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

Thomas J. Grindinger[†], Vidya N. Murali[†], Stephen Tetreault[‡], Andrew T. Duchowski[†], Stan T. Birchfield[†], and Pilar Orero[*]

[†]School of Computing, Clemson University,
[‡]Computer Science, Rhode Island College
[*]Traducció i Interpretació, Universitat Autònoma de Barcelona

IWGSI 2010 in conjunction with ACCV 2010,
8-12 November, Queenstown, New Zealand

Collaborators

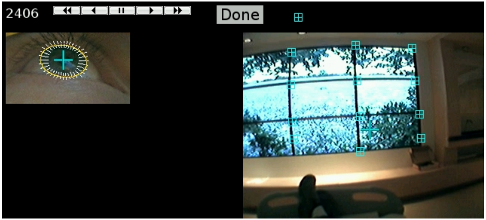
- **Thomas Grindinger:** My former PhD student on whose work this paper is based.
- **Vidya Murali:** Joined us during the summer to test collected eye movements against those predicted by the computational saliency model.
- **Stephen Tetreault:** One of my Research Experience for Undergraduates (REU) students who ported Thomas' code to C++ and helped with analysis of this study (among other things; he also implemented the Qt interface).
- **Stan Birchfield:** Prof. Birchfield is Vidya's advisor in the ECE department at Clemson.
- **Pilar Orero:** Prof. Orero is my colleague at the Universitat Autònoma de Barcelona (Traducció i Interpretació), who got me started on video and who provided the stimulus and one of the task suggestions that went with it.

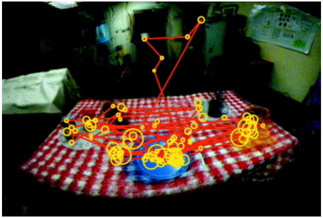
Abstract

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).

Motivation



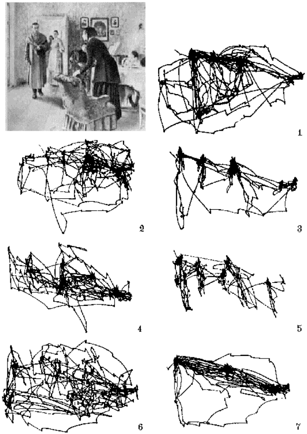


Wearable tracker (Ryan et al., 2010); making PB&J sandwiches (Land & Hayhoe, 2001).

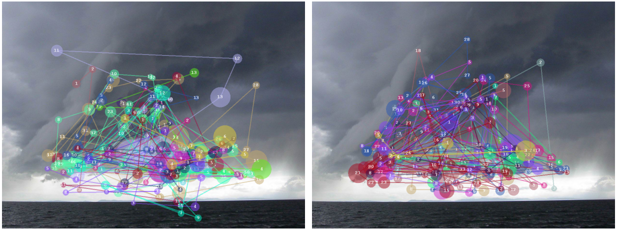
- Aggregate eye movements over video to indicate saliency
- **Classify *scanpaths* quantitatively and automatically**
- Use forward-facing camera or other video media
- Why not use saliency models, e.g., Itti et al. (1998)?
  - bottom-up vs. top-down cognition (Land & Tatler, 2009)
- **Identify *perceptually salient* image elements**

Novelty of Approach & Empirical Validation

- Eye movement classification operates over video sequences
- Approach accumulates gaze points over video frames, but can apply to any sampling rate (Grindinger et al., 2010)
- Classification distinguishes classes of viewers
- Distinction may be indicative of viewers' cognitive intent—eye movements are task-dependent as shown by Yarbus (1967)
  - **instructions influence scanpaths**



Background: Scanpath Comparison



Experts' and novices' scanpaths over weather image

- Scanpaths are vivid visualizations (Noton & Stark, 1971)
- Not yet fully exploited for quantitative potential
- Comparison can either be *content-* or *data-driven*
  - latter applied directly to (x, y, t) eye movement stream
- Two recent approaches exemplify the distinction
  - Jarodzka et al. (2010) propose a vector-based similarity measure quantizing stimulus frame 5×5 grid
  - Duchowski et al. (2010) revisit Privitera and Stark's (2000) string-editing approach operating directly on the scanpaths

Background: Heatmap Visualization

Wooding's (2002) heatmaps

- *Heatmap* visualizations, introduced by Pomplun et al. (1996), were popularized by Wooding (2002)
- Here, heatmaps are used for analysis and for visualization
- Other similar approaches include Hembrooke et al.'s (2006) "average scanpath" and scanpath distance via the Earth Mover's Distance (Dempere-Marco et al., 2006)
- Classification is similar in spirit to Torstling's (2007) machine learning classification of content

Classification Framework

- Using Airola et al.'s (2009) nomenclature, let:
  - D be a probability distribution over sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
  - $\mathcal{X}$ be the input space
  - $\mathcal{Y} = \{-1, 1\}$ be the output space

