Algorithm for Discriminating Aggregate Gaze Points: Comparison with Salient Regions-Of-Interest

Thomas J. Grindinger[1], Vidya N. Murali[1], Stephen Tetreault[2],
Andrew T. Duchowski[1], Stan T. Birchfield[1], and Pilar Orero[3]

[1] Clemson University
[2] Rhode Island College
[3] Universitat Autònoma de Barcelona

Abstract. A novel method for distinguishing classes of viewers from their aggregated eye movements is described. The probabilistic framework accumulates uniformly sampled gaze as Gaussian point spread functions (heatmaps), and measures the distance of unclassified scanpaths to a previously classified set (or sets). A similarity measure is then computed over the scanpath durations. The approach is used to compare human observers's gaze over video to regions of interest (ROIs) automatically predicted by a computational saliency model. Results show consistent discrimination between human and artificial ROIs, regardless of either of two differing instructions given to human observers (free or tasked viewing).

1 Introduction

A compelling means of analysis of human visual perception is drawn from the collection of eye movements over dynamic media, *i.e.*, video. The video stream can either be a scene captured by a forward-facing camera worn during the performance of some natural task [1], or of film presented to the viewer [2]. Analysis of the former leads to improved understanding of how humans function in the world, and in particular, how vision is used in concordance with basic motor actions such as walking or reaching [3]. Analysis of the latter leads to better understanding of how artistic media is perceived, and in turn, how its design and production can be altered to affect its perception.

Analysis of eye movements over dynamic media has largely been performed manually, *e.g.*, by hand-coding saccadic events as they occur in relation to events present in the media such as scene cuts [4]. What is needed, and what this paper addresses, is an automatic means of classification of disparate viewing patterns, or *scanpaths*—defined as the temporal sequence of gaze or fixation coordinates cast over the stimulus.

This paper contributes a means of classification of scanpaths accumulated over temporal event samples. Event samples happen to coincide with video frames in this instance, but the technique can assume any sampling rate and is thus also applicable to still imagery presented for extended viewing durations [5]. Applications of the approach include gaze-based discrimination between classes of human viewers (*e.g.*, experts from novices—eye movements are known to be task-dependent [6]), or discrimination between human gaze and artificially predicted regions of interest, or ROIs. The paper focuses on the latter, in a manner differing from previous work with images [7], distinguishing between *perceptually salient* and *computationally salient* gaze coordinates.

2 Background

Scanpath comparison can be classified as either *content-* or *data-driven*. The former is largely based on regions of interest, or ROIs, identified *a priori* in the stimulus and subsequently by associating those regions with fixations, leading to analysis of image regions or elements fixated by the viewer. The latter approach, in contrast, is made on scanpaths directly, independent of whatever was presented as the stimulus. An important advantage of the latter is that it obviates the need for establishing a reference frame within which the ROI stipulation must take place.

Consider two recent approaches to the scanpath comparison problem. The vector-based similarity measure is content-driven, as it relies on the quantization of the stimulus frame into an arbitrarily-sized 5×5 grid which serves as the method's source of ROI labeling [8]. A label is added to the scanpath stream whenever a fixation is present within a grid cell. In contrast, the revisited string-editing approach is data-driven, as it operates directly on scanpaths [9]. String (ROI) labels are determined by overlapping fixation clusters. Both approaches consider fixation durations and are therefore potentially suitable for analysis of gaze collected over dynamic media, however, their means of scanpath aggregation are derived from pairwise vector or string comparisons. For groups of viewers, considerable additional organization is required.

As an alternative to string-editing approaches, heatmaps have become a common tool for visualization of eye tracking data [10,11]. To our knowledge, to date they have not been successfully used for quantitative classification of aggregate eye movements.

Perhaps most similar to the present work are two previous efforts of calculation of the "average scanpath" [12] and of the computation of the scanpath distance via the Earth Mover's Distance [13]. The former was based on string-based multiple sequence alignment, although the derivative notion of variance (distance from the average) was omitted. The latter relied on the conceptualization of a scanpath composed of "piles of earth", with a comparison scanpath represented by "holes". The minimum amount of energy required to move earth from piles to holes gave the scanpath similarity.

