Journal of Eye Movement Research
7(3):5, 1-9

Image preference estimation with a data-driven approach: A comparative study between gaze and image features

Yusuke Sugano
The University of Tokyo, Japan

Yasunori Ozaki
The University of Tokyo, Japan

Hiroshi Kasai
The University of Tokyo, Japan

Keisuke Ogaki
The University of Tokyo, Japan

Yoichi Sato
The University of Tokyo, Japan

Understanding how humans subjectively look at and evaluate images is an important task for various applications in the field of multimedia interaction. While it has been repeatedly pointed out that eye movements can be used to infer the internal states of humans, not many successes have been reported concerning image understanding. We investigate the possibility of image preference estimation based on a person's eye movements in a supervised manner in this paper. A dataset of eye movements is collected while the participants are viewing pairs of natural images, and it is used to train image preference label classifiers. The input feature is defined as a combination of various fixation and saccade event statistics, and the use of the random forest algorithm allows us to quantitatively assess how each of the statistics contributes to the classification task. We show that the gaze-based classifier had a higher level of accuracy than metadata-based baseline methods and a simple rule-based classifier throughout the experiments. We also present a quantitative comparison with image-based preference classifiers, and discuss the potential and limitations of the gaze-based preference estimator.

Keywords: image preference, eye movements, machine learning

Introduction

The subjective values and meanings of images often receive considerable attention from the research community in the field of image understanding. However, this is in the eye of the beholder, so to speak, and is quite difficult to be assessed from images. Recent advantages in machine learning techniques allow us to tackle such an ambiguous task in a data-driven manner, and there have been several research attempts to estimate the subjective values of images, such as the aesthetic quality (Datta, Joshi, Li, & Wang, 2006; Ke, Tang, & Jing, 2006; Luo & Tang, 2008; Nishiyama, Okabe, Sato, & Sato, 2011; Marchesotti, Perronnin, Larlus, & Csurka, 2011), using human-labeled datasets. However, while these approaches have achieved a certain level of success, it is not clear whether such an objective ground-truth measure actually exists for subjective values.

On the other hand, there is a long history of research focusing on eye movement and its relationship to the human mind. The use of gaze inputs is recently receiving more and more attention amid the increasing demand for natural user interfaces, and casual gaze sensing techniques are becoming readily available. However, a gaze is simply considered an alternative pointing input of a different modality in most of the application scenarios. Several research attempts incorporating the concept of cognitive state recognition have recently been proposed to extend the potential of gaze interaction. In these works, the eye movements are indirectly used to infer the cognitive states of the users, *e.g.*, the task and contextual cues (Bulling & Roggen, 2011; Bulling, Weichel, & Gellersen, 2013), intention (Bednarik, Vrzakova, & Hradis, 2012), user characteristics (Toker, Conati, Steichen, & Carenini, 2013; Steichen, Carenini, & Conati, 2013), cognitive load (Bailey & Iqbal, 2008; Chen, Epps, & Chen, 2013), and memory recall (Bulling, Ward, Gellersen, & Troster, 2011). While the view that the eye movement patterns of a person while viewing images can reflect his or her complex mental state has been widely shared among researchers (Yarbus & Riggs, 1967), it has also been pointed out that a classification task based on eye

This work was supported by CREST, JST.

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

movements is often very challenging (Greene, Liu, & Wolfe, 2012). Therefore, it is still important to investigate what can be practically inferred from the eye movements.

We focus on a preference estimation in this work in which a user is comparing a pair of natural images. Shimojo *et al*. (Shimojo, Simion, Shimojo, & Scheier, 2003) reported on the cascade effect of a gaze, and they showed that people tend to fixate on a preferred stimulus longer when they are asked to compare two stimuli and make a two-alternative forced choice. Several methods have been proposed to predict the preferences from the eye movements (Bee, Prendinger, Nakasone, André, & Ishizuka, 2006; Glaholt, Wu, & Reingold, 2009) based on this study. However, the main focus of these studies is a comparison between the same categories of stimuli such as the faces and product images, and more importantly, the target task is the early detection of decision making events. The estimation is done while the users are making preference decisions, and therefore, it is unclear whether it is also possible to estimate their preference between two natural images during free viewing. Although eye movements during comparative visual searching have also been widely studied (Pomplun et al., 2001; Atkins, Moise, & Rohling, 2006), a comparison between two unrelated images has not been fully investigated.

