3D Wavelet Analysis of Eye Movements*

Andrew T. Duchowski†
andrewd@cs.tamu.edu
Department of Computer Science
Texas A&M University
College Station, TX, 77843-3112

ABSTRACT

Awareness of the viewer's gaze position in a virtual environment can lead to significant savings in scene processing if fine detail information is presented "just in time" only at locations corresponding to the participant's gaze, i.e., in a *gaze-contingent* manner. 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. In this paper, a conceptual Piecewise Auto-Regressive Integrated Moving Average (PARIMA) model of conjugate eye movements is 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. An off-line video frame-based 3D wavelet analysis technique is proposed for classification of eye movements into smooth pursuits, fixations, and saccades.

Key words: Gaze-contingent, eye movements, wavelets.

1 INTRODUCTION

Although it is known that the oculomotor system is inherently nonlinear, 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 of this paper is to identify eye movements in terms of their signal characteristics, that is, to specify filters describing the observed (external) signal characteristics of eye movements. The approach taken for the specification of these filters 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. Due to the Slutzky effect if such a filter can be found,

* 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.

† As of 01/98 the author can be reached at *andrewd@cs.clemson.edu*, Department of Computer Science, 451 Edwards Hall, Clemson University, Clemson, SC 29634-1906.

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.[7] Hence, the goal of the signal analysis methodology presented here is to model the observed signal by an inverse of the filters used to represent the oculomotor system. This modeling strategy is represented in Figure 1 (note the dual use of symbols s_t and x_t).

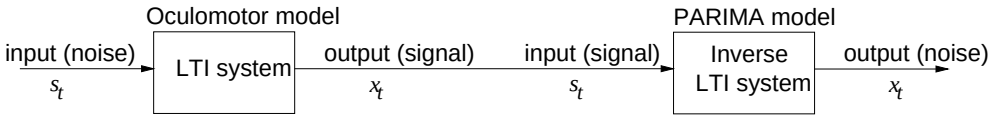


Figure 1: Linear filter modeling strategy.

2 CONCEPTUAL SPECIFICATION OF THE PARIMA MODEL

In the proposed Piecewise Auto-Regressive Integrated Moving Average (PARIMA) eye movement model, three principal types of eye movements, namely 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 is based on reported saccade characteristics.[6] 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.

2.1 Fixations

Fixations are assumed to generate a constant signal (w.r.t., 2D position) perturbed by the miniature eye movements tremor, drift, and microsaccades, corresponding to the variance of the system. In the PARIMA model representation, the stochastic signal representing fixations is described by the linear time-invariant filter h_f. The output of the system x_t is assumed to be white noise. In the z-domain, fixations are modeled by the following equation

$$X(z) = H_f(z)(S(z) - X(z)). \tag{1}$$

Supposing that the filter $H_f(z)$ represents a stable and causal system, then the filter can be represented by a rational function in z or ARMA model of the form $H_f(z) = B_f(z)/A_f(z)$. Substituting the rational approximation of $H_f(z)$ in Equation (1) gives

$$\begin{aligned} X(z) &= H_f(z)(S(z) - X(z)) \\ &= \frac{B_f(z)}{A_f(z)}(S(z) - X(z)). \end{aligned} \tag{2}$$

Expanding Equation (2) in the time domain,

$$\begin{aligned} &a_{f_0}x_t + a_{f_1}x_{t-1} + \cdots + a_{f_p}x_{t-p} = \\ &\quad b_{f_0}(s_t - x_t) + b_{f_1}(s_{t-1} - x_{t-1}) + \cdots + b_{f_q}(s_{t-q} - x_{t-q}), \end{aligned}$$

gives

$$\sum_{k=0}^{p} a_{f_k}x_{t-k} + \sum_{k=0}^{q} b_{f_k}x_{t-k} = \sum_{k=0}^{q} b_{f_k}s_{t-k}, \tag{3}$$

