Conditionally approved (see remarks below)
___ Tabled for future considerations
___ Disapproved (see remarks below)
___ Not Considered

by the Institutional Review Board - Human Subjects in Research.

The study is approved for one year. As stipulated in the IRB Guidelines all protocols are subject to annual review and any changes must be approved by the Board.

E. Murl Bailey, Chair
Institutional Review Board -
Human Subjects in Research

Fig. 103. Experiment 3 approval.

INFORMED CONSENT FORM

SUBJ. ID E 1 2 3 –
(to be filled in by experimenter)

Please read the following information pertaining to the experiment. If you choose to participate, please sign and date this document.

1. **Purpose of Study**
   The objective of the current research is just-perceptible processing video. That is, a method of video processing is sought where the results of the method is imperceptible. All experiments are conducted in the Virtual Environments Laboratory of the Computer Science Department (H. R. Bright Building, room 322).
2. **Number of Subjects**
   Approximately 30 subjects are requested to participate in this study.
3. **Experimental Procedure**
   I will be asked to view visual stimulus (3 short video clips, 8 second duration) while my eye movements are being monitored. Eye movements are recorded with the use of a video eye tracker which works by shining a low-intensity infra-red light at the eye so that the camera may be able to track the location of the pupil. The infra-red light poses no danger. Before displaying the video sequence I will be told to look for a particular object in the video and follow it with my eyes. Three such sequences will be shown. After viewing all sequences I will be asked to judge each sequence with respect to its quality.

   I will not need to wear anything, ingest anything, nor provide any fluid samples. However, during calibration and stimulus viewing, **I must try to keep my head perfectly still**. This is necessary so that the eye tracker does not lose its focus on my eye. To prevent fatigue, 7-minute breaks will be given between video sequences. The entire experiment should take about half an hour.
4. **Compensation and Benefits of Participation**
   I will receive neither compsenation nor benefits from this study except for 1 course credit as designated by the Psychology Department. If I choose to withdraw from the study I will not receive the credit. I may still earn the course credit by participating in other psychology experiments, or by writing a short paper, or whatever is required by the Psychology Department (please consult the Department for further explanation).
5. **Questionnaire**
   I will be asked to fill out a short questionnaire during the study. I may refuse to answer any question that makes me uncomfortable without withdrawing from the study.
6. **Experiment Anonymity**
   If I agree to participate in this study, all recorded data will be kept anonymous. My eye movement data will be encoded by *subject number*. My name will not appear in any reports or papers resulting from this research. A separate videotape release form is provided in case the experiment is to be recorded on videotape. If I choose to withdraw from the study, no record of my participation will be kept. I may withdraw at any time during the experiment.

This research study has been reviewed and approved by the Institutional Review Board–Human Subjects in Research, Texas A&M University. For research-related problems or questions regarding subjects' rights, the Institutional Review Board may be contacted through Dr. Richard E. Miller, IRB Coordinator, Office of Vice President for Research and Associate Provost for Graduate Studies at (409) 845-1811.

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. I have been given a copy of this consent form.

Subject_____Date_____

Principal Investigator/or Authorized Representative_____Date_____
If you have any further questions concerning this research study, please contact Andrew Duchowski (H. R. Bright office 414A) at 845-9980, 775-0986, or by email at *andrewd@cs.tamu.edu*.

Fig. 104. Experiment 3 Informed Consent Form.

of median data between viewing conditions is given in Table 25. One-way ANOVA of pairwise comparisons

TABLE 25
Two-way ANOVA of gaze error between viewing conditions.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	1.655	2	0.8273	0.3334
Rows	97.14	3	32.38	13.05
Interaction	10.34	6	1.724	0.695
Error	148.9	60	2.481	
Total	258.0	71		

p = 0.7178 0.0000 0.6546

of median means between resolution mappings are given in Tables 26, 27, and 28. One-way ANOVA of

TABLE 26
One-way ANOVA of gaze error (LIN mapping vs. HVS mapping).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	14.15	1	14.15	1.434
Error	533	54	9.87	
Total	547.1	55		

p = 0.2364

TABLE 27
One-way ANOVA of gaze error (LIN mapping vs. ORG mapping).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	14.84	1	14.84	1.609
Error	498.2	54	9.226	
Total	513	55		

p = 0.2101

pairwise comparisons of median means between viewing conditions are given in Tables 29, 30, 31, 32, 33, and 34.

