RedFame

Journal of Education and Training Studies
Vol. 10, No. 4; October 2022
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Research on Image Retrieval Optimization Based on Eye Movement Experiment Data

Tianjiao Zhao[1], Mengjiao Chen[2], Weifeng Liu[2], Jiayi Jia[2]

[1]Tianjin University, College of Intelligence and Computing, China

[2]Tianjin University School of Mechanical Engineering, China

Correspondence: Weifeng Liu, Room A139, 37 Building, Tianjin University, Tianjin, China.

Received: October 18, 2021 Accepted: June 29, 2022 Online Published: July 1, 2022

doi:10.11114/jets.v10i4.5380 URL: https://doi.org/10.11114/jets.v10i4.5380

Abstract

Satisfying a user's actual underlying needs in the image retrieval process is a difficult challenge facing image retrieval technology. The aim of this study is to improve the performance of a retrieval system and provide users with optimized search results using the feedback of eye movement. We analyzed the eye movement signals of the user's image retrieval process from cognitive and mathematical perspectives. Data collected for 25 designers in eye tracking experiments were used to train and evaluate the model. In statistical analysis, eight eye movement features were statistically significantly different between selected and unselected groups of images ($p < 0.05$). An optimal selection of input features resulted in overall accuracy of the support vector machine prediction model of 87.16%. Judging the user's requirements in the image retrieval process through eye movement behaviors was shown to be effective.

Keywords: image retrieval, eye tracking, visual search, support vector machine, user requirements

1. Introduction

The emergence of search engines has made daily life more convenient. However, as the volume of available image data grows dramatically, how to determine a user's requirements for image information more effectively and quickly becomes a question that needs to be answered urgently (Ai et al., 2013; Sugano et al., 2014). Currently, there are two main methods of image retrieval: text-based retrieval and content-based retrieval (Liu et al., 2007). Text-based image retrieval allows users to describe information requirements naturally and intuitively using advanced features (Chai et al., 2007; Liu et al., 2007). Content-based image retrieval (CBIR) uses computer vision technology to extract low-level features (e.g., shapes, colors, and textures) automatically in organizing digital image archives for an image-based search (Datta et al., 2008; Liu et al., 2007).

The past two decades have seen much research on image retrieval technology (Zhang et al., 2012). However, most studies eagerly pursue algorithms that are more optimized for analyzing users' input information to perform retrieval more efficiently without considering whether the input information fits the users' true underlying needs sufficiently. For example, when the user's actual desire is for an image of a white, high-performance sneaker that follows the current fashion, the keyword or picture that he or she inputs may only contain a white sneaker. There are two reasons for this phenomenon. First, inputting complex keywords or finding a picture more similar to the desired one can be too much work for the user. Second, the user may not know exactly what his or her target is when entering information. This problem is not exactly the same as that of the current semantic gap in mainstream CBIR research, which mainly concerns the limited ability of low-level image features to describe a user's high-level semantics (Chen et al., 2003; Smeulders et al., 2000).

In current human–computer interaction scenarios, people tend to prefer contactless interactions most (Chen et al., 2013). Contactless interactions are considered to be natural (Su et al., 2014) and have three advantages. First, they are simple and have low learning costs (Hariharan et al., 2014; Motta et al., 2012). Second, they are efficient in terms of time and manpower. Third, they are safe and hygienic in terms of avoiding the contact transmission of bacteria and viruses. As a key element of contactless interaction, vision is considered to be a physiological signal that indicates the user's cognitive state and intention directly (Anagnostopoulos et al., 2017; Han, Pereira, 2013; Haosheng Huang, George Gartner, 2012). Researchers have long used eye movements to study expertise (Orquin, Mueller Loose, 2013) and cognitive processes such as scene viewing, problem solving, the completion of natural tasks, and visual searching (Tseng, Howes, 2015).

Not until recently has eye tracking technology been applied in the field of information retrieval (Li et al., 2016; Papadopoulos et al., 2014; Vrochidis et al., 2011). Oyckoya and Stentiford (2005) compared visual input with mouse input and found that in a target recognition task, eye tracking was faster than the use of the mouse. They then _ENREF_33conducted an analysis of variance of eye movement data for image retrieval and found that combining eye tracking data with a similarity calculation provides better retrieval performance than random selection based on the same similarity information (Oyekoya, Stentiford, 2007). Zhang et al. (2010) pointed out a method of classifying images by training decision trees using features of eye movement. Papadopoulos et al. (2014) suggested an image retrieval method based on gaze feedback using a single-camera image to process eye movement features; however, compared with the use of eye trackers, their method has a lower sampling rate and yields less precise data of eye movement. To address the semantic gap in CBIR, Li et al. (2016) claimed a threshold strategy and included an analysis of the gaze time in the analysis of eye movement features. Kim et al. (2018) proposed that two eye-tracking measures, namely the saccade duration and time-window-based inter-fixation duration, can be used as indicators to evaluate the performance of visual search tasks in a computer-based environment. Among the above studies, Papadopoulos et al. (2014) and Li et al. (2016) used only a single feature variable of eye movement as feedback for the image retrieval. They did not systematically study eye movement features of image retrieval or create a predictive model with broad applicability. Additionally, in their experiments, they asked the participants to perform retrieval tasks with given images. Therefore, their experimental tasks did not reflect the underlying needs of the participant and they may not have engaged the participant in a realistic scenario of image retrieval.

Extending on previous work, this paper proposes a user-oriented image retrieval method based on visual feedback. We used statistical methods to analyze the relationship between eye movement data and subjective selection in scenarios of real-image retrieval. We then used machine learning to predict the subjective selection results. To this end, we collected eye movement records for 25 participants (i.e., 1835 eye movement data). Moreover, we generalized the eye movement data for the purpose of obtaining data with broad applicability. Using the data to train a support vector machine (SVM) classifier and using the G-mean to verify the validity of the classifier, we obtained a high-precision prediction model that can be used for feedback in an image retrieval system. This paper makes three notable contributions to the literature.

1) The paper proposes a contactless feedback method for image retrieval that uses eye-movement signals instead of manual input.
2) The paper constructs a model that can infer retrieval requirements from multidimensional features of eye movement.
3) The paper clarifies the meaning of eye movement features by establishing the correlation between eye movement signals and subjective selection in the process of image retrieval.

Our methods and findings can help clarify the role that visual signals play in image retrieval. As eye tracking technology becomes more readily available, it will provide critical support for the development of emerging applications in image retrieval. By then, intelligent, clean, and efficient human–computer interactions will be mainstream.