The goal of this research is to explore the possibility of gaze-based image preference estimation, and we make two contributions in this paper. First, we take a data-driven approach to the image preference estimation task using eye movements. A classifier that outputs image preference labels is trained by using a dataset of eye movements recorded while users are comparing pairs of images. The training is done by using an algorithm that can exploit the beneficial features for the classification task. In this way, we can identify the important features for preference estimation and to assess how they differ among different people. More importantly, we also investigate whether or not preference estimation based on eye movements is still possible in a scenario in which the users are freely viewing image pairs with no instruction. While most of the prior work focused on preference decision making, its application scenario is indeed quite limited and investigating the free-viewing scenario is of practical importance.

Second, we present a quantitative comparison with an image-based preference estimation technique. As briefly mentioned above, it has been demonstrated that the aesthetic image quality can be estimated in a data-driven manner. However, it is not yet clear if the same approach can be taken for highly subjective values such as personal preference. Another purpose of this work is to validate whether the standard framework of aesthetic quality classification is also beneficial for image preference estimation. It is quite unclear particularly in the case of the free-viewing scenario whether the gaze-based classification is still comparative with image-based classification. In this study, we quantitatively compare the classification performances of these two approaches.

Figure 1. Experimental setup

Data Collection

We assume a situation in this study in which the users are viewing a pair of natural images displayed side by side. We address the task of preference estimation in the supervised manner that we mentioned above. A binary classifier is trained based on the training data with the ground-truth labels to output the preference labels, *i.e.*, which image the user prefers, from the eye movement patterns. In this section, we first describe our experimental setting for the data collection.

Experimental Setting

We used a Tobii TX300 eye tracker, which is shown in Figure 1, for our data collection. The image pairs were displayed on the 23″ full HD TFT monitor of the tracker, and the eye movements were recorded at 60 Hz. The display areas were separated in the middle of the monitor, and each image was displayed in a 960×1080 pixel region. A chin rest was used to stabilize the viewing position at about 60 cm from the tracker.

The experiment had two phases: free viewing and preference labeling. After calibration, 14 novice participants were first asked to freely view 80 pairs of images without any specific instruction. Each pair was displayed for 10 seconds, and a white cross was displayed at the center of the monitor to control the fixation location for 3 seconds of intermission between the image pairs. Next, we showed 400 pairs of images in the same way, and instructed the participants to answer which image was preferred. After each pair was displayed, the participants were asked to press a number key corresponding to the side that he/she preferred. After a key was pressed, the next pair was displayed following the white cross targets. At the end, the first

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

80 pairs were displayed on the monitor again and the participants were instructed to answer with their preferences in the same way as in the labeling phase. Data was discarded if the participant mistakenly pressed the wrong key or a saccade event happened on only one side throughout the experiments.

Stimulus Images

We collected stimulus images from the Internet because our primary interest was whether objective measures such as user-provided metadata can be used to infer subjective image preference. More specifically, we collected images given high *interestingness* from the Flickr[1] website. This implies all of the stimulus images had a certain level of quality, and there was no obvious quality difference between the paired images. At the same time, two kinds of metadata, the number of comments and user favorites, were downloaded from the website to infer the popularity of the images. The downloaded images were restricted to having almost the same aspect ratio (from 1 : 1 to 8 : 9) for the display area and letterboxed to fit 8 : 9 to avoid cropping and any concomitant change in image composition. They were randomly combined to create the 480 image pairs described above.

Methodology

Gaze-based Preference Estimation

The input to our method is a gaze data sequence $\{(\boldsymbol{g}_n, t_n)\}$, *i.e.*, N gaze positions $\boldsymbol{g}_n$ associated with their time stamp values t_n. $t_0 = 0.0$ indicates the time when the image pair appeared on the display, and $t_{N-1} = 1.0$ is the time when the pair disappeared. Our goal is to classify which image the user prefers from the eye movement patterns during the comparative viewing.

As discussed earlier, it has been pointed out in prior work that humans tend to look at the preferred stimulus longer (Shimojo et al., 2003). In this study, we are interested in investigating whether any other kinds of features beneficial to the preference estimation task exist. Therefore, various fixation and saccade statistics are considered as the input features in a similar way as in (Castelhano, Mack, & Henderson, 2009; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011; Greene et al., 2012). The use of a random forest algorithm (Breiman, 2001) allows us to automatically select the more efficient features for the classification task, and their contribution can be quantitatively evaluated as feature weights.

Eye Movement Features. We first follow a standard procedure to extract the fixation and saccade events from these data; *i.e.*, if the velocity exceeds the threshold 30 [degrees/second], the gaze data is classified as saccades. We regard $\{(\boldsymbol{g}_n, t_n), \ldots, (\boldsymbol{g}_m, t_m)\}$ as data during a fixation if their angular velocities are below a predefined threshold. The first fixation is discarded because its position is highly affected by the previous stimulus. We define three attributes for each fixation event F, the position $\boldsymbol{p}$, duration T, and time t. If the i-th fixation F_i happens from t_n to t_m, $\boldsymbol{p}_i$ is defined as a median of the gaze positions, $T_i = t_m - t_n$ and $t_i = t_n$. Assuming that the areas in which each of the paired images is displayed are known, fixations $\{(\boldsymbol{p}_i, T_i, t_i)\}$ can be divided into two subsets, *i.e.*, fixations on the image on the left $\mathcal{F}_L$ and that on the right $\mathcal{F}_R$. At the same time, the fixation positions are normalized according to the display area of each image so that the x and y coordinates are at $[0, 1]$.