TABLE 28
One-way ANOVA of gaze error (HVS mapping vs. ORG mapping).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.0083	1	0.0083	0.0005
Error	921.7	54	17.07	
Total	921.7	55		

p = 0.9825

TABLE 29
One-way ANOVA of gaze error (*flight* (ideal) vs. *flight* (preat)).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	99.92	1	99.92	0.0000
Error	682.0	40	17.05	
Total	781.9	41		

p = 0.0201

TABLE 30
One-way ANOVA of gaze error (*flight* (ideal) vs. *brain2* (ideal)).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	84.33	1	84.33	7.995
Error	358.6	34	10.55	
Total	443.0	35		

p = 0.0078

TABLE 31
One-way ANOVA of gaze error (*flight* (ideal) vs. *cnn* (agg)).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	13.32	1	13.32	0.735
Error	724.8	40	18.12	
Total	738.1	41		

p = 0.3964

TABLE 32
One-way ANOVA of gaze error (*flight* (preat) vs. *brain2* (ideal)).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	16.56	1	16.56	7.604
Error	74.07	34	2.179	
Total	90.64	35		

p = 0.0093

TABLE 33
One-way ANOVA of gaze error (*flight* (preat) vs. *cnn* (agg)).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	186.2	1	186.2	38.87
Error	191.6	40	4.79	
Total	377.8	41		
p = 2.2103e-07				

TABLE 34
One-way ANOVA of gaze error (*brain2* (ideal) vs. *cnn* (agg)).

ANOVA TABLE				
Source	SS	df	MS	F
Columns	281.6	1	281.6	81.92
Error	116.9	34	3.438	
Total	398.5	35		
p = 1.4064e-10				

E.4 Impairment Perception Analysis

Video sequence impairment ratings (5-point scale, 1 = imperceptible, 5 = very annoying) over all conditions are given in Table 35.

E.4.1 Impairment Perception Analysis Between Conditions

Overall two-way analysis of variance (video sequence & resolution mapping) is given in Table 36. Rows represent viewing conditions (ideal, preat, agg) and columns represent resolution mappings (LIN, HVS, ORG).

Pairwise two-way ANOVA between viewing conditions is compiled in Tables 37, 37, 38, 39, and 40.

E.4.2 Impairment Perception Analysis Within Conditions

One-way analysis of variance of perception impairment within viewing conditions (different resolution mappings) is given in Tables 41, 42, 43, and 44. Columns represent resolution mappings (LIN, HVS, ORG).

E.4.2.1 Impairment Perception Analysis Within the Aggregate VOI Viewing Condition

Pairwise one-way ANOVA between resolution mappings within the aggregate VOI condition is compiled in Tables 45, 46, and 47.

TABLE 35
Experiment 3 video sequence subjective ratings.

Video sequence	Exp.#-Subj.#	Sequence order	Scores		
			L	H	N
flight (ideal)					
	03-05	L,H,N	1	2	2
	03-06	L,H,N	2	2	2
	03-07	H,N,L	1	2	2
	03-08	L,H,N	2	2	2
flight (preat)					
	04-05	L,H,N	1	2	1
	04-07	N,L,H	2	2	1
	04-09	L,N,H	2	1	1
	04-11	L,N,H	1	2	3
brain2 (ideal)					
	03-11	L,N,H	2	1	2
	03-15	L,N,H	3	3	2
	03-18	H,N,L	2	2	2
	03-19	H,L,N	2	1	2
cnn (agg)					
	03-22	N,L,H	3	2	2
	03-23	L,N,H	4	3	2
	03-24	H,N,L	5	3	3
	03-25	N,L,H	4	3	2

TABLE 36
Overall two-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	1.625	2	0.8125	2.34
Rows	13.42	3	4.472	12.88
Interaction	5.708	6	0.9514	2.74
Error	12.5	36	0.3472	
Total	33.25	47		
p = 0.1180 0.0000 0.0269				