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.[5] Fixations 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 (3) in the z-domain

$$A(z)X(z) + B(z)X(z) = B(z)S(z) \tag{4}$$

$$\frac{X(z)}{S(z)} = \frac{B(z)}{A(z) + B(z)}. \tag{5}$$

The collection of terms in Equation (4) 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 (5) gives

$$\begin{aligned} \frac{X(z)}{S(z)} &= \frac{\frac{B(z)}{A(z)}}{1 + \frac{B(z)}{A(z)}} \\ &= \frac{H(z)}{1 + H(z)}, \end{aligned} \tag{6}$$

which represents the noise-to-signal ($X(z)/S(z)$) ratio of the system. This is the inverse filter of an often used linear feedback model of smooth pursuits.[5]

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.

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 (6).

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 parameter d_f 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 classical 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.[1]

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 z-domain,

$$X(z) = G(z)S(z). \tag{7}$$

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 a filter commonly used to model saccades generated by the oculomotor plant.[5] In the time domain, the filter modeling the observed signal is specified by the linear moving average (MA) model,

$$\begin{aligned} x_t &= g_0 s_t + g_1 s_{t-1} \\ &= \frac{1}{\sqrt{2}} s_t - \frac{1}{\sqrt{2}} s_{t-1}, \end{aligned} \tag{8}$$

or an ARMA(0,1) sequence. In practice, Equation (8) 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.

3 WAVELET MODEL OF TEMPORAL TIME SERIES

The proposed framework for temporal analysis of eye movement time series is the Discrete Wavelet Transform (DWT), 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. A frame-based implementation strategy is presented in the following section.

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.[8] This criterion is satisfied with the choice of the Haar wavelet. At each scale j, local modulus maxima $M\{s_t\}(j)$ 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[9] (see §5 step 3). 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) prior to reconstruction 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.

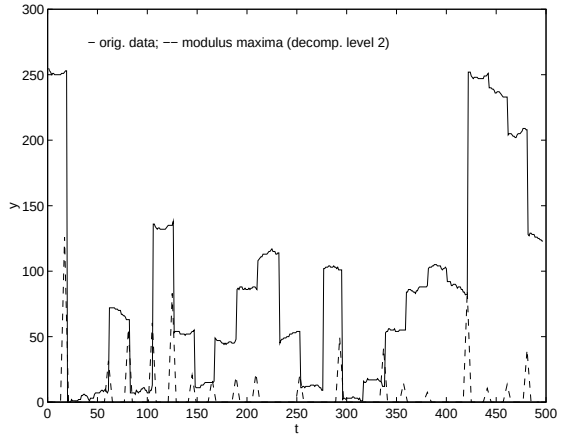
To complete the specification of the saccade detection model, the Haar scaling function is used for temporal decomposition, and the wavelet transform decomposition level j is derived from the eye movement sampling rate. The current experimental apparatus operates with an average eye movement sample period of $s_p = 18$ms, giving a minimum temporal decomposition level of

$$\begin{aligned} j &> \log_2 (\frac{s_p}{T_{min}}) + 1 \\ &> 1 \\ &\geq 2, \end{aligned}$$

where $T_{min} = 10$ms is the minimum expected duration of a saccade.

A hypothetical time series is plotted in Figure 2(a) with overlayed detected interventions (dashed lines) at two decomposition levels (after thresholding). The reconstructed time series is plotted in Figure 2(b) where isolated segments are presumed to be ARIMA sequences and are subject to individual ARIMA model specification.

(a) Time series with detected modula maxima.



(b) Partitioned time series.

