Robust Clustering of Eye Movement Recordings for Quantification of Visual Interest

Anthony Santella Doug DeCarlo

Department of Computer Science
Center for Cognitive Science
Rutgers University

Abstract

Characterizing the location and extent of a viewer's interest, in terms of eye movement recordings, informs a range of investigations in image and scene viewing. We present an automatic data-driven method for accomplishing this, which clusters visual point-of-regard (POR) measurements into gazes and regions-of-interest using the mean shift procedure. Clusters produced using this method form a structured representation of viewer interest, and at the same time are replicable and not heavily influenced by noise or outliers. Thus, they are useful in answering fine-grained questions about where and how a viewer examined an image.

Keywords: eye movement analysis, measures of visual interest, clustering, mean shift

1 Introduction

Human eye movements provide strong evidence about the location of meaningful content in an image or scene [Mackworth and Morandi 1967; Just and Carpenter 1976; Henderson and Hollingworth 1998]. And while the location of the meaningful content changes based on the viewer's task, so do their eye movements [Yarbus 1967; Just and Carpenter 1976]. This fundamental fact about human vision explains and motivates the need for algorithms that concisely quantify areas of a viewer's focus, so called regions-of-interest, in terms of a recording of their eye movements. Such methods are effective and time-saving tools for conducting psychological research on eye movements, and are a necessary ingredient in human-machine interfaces that use eye tracking technology.

This paper describes a clustering algorithm that processes point-of-regard (POR) measurements from an eye-tracker into collections that are either:

- grouped spatially as in Figure 1(d), indicating the viewer's *regions-of-interest*; or
- grouped spatially and temporally—ranging from individual *fixations* as in Figure 1(b), to *gazes*—sets of sequential fixations in a confined part of the viewing area [Just and Carpenter 1980], as in Figure 1(c).

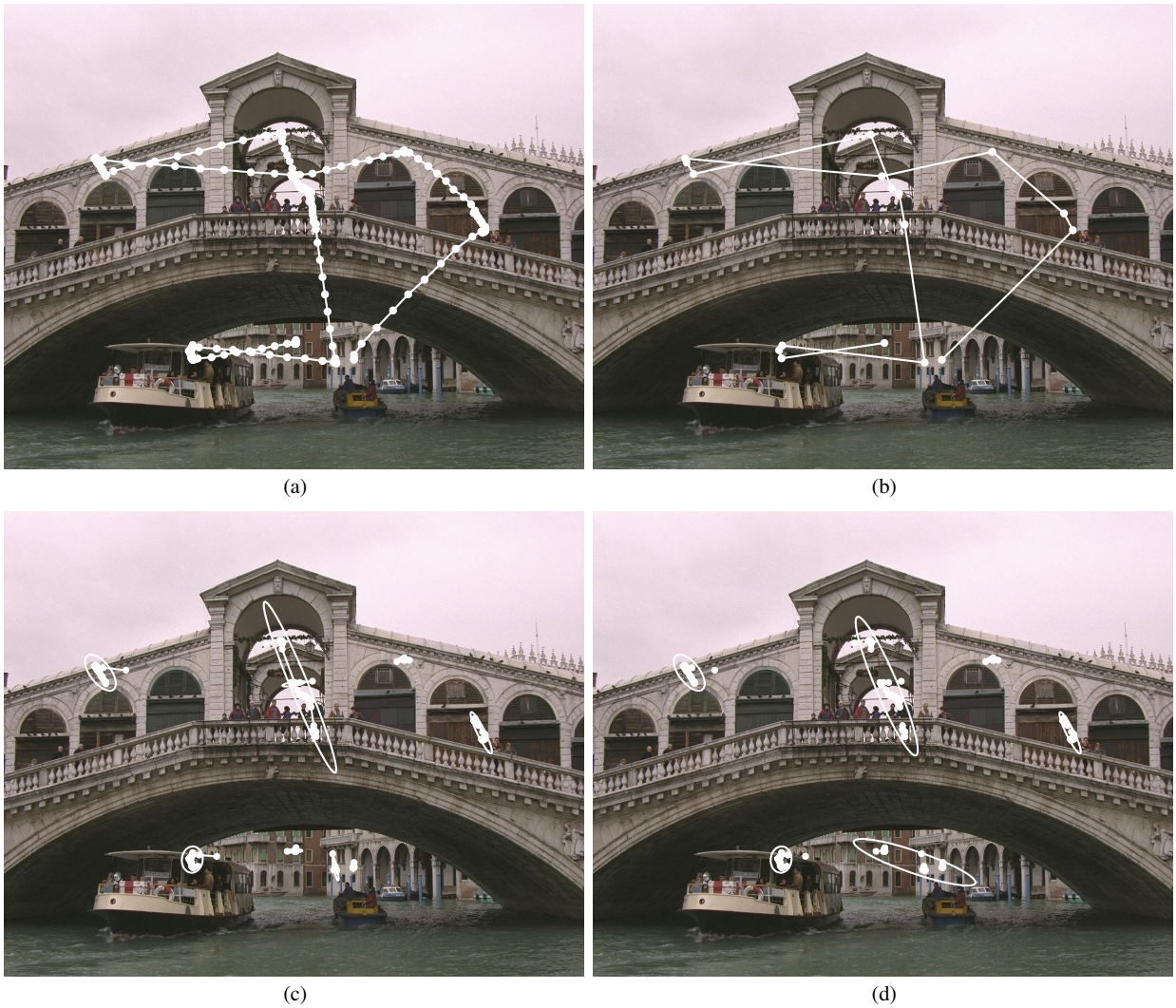
These clusters can inform investigations into higher-level questions about how an image is examined. The analysis is based upon an existing algorithm that can be used for robust clustering—the *mean shift* procedure [Fukunaga and Hostetler 1975; Comaniciu and Meer 2002]—adapted for eye movement data. This procedure decides the number and arrangement of clusters deterministically, and hence is entirely *data-driven*. It is robust because the results are not adversely affected by noise or outliers. We compare this approach to methods based on other clustering algorithms.