2. Research Methods

2.1 Participants

In visual search tasks, experts, compared with novices, generally have faster average eye movement (Zangemeister et al., 1995), higher information processing efficiency (Chapman,Underwood, 1998), and eye movements that are more effective (Kim et al., 2018). In this study, we improved the efficiency of image retrieval by designers as well-trained image retrievers. As a group that searches for images frequently, designers consider that image-based materials have simple and intuitive features that can inspire their creative process (Malaga, 2000). Designers therefore view a large number of images when designing (Lang et al., 2001). Taking a designer's image retrieval behavior as a research object may help us to identify visible features of eye movement, such as more attention being paid to relevant areas and less attention to irrelevant areas (Gegenfurtner et al., 2011). Furthermore, this study takes designers as an example group of users in seeking approaches to improving the image retrieval process for general users. This is a convenient approach for the early exploration of the potential relationship between image retrieval and eye movement data.

We enrolled 25 participants who are undergraduates majoring in industrial design or visual art design. These participants included 11 males and 14 females and had an average age of 21.6 years. All participants had at least 2 years of design experience. To reduce measurement error, we explicitly required participants without astigmatism, high myopia, or other eye problems when recruiting them. All participants gave written informed consent in accordance with the guidelines of our ethical board committee. The study complied with the Declaration of Helsinki for human experimentation.

2.2 Setup

Current eye tracking devices track and detect eye movements through corneal reflection. Corneal reflection refers to the use of a light source to illuminate the eye and generate an image, and a high-resolution camera detects the image from the cornea and pupil. After the image is obtained, the eye's fixation point on the stimulus can be determined using advanced image processing algorithms.

The eye tracking device used in the present experiment was a Tobii Pro TX300 screen-type eye tracker having a sampling rate of 300 Hz, a plug-and-play 23-inch display, and Tobii studio eye tracking data acquisition and processing software. Additionally, cameras recorded the process of the experiment and the experimental results were viewed on a laptop computer. The experiment was carried out in a room without environmental interference, such as that of strong light, noise, and other distractions. After the experiment, the eye tracking data were exported by Tobii studio software directly and read by Excel software.

Excessive communication was avoided as additional workload during the experiment would have distorted the eye movement data (N. Liu, Yu, 2017). We recorded the experiment and conducted post-event interviews.

2.3 Material Collection

We collected one sketch drawn by each participant for different topics, thus obtaining a total of 25 sketches. We then retrieved similar images on a 200,000-image database recently published by a national patent gallery. We here used CBIR technology because it relies on computers to perform automatic feature extraction and indexing, especially in the case of big data, and the accuracy and efficiency are much better than those of manual identification (Arandjelović et al., 2018; Ren et al., 2015). We thus retrieved approximately 90 images for each sketch. The image results were used as a material gallery for the selection experiment (Fig. 1). There were two advantages to this approach. First, we obtained the experimental gallery by retrieving images based on the participant's own sketch, which conformed to the designer's thinking process and helped the participants to engage with the real-image retrieval task quickly. Second, the retrieved gallery contains more useful images for designers than a random gallery. Using the retrieved gallery in subsequent experiments increases the proportion of forward samples (i.e., samples selected by the participant), allowing us to obtain more data from the participants and improving the accuracy of the prediction model.

Figure 1. Creation of galleries

In the selection experiment, there was one task per participant. For each task, we chose 84 images from the gallery and arranged them in a 7 × 4 grid on each of three screen pages (i.e., interfaces) (Fig. 2). The images on the interfaces were uniform in size (Parush et al., 2005). There are three reasons for adopting this arrangement. First, when designers browse a large number of images on a web page, they tend to browse multiple images on the same interface instead of browsing through the images one by one. This arrangement thus reflects the actual browsing process. Second, this arrangement enables us to clarify the browsing habits of the participants and obtain more eye movement information. Third, unifying the surface size reduces the effect of bottom-up attention on eye movement (Chandon et al., 2009; Orquin, Mueller Loose, 2013).

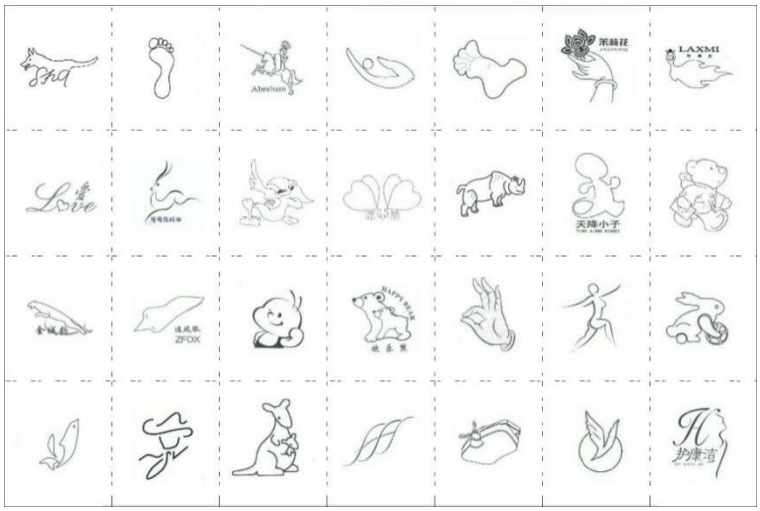

Figure 2. Image arrangement on each interface

2.4 Process

To avoid the participant's expectations affecting the results (Smith, Kosslyn, 2007), the participants completed the task in accordance with the researcher's guidance step by step, rather than being informed of all task details at the beginning. We first asked the participants to think about how to improve the design of the previous sketch while displaying their previous works on screen. After 5 minutes, we asked the participants to browse the galleries obtained by searching through their sketches with the eye tracker program on. To adapt to the browsing habits of the participants, we did not limit the browsing time for each interface. The participants clicked on a mouse to indicate that they had completed browsing on an interface. The browsing time is denoted t_1. The participants next marked useful pictures on the first interface with the mouse. The marking time is denoted t_2 (Fig. 3). This process was repeated on the second and third interfaces. The screen-type eye tracker had little presence and the participant's browsing process thus felt natural and realistic. We finally interviewed the participants while replaying the experimental process through the laptop. We asked about the selection, recalling images that had been fixated on for a long time, and the problems encountered during the selection process.