Saccade events are defined only when two successive fixations F_i and F_{i+1} happen on one side of the image pair. Four attributes are defined for each saccade event: direction $\boldsymbol{d}$, length l, duration T, and time t. Given a saccade vector $\boldsymbol{s} = \boldsymbol{p}_{i+1} - \boldsymbol{p}_i$, length l is defined as its norm $|\boldsymbol{s}|$ and the direction $\boldsymbol{d}$ is defined as a normalized vector $\boldsymbol{s}/l$. The duration and time are defined in the same way as the fixation events. As a result, two sets of saccade events $\mathcal{S}_L$ and $\mathcal{S}_R$ are defined for each side of the image pair.

We compute various statistics for each attribute from these fixation and saccade sets. Table 1 summarizes the attribute and statistical operation combinations. The means and variances are computed for all the attributes, and the covariances between x and y are additionally computed for the vector attributes (fixation position and saccade direction). The sums are computed for the scalar quantities other than time t, and the total counts of the fixation and saccade events are also computed and normalized so that the sum between the left and right images becomes 1.0. There are a total of 25 computed values for each side (11 from the fixations and 14 from the saccades), and they are concatenated to form a 50-dimensional feature vector $\boldsymbol{x}_f = (\boldsymbol{f}_L^{\mathrm{T}}, \boldsymbol{f}_R^{\mathrm{T}})^{\mathrm{T}}$ of a paired image.

Preference Classification. The task is to output preference label $y \in \{1, -1\}$, which indicates whether the preferred image is the one on the left (1) or right (-1) from the input feature vector $\boldsymbol{x}_f$. As discussed above, we assume that the ground-truth labels of the image preference are given, and train a classifier that maps $\boldsymbol{x}_f$ into y using the labeled data.

Due to the symmetric nature of the problem definition, a labeled pair of images and its corresponding eye movement data can provide two training data. If the user prefers the image on the left, for example, feature vector $\boldsymbol{x}_f = (\boldsymbol{f}_L^{\mathrm{T}}, \boldsymbol{f}_R^{\mathrm{T}})^{\mathrm{T}}$ is associated with label $y = 1$, while the left-right flipped feature vector $\boldsymbol{x}_f = (\boldsymbol{f}_R^{\mathrm{T}}, \boldsymbol{f}_L^{\mathrm{T}})^{\mathrm{T}}$ can also be used with label $y = -1$ for training.

[1] http://www.flickr.com

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

Table 1
Combinations of event attributes and statistical operations used to compute features for our classifier.

Fixation	Position $\boldsymbol{p}$	Mean (×2)
		Variance (×2)
		Covariance
	Duration T	Mean
		Variance
		Sum
	Time t	Mean
		Variance
	Count	
Saccade	Direction $\boldsymbol{d}$	Mean (×2)
		Variance (×2)
		Covariance
	Length l	Mean
		Variance
		Sum
	Duration T	Mean
		Variance
		Sum
	Time t	Mean
		Variance
	Count	

Random forest (Breiman, 2001) is a supervised classification method using a set of decision trees. Given a set of training samples, the random forest algorithm trains the decision trees using random sample subsets of the samples. Each tree is grown in a way to determine the threshold value for an element in the feature vector that most accurately splits the samples into correct classes. After the training, the classification of an unknown input feature is done based on a majority vote from these trees. In addition to its accuracy and computational efficiency, the random forest algorithm has an advantage in that it can provide feature importance by evaluating the fraction of the training samples that are classified into the correct class using each element. The classifiers used in the experiments are implemented using the scikit-learn library[2] (Pedregosa et al., 2011). The number of trees was empirically set to 1000, and the depth of each tree was restricted to 3.

Image-based Preference Estimation

An alternative approach for image preference estimation is to use features extracted from the image pairs. In addition to the method using eye movement features described above, we also examine the image features in the same classification framework for comparison. In this section, we briefly describe the details of the image features defined following a state-of-the-art method for aesthetic quality estimation. A pair of images, *i.e.*, image $\boldsymbol{I}_L$ displayed on the left side and $\boldsymbol{I}_R$ on the right side, is input. We can use the concatenated feature vector $\boldsymbol{x}_v = (\boldsymbol{v}_L^{\mathrm{T}}, \boldsymbol{v}_R^{\mathrm{T}})^{\mathrm{T}}$ in the same way as for the classification using the eye movement features by extracting image features $\boldsymbol{v}_L$ and $\boldsymbol{v}_R$ from each of the images.

