Published in 10th ACM Symposium on Eye Tracking Research and Applications, ETRA 2018. Warsaw: Association for Computing Machinery, 68. ISBN 978-1-4503-5706-7. http://dx.doi.org/10.1145/3204493.3204591

Useful Approaches to Exploratory Analysis of Gaze Data: Enhanced Heatmaps, Cluster Maps, and Transition Maps

Poika Isokoski
University of Tampere,
Faculty of Communication Sciences
Tampere, Finland
poika.isokoski@uta.fi

Jari Kangas
University of Tampere,
Faculty of Communication Sciences
Tampere, Finland
jari.kangas@uta.fi

Päivi Majaranta
University of Tampere,
Faculty of Natural Sciences
Tampere, Finland
paivi.majaranta@uta.fi

ABSTRACT

Exploratory analysis of gaze data requires methods that make it possible to process large amounts of data while minimizing human labor. The conventional approach in exploring gaze data is to construct heatmap visualizations. While simple and intuitive, conventional heatmaps do not clearly indicate differences between groups of viewers or give estimates for the repeatability (i.e., which parts of the heatmap would look similar if the data were collected again). We discuss *difference maps* and *significance maps* that answer to these needs. In addition we describe methods based on automatic clustering that allow us to achieve similar results with cluster *observation maps* and *transition maps*. As demonstrated with our example data, these methods are effective in highlighting the strongest differences between groups more effectively than conventional heatmaps. .[1]

CCS CONCEPTS

• **Human-centered computing → Interaction techniques**; **Heat maps**; *Empirical studies in HCI*; *Visual analytics*;

KEYWORDS

Eye Tracking, inter-group differences, Visualization, t-test, clustering, difference map, significance map

1 INTRODUCTION

The origins of exploratory data analysis are often tracked to Tukey's work, much of which is summarized in the 1977 book [Tukey 1977]. Tukey avoided a clear definition for the term possibly because it can have multiple meaningful uses in different contexts [Brillinger 2011]. In this paper we understand exploratory data analysis as a process where the goal is to understand the collected data as well as possible with the aim of constructing hypotheses for future experiments. After Tukey's pioneering work, a number of gaze data visualization methods have been developed, for a survey, see e.g. [Blascheck et al. 2017] or [Coutrot et al. 2018].

We will describe and discuss the use of two methods, heatmaps and clustering, in the context of a gaze data set collected from Finnish and Korean viewers while they completed a novel experimental task. The task was to write a short description of what was seen while viewing photos on a computer display. While heatmaps and clustering of gaze data are not novel as such, some of the visualizations that we constructed such as temporally segmented significance maps and highlighted cluster transition maps have not been described before in gaze data context.

A central theme in this paper is that ordinary PCs have computing power and memory resources to work with computations that were previously unimaginable (see e.g. the work by [Duchowski et al. 2012]). Using a large number of bitmaps with millions of pixels for computation and running clustering algorithms on hundreds of thousands of gaze points is no longer prohibitively expensive. Because such computation can save human labor, it should be utilized to the fullest.

2 HEATMAPS

One of the first heatmaps for visualizing gaze data was produced by [Pomplun et al. 1996]. Their goal was to visualize the area where viewers directed their visual attention in picture viewing. Consequently, they called their visualizations attentional landscapes. Such landscape is formed by dropping Gaussian distributions at the location of fixations and summing them up. The result is a landscape with the highest peak where the most fixations landed and possibly other peaks in other parts of the scene.

The attentional landscape can be visualized as a 3D height field or by modifying color or transparency of a layer on top of the original scene. Modern operating systems and programming environments include tools for creating images with layers that can vary in transparency. Thus the most straight-forward method of implementing attention maps is to utilize multiple layers of bitmaps with the top layers partially transparent. Thus, instead of following the luminance terminology of Pomplun et al. in this paper we will talk about transparency.

The approach we took was to consider a pixel in heatmap as the basic unit of computation and see what useful metrics could be generated by combining the information from several heatmaps.

The outcome of our exploration included many heatmap visualizations that are familiar from earlier work. However, we contribute the notion that only the "combining function" used in combining the data from heatmaps drawn for individuals needs to change in producing the majority of different heatmap visualizations. As an additional dimension, if the visualization pipeline stays the same but the segment of the data that is fed to it changes, one can produce temporal visualizations that add the much needed [Holmqvist et al. 2011] temporal dimension to heatmap-based visualizations.

Next we will describe four variants of heatmaps that are useful in explorative analysis of gaze data. The computational details are described in a separate section after the visualization descriptions.

2.1 Attention Map

The original goal of [Pomplun et al. 1996] was to visualize the distribution of the viewers attention on the image. This happened under the eye-mind hypothesis that assumes that the point of gaze also reflects the focus of visual attention.

Pomplun et al. chose the Gaussian distribution to visualize the focus of attention. The justification for this choice was not explicitly described, but we can hypothesize multiple reasons including, that it is a good approximation for random noise, that may be present because of measurement error or eye orientation error. Also, the standard deviation of the distribution (1°) corresponds approximately to the size of the part of the visual scene that is projected on the fovea. Thus, the attention map is also a focal map showing the parts of the image that have been projected onto the foveas of the viewers.