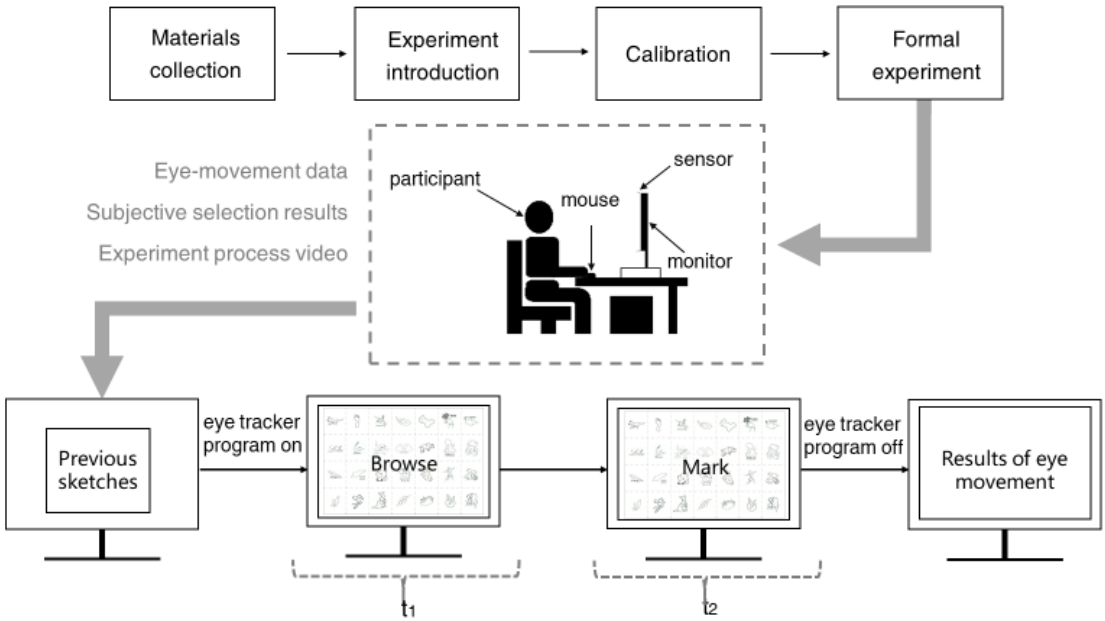

Figure 3. Experimental process

2.5 Data Collection

More than 20% of the samples of the eye-movement data were lost for each of three participants because of the limited accuracy of the eye tracker, head wobble, and reflections in the external environment. Data for the other 22 participants were higher in quality and thus retained for follow-up processing. The set of data for each participant included a series of eye-movement data, subjective selection results, the duration of the experimental process, and the interview record.

3. Relationship Between Eye-Tracking Signals and Subjective Selection

We obtained the order and duration data of each gaze point during the participant's browsing process (t_1). In Fig. 4, gaze points are indicated by dots whereas saccades are indicated by lines that connect the dots. The labels of the gaze points indicate the sequence of the gaze, and the size of the gaze points indicates the duration of the gaze. Processing data through statistical analysis can help us discover the connection between eye-tracking signals and subjective selection accurately and distinguish invalid data.

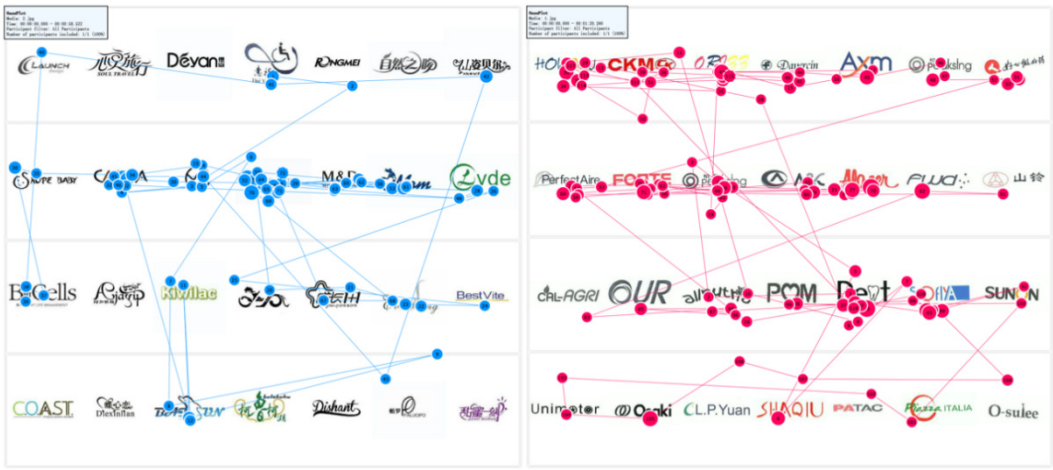

Figure 4. Illustration of eye tracking data

3.1 Areas of Interest

An area of interest (AOI) refers to setting a specific area of the imagery as an area of attention. Researchers can then obtain data of eye movement for each AOI. We kept the size of each AOI consistent. In addition, each AOI is centered on the geometric position of the image (Fig. 5). We exported five features for each AOI: the time to first fixation (TFF), fixation before (FB), first fixation duration (FFD), total fixation duration (TFD), and fixation count (FC).

Figure 5. AOIs

3.2 Data Processing

We exported 1825 pieces of valid data. Each participant had a different method of browsing, and summations of their TFDs and FCs for a single interface were therefore different. We thus denote the summation of the fixation duration on an interface as $Sum_{(TFD)}$, and the summation of fixation counts on an interface as $Sum_{(FC)}$. We calculated the percentage of each feature relative to $Sum_{(TFD)}$ or $Sum_{(FC)}$ to obtain five additional features. This negates the effects of individual participants' thought processes and browsing modes on the data, such that each individual's results better reflect the common trends. Following this processing, we consider that the processed data have the same meaning as the two sets of original data in Table 1.

Table 1. Five features of the two sets of data having the same meaning after removing the interference of $Sum_{(TFD)}$ and $Sum_{(FC)}$

AOI	TFF (Seconds)	FB (Counts)	FFD (Seconds)	TFD (Seconds)	FC (Counts)
001					
$*Sum_{(TFD)}$ =5s	1	1	1	1	1
$Sum_{(FC)}$ =5c					
002					
$*Sum_{(TFD)}$ =50s	10	10	10	10	10
$Sum_{(FC)}$=50c					

We do not, however, consider the original five data features to be useless. They may contain deeper connections and information that need to be understood. In addition to the above processing, images that were not selected by the participants were classified into group A whereas those that were selected were classified into group B.

3.3 Relationship Between Feature Variables and Selection Results

We analyzed the distribution of the 10 feature variables (in Table 2) in groups A and B. We first conducted a Kolmogorov–Smirnov test to determine the normality of all feature variables in groups A and B. Their significances are less than 0.05 ($p < 0.05$), rejecting the null hypothesis; that is, no feature variables obey a normal distribution. A nonparametric test is required for non-normally distributed data. We hence used the Man–Whitney test to compare the samples. In the results, $p > 0.05$ for the TFF and FFD, whereas $p < 0.05$ for the other variables. Therefore, the TFFP, FB, FBP, FFDP, TFD, TFDP, FC, and FCP have statistically significant differences between groups A and B. This shows that the processed data features are suitable for use in the prediction model.