It has traditionally been considered that there are several important rules defining the aesthetic quality of images, such as the color harmony theory and the *Rule of Thirds*. While such features can serve as a rough guideline, it is not an easy task to quantify the subjective image quality measurement. However, a data-driven learning approach for aesthetic quality estimation has recently become popular. An aesthetic quality estimator learns from a large dataset of images obtained from websites such as Photo.Net (Datta et al., 2006) and DPChallenge.com (Ke et al., 2006) with community-provided image quality scores in these works. They used several image-related features including generic image descriptors that are not explicitly related to image quality and showed that the community scores can be well predicted using the learned estimator.

Image Features. Following (Marchesotti et al., 2011), we also adopt two generic image features. The first one is the GIST feature (Oliva & Torralba, 2001; Douze, Jégou, Sandhawalia, Amsaleg, & Schmid, 2009), which is commonly used in scene recognition tasks. With the GIST feature, the overall layout and structure of an image is represented as a set of local histograms of Gabor filter responses. In our setting, an input image is resized to 64×64 pixels and then divided into 4×4 regular grids. The filter responses at six orientations are computed at each level of the two-level image pyramid, and histograms are extracted from each grid for each color channel to form the 192-dimensional GIST feature vector.

The second feature is based on the bag-of-features (BoF) representation of local descriptors (Sivic & Zisserman, 2003). Inspired by the bag-of-words representation used in natural language processing, an image in the BoF representation is described by the visual codewords frequency. Local descriptors are first extracted from the training images, and a visual codebook, *i.e.*, a discrete set of representative descriptors, is learned. Then, each local descriptor of an input image is assigned to one of the codewords, and the image is represented as a histogram of codewords.

The BoF representation, which is based on scale and rotation invariant local descriptors such as the scale-invariant feature transform (SIFT) (Lowe, 2004), is widely used in various image recognition tasks. We use two local descriptors, SIFT and color, just like in (Marchesotti et al., 2011). The SIFT descriptor is a rotation-invariant histogram of local gradients defined as relative to the most prominent orientation in the lo-

[2] http://scikit-learn.org/

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

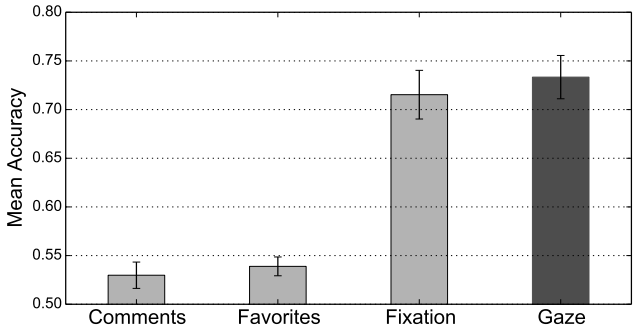

Figure 2. Comparison with baseline methods. The graphs show the mean accuracies from the 14 participants and the error bars indicate the standard errors. The first two graphs show the classification results using the objective metadata, which was the number of comments and favorites on the Flickr website. The third graph shows the accuracy of the simple classification using only the sum of the fixation duration, and the last graph corresponds to the proposed gaze-based classifier.

cal region. Unlike the original method (Lowe, 2004) that extracts SIFT descriptors at sparse keypoint locations, these descriptors are densely extracted on regular grids (Jurie & Triggs, 2005). The grids are placed every 64 pixels, and 64×64 local image patches are extracted. The 128-dimensional SIFT descriptors are computed from the local patches at four scales. The color descriptors are also extracted from the same local patches. Each patch is divided into 4×4 grids, and the mean and standard deviations per color channel are computed as the 96-dimensional color descriptor.

The dimensions of these two descriptors are reduced to 64 by principal components analysis (Jolliffe, 2005). Then, the codebooks of the two descriptors are obtained by clustering the features extracted from the training data into 100 clusters. The clustering is done by fitting Gaussian mixture models using the EM algorithm (Dempster, Laird, & Rubin, 1977). The original descriptors extracted from the input image are assigned with their nearest codeword, and 100-dimensional histograms of both features are concatenated to form the 200-dimensional BoF feature vector.

Results

We discuss our experimental results in this section to validate the gaze-based preference estimation method. There are three purposes for the experiments: 1) to see whether or not the data-driven training in an improvement over the simple classification approaches, 2) to assess the difference between gaze-based and image-based classifiers, and 3) to test the performance of the gaze-based estimation in a free-viewing scenario.