The distribution parameters chosen by Pomplun et al. have not been seriously challenged in subsequent attention mapping work. We do not wish to do this either. In our visualizations we used the eye-display distance reported by the eye tracker to scale the standard deviations of the Gaussian distributions to be exactly one degree of visual angle regardless of the participant's head movements.

Once the attention map has been computed, it can be displayed in many ways. A 3-dimensional height field is sometimes seen (e.g. [Sprenger et al. 2013; Wooding 2002]). Transparency map or a color display on top of the original image are also frequently seen. Figure 1 shows our example data[2] as transparency in a black layer on top of the original scene.

2.2 Deviation Map

In [Schiessl et al. 2003] different colors were mapped to different viewer groups. However, there are situations where the setup does not have two groups a priori. Then the question is how do we find the areas that were viewed differently? One approach is to use a measure of dispersion as the combining function. Variance and standard deviation are the commonly used measures of dispersion in statistics. Thus they are the obvious choices for combination functions in heatmaps that visualize differences in viewing patterns. We call these kinds of heatmaps deviation maps. Figure 2 shows a deviation map drawn based on standard deviation of the pixel values in per-participant heatmaps.

Deviation maps can be useful in exploratory data analysis. A region of high dispersion may indicate differences in the distribution of attention between participants.

2.3 Difference Map

In [Wooding 2002] it was suggested to subtract heatmaps from each other to create difference maps that have highest points in places where one heatmap is "hot" and the other is "cold". These regions show areas that are of interest in comparisons of two groups of users or two scenes. A related approach was used by [Schiessl et al. 2003] in generating heatmaps with two different colors showing the distribution of attention of two groups of viewers on the same image.

Our version of this visualization is shown in Figure 3. In comparison to the attention and deviation maps, the *difference map* more clearly suggests an interpretation of the data. It seems that dominant Korean focus was on people (especially the female in the

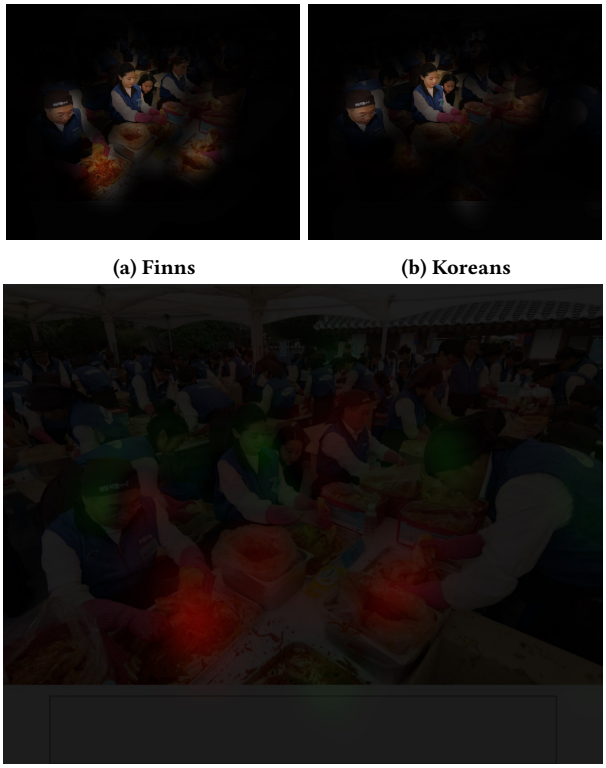
Figure 1: A heatmap with transparency of a black overlay showing the most attended areas. The displayed data is from the first two seconds of viewing by a group of 16 Finnish viewers and 16 South-Korean viewers. The photo has been released by Samsung under the Creative Commons "Attribution-NonCommerical-ShareAlike 2.0" license. Photo source: https://flic.kr/p/gt2ju4

Figure 2: Deviation map for the data in Figure 1.

[2]The data were collected while viewing pictures on a 17" LCD display. In South Korea a Tobii X2-60 tracker was used and in Finland a Tobii T-60, both with a capture rate of 60 Hz. The participants were recruited from a university community. In Korea the average age of the participants was 21.9 years (SD=3.4). In the Finnish group the average age was 23.4 years (SD=4.4).

The task of the participants was to write an answer to a question presented before the image into the text box below the image. 8 in each country answered to "What do you see in the image?" and the other 8 to "Why was this picture originally taken?". The minimum length of the answer was 2 rows of text. There was no time limit.

middle) whereas dominantly Finnish focus was on the Kimchi containers and the male on the right of the female dominantly attended by the Koreans.



(a) Finns **(b) Koreans**

(c) Difference map

Figure 3: Two attention maps and a difference map for the same data. The red areas were viewed more by the Finnish viewers and green areas were viewed more by the Korean viewers.

2.4 Significance Map

Interpreting difference maps can be deceptive. Based on a difference map only, we do not know how red or green an area needs to be in order for the difference to be large enough to be worth further investigation. For example, an area might be colored because only one or two participants attended to it excessively. We would like to know which differences we can expect to repeat if we collected the same kind of data again.

