A Novel Image Retrieval System with Real-Time Eye Tracking

Qingyong Li, Mei Tian, Jun Liu, Jinrui Sun
Beijing Key Lab of Transportation Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
{liqy, mtian, 12125123, 12120448}@bjtu.edu.cn

ABSTRACT

Relevance feedback is one of approach to improve the performance of content-based image retrieval system, and implicit feedback approaches, which gather users' feedback by biometric devices (e.g. eye tracker), are extensively investigated in recent years. This paper proposes a novel image retrieval system with eye tracking (IRSET). IRSET is composed of three modules: image retrieval module based on standard bag-of-words, eye tracking module to obtain a user's fixation data and to infer feedback information, and query expansion module that fuses the user's feedback and the input query to form a richer latent query. The implicit feedback of IRSET is implemented online and real-time, which makes IRSET remarkably distinguish from other systems with implicit feedback. We conduct experiments on the dataset of Oxford building for ten participants. The experimental results demonstrate that IRSET is an attractive interface to image retrieval and improves the retrieval performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback*

General Terms

Algorithms, Performance

Keywords

Content based image retrieval, Visual attention, Bag of words, Relevance feedback

1. INTRODUCTION

Numerous digital images are being produced everyday and everywhere with the rapid development of digital imaging technologies and the growth of communication networks. It has emerged as a big challenge to retrieve a desired picture even from a photo album on a personal computer because of the exponential increase in the number of images. Depending on the query modes, image retrieval can be roughly divided into two categories: text-based approaches and content-based methods. The text-based approaches associate keywords with each image in advance and users can search images with keywords. Researchers, however, find that it is laborious and subjective to describe image content with keywords.

On the contrary, content-based image retrieval (CBIR) methods search images based on visual features, such as color, texture, shape and spatial features, which are automatically extracted from images themselves[1]. Global features describe a picture in a holistic way (for example color histogram), while local descriptors like scale invariant feature transform (SIFT) [2] provide a way to describe several salient patches around key points within an image. Local descriptors demonstrate great discriminative power and they are extensively applied in CBIR. The most popular approach today relies on bag-of-words (BoW) or bag-of-features model[3]. BoW method first quantizes local descriptors into "visual words", and then represents each image as a vector of words like a text document. Lastly, it applies scalable indexing to search images.

CBIR also faces many challenges, such as semantic gap that means the big difference between low-level visual features and high-level users' concepts. Relevance feedback is one of the main methodologies adopted to overcome this problem[4]. The main idea of relevance feedback is to take into account users' interaction to establish the association between low-level features and semantics of images, Many research demonstrated that relevance feedback can help to improve retrieval performance[5].

In literature two types of feedback are defined: explicit feedback and implicit feedback. Explicit methods[5] require users to provide explicit statements regarding the relevance of returned images, such as submitting positive and negative samples of the retrieved results, or selecting areas-of-interest (AOIs) from the query image or result images. On the other hand, implicit approaches utilize information that is obtained from users in an unobscured and non-invasive way (e.g. gaze-tracking and heart beat rate), and then predict the relevance of the returned images[6, 7].

This paper provides an image retrieval system with eye tracking (IRSET). Compared with the related work, IRSET has three notable contributions as follows:

1. IRSET integrates state-of-the-art BoW architecture and eye tracking technology to form a novel image retrieval framework.

ICIMCS'14, July 10–12, 2014, Xiamen, Fujian, China.
Copyright 2014 ACM 978-1-4503-2810-4/14/07 ...$15.00.

2. IRSET implements the real-time implicit feedback mechanism, though most systems with implicit feedback is offline[8, 9].

3. IRSET brings forward a new query expansion method that applies the weighting strategy based on fixation information.

2. RELATED WORK

This section briefly introduces implicit relevance feedback that is gathered with biometric devices such as electroencephalography, electrocardiography sensors and eye trackers. This work will focus on exploiting the implicit user feedback based on eye movements.

Eye-tracking has been used extensively in the psychology literature, and recently applied in information retrieval tasks as well [10, 11, 6]. Studies, which utilize eye movements to investigate cognitive processes, started to appear three decades ago. Subsequently, eye movement data have proven to be very valuable in studying information processing tasks[12]. More specifically, eye tracking methods were mostly used for identifying items of interest or understanding the users' behaviour in information retrieval tasks.