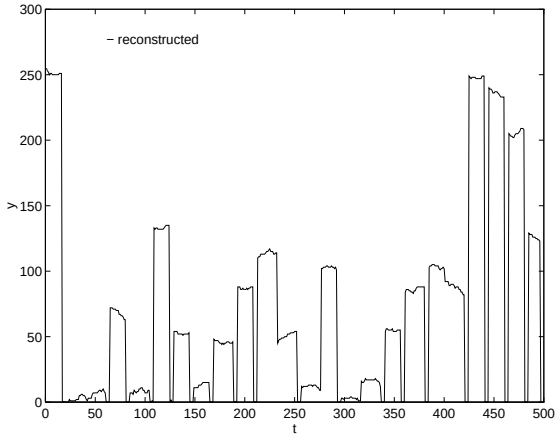


Figure 2: PARIMA time series model.

4 FRAME-BASED IMPLEMENTATION OF THE PARIMA MODEL

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 (e.g., approximation by B-Splines). In the current implementation, a frame-based approach is chosen in order to synchronize eye movement sample data with stimulus video frames.

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:

$$f(x_i, y_i, \left\lfloor \frac{t_i}{t_f} \right\rfloor) = \begin{cases} 255 & \forall\ x_i, y_i, t_i, \quad i \in [1, n] \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

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 $t_f = \{0, 0, 0, 0, 1, 1, \ldots\}$.

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). Note that under these implementational parameters temporal resolution is sacrificed for ease of representation of the eye movement data in stimulus video frame coordinates. In video format, eye tracker data is submitted to anisotropic 3D wavelet transform analysis.[1] 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 (9) above. Equation (9) specifies the mapping of raw POR data onto initially empty video frames composed of 0-intensity pixels.

5 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] The 3D analysis is anisotropic in terms of varying decomposition levels along the x, y, and t dimensions.

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. This gives horizontal and vertical display dimensions of 16.8×12.6 width $\times$ height, in inches, with effective resolution of about 30×20 (width $\times$ height) dots per inch (dpi).

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 $D = 23.622$ 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 $\theta = 2\tan^{-1}(r/D) = 0.32°$. Extending these calculations over successive dyadic decompositions produces values given in Table 1.

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

Table 1: Resulting subtended visual angle of POR at dyadic spatial subsampling levels.

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 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 1, 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}.[2] Since the eye movement video sequence is composed at 16fps, 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 then $1.25° - 6.25°$. 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.

6 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**
   Each frame is decomposed to 3 levels. Expressing this decomposition concisely,

$$\begin{aligned} f^{j-1}_{\phi\phi\cdot}(x,y,t) &= \sum_{k,m}(h_k \otimes h_m) f^{j}_{\phi\phi\cdot}(2x+k, 2y+m, t) \\ f^{j-1}_{\psi\phi\cdot}(x,y,t) &= \sum_{k,m}(g_k \otimes h_m) f^{j}_{\phi\phi\cdot}(2x+k, 2y+m, t) \\ f^{j-1}_{\phi\psi\cdot}(x,y,t) &= \sum_{k,m}(h_k \otimes g_m) f^{j}_{\phi\phi\cdot}(2x+k, 2y+m, t) \\ f^{j-1}_{\psi\psi\cdot}(x,y,t) &= \sum_{k,m}(g_k \otimes g_m) f^{j}_{\phi\phi\cdot}(2x+k, 2y+m, t) \end{aligned}$$

   where $\{h_k\}$, $\{g_k\}$ are the Haar low and high pass filters, respectively, 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.

2. **Temporal 1D (Inter-Frame) Wavelet Decomposition**
   The entire sequence, treated as a now one-dimensional signal, is decomposed temporally to 2 levels,

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

   giving the 1D temporal DWT:

$$\begin{aligned} &\{Wf(x,y,t)\}_t(j) = \\ &\quad \{f^{j}_{\cdot\cdot\phi}(x,y,1), f^{j}_{\cdot\cdot\psi}(x,y,2), \ldots, f^{j}_{\cdot\cdot\phi}(x,y,n-1), f^{j}_{\cdot\cdot\psi}(x,y,n)\}. \end{aligned}$$

   High- and low-pass filtered frames are rearranged forming two $n/2$ sequences.