In statistical hypothesis testing, we compute statistics such as t and F that take into account the difference between group means and also the variability within each group. Since we can compute any statistic for each pixel of a heatmap, we can also compute the t statistic. In fact this method has been used for a long time in other fields of science, e.g. in functional brain imaging (see e.g. [Bennett et al. 2009]) where the imaging results consist of blocks of voxels (pixels in three dimensions). The idea is to find corresponding voxels in images from different brains and then highlight those voxels that show different levels of activity under the conditions that are being compared.

Figure 4: Significance map with the data shown in Figure 1. The area with $p < .01$ is highlighted in red.

Figure 4 shows the now-familiar data from the earlier figures divided according to two groups of viewers. Comparing Figures 4 and 3 we see that t statistics corresponding to p<0.01 emerged only in one of the areas highlighted in Figure 3.

There is an effort by [Lao et al. 2016] to produce significance maps that can be used directly for hypothesis testing. Lao et al. have published a number of iterations of their computing framework in response to criticism regarding the validity of inference based on it. The 2016 iteration utilizes randomization tests for finding the per-pixel p-threshold. We agree that this is probably the best approach. However, randomization tests require thousands of randomization iterations per pixel whereas a t-test requires only a single pass variance computation and a table lookup. Thus, in the interest of making our visualization software interactive enough, we chose to use t-tests. This we can do because we are producing visualizations for hypothesis generation instead of hypotheses testing.

2.5 Generation of Heatmaps

The outline of our visualization software is shown in Figure 5. On the left we have gaze data recorded for each participant in the study. The next step is to pick the variable to visualize. For example, to visualize the distribution of the visual attention we pick raw data points. Data for each participant is processed separately until a normalized 32 bit floating point heatmap exists for each participant. Then the data is fed to a ”combining function” that utilizes the per-participant heatmaps to produce a bitmap. Depending on the visualization needs, the bits on the alpha or the color channels of the bitmap are manipulated creating the final resulting visualization layer.

Most of the computation takes place in the heatmap generation. Once the per-participant heatmaps exists, one can quickly explore the effect of the different combining functions on the visualization.

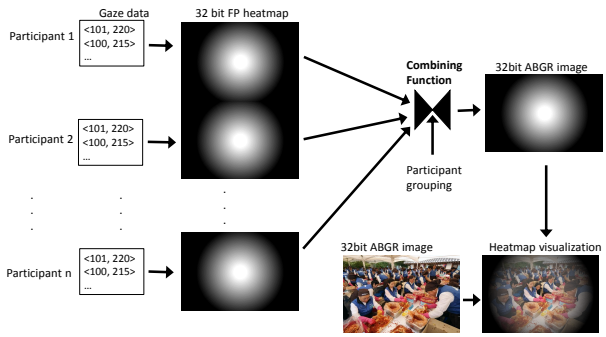


Figure 5: Outline of the computations we used to produce the heatmaps discussed int his paper. The process begins on the left and ends in the creation of the heatmap visualization on the lower right corner.

The conventional method of generating heatmaps has been to first find fixations in eye tracking data and then draw the heatmap based on those. This approach was recommended by [Bojko 2009] and [Holmqvist et al. 2011]. However, we have drawn every sample that the tracker recorded. The downside of this approach is that some samples were recorded while the eyes were moving so fast that saccadic suppression stopped the processing of what the eye saw. I.e. the gaze coordinates were not indicative of attention at that location. Vast majority of the samples, however, were recorded during fixations. Another disadvantage in this approach is that there is much more to compute because each fixation consists typically of 10-60 samples. Drawing the heatmaps based on samples rather than fixations requires more computing.

The positive sides of using samples include that we do not need to get into the argument of which fixation detection algorithm does the best job for us. Further, our approach works for smooth pursuit and other eye movements that do not consist of fixations and saccades. Especially on low-frequency trackers the exact starting and ending points of fixations can be difficult to measure. Yet, the data can be quite useful for heatmapping purposes.

In the past, further savings in the computation effort have been achieved by truncating the distributions e.g. at 2 or 3 standard deviations (e.g. [Holmqvist et al. 2011; Špakov 2008]). Nowadays, however, a PC can compute static heatmap visualizations from sizable datasets thanks to the powerful parallel processing capabilities in the main CPU and the display controller [Duchowski et al. 2012]. Thus such optimizations are no longer necessary except when computing visualizations for very large data sets or when doing real-time computations for dynamic visualizations.

The per participant heatmaps were in 32 bit floating point format. After adding up the distributions for each sample the heatmap was normalized by finding the largest value of the heatmap and dividing all pixels with it. These per-participant maps were kept in memory and used for further computations such as adding up the pixels to produce the attention map for the whole group.

To visualize the floating point numbers a black bitmap was taken as the starting point. Then its alpha channel was modified at each pixel by scaling the heatmap value to the range of 0-255 and inverting it (high opacity is wanted at 0 heat).

For a deviation map the process is the same except that instead of summing the maps the standard deviation is computed. Similarly, other statistical measures were computed and normalized to the range of the chosen output channel.