For textual document retrieval community, eye tracking methodologies are successfully applied in diverse way. Granka *et al.* [13] investigate how users interact with the results of a web search engine by employing eye-tracking techniques. Puolamaki *et al.* [14] propose the proactive information retrieval that combines implicit relevance feedback and collaborative filtering. In this method, implicit feedback is inferred from eye movements data with discriminative hidden Markov models. After that, Hardoon *et al.* [15] introduce a search strategy, in which a query is inferred from information extracted either from eye movements data or from a combination of eye movements and explicit relevance feedback. Buscher *et al.* [16] provide a different application of gaze analysis. That is, they model users' reading behaviour from eye movement data, and regard the inferred models as implicit feedback for query expansion and reranking. Furthermore, Cole *et al.* [17] recently report that users' domain knowledge can be inferred from their interactive search behaviours without considering the content of queries or documents, but using only measurements of eye movement patterns.

For multimedia retrieval domain, eye tracking techniques are used for studying user behaviour and identifying visual AOIs. An eye-tracking study is first conducted to investigate whether it is textual or visual representation of video [18]. Oyekoya and Stentiford [19] propose an interactive interface for image retrieval, in which input query is given by users' eye movements captured by an eye-tracker. Moreover, they conduct experiments to explore the relationship between gaze behaviour and a visual attention model [20]. This research reports that gaze behaviours could change with the content of images. Kozma *et al.* [21] bring forward the real time interface named GaZIR for browsing and searching images. GaZIR predicts the relevance of viewed images with classical logic regression based on fixation and saccade features. Similarly, several features based on eye movement are proposed and a decision tree is trained to classify positive and negative images in [8]. In another work [9], pupil responses play as a complementary modality with EEG signals, and the fused information is applied to learn a two-level linear classifier for visual detection. More recent research, which combine image features with eye movements for reranking or extracting local visual features, can be found in [22, 23, 24, 11, 6].

3. THE IMAGE RETRIEVAL SYSTEM WITH EYE TRACKING

The proposed image retrieval system adopts the bag-of-visual-words architecture which has been proven successful in achieving high performance; and integrates an unobtrusive eye tracker that captures users' eye gaze data to infer their interested images; finally, the initial query is expanded to refine the retrieval result.

3.1 Overview of the System

The outline of IRSET is illustrated as Fig.1, and includes five steps as follows:

1. Input a query image, and retrieve a set of images based on the BoW architecture.

2. Observe the retrieved images, and an eye tracker implicitly captures users' focused images.

3. Combine the user's focused images and the original query to form a richer latent query.

4. Search the image collection using the expanded query to refine retrieval results.

5. Repeat the process as necessary, alternating between retrieval refinement and querying expansion.

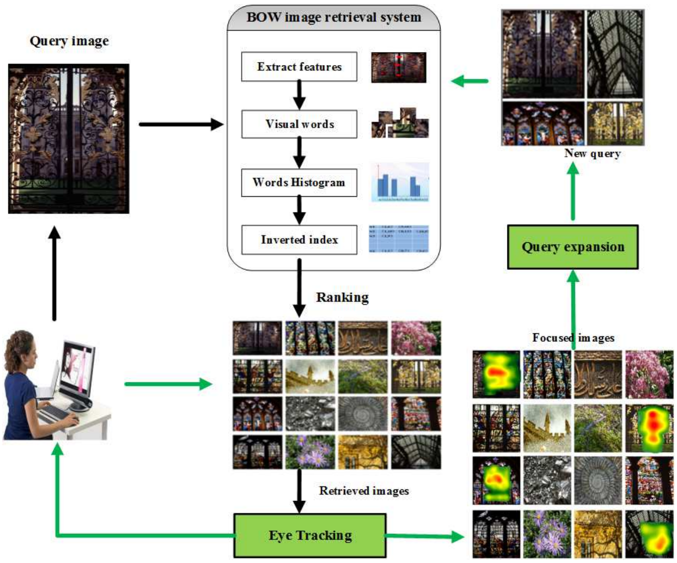

Figure 1: The image retrieval framework based on eye tracking technology

The proposed IRSET is composed of three main modules: image retrieval based on BoW, eye tracking and query expansion.

3.2 Image retrieval based on BoW

This module is constructed based on standard BoW architecture. Its key aspects are described as follows, and further details can be found in [25].