A wealth of different measurements of eye movements are possible [Inhoff and Radach 1998]. Measuring the location of particular objects or features that attracted the viewer's focus is among the more common objectives. Hand-coding of eye movement data to realize such measurements produces sound results (i.e. [Harris et al. 1988]); a trained observer easily parses eye movement recordings into fixations, gazes, or regions-of-interest (in the context of a particular experiment). The time-consuming nature of manual coding, however, has led to the development of automated approaches.

The most common automatic techniques simply extract fixations by removing saccades from a set of point-of-regard measurements. The constrained nature of eye movements invites simple algorithms based on applying velocity thresholds and enforcing lower-bounds on fixation durations. These algorithms are quite effective in many situations—see [Salvucci and Goldberg 2000] for a thorough description and comparative evaluation. Careful modeling of the processes involved can lead to more robust estimation procedures. For instance, modeling eye movements using a state variable that declares each moment as either a fixation or saccade leads to the use of Hidden Markov models (HMMs) as an estimation tool [Salvucci and Goldberg 2000]. The success of all of these algorithms comes from their dependence on the structure of fixations and saccades; but this dependence is also what prevents their more general application in detecting higher-level features like regions-of-interest.

Higher-level measurements of viewer interest have remained, on the whole, relatively simple. This hasn't been a serious limitation—the design of psychological experiments is often guided by the need to take simple measurements. For instance, one methodology divides the viewing area into a regular grid, and tallies the time spent inside each square [Mackworth and Morandi 1967]. Another common technique defines rectangular target regions (also described in [Salvucci and Goldberg 2000]), and records the fixations by the viewer inside each region. Such a method is well suited to situations such as reading, where each word is bounded by a rectangle [Just and Carpenter 1980]. (Several commercial eye tracking systems provide this functionality.)

Data-driven algorithms work in terms of the point-of-regard locations themselves to define the regions-of-interest; these are more flexible than measurements performed using a pre-defined grid or box, and hence admit fine-grained analysis. Data-driven algorithms have seen little deployment in the analysis of eye movements. But they are required when there is no a priori structure that describes the stimulus. In these cases, they can be used to simultaneously locate areas of importance and quantify interest in them.



(a) (b) (c) (d)

Figure 1: Clustering example (a) original image with recorded point-of-regard (POR) locations; (b) with estimated fixation locations (connected by lines); (c) with gazes (with lines connecting POR locations within each gaze); (d) with regions-of-interest (with POR)

One common data-driven method of clustering uses a distance threshold, and considers two points to be in the same cluster when they are closer than this distance. A specific number of clusters is obtained by simply choosing the appropriate distance threshold that realizes it. For purposes of comparing real and simulated fixation sequences, Privitera and Stark [2000] use such a method, choosing their threshold via k-means clustering [Duda et al. 2001]. The clusters provide a means to relate the real and simulated fixations. [Turano et al. 2003] proceeds similarly, but clusters quantities that measure performance of salience models for purposes of comparison (and not the fixation locations themselves). While this simple clustering algorithm is sufficient in situations where the clusters are clearly delineated, its performance in more general settings would be quite poor. Distance thresholds fail to resolve two dense but separate clusters if only two of their points happen to be close. As a result, they cannot finely characterize gazes or regions-of-interest.

More intricate algorithms exist that measure the regions-of-interest. Latimer [1988] partitions data by forming a histogram of fixation durations over the viewing area, and finds clusters of that histogram using k-means. This approach is less sensitive to noise. Latimer also presents an more intricate approach based on an information theoretic criterion. Latimer notes the difficulty in producing consistent results automatically with k-means and similar algorithms. (We discuss this further in Section 3.1.)

Clustering methods have also been used to extract gazes: spatial clusters of successive fixations. This is an easier problem, akin to identifying fixations, so simpler techniques are reasonably effective. Successive fixations can be added to a cluster until the next is too distant [Scinto and Barnette 1986]. Applying a threshold on the distance between the next fixation location and the mean of fixation points already in the cluster is more effective [Nodine et al. 1992]. This technique tolerates points leading away from the cluster (which was one difficulty with using simple thresholds). One might imagine adapting this method for regions-of-interest by choosing the spatially, rather than temporally closest neighbor as the next candidate for addition. However, this would introduce a dependence on the order of the fixations—an undesirable property for data analysis.

Other techniques provide *visualizations* of regions-of-interest by adapting the original imagery. Latimer [1988] plots the distribution of fixations over stimuli. Wooding [2002] presents a technique which darkens the image in uninteresting areas with a function defined using a mixture of Gaussians, each centered at a fixation. Other systems create stylized transformations of images by modulating the level-of-detail in the image based on where the viewer looked [DeCarlo and Santella 2002]. Indeed, such notions form the basis of gaze-contingent displays that omit unnoticeable detail [Baudisch et al. 2003].