Depending on the amount of data being processed, the generation of the heatmaps can take a while (from minutes to hours on a typical PC with 4-8 CPU cores). However, the rest of the visualization work is fast enough to be used interactively. The heatmap generation could be speeded up significantly by utilizing the parallel computing resources available in the graphics processors included in typical PC configurations.

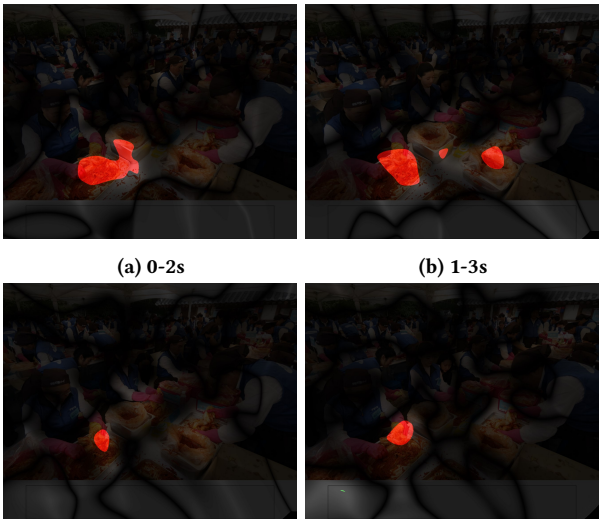
3 FILTERING LARGE DATASETS FOR INTERESTING SEGMENTS

The ability of the heatmap to give an overview of the distribution of attention for the whole viewing session for multiple participants at one glance is wonderful. However, such massive aggregation of the data also hides many details that are sometimes more interesting than the overview. For example, if there are events that take place only in some parts of the tracking session, it is possible that they are not noticeable in the aggregate heatmap.

For this reason it is sometimes important to look at the data in smaller slices. If it is not known beforehand when the interesting events will take place, it is interesting to automate the detection of potentially interesting segments. As described above, this can be done by generating heatmaps for short segments of the data and then inspecting the heatmap collection. This can be done with conventional attention maps, but the problem is, as demonstrated above, that attention maps always show something. Thus visually inspecting them is slow and leads to much guesswork. This is where the significance map is useful in highlighting the areas that show the strongest differences.

In Figure 6 the sub-figures are significance maps for 4 seconds of the same data from where the 2 first seconds were used in the earlier examples. Each image represents a snapshot of a 2 second sliding window passing over the data. We can see that in comparison to conventional heatmaps and even in comparison to the difference map, the significance map helps in focusing the attention to the most robust differences between the groups. Such 2-second slices of the data can be inspected quickly. Using a video or a matrix presentation of these images, spotting the interesting segments from a display with several minutes worth of data takes only a few seconds of human labor.

Depending on the number of participants, the task type, and possibly some other considerations it may be necessary to change the p-value that is used for flagging the areas with clearest differences. In this case we had 8 participants per group and p threshold of 0.01 seemed appropriate as it was large enough to still produce highlighted areas, but small enough to not produce too many of them. We are free to tune the threshold as we see fit because the purpose is not to use the result for statistical inference. We only want to identify the most robust differences in the data. Figuring out what, if anything, these differences mean must be done later with better evidence. In the study that contributed the data for this



(a) 0-2s (b) 1-3s

(c) 2-4s (d) 3-5s

Figure 6: Four frames in a series of two second segments of the gaze data with one second overlap. The red highlighting shows a potentially interesting event with the Finnish viewers focusing on the Kimchi containers. There were 348 other frames (not shown) most with no highlights. All other highlights in the data for this image were small and isolated.

visualization there was also the text written by the participants. Evidence for the meaning of the differences in viewing behavior was found in the texts where the Finns explicitly expressed their unfamiliarity with Kimchi and Koreans explicitly named Kimchi. This led us to the hypothesis that unfamiliarity with Kimchi and the Korean Kimpchi preparation tradition was the underlying cause for Finns paying so much more attention to the highlighted area.

4 AUTOMATIC CLUSTERING OF GAZE POINTS FOR EXPLORATION

While useful, the heatmap process described above has some unsatisfactory properties:

(1) the data is represented by gaussian distributions rather than the authentic gaze points measured with the eye tracker. In essence the heatmap is produced by running a smoothing filter over the gaze data and in the process some precision is lost.
(2) the gaze data is multiplied by distributing each gaze point over hundreds or thousands of pixels. There is much extra computation later when computations are done on each pixel of the scene instead of just the gaze coordinates originally measured.
(3) The sequence of the gaze movements is not a part of the analysis. Fixations at A and B create exactly the same heatmaps as fixations in the reverse order (B then A).