Image description. For each image in the dataset, we find multi-scale interest points and computer their 128-dimensional SIFT descriptors [2].

Visual words. A visual vocabulary is generated using an approximate K-means clustering method based on randomized trees. Each visual descriptor is assigned to a single cluster via approximate nearest neighbour search. These quantized visual features are regarded as words in a textual document, and then used to index images for the search engine.

Words histogram. The query and each image in the collection are represented as a visual word histogram or word occurrences. Furthermore, the standard tf-idf weighting scheme is applied to reduce the contribution of commonly occurring visual words in the corpus. The similarity between the query histogram and that of each image in the collection is calculated, and the most similar images are retrieved.

Inverted index. Inverted index is applied to improve the computational speed of similarity calculation. The engine stores word occurrences in an index, which maps individual words to the images in which they occur. The scores for each image are accumulated rather than explicitly computed through histograms.

3.3 Eye tracking

Eye tracking is the process of measuring the point of users' gaze, and the eye tracking module captures users' real-time gaze information after retrieved images are displayed, so it can be inferred that which images are positive and which images are negative.

Raw gaze data are obtained by finding the fixations with the built-in filter provided by SMI eye tracker. This filter maps a series of raw coordinates to a single fixation if the coordinates stay sufficiently long within a sphere of a given radius. Furthermore, SMI provides some APIs (Application Programming Interface) that retrieve user's fixation data, which include position and duration.

The design and examples of the IRSET interface is shown in Fig.2. The top left presents the query image, and the right is sixteen candidate retrieved images that are arranged as a 4x4 grid display. Each grid is defined as an AOI, and the fixation belonging to the AOI would be accumulated. Users' fixation during 20s-30s is captured after retrieved images by BoW are displayed.

The feature used in our eye tracking module is fixation duration, which counts the total fixation time that a user focuses on a retrieved image. Fixation duration ($FD(i)$) is defined as:

$$\text{FD(i)} = \sum_{e\in AOI(i)} T(e) \tag{1}$$

where i refers to the index of AOIs or retrieved images, e is a fixation event and $T(e)$ denotes the fixation time of e.

Positive images can be inferred based on FD of each AOI and a thresholding strategy. Given a threshold t, an image is regarded as positive if its fixation duration is greater than t, otherwise it is negative or neutral. So a user's implicit relevance feedback is deduced as follows:

$$P(i) = \begin{cases} 1, & if\ FD(i) \geq t \\ 0, & others \end{cases} \tag{2}$$

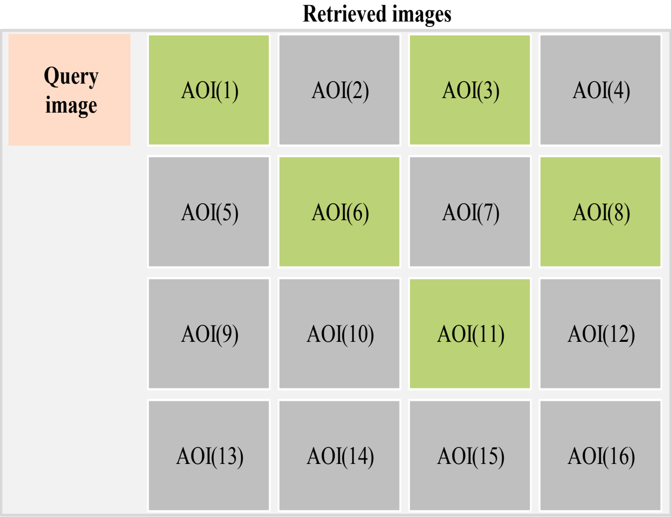

Figure 2: The user interface of IRSET. The grids filled with green color represent positive images that is decided based on users' fixation.