Table 2. Feature variables

Feature variables	Full name	Meaning
TFF	time to first fixation	the fixation duration from the first appearance of the stimulus material to the gazing point in the AOI
TFFP	the percentage of TFF relative to $Sum_{(TFD)}$	$TFFP = TFF/Sum_{(TFD)} \times 100\%$
FB	fixation before	the counts of fixation from the first appearance of the stimulus material to the gazing point in the AOI
FBP	the percentage of FB relative to $Sum_{(FC)}$	$FBP = FB/Sum_{(FC)} \times 100\%$
FFD	first fixation duration	the duration of the first gazing point in the AOI
FFDP	the percentage of FFD relative to $Sum_{(TFD)}$	$FFDP = FFD/Sum_{(TFD)} \times 100\%$
TFD	total fixation duration	the sum of the fixation duration of all gazing points in the AOI
TFDP	the percentage of TFD relative to $Sum_{(TFD)}$	$TFDP = TFD/Sum_{(TFD)} \times 100\%$
FC	fixation counts	the sum of the fixation counts of all gazing points in the AOI
FCP	the percentage of FC relative to $Sum_{(FC)}$	$FCP = FC/Sum_{(FC)} \times 100\%$

We also calculated the mean, maximum, and minimum values for each set of variables in groups A and B (in Table 3). A comparison of the mean of the TFFP between the two groups ($M_A = 36.63$, $M_B = 29.94$) shows that the selected images tend to have characteristics that attract a designer's attention earlier than the characteristics of the unselected images. The mean, minimum, and maximum values of the FB and FBP in group B (FB: $M_B = 26.39$, $min_B = 0$, $max_B = 100$; FBP: $M_B = 29.33$, $min_B = 0$, $max_B = 97.73$) were significantly smaller than those for group A (FB: $M_A = 30.59$, $min_A = 0$, $max_A = 187$; FBP: $M_A = 36.41$, $min_A = 0$, $max_A = 129$). This supports the description of the TFFP from the aspect of fixation points. There is no obvious difference in the FFDP between groups A and B ($M_A = 1.06$, $min_A = 0.04$, $max_A = 8.32$, $M_B = 1.11$,

$min_B = 0.03$, $max_B = 8.05$). In terms of numerical values, images in group B (TFD: $M_B = 2.30$, $min_B = 0.13$, $max_B = 13.20$; TFDP: $M_B = 7.53$, $min_B = 0.49$, $max_B = 47.04$) usually yield a longer fixation duration than those in group A (TFD: $M_A = 0.76$, $min_A = 0.03$, $max_A = 7.11$; TFDP: $M_A = 2.79$, $min_A = 0.12$, $max_A = 23.45$) in the test results. The difference in values is significant, and we thus consider that the TFD and TFDP are discriminative feature variables. Similarly, FC and FCP (FC: $M_A = 3.06$, $min_A = 1$, $max_A = 18$, $M_B = 8.41$, $min_B = 1$, $max_B = 52$; FCP: $M_A = 3.66$, $min_A = 0.52$, $max_A = 25.40$, $M_B = 8.77$, $min_B = 1.14$, $max_B = 40.51$) are also discriminative feature variables.

Table 3. Statistical analysis of the data

Feature variable	Kolmogorov-Smirnov Test	Mann-whitney Test	Mean (M)		Minimum（min）		Maximum（max）	
			A	B	A	B	A	B
TFF	0.000	0.216						
TFFP	0.000	0.006	36.63	29.94	0.00	0.00	99.85	97.90
FB	0.000	0.047	30.59	26.39	0	0.00	187	100
FBP	0.000	0.002	36.41	29.33	0	0.00	129	97.73
FFD	0.000	0.688						
FFDP	0.000	0.040	1.06	1.11	0.04	0.03	8.32	8.05
TFD	0.000	0.000	0.76	2.30	0.03	0.13	7.11	13.20
TFDP	0.000	0.000	2.79	7.53	0.12	0.49	23.45	47.04
FC	0.000	0.000	3.06	8.41	1	1	18	52
FCP	0.000	0.000	3.66	8.77	0.52	1.14	25.40	40.51

3.4 Comparison of the Times Required for Browsing and Manual Marking

We recorded the browsing time and manual marking time of each participant on each interface during the experiment. The average and proportion of each participant's browsing time and marking time are shown in Fig. 6. For the overall experiment, the average browsing time was 32.28 s, the average marking time was 16.45 s, and the ratio of the marking time to the entire image retrieval time was 33.76%. For a single participant, the ratio of the marking time to the entire image retrieval time was a maximum of 53.23% and a minimum of 16.67%.

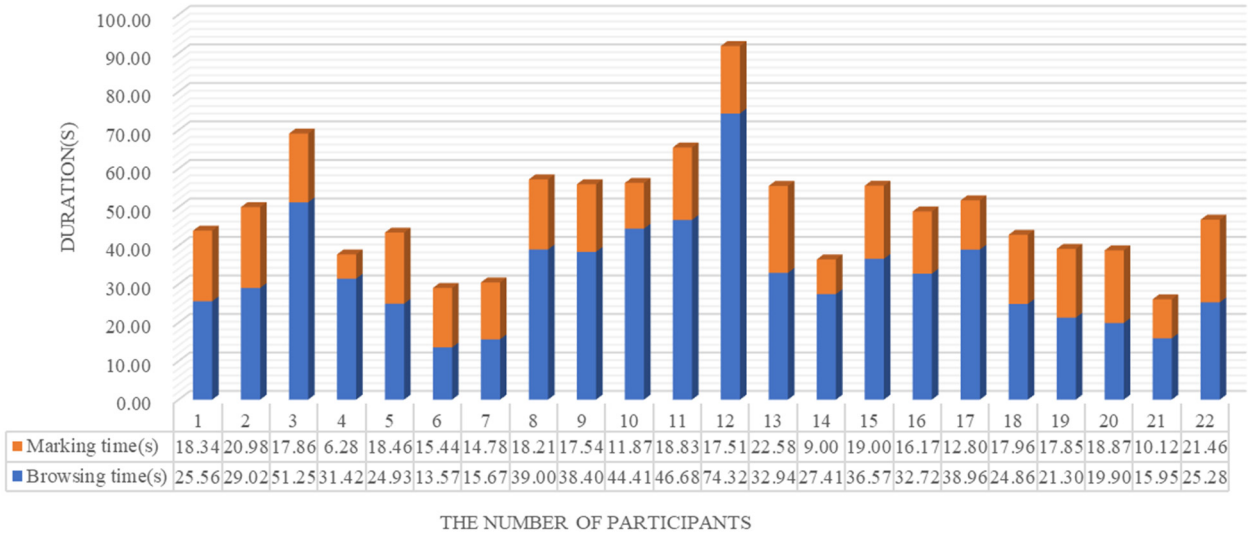

Figure 6. Statistics of the times required for marking and browsing during image retrieval

4. Designer Image Selection Model

4.1 Principle of the SVM

To establish a designer selection model with good predictive performance, we first need to solve the nonlinear correlation between the eye movement data and the subjective selection result. Conventional models of statistical predictive analysis, such as models of multiple regression analysis and multivariate analysis, often fail to predict nonlinear data well owing

to their use of linear assumptions. The SVM method is a general learning method derived from statistical learning theory. In contrast to neural networks, decision trees, and other traditional methods based on empirical risk minimization criteria, the SVM is based on structural risk minimization criteria. Moreover, it has a solid theoretical foundation and stronger approximation and generalization abilities, and it performs well at solving nonlinear dichotomy problems (Vapnik, 1995).