To complement the heatmap-based exploration, we wanted to find a complementary method that would preserve the discrete form of the gaze data including and emphasizing the possibility to

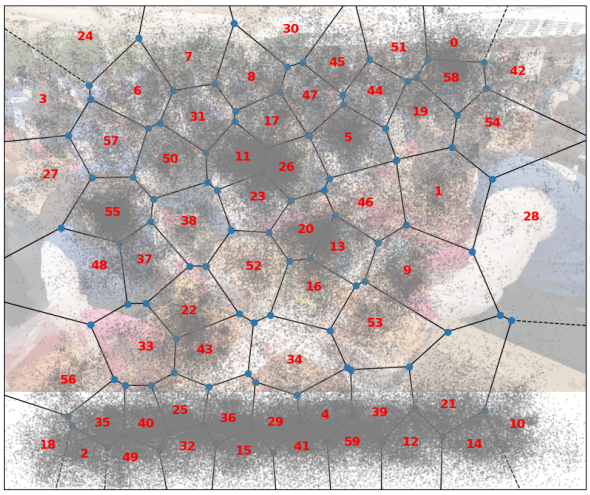


Figure 7: The cluster centers after clustering the gaze data into 60 clusters. Areas that were looked at a lot tend to have clusters with a smaller area. The numbers identify the clusters, they have no relation to the data. The Voronoi cell borders are overlaid on top of the gaze samples.

visualize gaze transitions between objects of interest. After some trials we ended up with the following automatic data clustering approach[3]:

(1) The gaze data was clustered to "natural" clusters by using the K-means algorithm[4] [MacQueen 1967] which minimizes the sum of distances of all data samples to the nearest cluster center. The cluster centers were the mean points of gaze data belonging into each cluster (see Figure 7). The number of cluster centers needs to be selected before running the algorithm. After some trials we ended up using 60 clusters out of which about 20 ended up in the text area at the bottom of the display.
(2) The image area was partitioned into separate Voronoi cells (see Figure 7), one for each cluster center. The algorithm partitions even those areas where there were only a few or no data.
(3) The sequences of gaze points for each participant was vector quantized, transforming them to code sequences where each gaze point was represented by the code of the Voronoi cell where the gaze point belonged to.
(4) The code sequences (for each participant) were filtered by a minimum code stretch length of 5. I.e. if the same code repeated at least 5 times it was kept, otherwise it was excluded. The reason for this filtering was to get rid of very short visits on a cell that were likely to be due to measurement noise or samples recorded during saccades. The final representation of the participants' gazing behavior was thus a sequence of

[3]The proposed data clustering is one possible method to perform a spatial transform of gaze positions as discussed in [Andrienko et al. 2012, chapter 5.2.2]
[4]The K-means algorithm has been used earlier for gaze data clustering by [Latimer 1988] and [Naqshbandi et al. 2016]. Also, [Privitera and Stark 2000] used K-means for gaze data processing.

Voronoi cell codes, one for each stretch of gazing the same Voronoi cell.

4.1 Observation and Transition Profiles

To visualize differences between the participants or groups of participants we calculated the histograms of gaze points on Voronoi cells for each participant. I.e., we counted how many times each Voronoi cell was looked at by a participant and normalized that number by dividing by the number of all cell sightings by the participant. The process created *observation profiles* where high values for some locations meant that participant was looking the area more often than some other areas.

As a measure of similarity we used the normalized dot-product $D_{a,b}$ of the profile vectors between two participants. The computation consists of calculating the dot-product of the profiles and then dividing the value by the product of the profile lengths:

$$D_{a,b} = \frac{\sum_{i=1}^{n} a_i \times b_i}{L_a \times L_b}, \tag{1}$$

where a, b are the profile vectors, n is the number of elements, and

$$L_a = \sqrt{\sum_{i=1}^{n} a_i^2}, \quad L_b = \sqrt{\sum_{i=1}^{n} b_i^2}. \tag{2}$$

In Figure 8 we show the distributions of similarities between participants of the same and different groups. For comparison we also calculated a distribution of the between groups similarities by randomizing the order of elements in the other profile vector (b in the equations 1 and 2). The purpose of this variant was to probe how big effect the order of the elements has in the distribution[5].

The results show that mostly the Finns and Koreans viewed the same areas for about the same duration as was already observed in the heatmaps. The small differences suggest that Finns were slightly more homogeneous than the Koreans. However, even a randomly ordered profile did not differ much from the others suggesting that this analysis method may not be very sensitive.

We also calculated a first order transition profiles based on the gaze transitions between Voronoi cells. The transitions were the sets of consecutive Voronoi cell pairs in the code sequences, each pair consisting of a starting cell and a target cell. A *transition profile* was then created by counting the number of each transition by a participant normalized by dividing by the total number of transitions.

In Figure 9 we show the distribution of similarities between participants of the same and different groups, using the transition profiles. Again, the first three distributions are similar, which indicates that there were no big differences between the groups in the transition profiles. However, the difference between the random order profile distributions and the other distributions is large. This shows that there is more structure in the transition profiles than in the observation profiles.

[5]The used data clustering algorithm usually leads to a balanced distribution of samples per cluster and the profiles tend to be flat on average. This is why the random order profile does not, on average, diverge much from the others.

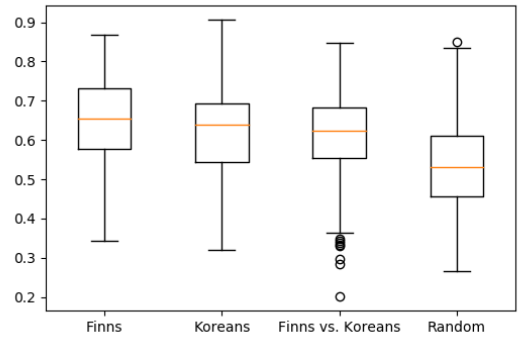


Figure 8: The distributions of similarity values between observation profiles of the participants within nationality, between nationalities, and in comparison to randomized cluster ordering.