3.4 Query expansion

Query expansion module integrates positive images that are obtained by eye tracking, and forms an expanded latent query representation. The diagram of our query expansion is demonstrated in Fig. 3

Supposed that N images are positive according to Equation 2 for the procedure of eye tracking, and the word histogram of image i is represented as $H(i)$:

$$H(i) = (w_1, w_2, \cdots, w_C)$$

where C denotes the number of visual word in the vocabulary, and w_i ,$i \in [1, C]$, is the weight defined by tf-idf rule. These positive images are merged together and formed an expanded query by a weighting strategy. The expanded histogram H^e, $H^e = (w_1^e, w_1^e, \cdots, w_C^e)$, is defined as follows:

$$w_i^e = \sum\nolimits_{n=1}^{N} \text{NFD}(n) \cdot w_i^n \tag{3}$$

where w_i^n, $1 \leq n \leq N$, refers to the weight of word i for the image n , and NFD(n) defined as the normalized fixation duration:

$$\text{NFD}\,(n) = FD(n) \Big/ \sum\nolimits_{j=1}^{N} FD(j).$$

The expanded histogram is regarded as a representation of the expanded query, and then it is used to further search relevant images from the corpus.

4. USER STUDY AND EXPERIMENTAL RESULTS

The system has been implemented based on the SDK (Software Development Kit) of SMI RED250 with Visual Studio C++ 2010. The SMI RED250 is fully remote, fiducial-free and contact-free. It equipped with a 22" monitor whose

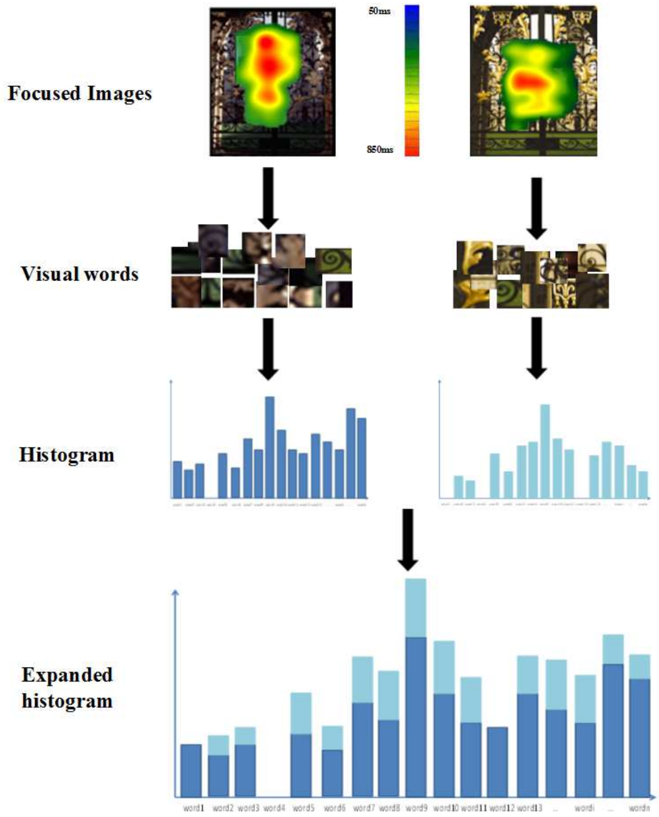

Figure 3: The query expansion method based on user's fixation duration

resolution is 1920 × 1280 pixels. Retrieved images are displayed on the screen with viewing distance of 600mm-800mm. RED250 has a sampling rate of 250Hz and high accuracy of 0.4°, which ensure a low error of gaze data.

Fig. 4 illustrates an example of the eye tracking information captured by IRSET. It can be figure out that the user pays more attention to the relevant images that are indicated with green rectangle, and sweeps the other images though some images (for example those in AOI 9 and 13) also get a little fixation. Note that there are also some noised fixation that locate in left-bottom area. Fig.5 shows the details of fixation data. We can find that positive images can attract more revisit or re-fixation. Much human computer interaction and usability research show that revisit on a target may be an indication of special interest on the target.

Ten participants took part in the study, six females and four males in an age range from 20 to 26. Their visions are either normal or correct-to-normal. More specifically, The participants can be divided into three categories:

1. postgraduates who study computer vision and are familiar with eye-tracker.

2. postgraduates who study computer vision and are not familiar with eye-tracker.

3. postgraduates who do not have any background about computer vision and eye-tracker.

The dataset used in this study is Oxford Building [25].