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:

$$|f^{j}_{\phi\phi\psi}(x,y,t-1)| \;\le\; |f^{j}_{\phi\phi\psi}(x,y,t)| \;\ge\; |f^{j}_{\phi\phi\psi}(x,y,t+1)|, \text{ and}$$

$$\begin{cases} |f^{j}_{\phi\phi\psi}(x,y,t)| \;>\; |f^{j}_{\phi\phi\psi}(x,y,t-1)|, & \text{or} \\ |f^{j}_{\phi\phi\psi}(x,y,t)| \;>\; |f^{j}_{\phi\phi\psi}(x,y,t+1)|. \end{cases}$$

   Temporal modulus maxima values correspond to step edges in time, or saccades. Wavelet coefficients at these locations are deleted through subsequent thresholding steps.

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 $\alpha = 0.05$.

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.

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.[10] 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 $f^j(x,y,t) = \{Wf(x,y,t)\}_{xy}(j)$.

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.,

$$\begin{matrix} f^j(2x+k,2y+m,t) \\ f^j(2x+k,2y-m,t) \\ f^j(2x-k,2y+m,t) \\ f^j(2x-k,2y-m,t) \end{matrix} \quad = \quad \begin{cases} f^{j-1}_{\phi\phi\cdot}(x,y,t) & \text{if } f^{j-1}_{\phi\phi\cdot}(x,y,t) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

for $k, m \in [0,1]$.

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 preceding 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 5° visual angle.

7 RESULTS

Eye tracking experiments were carried out in the Virtual Environments Laboratory, Department of Computer Science, Texas A&M University. Results of experiments pertaining to the proposed method of eye movement analysis through the three-dimensional wavelet transform are summarized here. Details of experimental procedures, including verification of eye tracker accuracy and slippage, are given elsewhere.[3]

The first experiment was used as a preliminary empirical evaluation of the proposed technique. Specifically, recorded eye movements were analyzed for correspondence to predicted and observed saccade locations. Three 8-second, 16fps video sequences were used as stimulus, each sequence composed of a single moving white dot (10-pixel diameter) subtending roughly 1° visual angle used to evoke eye movements. The moving dot randomly simulates fixations, saccades and smooth pursuit movements based on random distributions of a range of values corresponding to known eye movement characteristics. A total of 7 subjects participated in the experiment, all were instructed to visually track the moving dot.

The proposed wavelet-based analysis of eye movements was evaluated through comparison of detected and expected saccade locations. Mean "hit" and "correctness" rates were studied. Preliminary analysis of the results indicates only a modest success rate of the method. Particularly, the mean percent "hit rate" 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. The mean "correctness rate" of the method is estimated at 29%. This estimate was obtained by taking the ratio of correctly identified saccades versus the total number of saccades detected in the eye movement signal, per individual tested.

Evaluation of the proposed wavelet model was performed using a stringent evaluation of saccade detection. That is, no error tolerance was given for expected and observed saccade mismatches in time. For example, if a saccade between video frames 001-002 was expected, an observed (detected) saccade was labeled as a "hit" only if frames 001-002 were identified as ones containing a saccade (signal discontinuity). This stringent evaluation criterion may be too critical since it does not consider any delay between the stimulus and the detected saccade although there are several plausible sources of delay (e.g., the eye tracker). Consequently, the exact match criteria probably underestimates the power of the proposed method. A more robust analysis of the results is currently underway.

A second experiment was run utilizing the current analysis method as an indicator of expected dynamic fixation locations for the purposes of gaze-contingent video processing within the visual tracking paradigm.[4] In this experiment, peripheral regions of a monochromatic video sequence were degraded based on the present technique's identification of intra-frame fixation locations. Among several factors, a comparison was made between predicted fixation locations over an easily detectable target of an "ideal observer" and of a single individual's fixation locations, as identified by the present technique. The individual used as the test case (referred to as subject "hunter") provided eye movement patterns considered slightly better than average, with respect to visual target tracking performance (dynamic gaze position error) including small average eye tracker error and slippage.