In the next section, we start by describing our data-driven clustering method. This is followed in Section 3 by a demonstration of our method, and a comparison of our clustering strategy with alternative methods.

2 Clustering eye movements

When an expert encodes gazes or regions-of-interest by hand, they group nearby fixations together—they cluster the data. Although this is a deceptively easy task for a human observer, it is a difficult undertaking for a computer. There are many available algorithms for clustering [Duda et al. 2001] that can automate this process. A desirable algorithm for clustering eye movement data has three characteristics:

- it should produce consistent results (and not depend on a random initial guess);
- it should not need to know the number of clusters in advance;
- it should be robust (so the presence of outliers does not significantly affect distant clusters).

Commonly used techniques such as Expectation-Maximization (EM) and k-means clustering [Duda et al. 2001] are initialized randomly, and so different runs on the same data can produce different local minima. Both EM and k-means also require knowing the number of clusters in advance. While adaptations to these algorithms exist that aim to address these issues, they can only reduce the problem. Such arbitrary behavior would detrimentally affect any experimental analysis, and must be avoided.

Additionally, the technique should be robust. All of the places a viewer fixates are not really regions-of-interest. Sustained viewing will produce clusters of fixations around particularly interesting features, jittered by noise and microsaccades, this is the kind of noise all clustering methods seek to address. However, brief isolated fixations will also be present. These are outliers of the regions-of-interest—they are not just noise. They are isolated points that commanded some focus, perhaps just a distraction of some sort. Characterizing them might be valuable in particular situations. We want results for the major clusters to be unaffected by their presence, but also we want to describe these outliers, in order to analyze or discard them. Here again, EM and k-means are not robust. While their global solution characterizes outliers correctly, the local solutions provided by these algorithms in practice can include outliers inside larger clusters, which distorts their statistics.

2.1 Mean shift

Clustering based on the mean shift procedure meets the requirements above—it proceeds deterministically, does not require knowing the number of clusters in advance, and is robust to outliers. It has its origins in the pattern recognition community [Fukunaga and Hostetler 1975], but has only seen recent use—particularly in robust computer vision systems [Comaniciu and Meer 2002].

In general, a clustering process starts from a set of N points:

$$\{\mathbf{x}_j \mid j \in 1..N\},$$

and assigns one of K labels to each point:

$$\{c_j \in 1..K \mid j \in 1..N\}.$$

Some algorithms produce fuzzy labelings, and may also provide coarse descriptions such as means and covariances.

Perhaps the simplest clustering algorithm employs a distance threshold d, and considers any two points to be part of the same cluster when they are within distance d of each other. While this algorithm has the desirable property of not requiring the number of clusters to be known in advance, it is very fragile; any noise in the data causes apparently distinct groups of points to be clustered together.

Even so, this algorithm is salvageable. Mean shift clustering is a robust version of distance-based clustering that includes a preprocessing stage. The entire process involves two steps:

1. move the points into denser configurations until they can be easily separated into clusters
2. apply a clustering algorithm that employs a distance threshold

The first stage—known as the mean shift procedure—is crucial, as it makes the entire process robust. It proceeds by repeatedly moving a point $\mathbf{x}_i$ to a new location $\mathbf{s}(\mathbf{x}_i)$—the weighted mean of nearby points based on the kernel function k:

$$\mathbf{s}(\mathbf{x}) = \frac{\sum_j k\left(\mathbf{x}-\mathbf{x}_j\right)\,\mathbf{x}_j}{\sum_j k\left(\mathbf{x}-\mathbf{x}_j\right)} \tag{1}$$

The kernel k is typically a multivariate Gaussian with zero mean and covariance $\sigma^2\mathbf{I}$ [Comaniciu and Meer 2002]. (Note that this description of the kernel has been streamlined from [Comaniciu and Meer 2002].) Robustness to extreme outliers is achieved by simply limiting the support of the kernel, for instance by setting it to zero for distances greater than 2σ.

The parameter σ describes the spatial extent of the weighted mean computed by $\mathbf{s}$: it provides a *scale* control for the algorithm. For example, increasing the value of σ results in fewer, larger clusters which describe coarser structures in the data. Like the method using a simple distance threshold, the mean shift method is nonparametric: it make no assumptions about the global shape of the distribution of the data. This is typically an advantage, as such assumptions lead to garbled results when they do not apply.

The mean shift procedure proceeds as follows:

INITIALIZATION for $j \in 1..N$
$\mathbf{y}_j^0 = \mathbf{x}_j$

ITERATION n repeat until convergence
for $j \in 1..N$
$\mathbf{y}_j^n = \mathbf{s}(\mathbf{y}_j^{n-1})$

It works by interpreting the points $\mathbf{x}_j$ as samples from a distribution. The mean shift property, established in [Fukunaga and Hostetler 1975], estimates the gradient of the density of these samples at $\mathbf{x}$ as $\mathbf{s}(\mathbf{x}) - \mathbf{x}$. Hence, the mean shift procedure iteratively moves all points simultaneously towards locations of higher density (in the gradient direction): towards the eventual points of convergence, which are the *modes* of the distribution. (Proofs concerning convergence behavior are in [Comaniciu and Meer 2002].)