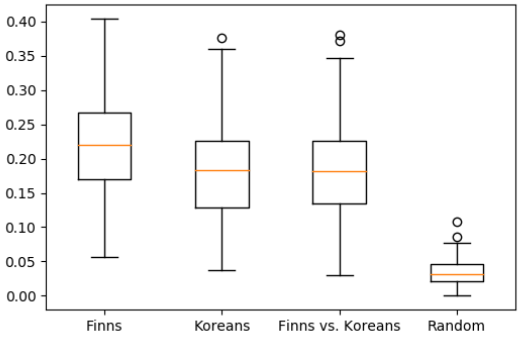


Figure 9: The distributions of similarity values between the transition profiles of participants in same groups and in different groups. The three first distributions are similar, but there is a clear difference between the random allocation and others.

4.2 Observation and Transition Differences

The preceding observation and transition analysis was focused on the aggregate results based on all observations and transitions. The results suggest that while the overall viewing behavior was similar among Finns and Koreans, there were also some differences between the groups. To see where the differences were, more detailed analysis on the individual clusters and transitions were needed.

Assuming that one group of participants was paying more attention to certain details in the image we would expect that the observation profile values related to the respective Voronoi cell(s) would be bigger for one group. To explore that idea we collected the observation profile values related to each Voronoi cell from both groups and tested if there were statistically significant differences between the groups. To restrict the number of comparisons we included only the 50 most often visited Voronoi cells.

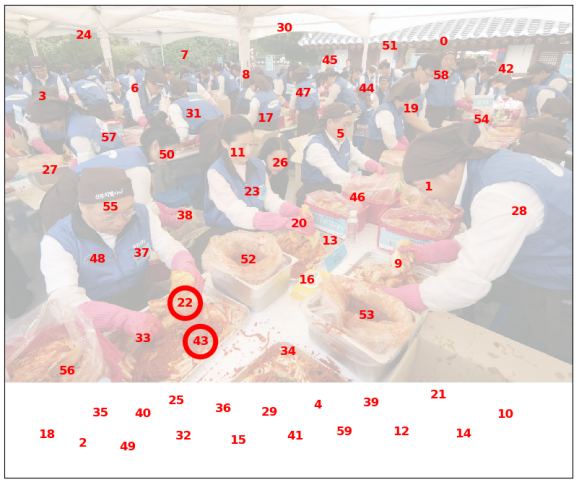


Figure 10: The red circles around the cells indicate that they were observed statistically significantly more by Finnish participants

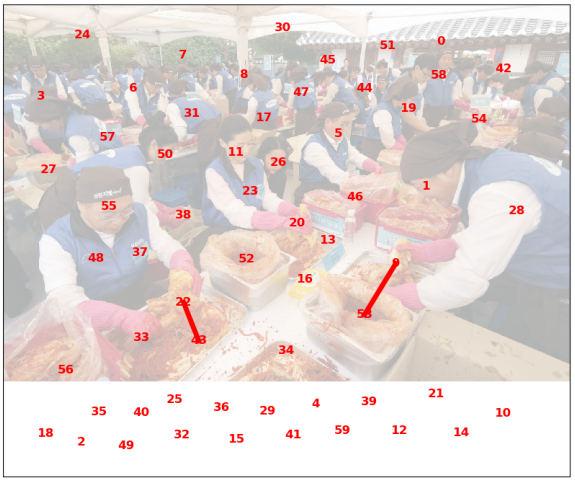


Figure 11: Transitions between cells: The red lines indicate transitions where Finnish participants had statistically significantly more transitions than the Koreans.

Since the number of comparisons was smaller than in the heatmaps we were able to use a Monte Carlo permutation test [Dugard 2014; Edgington and Onghena 2007; Howell 2007] to analyze possible statistically significant differences between the groups. The permutation test is not dependent on as many assumptions on the sample distribution as some other tests, such as t-test and ANOVA [Dugard 2014]. E.g. the test sample does not have to be normally distributed.

In Figure 10 we have separately marked those cells that had statistically significant differences between the groups ($p < 0.05$[6]). The circle markings were red if the Finnish group was visiting them more often and green if the Korean group was visiting them more often. For this data set red markings appeared. Clusters 22 and 43 are in the same area where the heatmap analysis found a difference.

A similar computation was run for the first order transition profiles, again using only the 50 most common transitions. In Figure 11 we have marked transitions with statistically significant differences between the groups. There were two such transitions, between clusters 22 and 43 and between clusters 9 and 53, where Finns transitioned more than Koreans. Figure 12 shows two partial raw gaze paths for illustration on what the raw-material of the clustering looked like.