As an example, we assume that there are two types of data in a two-dimensional space that need to be classified, as shown in Fig. 7, where different colors represent different data. If we want to separate the two types of data, we need to find a bisecting line. However, there are many lines, such as B1, B2, and other series of lines that divide the two types of data. The question then becomes how to evaluate which lines (hyperplanes) are good or bad. First, the bisecting line must separate the two types of data as much as possible, so that the training samples that we obtain are highly accurate. Second, we consider that B1 is better than B2 because of the principle of the maximum margin. Here, we draw upper and lower lines parallel to each of B1 and B2 that border the data and determine the margin between the upper and lower lines. The margin of B1 is clearly larger than that of B2. A model having a larger margin will have lower complexity and thus be more robust, leading to the accuracy and strong generalization ability of the training sample.

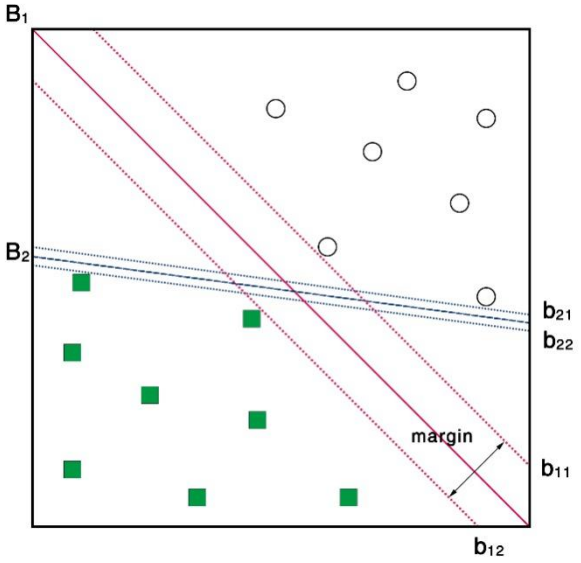

Figure 7. Example of a linear classifier

We assume that the middle line (hyperplane) is

$$\vec{w} \cdot \vec{x} + b = 0. \tag{1}$$

The upper and lower lines are then

$$\vec{w} \cdot \vec{x} + b = 1, \tag{2}$$

$$\vec{w} \cdot \vec{x} + b = -1, \tag{3}$$

where $\vec{x}_1$ and $\vec{x}_2$ exist on the upper and lower boundaries, respectively. Substituting these expressions into Eqs. (2) and (3) yields

$$\|\vec{w}\|\|\vec{x}_2 - \vec{x}_1\| \cos\theta = 2, \tag{4}$$

where θ is the angle between $\vec{x}_2 - \vec{x}_1$ and $\vec{w}$, . We thus obtain

$$\text{Margin} = \|\vec{x}_2 - \vec{x}_1\| \cos\theta. \tag{5}$$

Substituting Eq. (5) into Eq. (4) yields

$$\text{Margin} = 2/\|\vec{w}\|, \tag{6}$$

$$L(w) = \|\vec{w}\|^2/2. \tag{7}$$

From Eqs. (2) and (3), we have

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1. \tag{8}$$

We therefore find the minimum value of $L(w)$ when Eq. (8) is satisfied. Equation (8) satisfies the accuracy of classification whereas Eq. (7) guarantees that the model is as simple as possible and the generalization ability as good as possible. This optimization problem is a quadratic programming problem, and a globally optimal solution can thus be found.

We can transform the nonlinear problem into a linear problem in a high-dimensional feature space. We then find the optimal classification surface in the transformed space. When constructing the optimal hyperplane in the feature space, the training algorithm only needs to use the inner product in the feature space; namely, $\varphi(x_i) \cdot \varphi(x_j)$. We therefore use the appropriate inner product $K(x_i, x_j)$ in the optimal classification plane (Shieh,Yang, 2008) and ensure that it satisfies

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j). \tag{9}$$

We thus achieve a linear classification after a nonlinear transformation.

4.2 Model Training

We used the libsvm library to create a designer image selection prediction model. The libsvm library is a simple, easy-to-use, fast, efficient, and general-purpose SVM software package (Chang, Lin, 2011) used to solve classification problems and estimate distributions.

4.2.1 Class Imbalance of the Samples

According to the aims of the experiment, we did not impose requirements on the number of images selected by the participant. After the unrestricted selection of the participants, the number of negative samples was appreciably higher than the number of positive samples, leading to a class imbalance problem. In this case, if SVM predictive analysis is used directly, the separated hyperplane will be biased toward the positive class, which will eventually reduce the prediction accuracy.

In an SVM, the value C is a the variable responsible for penalizing misclassified data. One way to solve the class imbalance problem is to increase the weighted C value according to the class. The main idea is to increase the effect of the misclassification of the less common class, ensuring the correctness of its classifications and preventing it from being overwhelmed by another class.

4.2.2 Data Normalization

Data were normalized such that they fell within the interval [0,1]. The experimental data involve both the time and count, which have different units. Normalization removes the limits of the units of the data. The original data were converted into pure dimensionless values so that variables having different units or magnitude could be compared and weighted. The normalization formula is

$$X_i' = X_i - X_{min}/X_{max} - X_{min}.$$

Normalizing the data can improve the speed of convergence and the accuracy of the model. In addition, we digitized the dependent variables by labeling the unselected samples as 0 and the selected samples as 1.

4.2.3 Prediction Model

There were 1825 samples in total, of which 1200 (65.75%) were used for training and 625 (34.25%) were used for testing. The FB, TFD, FC, TFFP, FBP, FFDP, TFDP, and FCP, which were shown to be statistically significant, were used as input features. There were eight input features in total. The output was the subjective selection result. We built a prediction model by training an SVM using the samples, such that it was assigned the weight of each input feature. After several iterations, an optimal model containing all input features was obtained. The model had a prediction accuracy of 86.21% for class 1 and 82.38% for class 0.

The accuracy of the prediction model is not sufficiently sensitive to the positive classification results because of the class imbalance problem of the samples. The G-mean can consider the performance of the two classes at the same time, solving the model evaluation problem of the sample class imbalance effectively. The G-mean is calculated as

$$\mathrm{G-mean} = \sqrt{\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \times \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})},$$

where TN is the number of actual negative samples that are predicted as negative samples (i.e., true negatives), FP is the number of actual negative samples that are predicted as positive samples (i.e., false positives), FN is the number of actual positive samples that are predicted as negative samples (i.e., false negatives), and TP is the number of actual positive samples that are predicted as positive samples (i.e., true positives).

We used each feature as a separate input feature to construct several models. We concluded from the prediction results (Table 4) that the prediction effect of a single feature is not as good as that of all features together. The FFDP has the worst predictive performance, which is consistent with the analysis in Section 3.