With all of the points collected at modes, a clustering method that uses a distance threshold is both safe and successful. In this

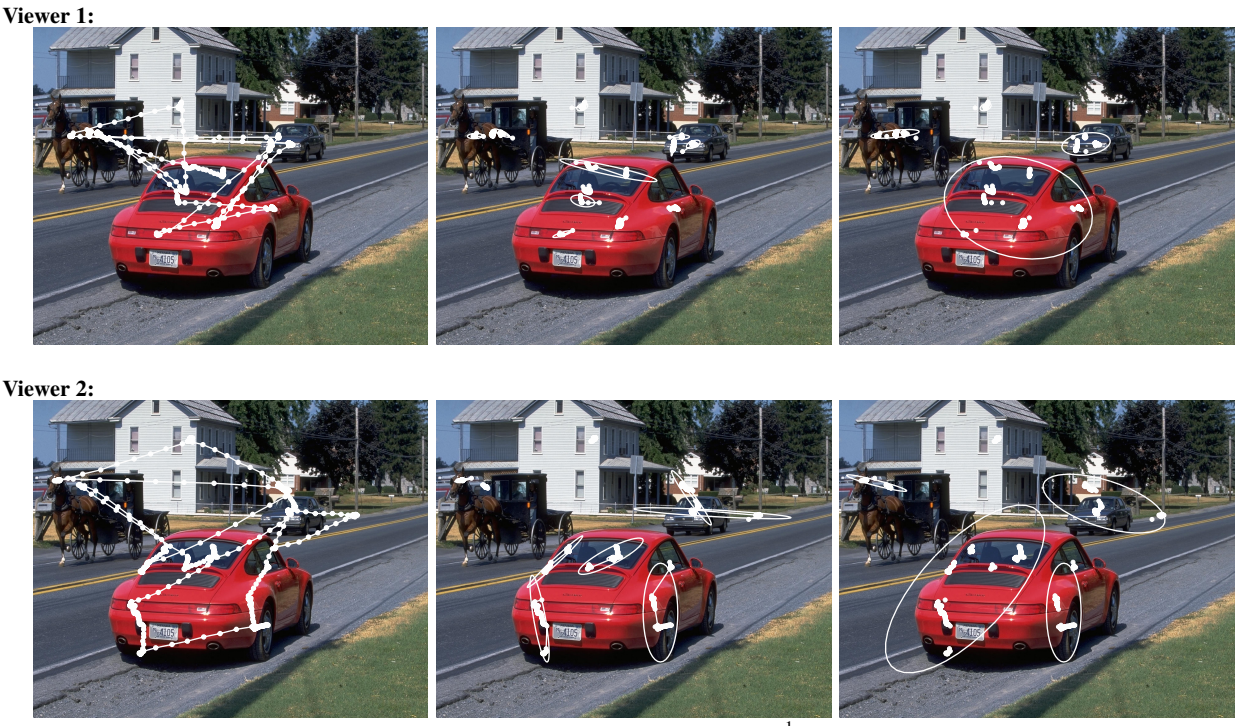


Figure 2: Gaze and region-of-interest clusters for two different viewers

case, a distance threshold of σ is used—the scale parameter used to specify the kernel. Finally, clusters containing a very small fraction of the data are optionally discarded, as they are typically outliers.

2.2 Formulations for POR data

Eye-trackers provide the location of a point on a particular planar surface (calibrated in advance), such as the pixel location on a screen, and the time the measurement was taken. In this case, $\mathbf{x}_i = (x_i, y_i, t_i)$. Clustering data spatially and temporally produces clusters that describe fixations, or gaze points (groups of nearby, sequential fixations), while spatial clustering produces regions-of-interest. Both easily fit within the framework described in Section 2.1—the kernel function k from (1) enables this control.

A zero-mean Gaussian kernel is specified using a covariance matrix—this matrix encodes the relative weighting of the dimensions when measuring distances. Diagonal matrices provide sufficient flexibility here, but independent control over the relative scale between spatial and temporal dimensions is important. A small temporal scale is called for when clustering points temporally (so as not to cluster across non-consecutive fixations), but the spatial scale is really application dependent. Using a spatial scale of σ_s and temporal scale of σ_t, the kernel k is:

$$k_{\text{spatiotemporal}}\left([x_i, y_i, t_i]\right) = \exp\left(-\frac{x_i^2 + y_i^2}{\sigma_s^2} - \frac{t_i^2}{\sigma_t^2}\right)$$

(Note that the kernel does not need to be normalized given the denominator in (1).) For an analysis that ignores temporal information, we can simply exclude the temporal dimension (effectively setting σ_t to infinity):

$$k_{\text{spatial}}\left([x_i, y_i]\right) = \exp\left(-\frac{x_i^2 + y_i^2}{\sigma_s^2}\right)$$

The spatial scale parameter σ_s can be varied to produce predictable clusters on coarser or finer scales. For instance, one can determine if a viewer looked at the eye of a person, or merely whether they looked at the entire face. A particular spatial scale choice does not control the size of the cluster itself, it ensures that no clusters exist which are closer than σ_s. Normally a σ_s below the distance between two image features will resolve them into separate clusters, assuming there is a drop-off in density in the space between them. If both features are part of a unimodal blob of interest it will be impossible to resolve them. Given a particular experimental setup and scale of question being posed, a single scale value will most likely be sufficient for all data across a set of stimuli.