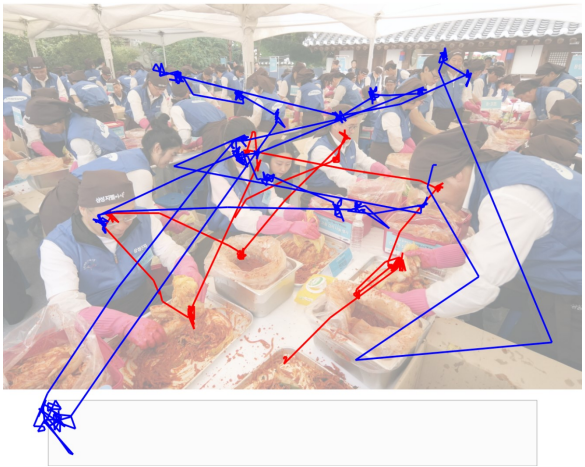


Figure 12: One Finnish (red) and one Korean (blue) example of raw gaze paths during the first 6 seconds of viewing.

5 DISCUSSION

5.1 Comparing Heatmaps and Clustering

In terms of the ability to highlight the most robust differences in the gaze data heatmaps and cluster-based analyses worked well. However, the ability to temporally segment the data is better for the heatmap-based techniques. Heatmaps tend to get saturated when long time-periods are inspected. When everybody has looked everywhere, there are no differences to see. Clustering, on the other hand is meaningful only if there is a lot of data to cluster. In very short time segments the data is so sparse that there are not many clusters to be found. The best way to combine the strengths of these methods is perhaps to use the cluster analysis as a starting point and then if temporal developments are of interest inspect the interesting segments with heatmaps.

5.2 Significance in Significance Maps

On the surface, it may appear as if significance maps are able to work around the heatmap criticism by [Holmqvist et al. 2011]. They show whether a hypothesis test is statistically significant or not while preserving the intuitive appeal of heatmaps. One can argue that a significance map as described above is the ultimate gridded area of interest (AOI) approach with pixel-sized AOIs. The problem with this is that because there are so many AOI:s, and there is no a-priori hypothesis on which ones should light up if a given phenomenon is

[6]We were using a Bonferroni corrected limit ($p < 0.05/50$) knowing that we were doing 50 comparisons.

at work, the probability of finding results due to random variability rather than an underlying interesting phenomenon is fairly large.

This is a form of the family-wise error rate (FWER) problem that always accompanies multiple statistical tests. Luckily, the heatmap data are not random on the pixel level. Instead larger continuous areas are highlighted in areas where the surfaces of the attention landscapes are, on average, far away from each other in comparison to the per-participant variability in that area. Thus, we are not really testing a million independent pixels, but rather a number of areas that are dynamically created based on the gaze data distribution.

Figuring out how exactly one should control for multiple comparisons under these conditions is not an easy task. For further information see the work by [Lao et al. 2016]. In this work we were interested in extending the powerful visualization capabilities of heatmaps. Thus, the t-tests used in the generation of the visualizations are not taken as hypothesis tests. Instead the t statistics is considered useful because it allows us to highlight areas with the most distinct differences between the sets of heatmaps.

5.3 Temporal Segmentation

A well-known weakness of conventional heatmaps is that they flatten the temporal dimension. However, as has been shown previously, heatmaps can be rendered in real time to get around this limitation (see e.g. [Duchowski et al. 2012]). All the further computations on the heatmaps are also possible to do in temporal segments in real time or viewed from videos that were computed off-line.

With our example data set it was no coincidence that the heatmap images used for illustrating this paper were from the first seconds of image viewing. As the task progressed, the looking and writing task alternated and over time different participants ended up in different phases of the work.

Generating a temporal series of images allows the investigation of sequences of attention focusing behavior. These series can be represented as a matrix of heatmaps, or on the same location serially (i.e. as a video). Viewing a video representation of the data would also be a good idea in studies that are not primarily exploratory. Unusual participant behaviors, such as failures to follow instructions, can sometimes be spotted before proceeding to AOI analysis or other means of testing hypotheses about the data.

5.4 Other Measures

As described above, heatmaps have conventionally been used as attention maps. That is, the point of gaze is assumed to reflect the focus of visual attention. The layman's interpretation of an attention map is that the highly attended areas are "important" for the goals that the viewer.

However, the amount of viewing that different parts of the image receive is not the only measure one can extract from gaze data. Measures such as fixation duration and saccade length are often measured in eye tracking studies. Also pupil size has been reported in many studies. One can build heatmaps of all these variables. Such heatmaps give an overview of the variable over the whole scene instead of data points for predetermined areas of interest. In future work all these variables can also be automatically mapped to clusters and differences between groups can be tested as we did with the transitions in this paper.

5.5 Other Considerations

Sometimes the gaze data and the visualization algorithms produce output that is not easy to interpret. For example, we have encountered situations where one group of participants tends to have lower precision data, both groups get high peaks on the center of the area, but because the precision in one group is high, the peak is narrower. This creates areas around the peak where one group seems to exhibit higher amount of attention. Uninformed interpretation of such images can lead to a false conclusion. These situations are not easily identifiable in difference maps.