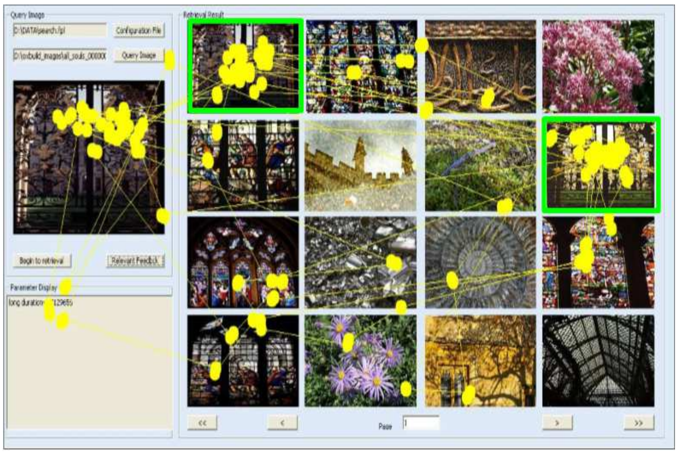

Figure 4: An example of eye tracking information in a search procedure of IRSET. Each yellow circle denotes a fixation, a line between circles refers to a tracking trace, and the green rectangle indicates positive images.

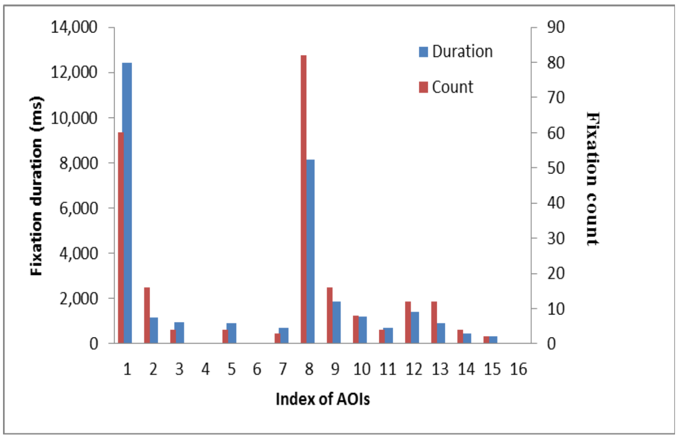

Figure 5: The fixation data captured by RED250 for the example shown in Fig.4.

Precision and *recall* of the top 16 returned images are used to evaluate the performance of IRSET. Precision and recall are defined as:

$$P = N_P/16$$

$$R = N_P/N_A$$

where N_P denotes the number of relevant images of the top 16 returned, and N_A is the total number of relevant images. Furthermore, a subjective test is carried out and results are grouped into three types according to the retrieved result after feedback:

1. *Very good*: Top N all are relevant images. N equals 8 in this experiment.

2. *Good*: Top one is relevant, and there are other relevant images in top 16.

3. *Bad*: Top one is not relevant.

Table 1: The comparison of retrieval performance. P_o and R_o denote the precision and recall without feedback; P_f and R_f denote the precision and recall with feedback. PI refers to participant index, and PT is participant type.

PI	PT	P_o	R_o	P_f	R_f	Subjective result
1	1	1.0000	0.2051	1.0000	0.2051	very good
2	1	0.1250	0.0800	0.0625	0.0400	bad
3	2	0.9375	0.2778	0.9375	0.2778	good
4	2	0.8125	0.2407	0.8750	0.2593	very good
5	2	0.1250	0.1667	0.1875	0.2500	very good
6	3	0.8125	0.1667	1.0000	0.2051	very good
7	3	0.9375	0.2778	0.9375	0.2778	very good
8	3	0.3125	0.0641	0.8125	0.1667	good
9	3	0.8750	0.1795	0.8750	0.1795	good
10	3	0.3125	0.2000	0.3125	0.2000	good

Table 1 demonstrates the retrieval performance of IRSET. It can be figured out that IRSET achieves an improvement on precision and recall for most cases. More specifically, nine participants are satisfied with IRSET except the second participant. More generally, participants without prior knowledge of eye tracker tend to get better result. We think that such participants are interested in eye tracker and pay more attention to the experiment, so the feedback information is more conformed to the participant's intention, and a better retrieval performance can be obtained.

5. CONCLUSIONS

This paper presents a novel image retrieval framework that integrates BoW architecture and eye tracking technology, and implements the image retrieval system with eye tracking. IRSET searches images with two stages: firstly, it retrieves images with standard BoW according to an input query. Secondly, IRSET captures the user's fixation data by SMI RED250, infers the user's feedback, and expands the input query to refine search results. The experiments, conducted on dataset of Oxford building with ten participants, demonstrate that IRSET achieves substantial improvement and eye tracking is highly beneficial for relevance feedback of CBIR.