The present paper extends a framework for multiple scanpath comparison and classification [5]. Although the previous approach was inspired by dynamic media, it was only implemented over still images viewed for very short durations. In this paper the analysis framework is applied to dynamic media for which it was originally conceived, namely video sequences. The resultant procedure may be conceptualized as a measure of deviation, over time, of one or more scanpaths of unknown classification from a set of scanpaths of known classification. This is similar to a prior effort based on machine learning, which was also intended to act as a classifier, although its aim was to classify content (*i.e.*, image regions) [14], whereas the present approach is directed at classification of the data (*i.e.*, scanpaths).

3 Classification Framework

Following Airola *et al.*'s nomenclature [15], let D be a probability distribution over the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with input space $\mathcal{X}$ and output space $\mathcal{Y} = \{-1, 1\}$, where $y \in \mathcal{Y}$ denotes the labeling of the input $x \in \mathcal{X}$ as a non-class ($x_- \in X_-$) or class

member ($x_+ \in X_+$), respectively. We define a classifier as a function $C_Z(x)$ that outputs a set of threshold-based decisions $Z = \{z_1, \ldots, z_m\} \in \mathcal{Z}^m$ where $z_i = (x_i, y_i)$, for the training set of m training examples $X = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$.

There are three steps to building and evaluating the real-valued prediction function C_Z produced by a learning algorithm developed with fixed training set Z. First, similarity scores are extracted from X. Second, a discrimination threshold h is computed from the similarity scores assigning the positive class X_+ to x if $C_Z(x) > h$ and the negative class X_- otherwise. Third, classifier reliability is gauged by the *conditional expected AUC*, or AUC, the area under Receiver Operating Characteristic (ROC) curve, $A(C_Z) = E_{x_+ \sim D_+, x_- \sim D_-}\left[H(C_Z(x_+) - C_Z(x_-))\right]$ where $H(a)$ is the Heaviside step function, which returns 1 when $a > 0$, $1/2$ when $a = 0$, and 0 when $a < 0$. In practice, because the probability distribution D cannot be accessed directly, the AUC estimate $\hat{A}$ is calculated *e.g.*, via cross-validation, or by the Wilcoxon-Mann-Whitney statistic:

$$\hat{A}(S, C_Z) = \frac{1}{|S_+||S_-|} \sum_{x_i \in S_+} \sum_{x_j \in S_-} H(C_{\overline{\{i\}}}(x_i) - C_{\overline{\{j\}}}(x_j))$$

where $S_+ \subset S$ and $S_- \subset S$ are the positive and negative examples of the set S, and $C_{\overline{\{i\}}}(x_i)$ is the classifier trained without the i^{th} training example.

Along with AUC, classifier accuracy is reported by evaluating $C_Z(w)$ on test data $w \in W$, assumed to be disjoint from X. Accuracy is defined as the ratio of correctly classified examples of W (true positives and true negatives) to all classified examples.

Accuracy and AUC measures can be seen to correspond to two different metrics of interest. The former is related to the quality of the learning algorithm, *i.e.*, how well on average C_Z generalizes to new test and training data. The latter addresses how well $C_Z(x)$ generalizes to future test examples once learned from the given training set. In the present context, the latter is more of interest as it provides a better indication of the discriminability of the given training data set against the test set or sets, *i.e.*, does a given scanpath class differ from another class or classes of scanpath sets.

3.1 Extracting Similarity Scores

The classifier's similarity measure computes a scanpath's deviation from a probabilistic model of one (or more) class(es) of scanpaths classified *a priori*. Scanpath classes can be operationalized arbitrarily, *e.g.*, based on some characterization of viewers. The classifier functions over dynamic stimuli, *i.e.*, video, which may be considered as a collection of static stimuli, *i.e.*, frames. Scanpath similarity metrics developed for static stimuli can thus be applied on a frame-by-frame basis and aggregated in some way (*e.g.*, averaged). The trouble with prior vector- or string-based approaches is their reliance on pairwise comparisons for aggregation. This leads to rather complicated bookkeeping requirements for pairwise organization, *e.g.*, labeling each pair as *local*, *repetitive*, *idiosyncratic*, or *global* based on the dyadic permutations of viewer and stimulus [7].