Table 4. Prediction results of the SVM model obtained using each variable as an input feature

Feature	Accuracy (0)	Accuracy (1)	G-mean	G-mean sorting
FB	0.1050	0.8965	0.3068	7
TFD	0.8675	0.7931	0.8294	2
FC	0.9160	0.6890	0.7944	4
TFFP	0.3860	0.6900	0.5161	6
FBP	0.3890	0.7580	0.5430	5
FFDP	0	1	0	8
TFDP	0.7886	0.8966	0.8409	1
FCP	0.7886	0.8276	0.8079	3
All variables	0.8305	0.8621	0.8462	

4.3 Optimization of the Prediction Model

It is uncertain whether a larger number of input features will lead to a more predictive model. The accuracy of the prediction model varies with the number of features and the data set. This is a phenomenon called the Hughes effect. It is necessary to choose input features that provide a better prediction model. Through the evaluation of each model in Table 4, the features with better performance are selected and combined to optimize the final model.

In the model optimization process, we removed the FFDP as input. The seven features (excluding the FFDP) were then sorted according to their G-mean values. We first constructed an SVM model with all seven input features and then excluded the variable with the lowest G-mean value to construct a new SVM model and repeated this process until there were only two features remaining (in Table 5).

Table 5. Prediction results of the SVM models after removing the poorly performing variables one by one

Feature	Accuracy (0)	Accuracy (1)	G-mean
FB.TFD.FC.TFFP.FBP.TFDP.FCP	0.8305	0.8621	0.8462
TFD.FC.TFFP.FBP.TFDP.FCP	0.8339	0.8276	0.8307
TFD.FC.FBP.TFDP.FCP	0.8372	0.8276	0.8323
TFD.FC.TFDP.FCP	0.8205	0.8621	0.8410
TFD.TFDP.FCP	0.8104	0.8621	0.8359
TFD.TFDP	0.8473	0.8966	0.8716

Among the eight models in Fig. 8, the prediction model having the TFD and TFDP as input features has an accuracy of 87.16% in the G-mean comprehensive evaluation and is substantially more accurate than the other models. Its prediction accuracy is 84.73% for class 0 and 89.66% for class 1. In this project, the SVM model was used to predict thousands of data in less than 1 second.

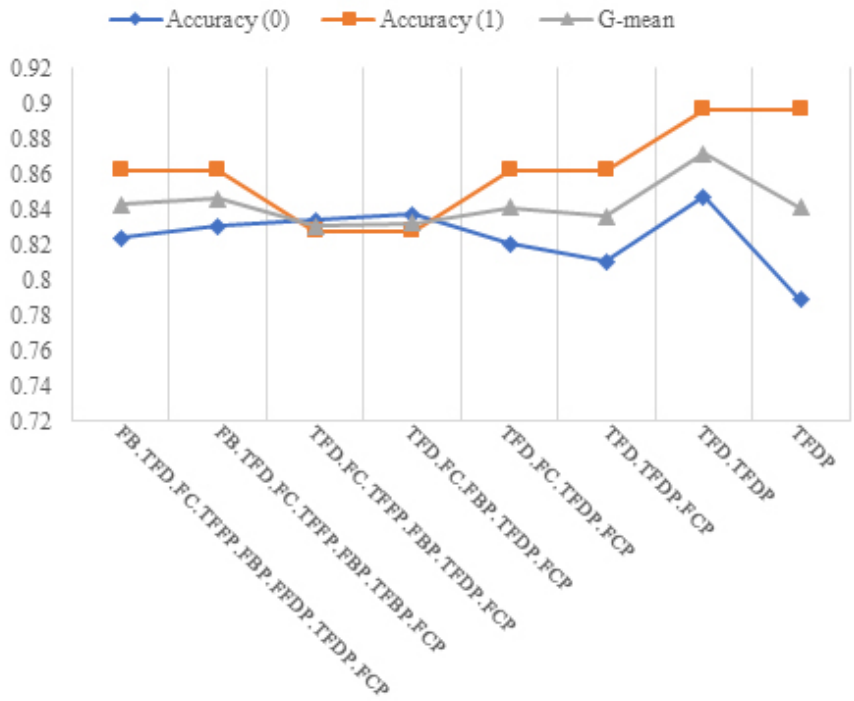

Figure 8. Accuracy of the prediction model when eliminating the worst performing features one by one

5. Discussion

Sections 3 and 4 confirmed the feasibility of using eye-movement signals instead of manual feedback. This contactless feedback method reduced the browsing time of users on an interface by approximately 16.45 s. Additionally, it improved the image retrieval efficiency by 33.76%. Among various eye-movement features, the TFD and TFDP are suitable for judging user retrieval requirements. We carried out an in-depth analysis of the results and found interesting phenomena.

5.1 Predictive Model with the TFD and TFDP as Input Features

Increasing the number of true positives reduces the number of images that meet the needs of designers but are not shown. This is a fundamental requirement of the model that must be guaranteed. Meanwhile, increasing the number of true negatives minimizes the scope of the provided gallery and improves the efficiency of the designers' image retrieval process, which is a high-level requirement of the model.

Evaluating and controlling the diversity of predictors and finding a balance between accuracy and diversity is not easy (Brown et al., 2005; Kuncheva, Whitaker, 2003). Among the SVM prediction models, the prediction model having the TFD and TFDP as input features performs best. This indicates that the TFD of the AOI is an important indicator for predicting a designer's image selection results (Bialkova, van Trijp, 2011; Lohse, 1997; Navalpakkam et al., 2012). A comparison of the model having the TFD as the input feature with the optimal model (see Fig. 8) reveals that the addition of the TFDP increases the number of true positives by 10% but reduces the number of true negatives by only 2%. This is consistent with our expectations of the prediction model; that is, true negatives can be sacrificed when ensuring that there are as many true positives as possible. In the model construction, the selection of the TFD and TFDP as input features does not mean that other features are meaningless. This selection can be explained in that the TFD and TFDP are the best combination of input features (Verikas et al., 2010). Although there are possibly other good features, in combination, they would reduce the prediction results of the model having TFD and TFDP features.

5.2 Analysis of Typical Features

The TFD includes the recognition and judgment times for the information in the image (Causse et al., 2019). The participants usually excluded some images after first browsing the entire interface, and then compared and excluded the remaining images one by one (Krajbich et al., 2010; Schotter et al., 2010; Shimojo et al., 2003). In this process, the TFD has the most direct and convincing connection with the image selections. This result is consistent with the results of statistical analysis and SVM models (see Table 6).

We cannot, however, predict a user's image selection results directly from the TFD and TFDP. We chose to use a kernel function in the SVM, and we thus solved a linearly indivisible problem in the original space by mapping the data to a high-dimensional space.