The value of the temporal scale σ_t has more to do with the average time separating fixations (we use a value of $1/3$ second). When clustering spatially and temporally, it is possible for non-consecutive POR data to be clustered together. Although we have seen no instances where this has occurred, these anomalies can easily be corrected during the second stage of mean shift clustering.

Given the choice of scale, the results of this algorithm are very intuitive and predictable. Distant outliers (where distant is defined by the scale choice) will become small clusters. Large regions of uniform density will collapse to one cluster. Where density is locally bimodal, one cluster will result if the modes are within the chosen scale of each other, and otherwise two clusters will result. (It is worth contrasting this behavior which depends on the distance of modes, with distance thresholding as in [Privitera and Stark 2000; Turano et al. 2003] which will collapse the clusters if any of their points are within the threshold distance of each other.) Mean shift

provides clusters of interest that are robust to both measurement noise and small outlying points-of-regard, and provides a consistent and interpretable estimate of the regions-of-interest. The results are replicable, and similar shaped distributions of POR data will produce similar clusters, making direct comparison possible.

This method should be particularly useful in characterizing the regions-of-interest in aggregate data from a number of viewers (over the same image). This is an interesting approach for trying to capture general, viewer independent patterns of interest over an image. In this case, data from multiple subjects is collapsed together. Robust clustering will extract patterns in the presence of noise and variations in calibration. In Section 3 we show some results of our method both on individual eye tracks and aggregate data, and argue their value as a quantitative measure.

2.3 Computational concerns

For off-line analysis of data, computation time is not critical. However this technique could be useful for identifying ROI in interactive systems—for example, in computer assisted examination of radiology slides [Mello-Thoms et al. 2002]. Here, computational concerns are more important. Though mean shift clustering is somewhat expensive, it is possible to adapt it to perform in near real-time on eye movement data, under a range of circumstances.

One iteration of the mean shift procedure takes $O(N^2)$. Although the precise behavior depends on the data, we have empirically determined that 5 to 10 iterations of the mean shift procedure are sufficient for convergence to within a 0.1% change across iterations. In this case, one minute of POR data ($N = 3600$) is processed in about 3 seconds on a 2.4GHz Pentium 4 PC. Full convergence is not necessary, however. What matters is that slightly overlapping clusters are sufficiently separated. This should be tuned for a particular application.

The limits on the spatial and temporal extents of the kernel (discussed in Section 2.1) suggest that only the closest points are required for the computation of $\mathbf{s}$; the use of a spatial hierarchy, or the use of spatial coherence across iterations would reduce the running time by a sizable constant.

But the structure of POR data lends itself to a particular optimization: the fixations are grouped quite tightly together. When determining gaze points or regions-of-interest, one approximation replaces each cluster of POR data representing a fixation, with a single point that is located at the mode of the cluster (the convergence point), weighted by the number of points contained in that fixation. This produces a set of fixations $\mathbf{f}_j$ and weights w_j, resulting in a weighted version of (1):

$$\mathbf{s}_{\text{fix}}(\mathbf{f}) = \frac{\sum_j w_j k\left(\mathbf{f}-\mathbf{f}_j\right)\ \mathbf{f}_j}{\sum_j w_j k\left(\mathbf{f}-\mathbf{f}_j\right)} \tag{2}$$

Iterating over fixations makes N effectively much smaller, and the quadratic complexity becomes more manageable. In this case, one minute of fixation data (with $N_{\text{fix}} = 190$) is processed in a under one tenth of a second. For region-of-interest detection, this approximation is exact when the original POR points within fixations are perfectly aligned. Noise and drift in fixation location measurements, as well as the temporal averaging that takes place for gaze detection are the sources of error.

3 Results

This section demonstrates the clustering algorithm from Section 2 on a set of images viewed by either a single viewer or a group of viewers. The data was picked at random from recordings of naive subjects who viewed the images as part of a larger experiment. The eye movement recordings were performed using an ISCAN ETL-500 table-top eye-tracker (with a RK-464 pan/tilt camera). The resulting point-of-regard data is expressed in pixel coordinates (the images measure 1024 units horizontally; the forthcoming values of σ_s are therefore expressed in pixel units). First, the POR data was pre-processed to remove saccades using the dispersion-based method of Widdel [1984] (also see [Salvucci and Goldberg 2000]). We have also successfully used a threshold on velocity between point samples to perform this filtering.

The examples here are representative of typical results of the proposed algorithm. The images in Figure 2 demonstrate gaze and region-of-interest clustering for two different viewers of the same image. On the left is the POR data superimposed on the image (recorded for 8 seconds). At center are gaze clusters, and at the right are regions-of-interest for this image. These locations, on the whole, correspond to objects and groups of objects in the image. The value of σ_t was $1/3$ second for the each of these examples.

Figure 3: The image viewed by the tracking subjects

The next set of examples involve viewings of the image in Figure 3, and demonstrate region-of-interest clustering. The clustering results are presented by drawing each cluster using a different color. (In examples with many clusters, there will be color re-use; and there is no color correspondence across related examples.) As some of the colors may appear similar on different displays, we additionally overlay covariance ellipses on large clusters (one for each cluster that contains more than 2.5% of the total data); they are centered at the mean of the cluster, and mark 99% of the variance. Data points in small clusters are colored grey. To show alignment with the viewed image, the clusters are drawn over a set of linear features that were extracted (semi-automatically) from the photograph in Figure 3.

