Detecting Regions of Interest using Eye Tracking for CBIR

Qingtao Ren[1], Yongqiang Chen[1], Peng Zhang[1], Dawei Song[1,2], Yuexian Hou[1],
[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]Department of Computing and Communications, The Open University, United Kingdom
{qtren, cyq, pzhang}@tju.edu.cn, dawei.song2010@gmail.com, yxhou@tju.edu.cn

ABSTRACT

Identifying Regions of Interest (ROIs) in images has been shown an effective way to enhance the performance of Content Based Image Retrieval (CBIR). Most existing ROI identification methods are based on salience detection, and the identified ROIs may not be the regions that users are really interested in. While manual selection of ROIs can directly reflect users' interests, it puts extra cognitive overhead to users. To alleviate these limitations, in this paper, we propose a novel eye-tracking based method to detect ROIs for CBIR, in an unobtrusive way. Experimental results have demonstrated that our model performed effectively compared with various state of the art methods.

Keywords: CBIR, Eye Tracking, ROI

1. INTRODUCTION

Identifying Regions of Interest (ROIs) to improve Content-Based Image Retrieval (CBIR) has recently drawn increasing attention [5]. Salience detection is a process of simulating human's detection of fixation points that a user would focus on at the first glance of an image by using various algorithms. Most existing ROI identification methods focus on salience detection and simply assume the detected ROIs are truly of users' interest. Such approach, however, neglects the fact that the salient regions of an image may not necessarily be what users are really interested in. Allowing users to manually select ROIs could solve this problems to some extent, but it brings a substantial cognitive burden to users.

Recently, advanced eye tracking technologies have been applied in various fields such as human computer interaction, for its convenience of capturing, in an unobtrusive and natural way, users' eye movements that largely reflect the users' interests. Eye tracking in IR has been studied to model user's implicit relevance feedback and to improve search experience. It has been shown by Rayner *et al.* [3] that different eye movement behaviors reveal different levels of user engagement during document reading. Cole *et al.* [2] investigated the relationships between eye movement patterns and different tasks, and utilized eye tracking data to examine users' information acquisition strategies. To our best knowledge, most existing studies focus on text retrieval rather than CBIR. There is also a lack of study on integeration of eye-tracking data into realtime retrieval models.

In this paper, we propose to combine image segmentation and eye tracking to automatically detect ROIs, which are then used to improve CBIR. Our hypothesis is that users tend to spend more time fixating on particular parts of the images they are interested in. This is consistent with various widely used assumptions in implicit relevance feedback, e.g., the SAT criteria (i.e., the documents clicked and viewed for above a certain amount of time are assumed relevant).

NeuroIR'15, August 13, 2015, Santiago, Chile.
.

2. METHODOLOGY

Formally, an image consists of N segmented regions:

$$\mathcal{R} = \{r_1, r_2, \ldots, r_N\} \tag{1}$$

where r_i is the i^{th} segmented region. Eye tracking data collected when users interact with the images, particularly the fixations on segmented regions, are used to identify ROIs. A fixation refers to a time of more than 250ms duration when the user is gazing at one point [1]. We select fixation duration as the main feature, by assuming a longer fixation can better reflect a user's interest. For each region r_i, an importance value C_i is defined based on the relative gaze duration d_i on r_i. C_i reflects the degree of the user's interest on the region r_i, and can be calculated as $C_i = d_i / \sum_{i=1}^{N} d_i$ where $\sum_{i=1}^{N} C_i = 1$. Its value will be 0 if there is no fixation on the region.

Consequently, an image is represented by a set of regions, some of which may be identified as ROIs. The color and texture features are extracted to represent regions. Specifically, a feature set including an 11x11x11 HSV color histogram and a 1x41 texture histogram of the edge map, is used to characterize a region. For similarity measurement between two images, the Euclidean distance between two feature vectors, denoted as $dis(\cdot,\cdot)$, is measured, and then the similarity between two images is defined as the minimum distance of regions. This allows retrieval of similar images at a finer semantic level according to the weights of different ROIs. Given a query image q, the ranking score $Score(q,d)$ for an image d can be computed as follows:

$$Score(q,d) = \min_{i\in[1,|\mathcal{R}^q|],j\in[1,|\mathcal{R}^d|]} \{dis((1+C_i^q)r_i^q, (1+C_j^d)r_i^d)\} \tag{2}$$