Presently, each frame is composed of a sampled set of gaze points (or fixations), sampled from as many sets as there are scanpath classes, with each set composed of scanpaths collected from multiple viewers. A per-frame similarity measure is then derived and averaged over the duration of the video sequence to compute the total similarity of an unclassified scanpath to the one or more sets of classified scanpaths.

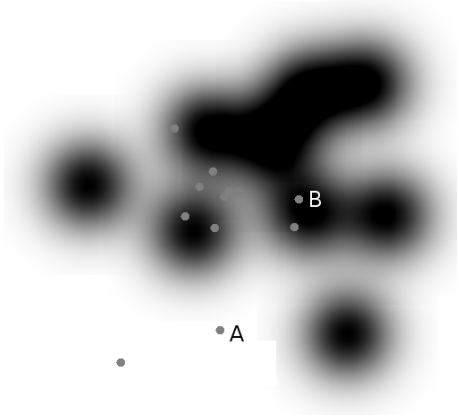


Fig. 1. Heatmap of a classified scanpath set S at a discrete timestamp. As yet unclassified scanpaths' (gray circles not used in heatmap generation) similarities are calculated as the average Gaussian similarity, *e.g.*, $d(A, S) < d(B, S)$ in this example.

With video acting as the temporal reference frame, a scanpath $s(t)$ is parametrized by the frame timestamp t, such that $s(t) = \{(i(t), j(t)) \mid t \in [t - w, t + w]\}$ for some window w, with $w = 0$ identifying a single frame, yielding the scanpath's 0^+ gaze points over a video frame at t.[4] This *event-driven* model, effectively samples a scanpath at a single point in time, and affords notational interchangeability between a gaze point, fixation, and scanpath, when considered on a per-event, or in this case per-frame, basis. A set of scanpaths $S(t) = \{s_1(t), s_2(t), \ldots, s_m(t)\}$ is similarly parametrized to define the combined set of gaze points over frame t from the scanpath set collected from m viewers. Over each frame, multiple sets are represented, e.g., S_+ member and S_- non-member sets (in the experiment described below, three such sets are established).

Modeling a classified scanpath s by a normally distributed point spread function $f(s) = 1/\sqrt{2\pi\sigma^2}\exp\left(-s^2/2\sigma^2\right)$ produces the well-known *heatmap* scanpath visualization (on a per-frame basis; see Fig. 1), typically visualized with the Gaussian kernel's support truncated beyond 2σ for computational efficiency [16]. Extending kernel support also defines the scanpath's first moment $\mu_s = \int_{-\infty}^{\infty} sf(s)ds$ so that the (Gaussian) similarity of an unclassified scanpath s' to s is estimated by its deviation

$$g(s', \mu_s) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(s' - \mu_s)^2}{2\sigma^2}\right)$$

with frame timestamp t made implicit and σ set to the expected eye tracker error, as illustrated in Fig. 1. In practice, the above model is necessarily discrete and s is under-

[4] With a 50 Hz eye tracking sampling rate and a common video refresh rate of 30 Hz, it is assumed that a scanpath will yield at most two gaze point samples per frame; alternatively, if operationalized by a sequence of fixations, a scanpath will yield a single fixation coordinate per frame (or none if the frame happened to sample an inter-fixation saccade).

stood to be two-dimensional, $s(t) = (i(t), j(t)), s(t) \in \mathbb{R}^2$, with t denoting the frame timestamp and (i, j) the image (video frame) coordinates.

The similarity of s' to a set of classified scanpaths S (at t) is defined as

$$d(s', S) = \frac{1}{|S|} \sum_{s \in S} g(s', \mu_s)$$

where the weighting factor $1/|S|$ is used for similarity score normalization. The measure $d(s', S)$ is averaged over the entire video sequence to estimate the mean similarity of an unclassified scanpath to the classified scanpath set, $\bar{d}(s', S) = 1/T \sum_t d(s', S)$ with $t \in T$, the sequence duration. The resultant mean similarity lies between 0 and 1, but tends to fall near 0. Its value, however, is not as important as the probability that the score lies within the expected distribution of scores for a specific class.