Classifier Performance

Figure 2 shows a comparison of the gaze-based preference classifier with the simple baseline methods. Accuracy scores were used for the evaluation because the positive and negative classes are symmetric in our problem setting. We compared the proposed classifier with three baseline methods. The first and second graphs show the classification results using the metadata obtained from the Flickr website. In these classifiers, the output label is the image with the higher metadata score (the greater number of comments or favorites). We additionally show the third classification result using only the sum of the fixation duration to evaluate the performance gain of the data-driven training. In this case, the sides with the longer fixation duration were treated as the output labels.

The proposed gaze-based classifier was trained and tested in a leave-one-out manner using the personal training datasets obtained during the labeling phase. For each image pair, a classifier was trained using the rest of the training data and the output label was compared with the ground-truth label to compute the classifier accuracy. The rightmost graph corresponds to the proposed classifier.

In all cases, the graphs show the mean accuracies of the 14 participants and the error bars indicate the standard errors. Not surprisingly, the accuracies of the first three classifiers based on the objective metadata were quite low and barely above the chance level. The mean accuracy of the proposed method was 73%, and was higher than all of the metadata-based methods (Wilcoxon signed-rank test: $p < 0.01$). While the simple classification based on the fixation duration also achieved a comparative level of accuracy, the performance was improved by using the proposed training approach (Wilcoxon signed-rank test: $p < 0.01$).

Image-based Estimation

Figure 3 shows a comparison between the gaze-based and image-based classifiers. The first and second graphs show the classification accuracies using the two image features, GIST and BoF (SIFT and color descriptors). As described earlier, the same random forest framework as for the gaze-based classifier (the third graph) was used for both features. The fourth graph additionally shows the mean accuracy of the classifier using a combined image and gaze feature. In this case, the BoF and gaze feature vectors were concatenated, and the random forest classifier was trained in the same way as above. All of the classifiers were evaluated by conducting a within-subject leave-one-out cross validation.

The image-based classifiers performed better than the metadata-based baseline methods discussed in the previous section; however, the gaze-based classifier significantly outperformed them all (Wilcoxon signed-rank test: $p < 0.01$). The results using the joint fea-

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

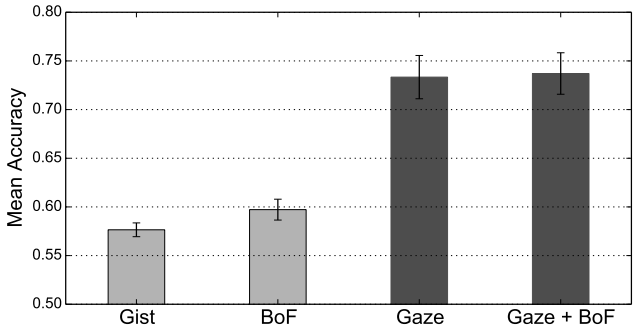

Figure 3. Comparison between gaze-based and image-based classifiers. The first and second graphs show the mean accuracies using two image features, GIST and BoF, respectively. The third graph corresponds to the gaze-based classifier, and the fourth graph shows the mean accuracy of a classifier using a combined image and gaze feature.

ture showed a slightly better level of accuracy, but we did not observe any significant difference. Although prior work claimed that the aesthetic image quality can be estimated in a similar data-driven manner, these results show that inferring personal image preference is a much more difficult task. Eye movements can tell us a lot about personal preferences, and this indicates the potential of gaze information in the context of media understanding.

Personal Differences

In the previous section, the training that was conducted used the personal datasets. While this follows the standard procedure for supervised classifications, it is not always possible to collect the most appropriate training data from the target user. The objective in this section is to confirm whether or not it is possible to use the training data obtained from different people for the classification task.

Figure 4 shows an accuracy comparison between the within-subject and cross-subject training conditions for both the image-based and gaze-based classifiers. The within-subject condition corresponds to the leave-one-out setting discussed in the previous section. In the cross-subject condition, the training and testing were done in a leave-one-subject-out manner; the classifier was trained for each person using the data from the other 10 participants. Each graph in Figure 4 corresponds to a participant ($s1$ to $s14$), and the rightmost graphs show the mean accuracy from among all the participants.

While the within-subject training improves the accuracies of some participants, such as for $s4$, the cross-subject training generally achieved a comparative level of accuracy and there was no statistically significant difference in the mean scores. This indicates that the learning-based framework could successfully capture discriminative eye movements that can be commonly observed among different people.

