Image Preference Estimation from Eye Movements with A Data-driven Approach

Yusuke Sugano
The University of Tokyo

Hiroshi Kasai
The University of Tokyo

Keisuke Ogaki
The University of Tokyo

Yoichi Sato
The University of Tokyo

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 pointed out over the years that eye movements can be used to infer the internal states of humans, there have not been many successes concerning image understanding. In this work, we investigate the possibility of image preference estimation based on a person's eye movements in a supervised manner. A data set of eye movements is collected when the participants are viewing pairs of natural images, and it is used to train an 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. The proposed classifier achieved a higher level of accuracy than the metadata-based baseline methods and a simple rule-based classifier.

Keywords: image preference, eye movements, machine learning

Introduction

Estimating the internal states of humans has been considered one of the most ultimate tasks for computers. In the field of image understanding, the subjective value and meaning of images often receive considerable attention from the research community. 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 values, such as the aesthetic quality (Luo & Tang, 2008; Nishiyama, Okabe, Sato, & Sato, 2011), using human-labeled data sets. 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 movements and their relationship to the human mind. Gaze input is achieving more and more attention recently amid the increasing demand for natural user interfaces, and casual gaze sensing techniques are becoming more increasingly available. However, in most of the application scenarios, a gaze is simply considered an alternative pointing input modality and the possibility of a gaze used as a cue to infer the mental states of the users has not been fully explored. While the view that the eye movement patterns of humans viewing images can reflect complex mental states has been widely shared among researchers (Yarbus & Riggs, 1967), very few studies have been conducted on actual classification tasks. It has even been pointed out that task classification based on eye movements is indeed a very challenging task (Greene, Liu, & Wolfe, 2012), it is still an open question as to what can be practically inferred from the eye movements.

We focus on preference estimation in this work for situations in which a user is comparing a pair of natural images. Shimojo *et al*. (Shimojo, Simion, Shimojo, & Scheier, 2003) reported the cascade effect of gaze. They showed that people tend to fixate on the preferred stimulus longer when they are asked to compare two stimuli and make a two-alternative forced choice. Following this study, several methods have been proposed to predict preferences from eye movements (Bee, Prendinger, Nakasone, André, & Ishizuka, 2006; Glaholt, Wu, & Reingold, 2009). However, the main focus of these studies is a comparison between the same categories of stimuli such as 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 the preference between two natural images during free viewing. Although the eye

This work is supported by CREST, JST.

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 mainly make two contributions in this paper. First, we take a data-driven approach to the preference estimation task. We train a classifier that outputs image preference labels using a data set of eye movements recorded while users are comparing image pairs. It becomes possible to assess the important features for preference estimation and how they differ among different people by using an algorithm that exploits the beneficial features for the classification task. Second, we quantitatively investigate the effectiveness of the preference estimation in a more realistic scenario in which the users are freely viewing image pairs without instruction. Using our method, image preference can be estimated from the natural eye movements, and this is expected to enhance the possible applications, such as for automatic shot selection and photo collection organization.

Method

We assume a situation in this study where the users are viewing a pair of natural images displayed side by side. Our goal is to classify which image the user prefers from the patterns of the eye movements during the comparative viewing. We address this task in the supervised manner that we mentioned above. A binary classifier is trained to output unknown preference labels from the eye movement patterns based on the ground-truth labels for the image preferences. Since it is not clear what kind of measures could contribute to the classification task, various fixations and saccades 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. We describe the details for the experimental setting, the input feature vector used in our method, and the learning procedure of the classifier in the following sections.

Experimental Settings

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.

Figure 1. Experimental setup

The experiment had two phases: free viewing and preference labeling tasks. After calibration, 11 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 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. Throughout the experiments, data was discarded if the participant pressed the wrong key by mistake or a saccade event happened on only one side.

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 have 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 make the 480 image pairs described above.

Eye Movement Feature

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

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

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.

We first follow a standard procedure that is based on the velocity threshold to extract the fixation and saccade events from these data. 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 from 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 sets of fixations and saccades. 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 $\boldsymbol{x} = (\boldsymbol{f}_L^{\mathrm{T}}, \boldsymbol{f}_R^{\mathrm{T}})^{\mathrm{T}}$ of a paired image.

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

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

Preference Classification

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

Because of 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 $\boldsymbol{x} = (\boldsymbol{f}_L^{\mathrm{T}}, \boldsymbol{f}_R^{\mathrm{T}})^{\mathrm{T}}$ is associated with label $y = 1$, while the left-right flipped feature vector $\boldsymbol{x} = (\boldsymbol{f}_R^{\mathrm{T}}, \boldsymbol{f}_L^{\mathrm{T}})^{\mathrm{T}}$ can also be used with label $y = -1$ for training.