TABLE 37
Flight (ideal) vs. *flight* (preat) two-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.25	2	.125	0.4091
Rows	0.6667	1	.6667	2.182
Interaction	0.08333	2	.04167	0.1364
Error	5.5	18	.3056	
Total	6.5	23		
p = 0.6703 0.1569 0.8734				

TABLE 38
Flight (ideal) vs. *brain2* (ideal) two-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.08333	2	0.04167	0.1765
Rows	0.04167	1	0.04167	0.1765
Interaction	0.5833	2	0.2917	1.235
Error	4.25	18	0.2361	
Total	4.958	23		

p = 0.8397 0.6794 0.3143

TABLE 39
Flight (ideal) vs. *cnn* (agg) two-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	2.333	2	1.167	4.941
Rows	7.042	1	7.042	29.82
Interaction	4.333	2	2.167	9.176
Error	4.25	18	0.2361	
Total	17.96	23		

p = 0.0195 0.0000 0.0018

TABLE 40
Brain2 (ideal) vs. *cnn* (agg) two-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	4.75	2	2.375	6.107
Rows	6.0	1	6.0	15.43
Interaction	2.25	2	1.125	2.893
Error	7.0	18	0.3889	
Total	20.0	23		

p = 0.0095 0.0010 0.0814

TABLE 41
Flight (ideal) one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.1667	2	0.08333	1.0
Error	0.75	9	0.08333	
Total	0.9167	11		

p = 0.4053

TABLE 42
Flight (preat) one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.1667	2	0.08333	0.1579
Error	4.75	9	0.5278	
Total	4.9167	11		

p = 0.8563

TABLE 43
Brain2 (ideal) one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.5	2	0.25	0.6429
Error	3.5	9	0.3889	
Total	4.0	11		

p = 0.5483

TABLE 44
Cnn (agg) one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	6.5	2	3.25	8.357
Error	3.5	9	0.3889	
Total	10.0	11		

p = 0.0089

TABLE 45
Cnn (agg) LIN vs. HVS mapping one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	3.125	1	3.125	6.818
Error	2.75	6	0.4583	
Total	5.875	7		

p = 0.0401

TABLE 46
Cnn (agg) LIN vs. ORG mapping one-way ANOVA.

ANOVA TABLE				
Source	SS	df	MS	F
Columns	6.125	1	6.125	13.36
Error	2.75	6	0.4583	
Total	8.875	7		

p = 0.0106

TABLE 47
Cnn (agg) HVS vs. ORG mapping one-way ANOVA.