Feature Importances

It is also important to visualize the differences between the within-subject and cross-subject conditions and to quantitatively assess how each element of the feature vector contributed to the classification task. The variable importances of the gaze features obtained using the random forest classifier training process are shown in Figure 5. In our implementation, the feature importances are computed as a fraction of the samples that each of the elements contributed to in the final prediction. A higher value thus means there was more contribution to the classification.

Our 50-dimensional gaze feature vector consists of 25 statistical measures computed from both sides of the paired image regions. However, as discussed earlier, the definition of the classification task is symmetric and the labeled training data was duplicated to create left-right flipped training samples. Therefore, two corresponding elements (*e.g.*, fixation counts on the left side and the right side) theoretically have the same importance throughout the training process, and the sums of the two values are shown in Figure 5. The graphs correspond to the importances of the 25 features listed in Table 1 and are color-coded according to the training data used. $s1$ to $s14$ indicate the within-subject training condition, *i.e.*, the feature importances obtained when personal training datasets were used. *All* indicates the case when all of the data from the 14 participants were used for training.

The three most contributing features are *fixation-count*, *fixation-duration-sum*, and *saccade-count* in most of the cases, and this agrees with the gaze cascade effect. Compared to these three elements, the contribution of *saccade-duration-sum* is not very high. The time stamp statistics (*time-mean* and *time-variance* of both the fixation and saccade) showed a certain amount of contribution, and *saccade-length-sum* also contributed for some participants. It can be seen that person $s4$, who showed the largest performance improvement from the within-subject training in Figure 4, had a unique distribution compared to the other participants, and the fixation position was the key to the improvement. The random forest algorithm can only assure that the combination of these features led to the performance gain, and cannot provide any reasoning behind each factor. Further study will be an important future work to gather feedback for better understanding the mechanism behind preference decision.

Free Viewing

The results discussed in the previous sections were based on the dataset obtained during the labeling phase, where the participants were instructed to assign preference labels. While this setting is the same as in

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

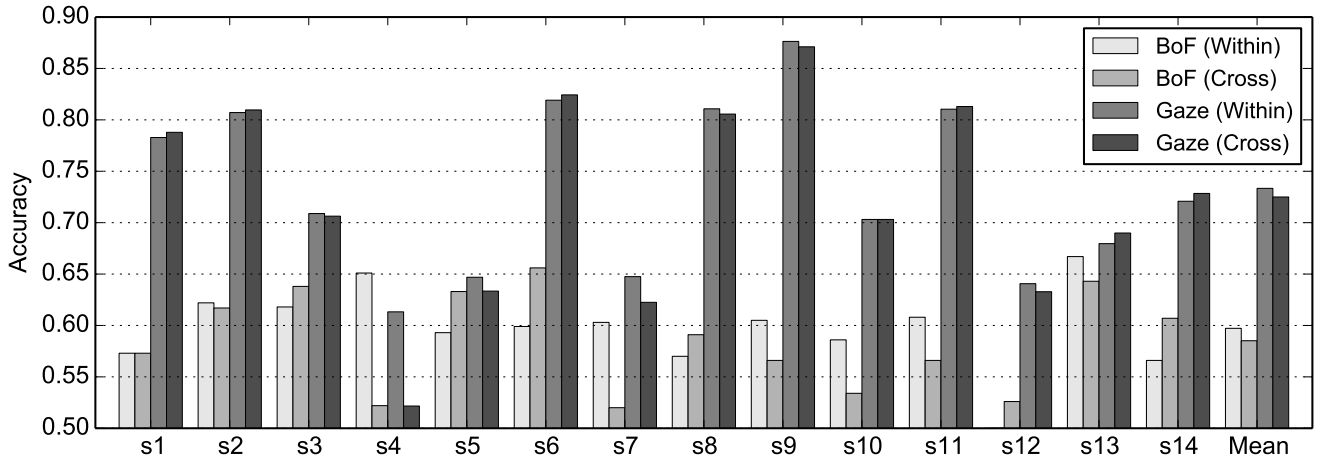

Figure 4. Cross-subject training. The within-subject condition corresponds to the leave-one-out training and testing for each participant. In the cross-subject condition, the classifier is trained for each person using the data from the other participants. Each graph corresponds to a participant ($s1$ to $s14$), and the rightmost graphs show the mean accuracy from among all the participants.