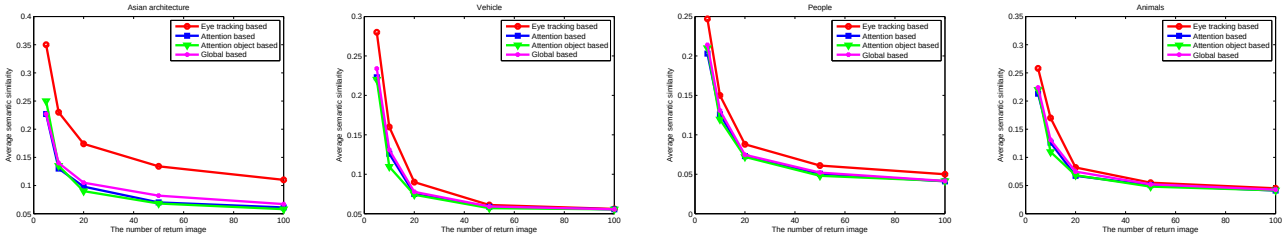

Figure 1: Average semantic similarity scores along with the number of returned images

where C_i^q and C_j^d are the importance values of the segmented regions r_i^q and r_j^d in q and d, respectively. $|\mathcal{R}^q|$ and $|\mathcal{R}^d|$) are the numbers of segmented regions in q and d respectively.

2.1 Experimental Setup

Our experiments are conducted on the public Hemmer color image database, which contains more than 7000 images with 8 categories, i.e., people, Asian architecture, animals, landscapes, tools, vehicles, fruits, and arts. All images have a resolution of 1920*1200 pixels in the 32-bit color mode. In this paper, we select 4 categories (Asian architecture, vehicle, people and animals), each of which contains 100 images. Each image in a category will be issued into the system as a query example to search over the whole collection.

In order to test the eye tracking based retrieval method, we first need to collect users' eye movement data and identify ROIs. This was done through a controlled task-based user study. We adopted a Tobii TX300 eye tracker in a user-friendly environment with a high accuracy (0.5 degree) at 300Hz sample rate. The freedom of head movement is 30x22x30 cm. Fixation is a widely used eye movement feature, which can be easily extracted by the Tobii Studio software. We asked users to complete 4 different retrieval tasks designed according to the 4 selected categories. Taking the "vehicle" category as an example, we requested users to go through all the 100 images to find vehicles and used the eye tracker to record all the fixations falling into different regions to identify ROIs. The more fixations a ROI contains, the higher weight (i.e., C_i) of the ROI will have. Once the ROIs are identified, they will be used to in the similarity measurement as described in the previous section.

We evaluate the retrieval performance by computing the average semantic similarity over all search rounds. Each image in the dataset has 20-30 keywords on average to describe what objects it contains. We define the semantic similarity S between a query image and a retrieved image as the ratio of matched keywords between their keyword descriptions: $S = \frac{2P}{M+N}$, where M and N are the numbers of keywords that the two images contain respectively, and P is the number of matched keywords.

Three baselines for comparison are: 1) Global feature based approach: retrieval based on the Euclidean distance of the global color and texture histograms of two images; 2) Attention based approach: retrieval based on the manually extracted salience objects (as ROIs) [4]; 3) Attention object based approach: Instead of using all the objects extracted in 2), Wang *et al.* used the first popped-out object only in the attention-driven image retrieval strategy [4].

2.2 Result Analysis

The performance of our proposed image retrieval model and the three baselines on different categories of query images are shown in Figure 2. We can see that for "Asian Architecture" and "People" image classes, our eye-tracking based method performs better than the others, in terms of average semantic similarity. For the other two image classes "Vehicle" and "Animals", our method is better when the number of returned images remains small (20 or less). The results demonstrate the robustness of our method in achieving better retrieval performance.

3. CONCLUSIONS

In this paper, we have presented a novel fixation-based method to detect regions of interest (ROIs) for content-based image retrieval. Compared with traditional methods, our proposed eye tracking based method generates encouraging retrieval performance without causing extra burden to users during the search process. In the future, we will validate our method in larger datasets and take into account more eye movement features such as pupil diameters, regressions and saccade trajectories.

4. ACKNOWLEDGEMENT

This work is funded in part by the Chinese National Program on Key Basic Research Project (973 Program, grant no. 2013CB329304 and 2014CB744604), the Chinese 863 Program (grant no. 2015AA015403), and the Natural Science Foundation of China (grant no. 61272265 and 61402324).

5. REFERENCES

[1] G. Buscher, A. Dengel, R. Biedert, and L. V. Elst. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):9, 2012.

[2] M. J. Cole, J. Gwizdka, C. Liu, R. Bierig, N. J. Belkin, and X. Zhang. Task and user effects on reading patterns in information search. *Interacting with Computers*, 23(4):346–362, 2011.

[3] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[4] W. Xiangyang, H. Fengli, and Y. Hongying. A novel regions-of-interest based image retrieval using multiple features. In *Multi-Media Modelling Conference Proceedings, 2006 12th International*, pages 4–pp. IEEE, 2006.

[5] Y. Zhang, H. Fu, Z. Liang, Z. Chi, and D. 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.