- A classifier is defined as a function $C_Z(x)$ that is used to output a set of threshold-based decisions $Z = \{z_1, \ldots, z_m\} \in \mathcal{Z}^m$, where $z_i = (x_i, y_i)$, with:
  - $X = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$, the training set of m training examples, and
  - $Y = \{y_1, \ldots, y_m\} \in \mathcal{Y}^m$, denoting the labeling of $x \in \mathcal{X}$ as a non-class ($x_- \in X_-$) or class member ($x_+ \in X_+$), respectively

Three steps to Building and Evaluating $C_Z(x)$

- First, similarity scores are extracted from X
- Second, a discrimination threshold h is computed from similarity scores so that x is assigned as member or non-member,

$$x \leftarrow \begin{cases} X_+ & \text{if } C_Z(x) > h \qquad\qquad \text{(member class)} \\ X_- & \text{otherwise} \qquad\quad \text{(non-member class)} \end{cases}$$

- 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,

$$H(x) = \begin{cases} 1 & x > 0 \\ 1/2 & x = 0 \\ 0 & x < 0 \end{cases}$$

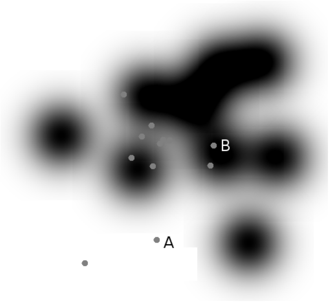
Step 1: Extract Similarity Scores

- A scanpath $s(t)$ is parametrized by timestamp t such that $s(t) = \{(i(t), j(t)) \mid t \in [t - w, t + w]\}$
- A set of scanpaths is parametrized similarly $S(t) = \{s_1(t), s_2(t), \ldots, s_m(t)\}$
- The similarity of s' to a set of classified scanpaths S (at t) is

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

with

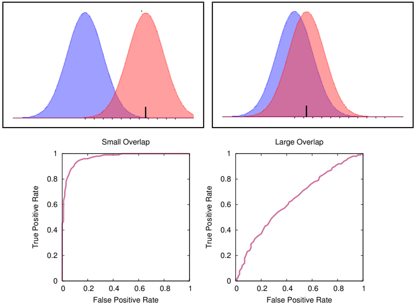
$$g(s', \mu_s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s' - \mu_s)^2}{2\sigma^2}\right)$$



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., $d(A, S) < d(B, S)$ here.

Step 2: Compute Classification Threshold h

- An unclassified scanpath is accepted by the classifier if its similarity score is higher than the computed threshold
- The threshold is selected as the value closest to $(0, 1)$ on the ROC curve, where the ratio of false positives to true positives is balanced



Small and large distribution overlap with Receiver Operating Characteristic curves

Step 3: Estimate Classifier Performance

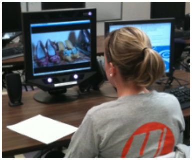
- In a small sample setting, estimate the performance, or reliability, of a classifier via cross-validation
- Leave-pair-out cross-validation, or LPOCV, is adopted since the intent is to estimate the conditional AUC, avoiding pitfalls associated with pooling and averaging of LOOCV (leave-one-out cross-validation) (Airola et al., 2009)
- With LPOCV, AUC is estimated as

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

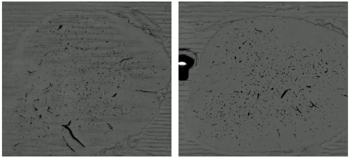
  where $X_+ \subset X$ and $X_- \subset X$ are the positive and negative examples of the training set X, and $C_{\overline{\{i,j\}}}(s_i)$ is the classifier trained without the i^{th} and j^{th} training examples
- The AUC estimate $\hat{A}(X, C_Z)$ is equivalent to the Mann-Whitney U statistic

Experimental Validation

- **Stimulus:**
  three video sequences
- **Participants:**
  27 college students,
  ages 18–21
- **Procedures:**
  free or tasked viewing
- **Apparatus:**
  Tobii ET-1750 tracker,
  17″ (1280×1024) display



Sequence A, with pair of modern sneakers



Sequence B, expected to be unfamiliar

Sequence C, with large number of faces

Computational Saliency

- Artificial gaze points over video were generated by the iLab Neuromorphic Toolkit
- The toolkit contains *ezvision* that can be executed on static images to predict human visual attention
- The model also operates on video, and forced to find a salient point within the frame's exposure duration (33 ms for *Marie Antoinette* or 40 ms for mouse video)
- Noise added to salient points simulating 27 hypothetical users

Results

	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

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

- Results show significantly consistent discriminability between perceptual (top-down) and computational (bottom-up) saliency
- Consistency refers to the evaluation of the Heaviside step function over all $m \times (m-2)$ cross-validation partitionings
- Recall estimated AUC is the Mann-Whitney U statistic
- Human scanpaths (tasked vs. free viewing) are not as easily distinguished

Heatmaps: Video Excerpts

Free and tasked scanpaths (artificial scanpaths in inset)

- Seq. A: Classifier does not discriminate human scanpaths
- Seq. B: Classifier distinguishes tasked human scanpaths
- Seq. C: Classifier shows best discriminative performance

Conclusion

- An eye movement classification algorithm was presented to distinguish scanpaths
- The algorithm successfully discriminated between perceptual and computational saliency
- Results illustrate the disparity between top-down visual processes and bottom-up computational models

Questions

- Thank you
- Comments, Questions?