3.2 Computing the Classification Threshold

Gaussian similarity scores serve as input to the classification mechanism that estimates an optimal discrimination threshold for scanpaths of unknown classification. An unclassified scanpath is accepted by the classifier if its similarity score is higher than the computed threshold.

The ROC curve plots the true positive response against the false positive response of the threshold at each threshold level and provides two convenient facilities. First, it facilitates the choice of an optimal threshold, by selecting the level at which the threshold is closest to $(0, 1)$, where the ratio of false positives to true positives is balanced. Second, AUC indicates the classifier's discriminative capability. Ideally, AUC should equal unity (1), while a completely random classifier yields AUC close to 0.5. AUC represents the probability of an arbitrarily-chosen class member obtaining a similarity score greater than some arbitrarily-chosen non-class member.

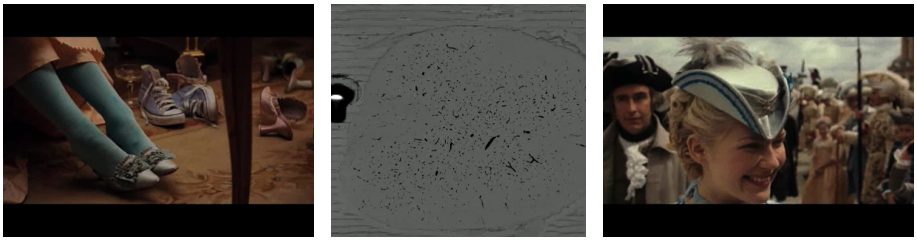
3.3 Estimating Classifier Performance via Cross-Validation

A typical strategy used for estimating the performance, or reliability, of a classifier, when operating in a small sample setting, is cross-validation.[5] Specifically, leave-pair-out cross-validation, or LPOCV, is adopted since the intent is to estimate the conditional AUC as an indicator of the classifier's performance while avoiding the pitfalls associated with pooling and averaging of LOOCV (leave-one-out cross-validation) [15].

Cross-validation is performed by repeatedly partitioning the data set into two non-overlapping parts: a training set and a hold-out set. For each partitioning, the hold-out set is used for testing while the remainder is used for training. Accuracy is computed as the percentage of hold-out sets successfully classified. For each partitioning, LPOCV leaves out at a time from the training set each possible positive-negative pair of training examples. With LPOCV, AUC is estimated as

$$\hat{A}(X, C_Z) = \frac{1}{|X_+||X_-|} \sum_{s_i \in X_+} \sum_{s_j \in X_-} H(C_{\overline{\{i,j\}}}(s_i) - C_{\overline{\{i,j\}}}(s_j))$$

[5] Scanpath data sets generally number in the tens, whereas classifiers tend to operate on data sets numbering in the thousands.



(a) Seq. A, chosen for its misplaced pair of modern sneakers.

(b) Seq. B, chosen for its unfamiliarity.

(c) Seq. C, chosen for its large number of prominent faces.

Fig. 2. Frames from stimulus sequences. Seqs. A and C were excerpts from Sofia Coppola's *Marie Antoinette* © 2006, Columbia Pictures and Sony Intl., obtained with permission for research purposes by the Universitat Autònoma de Barcelona. Seq. B shows the mouse vasculature in the spinal cord at $0.6\times0.6\times2$ μm resolution with blood vessels stained black, as obtained by a knife-edge microscope (courtesy of Texas A&M).

where $X_+ \subset X$ and $X_- \subset X$ are the positive and negative examples of the training set X, $C_{\overline{\{i,j\}}}(s_i)$ is the classifier trained without the i^{th} and j^{th} training examples, and $H(a)$ is the Heaviside step function. Because AUC estimate $\hat{A}(X, C_Z)$ is equivalent to the Wilcoxon-Mann-Whitney U statistic, AUC > 0.7 is generally considered a statistically significant indicator of discriminability, although a test of significance should be performed by computing the standardized value under assumption of normality of class distributions.

The training data generally consists of multiple classes, very often two, but possibly more. The current approach generates multiple classifiers, each trained to a single class, with all other classes acting as non-class training data. Generally, when there are more than two classes, a "one-to-many" comparison may be carried out first, with all non-class training data pooled into the negative class set. Should the classifier AUC be significant, "one-to-one" comparisons can then be performed, in a manner analogous to ad-hoc pairwise t-tests following ANOVA.