Peripheral video regions were degraded based on identified dynamic fixation locations, as prescribed by either the ideal observer's or subject "hunter"'s scan patterns. Human eye movements typically involve fixation discontinuities due to saccades or blinks, whereas the ideal observer provides a continuous dynamic fixation pattern. In the experiment, subject "hunter"'s fixation discontinuities were eliminated. Fixation discontinuities result in a lack of intra-frame region of interest (ROI) necessitating the spatial degradation of an entire video frame. Inclusion of such a frame in the video sequence results in a sudden, brief loss of resolution, potentially impeding perception. Hence, any such frames, corresponding to discontinuities, must be processed by forcibly including an ROI to prevent potential perceptual distraction. With respect to a given video frame containing such a discontinuity, this is done by either extending a previous fixation ROI into the future of a video stream, or extending a forthcoming fixation ROI into the past. Since the latter strategy anticipates fixations (within the visual tracking paradigm), it is adopted as the strategy for discontinuity elimination, and is dubbed *preattentive* due to its anticipatory nature.

The same video sequence was presented to two groups of 4 subjects, each group viewing the sequence processed according to either the ideal observer's or subject "hunter"'s expected fixation locations. Results between the ideal and preattentive conditions support the effectiveness of the proposed wavelet-based eye movement analysis method for prediction of eye movements in the visual tracking tasks. In terms of gaze position error over the intended ROIs, subjects' performance was better over the preattentive sequence than over the ideal observers'. This finding was surprising since viewers were expected to better foveally match the ideal target. It is interesting to note that in this experiment, subject "hunter"'s dynamic fixations, as identified by the present wavelet-based technique, provide a better prediction of scan patterns than the ideal observer.

8 CONCLUSION

In this paper 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 three-dimensional discrete wavelet transform (DWT). 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 DWT 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. 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.

Empirical evidence suggests the PARIMA model is an adequate representation of conjugate eye movements. Although aggressive statistical analysis exposes the current frame-based technique's susceptibility to error, the method's utility is demonstrated through eye movement prediction for gaze-contingent image representation under the visual tracking paradigm.

9 REFERENCES

[1] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day, Inc., Oakland, CA, 1976.

[2] R. H. S. Carpenter. *Movements of the Eyes.* Pion Limited, London, 1977.

[3] Andrew T. Duchowski. Incorporating the Viewer's Point-Of-Regard (POR) in Gaze-Contingent Virtual Environments. In *The Engineering Reality of Virtual Reality'98*, Bellingham, WA, January 1998. SPIE.

[4] Andrew T. Duchowski and Bruce H. McCormick. Gaze-Contingent Video Resolution Degradation. In *Human Vision and Electronic Imaging II*, Bellingham, WA, January 1998. SPIE.

[5] Andrew Ted Duchowski. *Gaze-Contingent Visual Communication.* PhD thesis, Texas A&M University, College Station, TX, 1997.

[6] John M. Findlay. Programming of Stimulus-Elicited Saccadic Eye Movements. In Keith Rayner, editor, *Eye Movements and Visual Cognition: Scene Perception and Reading*, pages 8–30. Springer-Verlag, New York, NY, 1992. Springer Series in Neuropsychology.

[7] John M. Gottman. *Time-Series Analysis: A Comprehensive Introduction for Social Scientists.* Cambridge University Press, Cambridge, 1981.

[8] Stephane Mallat and Wen Liang Hwang. Singularity Detection and Processing with Wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, March 1992.

[9] Stephane Mallat and Sifen Zhong. Wavelet Transform Maxima and Multiscale Edges. In Mary Beth Ruskai, Gregory Beylkin, Ronald Coifman, Ingrid Daubechies, Stephane Mallat, Yves Meyer, and Louise Raphael, editors, *Wavelets and Their Applications*, pages 67–104. Jones and Bartlett, Boston, MA, 1992.

[10] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, 2nd edition, 1992.