Results

Cross-Validation Results

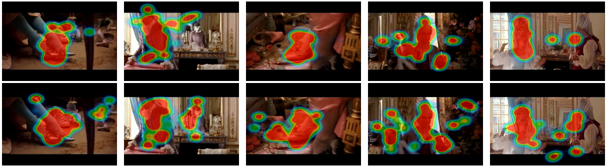
	Fixations			Gaze points		
	A	B	C	A	B	C
ACC	0.598	0.586	0.745	0.544	0.761	0.733
AUC	0.681	0.627	0.808	0.611	0.815	0.808
Z	-2.26*	-1.57	-3.77**	-1.39	-3.90**	-3.77**

* $p < 0.05$ ** $p < 0.01$ (two-tailed)

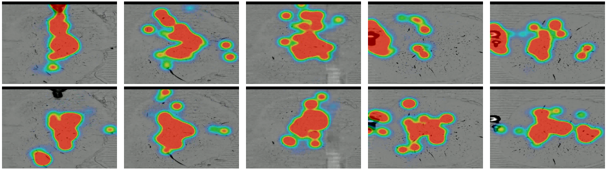
Results composed of classifier accuracy (ACC) and its conditional expected AUC for one-to-one comparison of two viewer classes for each of three video stimuli. AUC is reported with standardized score z. The left set of results pertains to fixations sampled at each video frame; the right set to accumulated gaze points.

- Results show significantly consistent discriminability between (human) tasked and free viewing over Seq. B and Seq. C, but not over Seq. A when comparing gaze points
- The difference between tasked and free viewing scanpaths over Seq. C was highly significant regardless of whether gaze points or fixations were examined

Heatmaps: Static Frames



Seq. A: free (above) and tasked viewing (below).



Seq. B: free (above) and tasked viewing (below).

Seq. C: free (above) and tasked viewing (below).

Selected References I

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., & Salakoski, T. (2009). A Comparison of AUC Estimators in Small-Sample Studies. In *Proceedings of the 3rd international workshop on machine learning in systems biology* (pp. 15–23).

Dempere-Marco, L., Hu, X.-P., Ellis, S. M., Hansell, D. M., & Yang, G.-Z. (2006, August). Analysis of Visual Search Patterns With EMD Metric in Normalized Anatomical Space. *IEEE Transactions on Medical Imaging*, *25*(8), 1011–1021.

Duchowski, A. T., Driver, J., Jolaoso, S., Ramey, B. N., Tan, W., & Robbins, A. (2010, March 22-24). Scanpath Comparison Revisited. In *Etra '10: Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 219–226). New York, NY: ACM.

Selected References II

Grindinger, T., Duchowski, A. T., & Sawyer, M. (2010, March 22-24). Group-Wise Similarity and Classification of Aggregate Scanpaths. In *ETRA '10: Proceedings of the 2010 Symposium on Eye Tracking Research & Applications* (pp. 101–104). New York, NY: ACM.

Hembrooke, H., Feusner, M., & Gay, G. (2006). Averaging Scan Patterns and What They Can Tell Us. In *ETRA '06: Proceedings of the 2006 Symposium on Eye Tracking Research & Applications* (p. 41). New York, NY: ACM.

Itti, L., Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *20*(11), 1254–1259.

Selected References III

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A Vector-Based, Multidimensional Scanpath Similarity Measure. In *ETRA '10: Proceedings of the 2010 Symposium on Eye Tracking Research & Applications* (pp. 211–218). New York, NY: ACM.

Land, M. F., & Hayhoe, M. (2001). In What Ways Do Eye Movements Contribute to Everyday Activities. *Vision Research*, *41*(25-26), 3559–3565.

Land, M. F., & Tatler, B. W. (2009). *Looking and Acting: Vision and Eye Movements in Natural Behavior*. New York, NY: Oxford University Press.

Noton, D., & Stark, L. (1971). Scanpaths in Saccadic Eye Movements While Viewing and Recognizing Patterns. *Vision Research*, *11*, 929–942.

Selected References IV

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating Complex Visual Information: Towards Communication of Personal Views of a Scene. *Perception*, *25*(8), 931–948.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *22*(9), 970–982.

Ryan, W. J., Duchowski, A. T., Vincent, E. A., & Battisto, D. (2010, March 22-24). Match-Moving for Area-Based Analysis of Eye Movements in Natural Tasks. In *Eye Tracking Research & Applications (ETRA).* Austin, TX: ACM.

Selected References V

Torstling, A. (2007). *The Mean Gaze Path: Information Reduction and Non-Intrusive Attention Detection for Eye Tracking*. Unpublished master's thesis, The Royal Institute of Technology, Stockholm, Sweden. (Techreport XR-EE-SB 2007:008)

Wooding, D. S. (2002). Fixation Maps: Quantifying Eye-Movement Traces. In *Etra '02: Proceedings of the 2002 symposium on eye tracking research & applications* (pp. 31–36). New York, NY: ACM.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York, NY: Plenum Press.