Random forest (Breiman, 2001) is a method of supervised classification using a set of decision trees. Given a set of training samples, the random forest algorithm trains the decision trees using random 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. In the experiments, the classifiers are implemented using the scikit-learn library (Pedregosa et al., 2011)[2]. The number of trees was set to 1000, and the depth of each tree was restricted to 3.

Results

Classifier Performance

Figure 2 shows a comparison of the preference classifier with the baseline methods. Accuracy scores

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

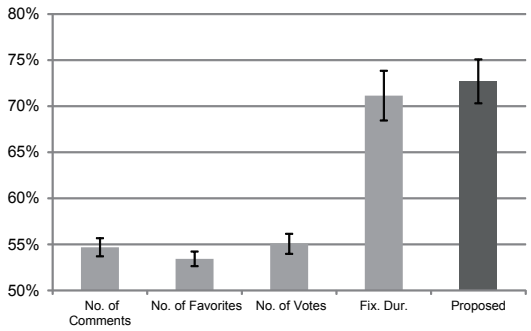

Figure 2. Comparison with baseline methods. The graphs show the mean accuracies from the 11 participants and the error bars indicate the standard errors. The first three graphs show the classification results using the objective metadata, which was the number of comments and favorites on the Flickr website, and the local votes from the other participants. The fourth graph shows the accuracy of the simple classification using only the sum of the fixation duration, and the last graph corresponds to the proposed method.

were used for evaluation because 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). The third graph shows the results when using more local metadata, which were the preferences of the other participants. Since the same image combinations were given to all the participants, the side that received more votes from the other participants was treated as the classification result. If the numbers of votes were the same, the pair was regarded as a misclassification.

We additionally show a simple classification result using only the sum of the fixation duration to assess the meaning of the data-driven training. The fourth graph shows the accuracy of the classification strategy where the sides with the longer fixation duration were treated as the output labels.

The proposed classifier was trained and tested in a leave-one-out manner using the personal training data sets 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 11 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. Although the third method using the local votes achieved the best accuracy, even the local choices can diverge and it is not easy to estimate the image preferences without observing the target person. The mean accuracy of the proposed method was 73%, and 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 accuracy, the performance was improved by using the proposed training approach (Wilcoxon signed-rank test: $p = 0.02$).

Cross-subject Training

In the previous section, the trainings were done using the personal data sets. While this follows the standard procedure for supervised classifications, it is not always possible to collect the 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 3 shows an accuracy comparison between the within-subject and cross-subject training conditions. 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 are done in a leave-one-subject-out manner; the classifier is trained for each person using the data from the other 10 participants. Each graph in Figure 3 corresponds to a participant ($s1$ to $s11$), 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 $s4$, the cross-subject training generally achieved a comparative accuracy and there was no statistically significant difference in the mean scores (Wilcoxon signed-rank test: $p = 0.91$). This indicates that the proposed framework could successfully capture discriminative eye movements that can be commonly observed among different people.

Feature Importances

The feature importances obtained through the random forest classifier training process are shown in Figure 4 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. In our case, the feature importances are computed as a fraction of the samples each of the elements contributed to the final prediction. The higher value thus means there was more contribution to the classification.

Our 50-dimensional 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

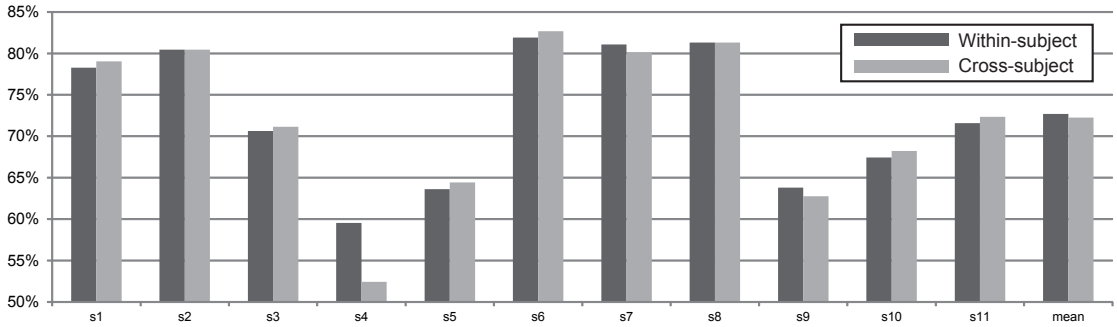

Figure 3. Cross-subject training. The within-subject condition corresponds to the leave-one-out training and testing for each participant. In the cross-subject condition, the classifier is trained for each person using the data from the other participants. Each graph corresponds to a participant ($s1$ to $s11$), and the rightmost graphs show the mean accuracy from among all the participants.