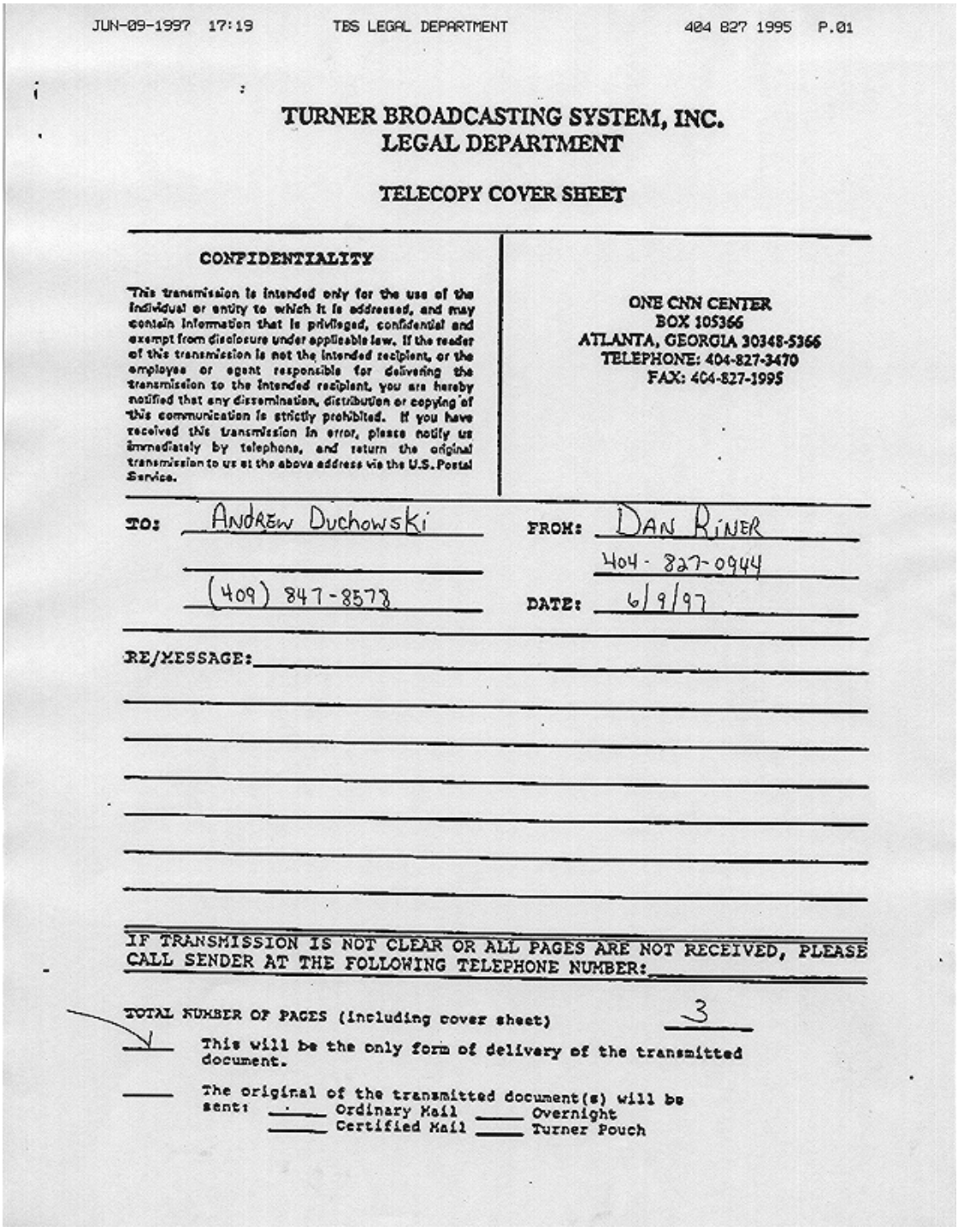
ANOVA TABLE				
Source	SS	df	MS	F
Columns	0.5	1	0.5	2.0
Error	1.5	6	0.25	
Total	2.0	7		

p = 0.2070

APPENDIX F

LETTER OF PERMISSION

Proof of permission for use of the *cnn* sequence was obtained from the Legal Department of Turner Broadcasting System, Inc. The official copyright permission letter (3 pages) is shown in a digitized reproduction in Figures 105– 107.



JUN-09-1997 17:19 TBS LEGAL DEPARTMENT 404 827 1995 P.01

TURNER BROADCASTING SYSTEM, INC.
LEGAL DEPARTMENT

TELECOPY COVER SHEET

CONFIDENTIALITY

This transmission is intended only for the use of the individual or entity to which it is addressed, and may contain information that is privileged, confidential and exempt from disclosure under applicable law. If the reader of this transmission is not the intended recipient, or the employee or agent responsible for delivering the transmission to the intended recipient, you are hereby notified that any dissemination, distribution or copying of this communication is strictly prohibited. If you have received this transmission in error, please notify us immediately by telephone, and return the original transmission to us at the above address via the U.S. Postal Service.

ONE CNN CENTER
BOX 105366
ATLANTA, GEORGIA 30348-5366
TELEPHONE: 404-827-3470
FAX: 404-827-1995

TO: Andrew Duchowski
(409) 847-8578

FROM: Dan Riner
404 - 827-0944

DATE: 6/9/97

RE/MESSAGE:

IF TRANSMISSION IS NOT CLEAR OR ALL PAGES ARE NOT RECEIVED, PLEASE CALL SENDER AT THE FOLLOWING TELEPHONE NUMBER:

TOTAL NUMBER OF PAGES (including cover sheet) 3

____ This will be the only form of delivery of the transmitted document.

____ The original of the transmitted document(s) will be sent: ____ Ordinary Mail ____ Overnight
____ Certified Mail ____ Turner Pouch

Fig. 105. Copyright permission for *cnn* sequence (page 1).