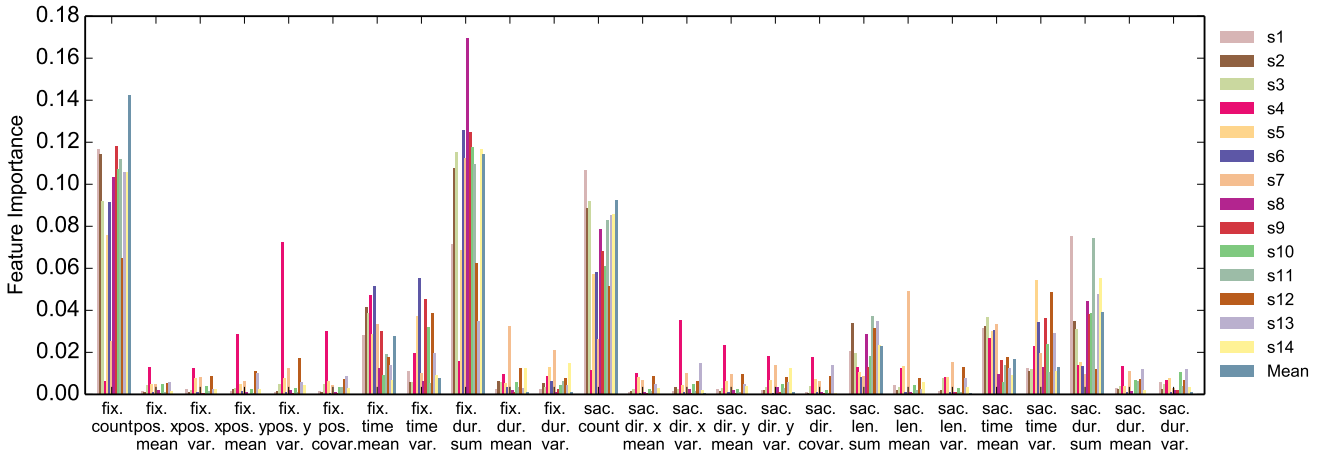

Figure 5. Feature Importances obtained through training process of random forest classifier. The graphs correspond to the importances of the 25 features listed in Table 1 and are color-coded according to the training data used.

prior works (Bee et al., 2006; Glaholt et al., 2009), as discussed in (Shimojo et al., 2003), the labeling task itself can affect the eye movements and the gaze cascade effect is not strongly observed during free viewing. From a practical point of view, its application is severely limited if the preference estimation can be done only when users are instructed to judge their preferences.

Figure 6 shows the performance of the gaze-based and image-based classifiers for the data recorded during the free viewing phase of the experiments. We used 400 pairs from the labeling phase as the training data for the target person, and the classifier was tested against 80 pairs from the free viewing phase. The first two graphs show the mean accuracy of the two image-based classifiers, and the rightmost graph shows the mean accuracy of the gaze-based classifier.

While it was less accurate than when using the test data from the labeling phase, the mean accuracy of the gaze-based classifier was 61% and still significantly higher than the results when using the metadata-based baseline methods (Wilcoxon signed-rank test: $p < 0.01$). However, it must be pointed out that the difference from the image-based classifiers was much smaller than in the previous cases and no statistically significant difference was found between the image-based and gaze-based classifiers (Wilcoxon signed-rank test: $p = 0.70$). This indicates there is an important limitation to the gaze-based preference estimation method; *i.e.*, the performance gain from image-based estimation method highly depends on the existence of a preference decision task and its performance is almost equivalent to the image-based estimation method in a free-viewing scenario.

For comparison, the third graph shows the results

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

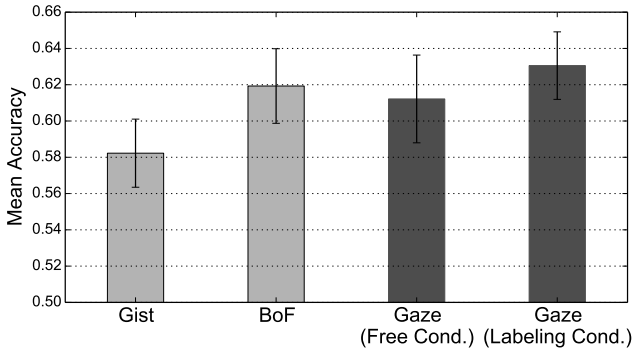

Figure 6. Performance comparison based on free viewing. The rightmost graph shows the mean accuracy, where the data from the labeling phase were used as the training data and the classifier was tested against the data from the free viewing phase. The first two graphs show the mean accuracy of the two image-based classifiers, and the third graph shows the results when using the data from the free viewing phase for both the training and testing.

when using the data from the free viewing phase for both the training and testing. The mean accuracy was evaluated by conducting a within-subject leave-one-out test. They are less accurate than when using the training data from the labeling phase; however, since the amount of training data from the free viewing phase was much lower than that from the labeling phase, a direct comparison was impossible. A detailed investigation using more training data will be an important future work.

Conclusion

We presented a data-driven approach for image preference estimation from eye movements. A labeled dataset of eye movements was collected from 14 participants that were comparing two images side by side under two conditions, free viewing and preference labeling. The feature vectors were composed of a set of fixation and saccade event statistics, and the random forest algorithm was used to build a set of decision trees. This allowed us to not only build image preference classifiers but also assess the contributions of each statistic element to the classification task.