4 Empirical Evaluation

The classifier was applied to scanpaths drawn from three classes: two from human observers distinguished by differing tasks, and the third from a bottom-up saliency model (simulating artificial observers), developed by Itti *et al*. [17]. The model is part of iLab's Neuromorphic Visual C++ Toolkit and is freely available online.[6] At the model's core is a neuromorphic simulation that predicts elements of a visual scene that are likely to attract the attention of human observers. This has wide applications in machine vision, *e.g*., automated target detection in natural scenes, smart image compression, *etc*. The model was compared to human scanpaths captured over video sequences.

[6] `http://ilab.usc.edu/bu/`, last accessed Aug., 2010.

Stimulus. Stimuli consisted of three video sequences, named A, B, and C, shown to human observers in Latin square counterbalanced order, with approximately each third of the viewers seeing the sequences in order $\{A, B, C\}$, $\{B, C, A\}$, or $\{C, A, B\}$. Seq. A contained a misplaced modern pair of sneakers in an 18^{th} century setting, while a modern popular song played in the background. Seq. C was from the same feature film, with scenes containing a large number of human faces. Seq. B was composed of CT-like scans of the mouse vasculature in the spinal cord. Select frames from the clips are shown in Fig. 2.

Apparatus. Eye movements were captured by a Tobii ET-1750 eye tracker, a 17 inch (1280×1024) flat panel with built-in eye tracking optics. The eye tracker is binocular, sampling at 50 Hz with 0.5° accuracy.

Participants. Twenty-seven college students volunteered in the study (seven male, twenty female). Participants' ages ranged from 18 to 21 years old.

Procedures. Participants sat in front of the eye tracker at about 60 cm distance. Following 9-point calibration, subjects were asked to naturally watch the first of two viewings of each of the three sequences (amounting to "free viewing"). They then received viewing instructions prior to the second viewing of the same sequence. For seq. A, they were asked to look for anything unusual (they were meant to notice the sneakers). For seq. B, they were asked to focus on the vascular stains (they were meant to avoid the aberrant artifacts at the top and sides of the frames). For seq. C, they were asked to avoid looking at faces (they were meant to simulate autism, since autistic viewers have been shown to exhibit reduced face gaze [18]).

Artificial gaze points over video were generated by the iLab Neuromorphic Toolkit. The toolkit contains a program called *ezvision* that can be executed on static images to produce a primary point of focus that is expected to match the visual attention of a human viewing the scene, followed by other salient points in the scene that are connected by a trajectory depending on the exposure time stipulated. However, the model also operates in video mode by extracting images from the video at the video frame rate. This causes the algorithm to be forced to find a salient point within the frame within the frame's exposure duration. For a typical video, this means the algorithm has only 33 ms to arrive at a salient viewpoint in the frame.

To compare the model's prediction with gaze points captured from human observers *ezvision* was run in video mode with the timestep set to 33 ms for the faces and shoes video, and 40 ms for the mouse video. Itti's algorithm is able to produce predictions with small amounts of noise added to the predictions [17]. This helped simulate results for 27 hypothetical users, by running *ezvision* on each video 27 times with random noise added to the predictions made each time.

5 Results

Classifier AUC and accuracy shows significantly consistent discriminability (AUC $>$ 0.7) between perceptual (top-down) and computational (bottom-up) saliency (see Tab. 1).

Table 1. Results composed of classifier accuracy (ACC) and area under ROC curve (AUC) for one-to-many and one-to-one comparisons of two classes of viewers ("free viewing" and tasked) vs. the computational model for each of the three video stimuli.