The analysis demonstrates the algorithm on data from two individual viewers, as well as on data combined from six separate viewers (all were 8 second recordings). Figure 4 shows the regions-of-interest gathered from two viewers, processed at three different spatial resolutions (σ_s). Small values of σ_s correspond largely to isolated fixations, grouping together only fixations that overlapped due to re-fixations of a particular feature. Clusters produced using larger values of σ_s capture coherent regions that the viewer examined repeatedly (and not necessarily sequentially, as with gaze clusters). Comparisons between viewers seem quite reasonable with the coarser scale clusters.

Similar behavior of the algorithm is seen for data collected together from six different viewers, which is displayed in Figure 5.

As expected, the patterns of information in the image become apparent—especially as σ_s is increased. Major clusters correspond to both salient features in the POR data, to features in the single user data in Figure 4, and for the most part, to features in the image. (This can be confirmed by comparing the cluster locations with the original image in Figure 3.)

It should be noted that not all clusters correspond to regions of significant interest. Particularly when data from multiple viewers is combined, it is the clusters that hold the largest percentage of the data that define the most important allocation of interest. This is because important features are likely to be viewed longer, and by more of the viewers. Very small clusters (colored in grey in the examples) are outliers: they represent brief fixations of some isolated feature. In some contexts, these may be of interest (in a study of distracting features, for instance). In others, they can be discarded. In either case, they are easily identified by the fact that they contain a tiny proportion of the data.

Figure 6 repeats the computations from Figure 5, but using the approximate solution from Section 2.3 that uses weighted fixations in place of the raw POR data. The similarity in the output between the approximation and full solution suggests this approximation holds up well in these conditions.

3.1 Comparison of clustering methods

Eye movement data is complex and noisy, both in terms of measurement and individual behavior. Other clustering algorithms such as k-means clustering or EM [Duda et al. 2001] tend to produce erratic results in this case. The sources of such irregular behavior include the choice of the number of clusters (K), randomization in the initial guess, and statistical properties of the data. Strategies for determining an optimal K exist, and may involve running EM many times (across a range of K values), and selecting the best fit. However, many difficulties remain. These are algorithms with a significant random component which makes the clusters difficult to reproduce: running the algorithm multiple times on the same data can produce significantly different results. To combat this, these algorithms are in practice always run multiple times with different random initializations and the best result selected. This encourages convergence to a global minimum, though it provides no guarantees.

Figure 7 shows three different results from such an EM based algorithm (which tested a range of values of K, using 5 different initializations for each, and selected the best result): it produced $K = 6$ once and $K = 7$ twice. The randomness inherent in the algorithm clearly shows through—different runs produced quite different answers. More initializations would make the results more consistent at increasing expense, with no principled way of knowing how many are sufficient.

We can compare the output from Expectation-Maximization, shown in Figure 7, with the mean shift clustering (of the same data) in Figure 5 (specifically, when $\sigma_s = 75$). Because these EM estimates represent local minima, the clusters do not always correspond to features or regions in the image as consistently as the mean shifted clusters do. Similar effects are observed when using k-means clustering. The most obvious failure in the EM results is robustness. This can be seen in the leftmost cluster in Figure 7 in both the $K = 6$ result and the second $K = 7$ result. Such clusters contain distant fixations (outliers), and are substantially distorted. Again, more random initializations may improve the results, but there are no guarantees.

All of this seriously impairs the interpretability and the usability of the EM clusters. More consistent and robust clusters are necessary to answer fine-grained questions, such as whether two viewers had similar interest in some set of image features.

4 Discussion and Conclusion

The clustering method presented here draws upon robust methods that have deterministic behavior. It is simple to understand and implement. As a result, we expect it will become a useful tool for detecting and analyzing the regions-of-interest of a viewer; both for use in experimental analysis, and to automate content decisions in interactive systems using eye trackers. This approach applies most in situations involving unconstrained image or scene viewing. More traditional approaches are still in order for specific domains, such as reading (which successfully employs boxes around words).

We have not performed a formal evaluation of this algorithm as it applies to analyzing eye movement recordings. Comparing the results to hand-labeled data (using a range of algorithms, including mean-shift) is the most obvious course of action. Alternatively, validation can proceed by replicating known effects established by other means. In this case, the benefits of robustness and order-independence would be demonstrated should lower variances be seen in measuring these effects (when compared to analyses using algorithms such as EM).

Improvements to the proposed clustering algorithm can come on a number of fronts. Automatic methods for scale-selection [Lindeberg 1994] would choose optimal values of σ_s—the extracted clusters would be very stable under a range of scale values around the selected value. Varying σ_s over the extent of the viewing area could accommodate stimuli containing varying scales. Formulating an on-line approximate version of the mean shift procedure, perhaps taking advantage of the structure in POR data, would further facilitate its introduction into interactive systems.

Perhaps most interesting are the new experimental protocols and interactive systems that will become possible with such an automated technique. The method seems to have the sensitivity and robustness to be applied to a wide variety of problems. The locations and properties of the estimated clusters can be used to compare the similarity of interest across viewers, tasks, or different versions of images. This could be done spatially by comparing the cluster centroids produced by different viewers or groups of viewers. Or, clusters could be used to label fixations for a string matching comparison [Privitera and Stark 2000]. The clusters themselves can also be analyzed further. The relative number of data points inside a cluster indicates its importance. We can examine the percentages of different viewers POR samples which are included in a cluster, and so determine if the location is of general interest or only interesting to one or a few idiosyncratic viewers. Alternatively the time stamps of gazes can be used to determine the percentage of interest that represents initial or reoccurring fixations. This flexibility should make the method a useful technique for quantifying visual interest.