The proposed classifier was more accurate than the metadata-based baseline methods, and the training process was shown to better improve the accuracy than a simple classification strategy using the fixation duration. While the training was shown to be effective even when using training data from different people, variations could be observed in the feature importances obtained during the training process.

We also compared the gaze-based preference estimation technique with the image-based methods based on generic image features. The classification performance of the gaze-based method was significantly better than the image-based methods, indicating the effectiveness of the data-driven approach for classification tasks that use eye movements. However, we observed a lower level of accuracy under the free viewing condition than under the labeling condition, and the performance was almost equivalent to the image-based estimation technique. This strongly suggests that characteristic eye movements are caused by the preference decision activity itself, and further investigation will be required to improve the accuracy of preference estimation under free viewing.

The image preferences when using our approach can be inferred from the eye movements during image browsing. This allows us to explore using the eye movements in new applications, *e.g.*, automatic image organization and summarization. Our future work will include extension of the learning-based preference estimation approach to single images. Since our experimental setting implies a two-item comparison task even without instruction, there can still be a task-related eye movement bias. On the other hand, the relationship between eye movements and subjective preference in the case of single images becomes more unclear and it will be increasingly important to more thoroughly look into the machine learning-based techniques.

References

Atkins, M. S., Moise, A., & Rohling, R. (2006). An application of eyegaze tracking for designing radiologists' workstations: Insights for comparative visual search tasks. *ACM Transactions on Applied Perception (TAP)*, *3*(2), 136–151.

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *14*(4), 21.

Bednarik, R., Vrzakova, H., & Hradis, M. (2012). What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the symposium on eye tracking research and applications* (pp. 83–90).

Bee, N., Prendinger, H., Nakasone, A., André, E., & Ishizuka, M. (2006). Autoselect: What you want is what you get: Real-time processing of visual attention and affect. In *Perception and interactive technologies* (pp. 40–52).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bulling, A., & Roggen, D. (2011). Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th international conference on ubiquitous computing (UbiComp 2011)* (pp. 455–464).

Bulling, A., Ward, J. A., Gellersen, H., & Troster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(4), 741–753.

Bulling, A., Weichel, C., & Gellersen, H. (2013). Eyecontext: Recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 305–308).

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3).

Chen, S., Epps, J., & Chen, F. (2013). Automatic and continuous user task analysis via eye activity. In *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 57–66).

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 288–301).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., & Schmid, C. (2009). Evaluation of gist descriptors for web-scale image search. In *International conference on image and video retrieval*.

Glaholt, M. G., Wu, M.-C., & Reingold, E. M. (2009). Predicting preference from fixations. *Psychology Journal*, *7*(2), 141–158.

Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*.

Jolliffe, I. (2005). Principal component analysis. In *Encyclopedia of statistics in behavioral science*. John Wiley & Sons, Ltd.

Jurie, F., & Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Proceedings of the tenth ieee international conference on computer vision (ICCV)* (pp. 604–610).

Ke, Y., Tang, X., & Jing, F. (2006). The design of high-level features for photo quality assessment. In *Proceedings of the ieee conference on computer vision and pattern recognition (CVPR)* (pp. 419–426).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the 10th european conference on computer vision (ECCV2008)* (pp. 386–399).

Marchesotti, L., Perronnin, F., Larlus, D., & Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of the ieee international conference on computer vision (ICCV)* (pp. 1784–1791).

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*(8).

Nishiyama, M., Okabe, T., Sato, I., & Sato, Y. (2011). Aesthetic quality classification of photographs based on color harmony. In *Proceedings of the 2011 ieee conference on computer vision and pattern recognition (CVPR2011)* (pp. 33–40).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pomplun, M., Sichelschmidt, L., Wagner, K., Clermont, T., Rickheit, G., & Ritter, H. (2001). Comparative visual search: A difference that makes a difference. *Cognitive Science*, *25*(1), 3–36.

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*(12), 1317–1322.

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the ninth ieee international conference on computer vision (ICCV)* (pp. 1470–1477).

Steichen, B., Carenini, G., & Conati, C. (2013). User-adaptive information visualization: Using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 317–328).

Toker, D., Conati, C., Steichen, B., & Carenini, G. (2013). Individual user characteristics and information visualization: Connecting the dots through eye tracking. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 295–304).

Yarbus, A. L., & Riggs, L. A. (1967). *Eye movements and vision* (Vol. 2) (No. 5.10). Plenum press New York.

DOI 10.16910/jemr.7.3.5

ISSN 1995-8692