JUN-09-1997 17:19 TBS LEGAL DEPARTMENT 404 827 1995 P.02

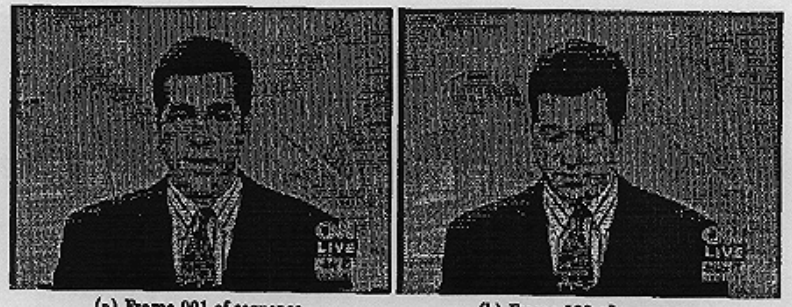
Texas A&M University
Department of Computer Science
College Station, TX 77843-3112
Phone: (409) 845-9980
Fax: (409) 847-8578

June 6, 1997

CNN
Legal Department
Attn: Dan Riner
Atlanta, GA 30303-3110
Phone: (404) 827-2600
Fax: (404) 827-1995

Dear Mr. Riner,

I am completing a doctoral dissertation at Texas A&M University entitled "Gaze-Contingent Visual Communication". Our library sends the dissertation to University Microfilms Inc. (UMI), for preparation of a microfilm copy of the document. I would like your permission to reprint in my dissertation excerpts from a CNN broadcast featuring anchor Miles O'Brien which aired on July 21, 1996 at approximately 8:02am ET. I used the video in a study of eye movements and video processing. The excerpt to be reproduced is a digitized sequence containing 128 images scanned at 16 frames per second (8 seconds of video). The first and last images of the sequence are shown below.