Acknowledgments

Thanks to Andrew Duchowski, Eileen Kowler and Peter Meer. This research was supported by the NSF (SGER 0227337).

References

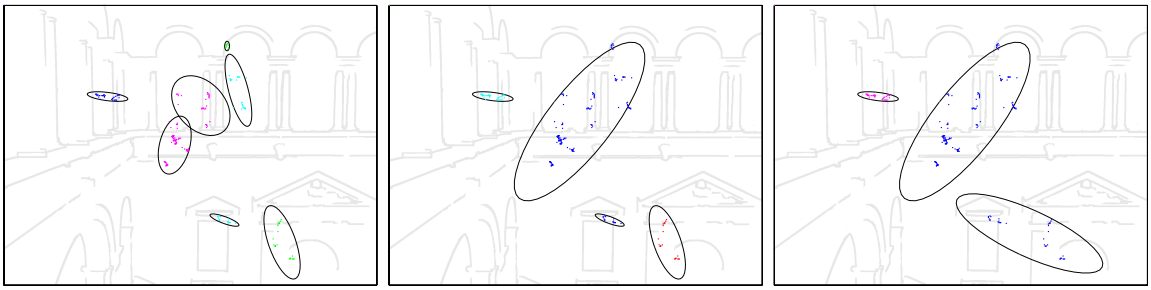
BAUDISCH, P., DECARLO, D., DUCHOWSKI, A. T., AND GEISLER, W. S. 2003. Focusing on the essential: Considering attention in display design. *Communications of the ACM 46*, 3, 60–66.

COMANICIU, D., AND MEER, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 5, 603–619.

DECARLO, D., AND SANTELLA, A. 2002. Stylization and abstraction of photographs. In *Proceedings of ACM SIGGRAPH 2002*, 769–776.

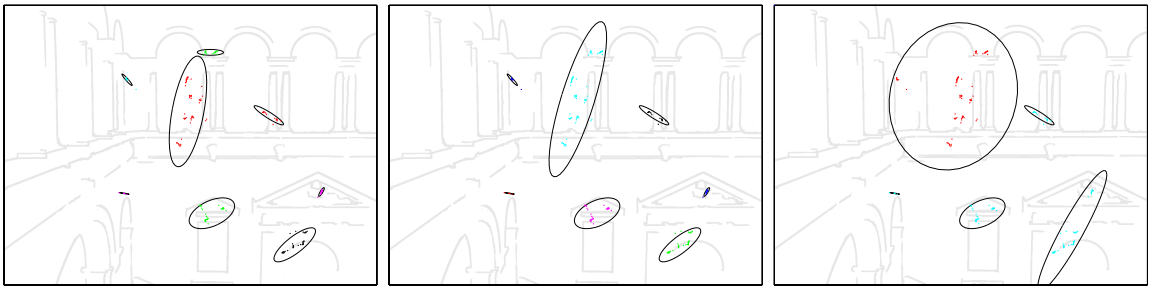
DUDA, R., HART, P., AND STORK, D. 2001. *Pattern Classification*. Wiley.

Viewer 1:



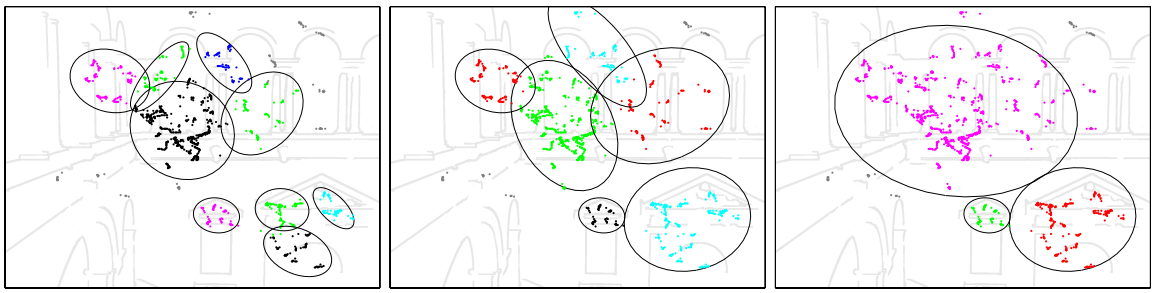
$\sigma_s = 50$ $\sigma_s = 75$ $\sigma_s = 100$

Viewer 2:



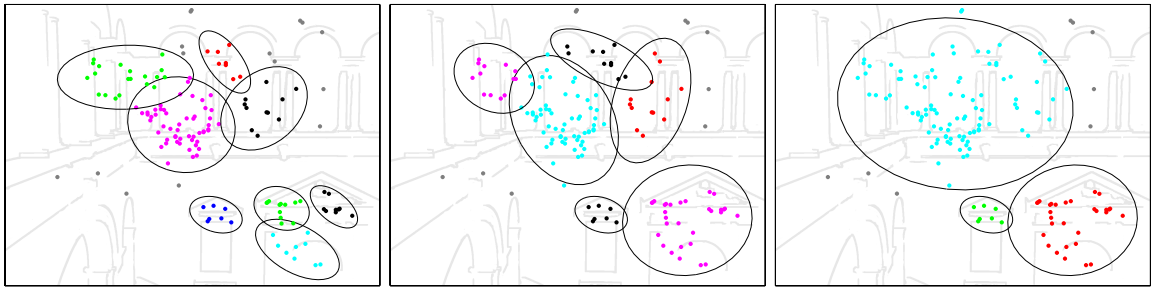
$\sigma_s = 50$ $\sigma_s = 75$ $\sigma_s = 100$