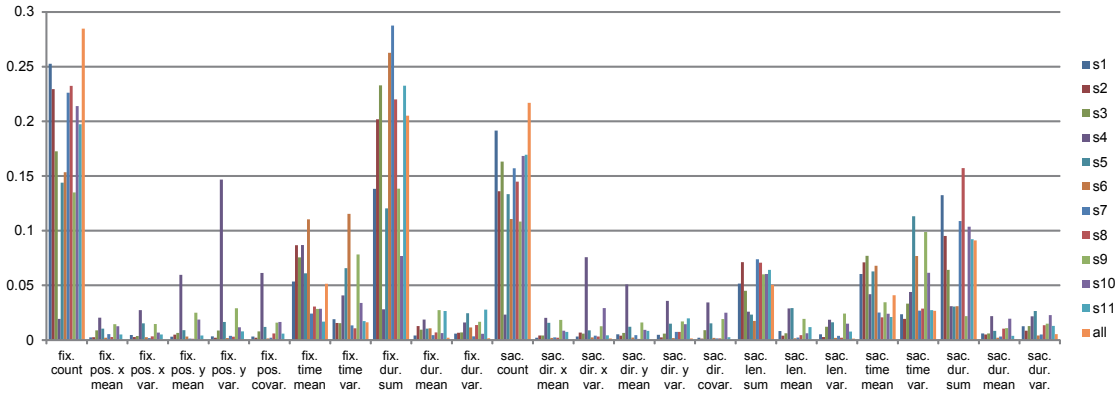

Figure 4. Feature Importances obtained through training process of random forest classifier. The graphs correspond to the importances of 25 features listed in Table 1 and are color-coded according to the training data used.

and the right side) theoretically have the same importance through the training process, and the sums of the two values are shown in Figure 4. 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 $s11$ indicate the within-subject training condition, *i.e.*, the feature importances obtained when personal training data sets were used. *All* indicates the case when all of the data from the 11 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 3, has a unique distribution compared to the other participants, and the fixation position was the key to the improvement.

Free Viewing

The results discussed in the previous sections were based on the data set obtained during the labeling phase, where the participants were instructed to assign preference labels. While this setting is the same as in the 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 5 shows the performance of the proposed classifier for the eye movements recorded during the free viewing phase of the experiments. The rightmost graph shows the mean accuracy when using the within-subject training data of the labeling phase. 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. While it was less accurate than when using the test data from the labeling phase, the mean accuracy was 61% and still significantly higher than the results based on the local

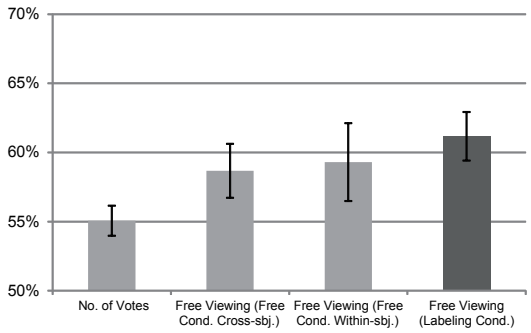

Figure 5. Performance comparison based on free viewing. The rightmost graph shows the mean accuracy, where the data from the labeling phase were used as the training data and the classifier was tested against the data from the free viewing phase. The other two graphs show the results when using the data from the free viewing phase for both the training and testing under within-subject and cross-subject training conditions.

votes (the leftmost graph, Wilcoxon signed-rank test: $p < 0.01$).

For comparison, the other two graphs show the results when using the data from the free viewing phase for both the training and testing. The mean accuracies of the within-subject leave-one-out test and cross-subject leave-one-subject-out test are both shown, and they are less accurate than the previous case 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 data set of eye movements was collected from 11 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 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. This indicates the effectiveness of the data-driven approach for classification tasks that uses eye movements.

The classification could be done under the free viewing condition. However, we observed lower accuracy than the test under the labeling condition. While it strongly suggests that the characteristic eye movements are caused by the preference decision activity, the performance gain obtained via the data-driven training is promising enough for further improvement.

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 proposed approach to single images and other tasks for estimating subjective values.

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.

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.

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).

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*.

Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *Proc. The 10th European Conference on Computer Vision (ECCV2008)* (pp. 386–399).

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 *Proc. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)* (pp. 33–40).

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.

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