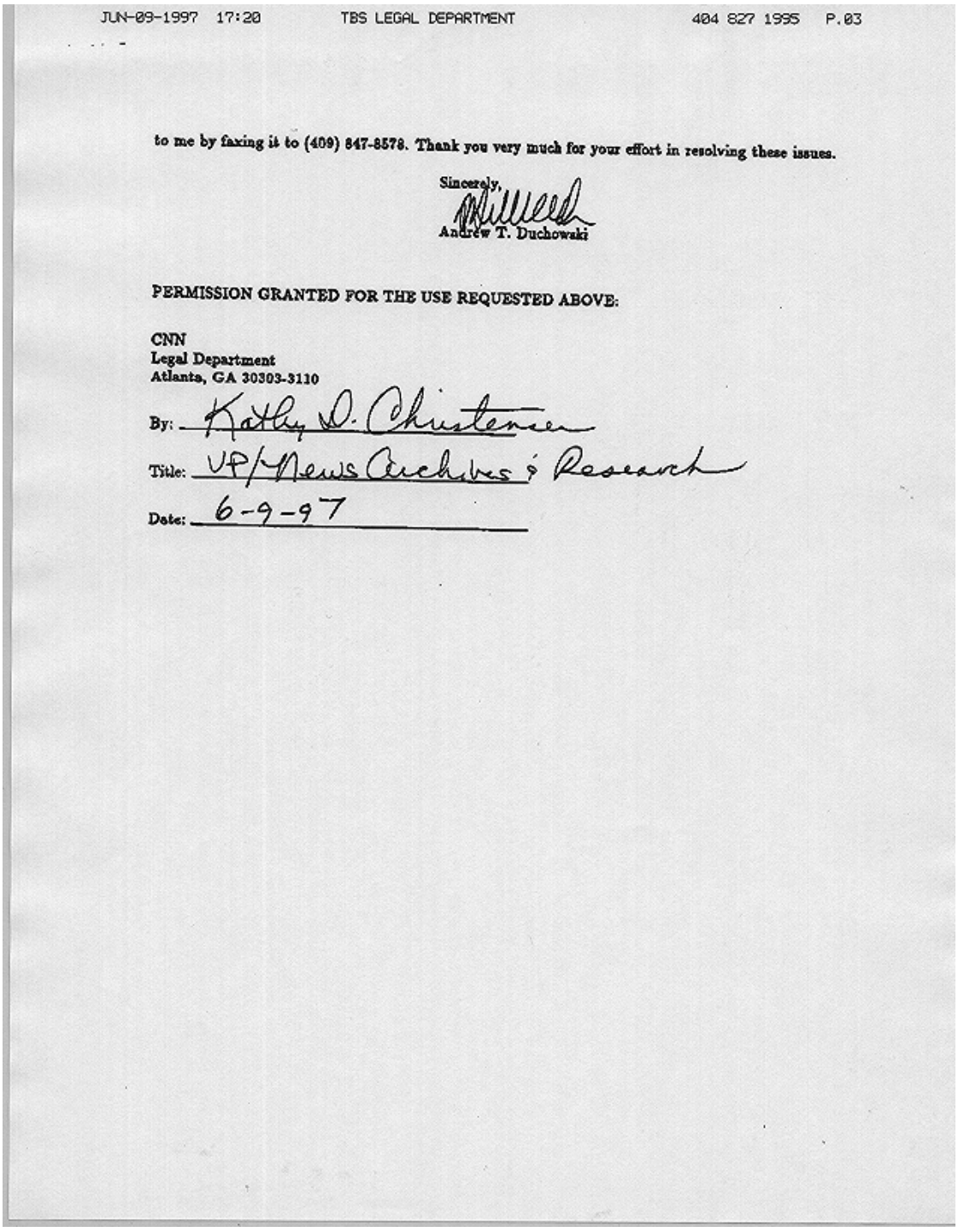


(a) Frame 001 of sequence. (b) Frame 128 of sequence.

The requested permission extends to any future revisions and editions of my dissertation, including non-exclusive world rights in all languages, and to the prospective publication of my dissertation by UMI Company. These rights will in no way restrict republication of the material in any other form by you or by others authorized by you. Your signing of this letter will also confirm that you own, or CNN owns, the copyright to the above-described material.

If these arrangements meet with your approval, please sign this letter where indicated below and return it

Fig. 106. Copyright permission for *cnn* sequence (page 2).



JUN-09-1997 17:20 TBS LEGAL DEPARTMENT 404 827 1995 P.03

to me by faxing it to (409) 847-8578. Thank you very much for your effort in resolving these issues.

Sincerely,

Andrew T. Duchowski

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

CNN
Legal Department
Atlanta, GA 30303-3110

By: Kathy D. Christensen

Title: VP/News Archives & Research

Date: 6-9-97

Fig. 107. Copyright permission for *cnn* sequence (page 3).

VITA

Andrew Duchowski was born on August 14, 1966 in Warsaw, Poland, to Helena L. Duchowska (mother) and Kazimierz ("Kaz") Duchowski. The family, together with John (author's brother) emigrated to Montréal, Québec, Canada, in 1974. In 1979 the family moved to Vancouver, BC, Canada where the author's parents still reside. In 1985 the author entered Simon Fraser University in Burnaby, BC, Canada and in 1990 he earned his B.Sc. (Co-op) degree in Computer Science. That same year Mr. Duchowski began graduate studies in the Department of Computer Science at Texas A&M University in College Station, TX. In 1992, he married Corey Lee Ferrier, of Coquitlam, BC, Canada.

From 1990 to 1994 the author worked as a Research Assistant with Drs. Deuermeyer and Curry in the Department of Industrial Engineering. From 1994 to 1995 he was a Graduate Teaching Assistant in the Department of Computer Science. From 1995 to 1997, as a Research Assistant in the Scientific Visualization Laboratory in the Department of Computer Science, he worked on the present topic of Gaze-Contingent Visual Communication.

Mr. Duchowski's general research interests include human visual perception and computer vision, signal and wavelet analysis, image and video processing, computer graphics, and human and computer interaction.

The author can be reached through the chairman of his committee, Dr. Bruce H. McCormick, Professor, Department of Computer Science, Texas A&M University, College Station, TX 77843-3112.

This dissertation was typeset with LaTeX by the author.[1]

[1]The LaTeX document preparation system was developed by Leslie Lamport as an interface extension to Donald Knuth's TeX computer typesetting program. TeX is a trademark of the American Mathematical Society.