Table 6. Prediction results of the SVM model with the TFD and TFDP as input features

		Predicted results	
		0	1
Actual results	0	505	91
	1	3	26

5.3 Discussion on the Connotation of Features

In goal-driven visual search behavior, the FFD for the stimulus tends to decrease as the participant becomes more experienced at the task (Jovancevic-Misic, Hayhoe, 2009). In this study, when generating the first fixation point, participants did not make a clearly conscious choice (Orquin, Mueller Loose, 2013; van der Laan et al., 2015). Therefore, using the FFD and FFDP to predict choices is meaningless. In addition, two different browsing styles, top-down and bottom-up during visual search, can affect the results (Orquin, Mueller Loose, 2013), for example, the positions of the images and participants' browsing habits result in the participants browsing the images in different orders. It is therefore difficult to arrive at the selection result directly from the FFD, TFF, and FB. This explains why the FFD and TFF, which were eliminated first, were not significantly different in the statistical analysis. For the same reason, the FFDP, FB, TFFP, and FBP, which were eliminated in turn, had poor performance in the construction of the SVM model.

Depending on the statistical analysis and SVM model, we found that the TFD, TFDP, FC, and FCP had better performance in prediction. These features represent the ability to attract a participant's attention continuously and are thus related to the selection result. In addition, the TFD and TFDP performed better than the FC and FCP in prediction. In terms of units, the features can be divided into duration-type indicators (TFD and TFDP) and frequency-type indicators (FC and FCP). We therefore conclude that the duration indicators are better than the frequency indicators. The answers given in the interviews indicate the participant's long-term gaze at an AOI is a process of further recognizing and generating a judgment. Attracting attention is a prerequisite for cognitive processing, but a quick attraction does not directly lead to selection. An FC that is too high may, in some cases, mean that the participant does not have a good cognition of the AOI (Djamasbi, 2014), remains suspicious of his or her judgment, and needs to confirm the judgment repeatedly.

5.4 Limitations

This study had several limitations. First, the study only evaluated the relationship between the eye-movement behavior and subjective selection, and other physiological signals, such as electroencephalography and emotion signals, should also be considered. Second, the study was conducted only for designers. We need to consider the problems encountered in promoting this new method to other users in the future and determine the particularities of other user groups, the commonality among different user groups, and the differences within a user group. Third, the study did not investigate the effect of the environment on the visual search in real image retrieval tasks.

6. Conclusion

To optimize the performance of the image retrieval system, this study used an SVM model based on eye-movement features to give feedback on the image retrieval results so that the retrieved images better satisfy the user's underlying needs. Using the features of an eye tracker's output, we proposed another set of features that were shown to be effective and meaningful through statistical analysis and the SVM model performance. Moreover, the study verified that eye-movement data have practical importance in the image selection process. This has serious implications for the generalization of results obtained from the design field to other areas. As a type of contactless interaction, the use of eye-movement signals for feedback makes the experience of human–computer interaction more direct, hygienic, and efficient. We will continue to optimize the prediction model in combination with an eye-movement heat map in future work.

Acknowledgements

The authors thank the designers and participants for their contributions to the study . This work was supported by the National Natural Science Foundation of China (5170050747).

References

Ai, L. F., Yu, J. Q., He, Y. F., & Guan, T. (2013). High-dimensional indexing technologies for large scale content-based image retrieval: a review. *Journal of Zhejiang University-Science C-Computers & Electronics, 14*(7), 505-520. https://doi.org/10.1631/jzus.CIDE1304

Anagnostopoulos, V., Havlena, M., Kiefer, P., Giannopoulos, I., Schindler, K., & Raubal, M. (2017). Gaze-Informed location-based services. *International Journal of Geographical Information Science, 31*(9), 1770-1797. https://doi.org/10.1080/13658816.2017.1334896

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2018). NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1437-1451. https://doi.org/10.1109/TPAMI.2017.2711011

Bialkova, S., & van Trijp, H. C. M. (2011). An efficient methodology for assessing attention to and effect of nutrition information displayed front-of-pack. *Food Quality and Preference, 22*(6), 592-601. https://doi.org/10.1016/j.foodqual.2011.03.010

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion, 6*(1), 5-20. https://doi.org/10.1016/j.inffus.2004.04.004

Causse, M., Lancelot, F., Maillant, J., Behren, J., Cousy, M., & Schneider, N. (2019). Encoding decisions and expertise in the operator's eyes: Using eye-tracking as input for system adaptation. *International Journal of Human-Computer Studies, 125*, 55-65. https://doi.org/10.1016/j.ijhcs.2018.12.010

Chai, J. Y., Zhang, C., & Jin, R. (2007). An empirical investigation of user term feedback in text-based targeted image search. *Acm Transactions on Information Systems, 25*(1), 25. https://doi.org/10.1145/1198296.1198299

Chandon, P., Hutchinson, J. W., Bradlow, E. T., & Young, S. H. (2009). Does In-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase. *Journal of Marketing, 73*(6), 1-17. https://doi.org/10.1509/jmkg.73.6.1

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*(3), 1-27. https://doi.org/10.1145/1961189.1961199

Chapman, P. R., & Underwood, G. (1998). Visual search of driving situations: Danger and experience. *Perception, 27*(8), 951-964. https://doi.org/10.1068/p270951

Chen, L., Wei, H., & Ferryman, J. (2013). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters, 34*(15), 1995-2006. https://doi.org/10.1016/j.patrec.2013.02.006

Chen, Y. X., Wang, J. Z., Krovetz, R., Ieee, & Ieee. (2003). *An unsupervised learning approach to content-based image retrieval*. New York: Ieee. https://doi.org/10.1109/ISSPA.2003.1224674

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *Acm Computing Surveys, 40*(2), 60. https://doi.org/10.1145/1348246.1348248

Djamasbi, S. (2014). Eye Tracking and Web Experience. *AIS Transactions on Human-Computer Interaction, 6*(2), 37-54. https://doi.org/10.17705/1thci.00060

Gegenfurtner, A., Lehtinen, E., & Saljo, R. (2011). Expertise Differences in the Comprehension of Visualizations: a Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review, 23*(4), 523-552. https://doi.org/10.1007/s10648-011-9174-7

Han, T. A., & Pereira, L. M. (2013). State-of-the-art of intention recognition and its use in decision making. *Ai Communications, 26*(2), 237-246. https://doi.org/10.3233/AIC-130559

HaoshengHuang, & GeorgGartner. (2012). Collective intelligence-based route recommendation for assisting pedestrian wayfinding in the era of Web 2.0. *Journal of Location Based Services, 6*(1), 1-21. https://doi.org/10.1080/17489725.2011.625302

Hariharan, B., Padmini, S., Gopalakrishnan, U., & Ieee. (2014). Gesture Recognition Using Kinect in a Virtual Classroom Environment. In *2014 Fourth International Conference on Digital Information and Communication Technology and It's Applications* (pp. 118-124). https://doi.org/10.1109/DICTAP.2014.6821668

Jovancevic-Misic, J., & Hayhoe, M. (2009). Adaptive Gaze Control in Natural Environments. *Journal of Neuroscience, 29*(19), 6234-6238. https://doi.org/10.1523/JNEUROSCI.5570-08.2009

Kim, J. H., Zhao, X., & Du, W. (2018). Assessing the performance of visual identification tasks using time window-based eye inter-fixation duration. *International Journal of Industrial Ergonomics, 64*, 15-22. https://doi.org/10.1016/j.ergon.2017.09.002

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292-1298. https://doi.org/10.1038/nn.2635

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51*(2), 181-207. https://doi.org/10.1023/A:1022859003006

Lang, S., Dickinson, J., & Buchal, R. (2001, 14-14 July 2001). *An overview of cognitive factors in distributed design.* Paper presented at the Proceedings of the Sixth International Conference on Computer Supported Cooperative Work

in Design (IEEE Cat. No.01EX472).