Figure 4: Region-of-interest clusters (analyzed at varying spatial scales) for recordings of two different viewers



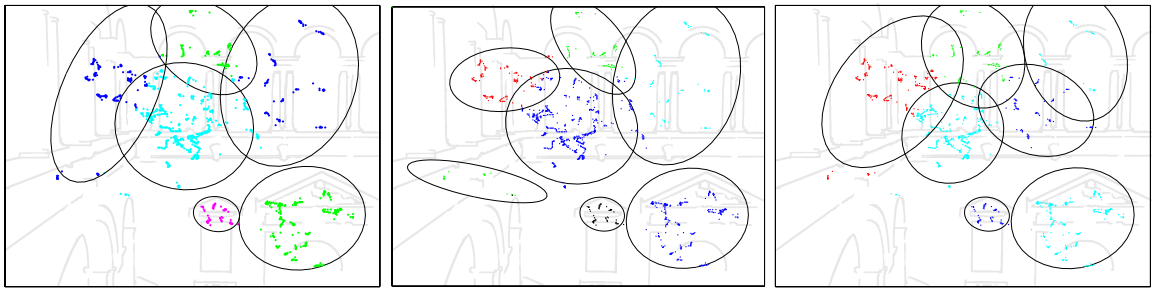
$\sigma_s = 50$ $\sigma_s = 75$ $\sigma_s = 100$

Figure 5: Region-of-interest clusters (analyzed at varying spatial scales) for recordings lumped together from six viewers



$\sigma_s = 50$ $\sigma_s = 75$ $\sigma_s = 100$

Figure 6: Region-of-interest clusters using fixation approximation in (2), for multiple viewer data (at varying spatial scales)



$K = 6$ $K = 7$ (first run) $K = 7$ (second run)

Figure 7: Region-of-interest clusters computed using Expectation-Maximization clustering (with automatic K selection)

FUKUNAGA, K., AND HOSTETLER, L. D. 1975. The estimation of the gradient of a density function, with applications to pattern recognition. *IEEE Trans. Information Theory*, 32–40.

HARRIS, C., HAINLINE, L., ABRAMOV, I., LEMERISE, E. A., AND CAMENZULI, C. 1988. The distribution of fixation durations in infants and naive adults. *Vision Research 28*, 419–432.

HENDERSON, J. M., AND HOLLINGWORTH, A. 1998. Eye movements during scene viewing: An overview. In *Eye Guidance in Reading and Scene Perception*, G. Underwood, Ed. Elsevier Science Ltd., 269–293.

INHOFF, A. W., AND RADACH, R. 1998. Definition and computation of oculomotor measures in the study of cognitive processes. In *Eye Guidance in Reading and Scene Perception*, G. Underwood, Ed. Elsevier Science Ltd., 29–53.

JUST, M. A., AND CARPENTER, P. A. 1976. Eye fixations and cognitive processes. *Cognitive Psychology 8*, 441–480.

JUST, M. A., AND CARPENTER, P. A. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review 87*, 329–354.

LATIMER, C. R. 1988. Cumulative fixation time and cluster analysis. *Behavior Research Methods, Instruments, and Computers 20*, 437–470.

LINDEBERG, T. 1994. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers.

MACKWORTH, N., AND MORANDI, A. 1967. The gaze selects informative details within pictures. *Perception and Psychophysics 2*, 547–552.

MELLO-THOMS, C., NODINE, C. F., AND KUNDEL, H. L. 2002. What attracts the eye to the location of missed and reported breast cancers? In *Proceedings of the Eye Tracking Research and Applications (ETRA) Symposium 2002*, 111–117.

NODINE, C. F., KUNDEL, H. L., TOTO, L. C., AND KRUPINSKI, E. A. 1992. Recording and analyzing eye-position data using a microcomputer workstation. *Behavior Research Methods, Instruments, and Computers 24*, 475–485.

PRIVITERA, C. M., AND STARK, L. W. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 9, 970–982.

SALVUCCI, D., AND GOLDBERG, J. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications (ETRA) Symposium 2000*, 71–78.

SCINTO, L. F. M., AND BARNETTE, B. D. 1986. An algorithm for determining clusters, pairs or singletons in eye-movement scan-path records. *Behavior Research Methods, Instruments, and Computers 18*, 41–44.

TURANO, K. A., GERUSCHAT, D. R., AND BAKER, F. H. 2003. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research 43*, 333–346.

WIDDEL, H. 1984. Operational problems in analysing eye movements. In *Theoretical and applied aspects of eye movement research*, A. G. Gale and F. Johnson, Eds. Elsevier Science Ltd., 21–29.

WOODING, D. S. 2002. Fixation maps: quantifying eye-movement traces. In *Proceedings of the Eye Tracking Research and Applications (ETRA) Symposium 2002*, 31–36.

YARBUS, A. L. 1967. *Eye Movements and Vision*. Plenum Press.