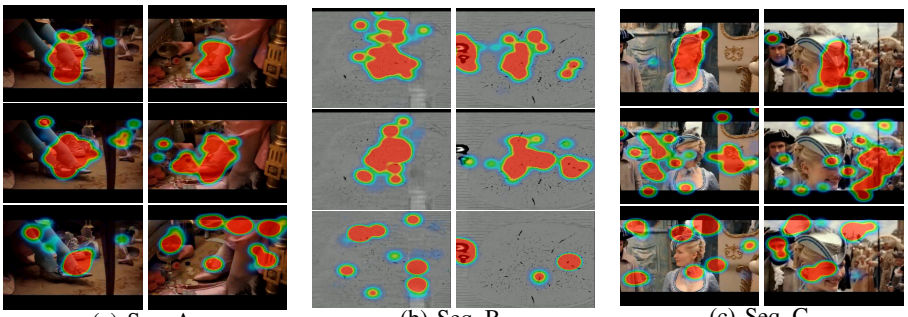
	One-to-many Cross-Validation			One-to-one Cross-Validation					
	Perceptual (pooled) vs. computational saliency			Perceptual "free viewing" vs. computational saliency			Perceptual tasked vs. computational saliency		
	A	B	C	A	B	C	A	B	C
ACC	1.000	1.000	0.997	1.000	0.999	0.999	0.999	1.000	1.000
AUC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Consistency refers to the evaluation of the Heaviside step function $H(a)$, where the classifier correctly discriminates between human and artificial scanpath classes in all of the $m \times (m-2)$ cross-validation partitionings, over all frames of each of the three video stimuli. The classifier is not as consistent in distinguishing between the two human scanpath classes, able only to distinguish between them in two of the three cases (Seq. B and C; these results are discussed at length elsewhere [19]).

Human observers tend to exhibit extreme preferential behavior over Seq. C, *i.e.*, when free viewing, heatmap visualization (see Fig. 3) suggests most viewers fixate faces, particularly in "close shots". Tasked viewers, in contrast, who were told to avoid faces, did so, but without apparent agreement on scene elements. Both strategies employ top-down directives that are apparently different from the strategy employed by the computational saliency model. The model fails to match human scanpaths over Seq. B even though it seems well suited to this stimulus (high contrast elements and sudden onset stimulus). Visualization suggests that both the model's and free viewers' gaze fell atop the sudden onset aberrant artifacts at the video frame edges. However, once humans were tasked to avoid these artifacts, they did so, whereas the model was not privy to this top-down goal-directed information. In either case, insufficient gaze overlap was detected over the length of this short video clip to diminish classifier output below unity. Seq. A yields similarly consistent discriminability results. Verbal instructions had little impact on perturbing human gaze (tasked scanpaths were not discriminable from free viewers' scanpaths by the classifier). Seq. A appears sufficiently complex to foil the saliency model from accurately predicting features selected by human observers.

6 Discussion

The saliency model works well on simple videos/images of traffic signals, or on tracks of single or multiple persons moving against fairly non-complex backgrounds, or in interactive visual environments [20]. However, for complex video segments with multiple objects of interest in the foreground and background and with rapid motion between the frames such as the *Marie Antoinette* videos, the bottom-up saliency model's gaze selection differs from that of natural viewing by humans. Two hypothetical parameters describe the extent of success/failure of the model: (1) the complexity of a single frame in the video, and (2) the amount of motion (apparent or real) between frames. When the two are low (simple images with small motion between frames), the model is



(a) Seq. A. (b) Seq. B. (c) Seq. C.

Fig. 3. Heatmap visualizations of two excerpted video frames viewed freely (top row), with task (middle row) or by the saliency model (bottom row).

likely to match human gaze points. However, when the complexity of the image and/or inter-frame motion increase(s), results diverge. The model could probably be used to describe the human visual system's tropism to salient points in a video, but only under fairly simple conditions. Once video complexity increases, bottom-up saliency can be clearly distinguished from tasked as well as natural viewing.

Given sufficiently clear instructions (*e.g.*, avoid looking at faces), the tropism of the human visual system, driven by top-down cognitive processes, differs from free viewing such that it can generally be distinguished by the classifier. The saliency model is, in contrast, task-independent and models bottom-up processes. Although it is possible to modify the relative feature weights in the construction of the saliency map with supervised learning to achieve some degree of specialization, it is at present unlikely that such specialization is sufficient to adequately model top-down visual processes.

7 Conclusion

A classification algorithm was developed to distinguish scanpaths collected over dynamic media. The algorithm successfully discriminated between perceptual and computational saliency over video sequences, illustrating the disparity between top-down visual processes and their bottom-up computational models.