Li, Q. Y., Tian, M., Liu, J., & Sun, J. R. (2016). An implicit relevance feedback method for CBIR with real-time eye tracking. *Multimedia Tools and Applications, 75*(5), 2595-2611. https://doi.org/10.1007/s11042-015-2873-1

Liu, N., & Yu, R. F. (2017). Influence of social presence on eye movements in visual search tasks. *Ergonomics, 60*(12), 1667-1681. https://doi.org/10.1080/00140139.2017.1342870

Liu, Y., Zhang, D. S., Lu, G. J., & Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40*(1), 262-282. https://doi.org/10.1016/j.patcog.2006.04.045

Lohse, G. L. (1997). Consumer eye movement patterns on yellow pages advertising. *Journal of Advertising, 26*(1), 61-73. https://doi.org/10.1080/00913367.1997.10673518

Malaga, R. A. (2000). The effect of stimulus modes and associative distance in individual creativity support systems. *Decision Support Systems, 29*(2), 125-141. https://doi.org/10.1016/S0167-9236(00)00067-1

Motta, T., Nedel, L., & Ieee. (2012). *Gestural Interaction for Manipulating Graphs in a Large Screen Using the Kinect Integrated to the Browser*. https://doi.org/10.1109/CLEI.2012.6427128

Navalpakkam, V., Kumar, R., Li, L., & Sivakumar, D. (2012) Attention and selection in online choice tasks. In*: Vol. 7379 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 200-211). https://doi.org/10.1007/978-3-642-31454-4_17

Orquin, J. L., & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica, 144*(1), 190-206. https://doi.org/10.1016/j.actpsy.2013.06.003

Oyckoya, O., & Stentiford, F. W. M. (2005, 30 Nov.-1 Dec. 2005). *A performance comparison of eye tracking and mouse interfaces in a target image identification task.* Paper presented at the The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005. (Ref. No. 2005/11099). https://doi.org/10.1049/ic.2005.0723

Oyekoya, O. W., & Stentiford, F. (2007). Perceptual image retrieval using eye movements. *International Journal of Computer Mathematics, 84*(9), 1379-1391. https://doi.org/10.1080/00207160701242268

Papadopoulos, G. T., Apostolakis, K. C., & Daras, P. (2014). Gaze-Based Relevance Feedback for Realizing Region-Based Image Retrieval. *Ieee Transactions on Multimedia, 16*(2), 440-454. https://doi.org/10.1109/TMM.2013.2291535

Parush, A., Shwarts, Y., Shtub, A., & Chandra, M. J. (2005). The impact of visual layout factors on performance in web pages: A cross-language study. *HUMAN FACTORS, 47*(1), 141-157. https://doi.org/10.1518/0018720053653785

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*. https://doi.org/10.1109/TPAMI.2016.2577031

Schotter, E. R., Berry, R. W., McKenzie, C. R. M., & Rayner, K. (2010). Gaze bias: Selective encoding and liking effects. *Visual Cognition, 18*(8), 1113-1132. https://doi.org/10.1080/13506281003668900

Shieh, M. D., & Yang, C. C. (2008). Classification model for product form design using fuzzy support vector machines. *Computers & Industrial Engineering, 55*(1), 150-164. https://doi.org/10.1016/j.cie.2007.12.007

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience, 6*(12), 1317-1322. https://doi.org/10.1038/nn1150

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349-1380. https://doi.org/10.1109/34.895972

Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive Psychology: Mind and Brain*: Pearson Schweiz Ag.

Su, C., Xia, X., Li, H., Liu, X., Kuang, C., Xia, J., & Wang, B. (2014). A penetrable interactive 3D display based on motion recognition. *Chinese Optics Letters, 12*(6). https://doi.org/10.3788/COL201412.060007

Sugano, Y., Ozaki, Y., Kasai, H., Ogaki, K., & Sato, Y. (2014). Image preference estimation with a data-driven approach: A comparative study between gaze and image features. *Journal of Eye Movement Research, 7*(3), 9.

Tseng, Y. C., & Howes, A. (2015). The adaptation of visual search to utility, ecology and design. *International Journal of Human-Computer Studies, 80*, 45-55. https://doi.org/10.1016/j.ijhcs.2015.03.005

van der Laan, L. N., Hooge, I. T. C., de Ridder, D. T. D., Viergever, M. A., & Smeets, P. A. M. (2015). Do you like what

you see? The role of first fixation and total fixation duration in consumer choice. *Food Quality and Preference, 39*, 46-55. https://doi.org/10.1016/j.foodqual.2014.06.015

Vapnik, V. N. (1995). *The nature of statistical learning theory*: Springer, New York, NY. https://doi.org/10.1007/978-1-4757-2440-0

Verikas, A., Gelzinis, A., Kovalenko, M., & Bacauskiene, M. (2010). Selecting features from multiple feature sets for SVM committee-based screening of human larynx. *Expert Systems with Applications, 37*(10), 6957-6962. https://doi.org/10.1016/j.eswa.2010.03.025

Vrochidis, S., Patras, I., & Kompatsiaris, I. (2011). *An eye-tracking-based approach to facilitate interactive video search*. Paper presented at the Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Trento, Italy. https://doi.org/10.1145/1991996.1992039

Zangemeister, W. H., Sherman, K., & Stark, L. (1995). Evidence for a global scanpath strategy in viewing abstract compared with realistic images. *Neuropsychologia, 33*(8), 1009-1025. https://doi.org/10.1016/0028-3932(95)00014-T

Zhang, D. S., Islam, M. M., & Lu, G. J. (2012). A review on automatic image annotation techniques. *Pattern Recognition, 45*(1), 346-362. https://doi.org/10.1016/j.patcog.2011.05.013

Zhang, Y., Fu, H., Liang, Z., Chi, Z., & Feng, D. D. (2010). *Eye movement as an interaction mechanism for relevance feedback in a content-based image retrieval system.* Paper presented at the Symposium on Eye-tracking Research & Applications. https://doi.org/10.1145/1743666.1743674