6. ACKNOWLEDGMENTS

This work is supported by National Nature Science Foundation of China (61105119), Fundamental Research Funds for the Central Universities (2014JBZ003), Beijing Higher Education Young Elite Teacher Project.

7. REFERENCES

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[2] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[3] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477. IEEE, 2003.

[4] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655, 1998.

[5] Xiang Sean Zhou and Thomas S Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6):536–544, 2003.

[6] G Papadopoulos, K Apostolakis, and Petros Daras. Gaze-based relevance feedback for realizing region-based image retrieval. *IEEE Transactions on Multimedia*, 16(2):440–454, 2014.

[7] S Navid Hajimirza, Michael J Proulx, and Ebroul Izquierdo. Reading users' minds from their eyes: A method for implicit image annotation. *IEEE Transactions on Multimedia*, 14(3):805–815, 2012.

[8] Yun Zhang, Hong Fu, Zhen Liang, Zheru Chi, and Dagan Feng. Eye movement as an interaction mechanism for relevance feedback in a content-based image retrieval system. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 37–40. ACM, 2010.

[9] Ming Qian, Mario Aguilar, Karen N Zachery, Claudio Privitera, Stanley Klein, Thom Carney, and Loren W Nolte. Decision-level fusion of eeg and pupil features for single-trial visual detection analysis. *IEEE Transactions on Biomedical Engineering*, 56(7):1929–1937, 2009.

[10] Arto Klami, Craig Saunders, Teófilo E de Campos, and Samuel Kaski. Can relevance of images be inferred from eye movements? In *ACM international conference on Multimedia information retrieval*, pages 134–140. ACM, 2008.

[11] Stefanos Vrochidis, Ioannis Patras, and Ioannis Kompatsiaris. An eye-tracking-based approach to facilitate interactive video search. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 43. ACM, 2011.

[12] Keith Rayner. Eye movements in reading and information processing. *Psychological bulletin*, 85(3):618, 1978.

[13] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *27th annual international ACM SIGIR*, pages 478–479. ACM, 2004.

[14] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski. Combining eye movements

and collaborative filtering for proactive information retrieval. In *The 28th annual international ACM SIGIR*, pages 146–153. ACM, 2005.

[15] David R Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolamäki, and Samuel Kaski. Information retrieval by inferring implicit queries from eye movements. In *International Conference on Artificial Intelligence and Statistics*, pages 179–186, 2007.

[16] Georg Buscher, Andreas Dengel, and Ludger van Elst. Query expansion using gaze-based feedback on the subdocument level. In *The 31st annual international ACM SIGIR*, pages 387–394. ACM, 2008.

[17] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075–1091, 2013.

[18] Anthony Hughes, Todd Wilkens, Barbara M Wildemuth, and Gary Marchionini. Text or pictures? an eyetracking study of how people view digital video surrogates. In *International Conference on Image and Video Retrieval*, pages 271–280. Springer, 2003.

[19] OK Oyekoya and Fred Stentiford. Eye tracking as a new interface for image retrieval. *BT Technology Journal*, 22(3):161–169, 2004.

[20] Oyewole Oyekoya and Fred Stentiford. Exploring human eye behaviour using a model of visual attention. In *17th International Conference on Pattern Recognition*, volume 4, pages 945–948. IEEE, 2004.

[21] László Kozma, Arto Klami, and Samuel Kaski. Gazir: gaze-based zooming interface for image retrieval. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 305–312. ACM, 2009.

[22] David R Hardoon and Kitsuchart Pasupa. Image ranking with implicit feedback from eye movements. In *2010 Symposium on Eye-Tracking Research & Applications*, pages 291–298. ACM, 2010.

[23] Zhen Liang, Hong Fu, Yun Zhang, Zheru Chi, and Dagan Feng. Content-based image retrieval using a combination of visual features and eye tracking data. In *Symposium on Eye-Tracking Research & Applications*, pages 41–44. ACM, 2010.

[24] Alberto Faro, Daniela Giordano, Carmelo Pino, and Concetto Spampinato. Visual attention for implicit relevance feedback in a content based image retrieval. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 73–76. ACM, 2010.

[25] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 1–8. IEEE, 2007.