References

1. Land, M.F., Tatler, B.W.: Looking and Acting: Vision and Eye Movements in Natural Behavior. Oxford University Press, New York, NY (2009)
2. Smith, T.J., Henderson, J.M.: Edit Blindness: The Relationship Between Attention and Global Change Blindness in Dynamic Scenes. Journal of Eye Movement Research **2** (2008) 1–17
3. Franchak, J.M., Kretch, K.S., Soska, K.C., Babcock, J.S., Adolph, K.E.: Head-Mounted Eye-Tracking of Infants' Natural Interactions: A New Method. In: ETRA '10: Proceedings of the 2010 Symposium on Eye Tracking Research & Applications, New York, NY, ACM (2010) 21–27

4. d'Ydewalle, G., Desmet, G., Van Rensbergen, J.: Film perception: The processing of film cuts. In Underwood, G.D.M., ed.: Eye guidance in reading and scene perception. Elsevier Science Ltd., Oxford, UK (1998) 357–367
5. Grindinger, T., Duchowski, A.T., Sawyer, M.: Group-Wise Similarity and Classification of Aggregate Scanpaths. In: ETRA '10: Proceedings of the 2010 Symposium on Eye Tracking Research & Applications, New York, NY, ACM (2010) 101–104
6. Yarbus, A.L.: Eye Movements and Vision. Plenum Press, New York, NY (1967)
7. Privitera, C.M., Stark, L.W.: Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **22** (2000) 970–982
8. Jarodzka, H., Holmqvist, K., Nyström, M.: A Vector-Based, Multidimensional Scanpath Similarity Measure. In: ETRA '10: Proceedings of the 2010 Symposium on Eye Tracking Research & Applications, New York, NY, ACM (2010) 211–218
9. Duchowski, A.T., Driver, J., Jolaoso, S., Ramey, B.N., Tan, W., Robbins, A.: Scanpath Comparison Revisited. In: ETRA '10: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, New York, NY, ACM (2010) 219–226
10. Pomplun, M., Ritter, H., Velichkovsky, B.: Disambiguating Complex Visual Information: Towards Communication of Personal Views of a Scene. Perception **25** (1996) 931–948
11. Wooding, D.S.: Fixation Maps: Quantifying Eye-Movement Traces. In: ETRA '02: Proceedings of the 2002 Symposium on Eye Tracking Research & Applications, New York, NY, ACM (2002) 31–36
12. Hembrooke, H., Feusner, M., Gay, G.: Averaging Scan Patterns and What They Can Tell Us. In: ETRA '06: Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, New York, NY, ACM (2006) 41
13. Dempere-Marco, L., Hu, X.P., Ellis, S.M., Hansell, D.M., Yang, G.Z.: Analysis of Visual Search Patterns With EMD Metric in Normalized Anatomical Space. IEEE Transactions on Medical Imaging **25** (2006) 1011–1021
14. Torstling, A.: The Mean Gaze Path: Information Reduction and Non-Intrusive Attention Detection for Eye Tracking. Master's thesis, The Royal Institute of Technology, Stockholm, Sweden (2007) Techreport XR-EE-SB 2007:008.
15. Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., Salakoski, T.: A Comparison of AUC Estimators in Small-Sample Studies. In: Proceedings of the 3rd International workshop on Machine Learning in Systems Biology. (2009) 15–23
16. Paris, S., Durand, F.: A Fast Approximation of the Bilateral Filter using a Signal Processing Approach. Technical Report MIT-CSAIL-TR-2006-073, Massachusetts Institute of Technology (2006)
17. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **20** (1998) 1254–1259
18. Leigh, R.J., Zee, D.S.: The Neurology of Eye Movements. 2nd edn. Contemporary Neurology Series. F. A. Davis Company, Philadelphia, PA (1991)
19. Grindinger, T.J.: Event-Driven Similarity and Classification of Scanpaths. PhD thesis, Clemson University, Clemson, SC (2010)
20. Peters, R.J., Itti, L.: Computational Mechanisms for Gaze Direction in Interactive Visual Environments. In: ETRA '06: Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, New York, NY, ACM (2006) 27–32