As [Holmqvist et al. 2011] point out, heatmaps can only tell what was looked at, but not why. In the study that served as our data source the why question was answered by examining the text written by the participants. In our example image, the familiarity with the Kimchi was identified as the cause for the culturally determined difference in viewing based on the texts. Finnish viewers expressed their unfamiliarity directly or through (wrong) guesses on the nature of the material and Korean viewers explicitly named Kimchi or Kimchi making practices.

Heatmap generation and the clustering approach produce different results depending on what values are set for the various parameters such as the standard deviation of the smoothing distributions, temporal segments, the number of clusters, etc. This is a strength and a curse of exploratory data analysis. The methods need to be flexible in order to be able to address many circumstances. However, the number of parameters and the different results one gets by changing their values can feel overwhelming to a user and cause confusion for readers if reports with such visualizations are not detailed enough.

6 CONCLUSION

We described two approaches to exploratory data analysis of gaze data. The first approach relied on computations on the pixels of heatmaps drawn for individual viewers in the time-frame of interest. The second approach relied on clustering of the gaze data based on location and follow-up analyses based on gaze transitions between clusters. Both techniques were found to be useful. The vast majority of the data could be quickly discarded from further analysis. This allows efficient use of human labor in the parts of the data where interesting differences between groups of viewers emerge.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland under grants Mind Picture Image (decision 266285) and Haptic Gaze Interaction (decisions 260179 and 260026).

REFERENCES

Gennady Andrienko, Natalia Andrienko, Michael Burch, and Daniel Weiskopf. 2012. Visual Analytics Methodology for Eye Movement Studies. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec 2012), 2889–2898. https://doi.org/10.1109/TVCG.2012.276

Craig M Bennett, George L Wolford, and Michael B Miller. 2009. The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience* 4, 4 (2009), 417–422.

T Blascheck, K Kurzhals, M Raschke, M Burch, D Weiskopf, and T Ertl. 2017. Visualization of eye tracking data: A taxonomy and survey. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 260–284.

Agnieszka (Aga) Bojko. 2009. *Informative or Misleading? Heatmaps Deconstructed.* Springer Berlin Heidelberg, Berlin, Heidelberg, 30–39. https://doi.org/10.1007/

978-3-642-02574-7_4

David Brillinger. 2011. Data Analysis, Explorative. In *International encyclopedia of political science*, Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino (Eds.). Vol. 1. Sage, 530–537.

Antoine Coutrot, Janet H. Hsiao, and Antoni B. Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (01 Feb 2018), 362–379. https://doi.org/10.3758/s13428-017-0876-8

Andrew T. Duchowski, Margaux M. Price, Miriah Meyer, and Pilar Orero. 2012. Aggregate Gaze Visualization with Real-time Heatmaps. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, NY, USA, 13–20. https://doi.org/10.1145/2168556.2168558

Pat Dugard. 2014. Randomization tests: a new gold standard? *Journal of Contextual Behavioral Science* 3, 1 (2014), 65–68. https://doi.org/10.1016/j.jcbs.2013.10.001

Eugene Edgington and Patrick Onghena. 2007. *Randomization tests*. CRC Press, Boca Raton, FL, USA.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Rrichard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

David C Howell. 2007. *Statistical methods for psychology*. Thomson Wadsworth, Belmont, CA, USA.

J. Lao, S. Miellet, C. Pernet, N. Sokhn, and R Caldara. 2016. iMap4: An Open Source Toolbox for the Statistical Fixation Mapping of Eye Movement data with Linear Mixed Modeling. *Behavior Research Methods* (2016). https://doi.org/DOI10.3758/s13428-016-0737-x

CR Latimer. 1988. Eye-movement data: Cumulative fixation time and cluster analysis. *Behavior Research Methods, Instruments, & Computers* 20, 5 (1988), 437–470.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, Berkeley, Calif., 281–297. https://projecteuclid.org/euclid.bsmsp/1200512992

Khushnood Naqshbandi, Tom Gedeon, and Umran Azziz Abdulla. 2016. Automatic clustering of eye gaze data for machine learning. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. IEEE, 001239–001244.

Marc Pomplun, Helge Ritter, and Boris Velichkovsky. 1996. Disambiguating complex visual information: Towads communication of personal views of a scene. *Perception* 25, 8 (1996), 931–948.

Claudio M. Privitera and Lawrence W. Stark. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence* 22, 9 (2000), 970–982.

Michael Schiessl, Sabrina Duda, Andreas Thölke, and Rico Fischer. 2003. Eye tracking and its application in usability and media research. *MMI Interaktiv - Eye Tracking* 1, 06 (mar 2003), 41–50.

Andreas Sprenger, Monique Friedrich, Matthias Nagel, Christiane S. Schmidt, Steffen Moritz, and Rebekka Lencer. 2013. Advanced analysis of free visual exploration patterns in schitzophrenia. *Frontiers in Psychology* 4, 737 (2013).

J.W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Company. https://books.google.fi/books?id=UT9dAAAAIAAJ

Oleg Špakov. 2008. *iComponent - Device-Independent Platform for Analyzing Eye Movement Data and Developing Eye-Based Applications*. Ph.D. Dissertation. University of Tampere.

David S. Wooding. 2002. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 518–528. https://doi.org/10.3758/BF03195481