EYE-TRACKER BASED PART-IMAGE SELECTION FOR IMAGE RETRIEVAL

Christian Schulze[†] *Robby Frister*[⋆] *Faisal Shafait*[†,‡]

[†]German Research Center for Artificial Intelligence, Germany, christian.schulze@dfki.de
[⋆]Helmholtz Center for Environmental Research, Germany, robby.frister@ufz.de
[‡]School of Computer Science and Software Engineering, University of Western Australia, faisal.shafait@uwa.edu.au

ABSTRACT

Given the growing amount of very large image databases, content based image retrieval (CBIR) is becoming more and more important. One of the major challenges in CBIR is the semantic gap – commonly used feature-based algorithms are not able to identify what really draws human attention in an image. This problem is more crucial for localized CBIR, where certain regions / parts of the image are what the user is really interested in. This paper explores how human gaze can be utilized to extract regions of interest (ROIs) of an image to perform attention based image retrieval. Using eye tracking data and knowledge about foveal and peripheral vision of humans, we present a *foveal fixation clustering algorithm* that automatically generates ROIs in an image while a person is viewing it. To objectively set different parameters of the algorithm, a small user study was conducted. The method was evaluated for use in a localized CBIR system. Image retrieval results using the publicly available SIVAL dataset were scored using mean average precision (MAP). Comparison to a saliency-based visual attention algorithm as well as to manually labeled regions showed that the retrieval results of the developed algorithm are nearly two times better than the saliency-based visual attention algorithm and very close to the results using hand-labeled regions.

Index Terms— eye tracking, CBIR, image retrieval, gaze based interaction

1. INTRODUCTION

To search for digital images of landscapes or specific objects, we want to apply search engines. Such an engine needs information about the contents of the images on which the search can be performed. One way to achieve this is to annotate a digital image with meta information, i.e., the caption, date of creation, size, resolution, or color depth. Furthermore, additional information about the actual content of the image can be added. Text-based retrieval engines can utilize this meta information to retrieve images based on the query terms. Normally, a ranked list of images with respect to the relevance between query terms and tags is returned as the result. This approach to image retrieval suffers from a major constraint: the image content is not necessarily related to the provided meta information. A further concern is the huge amount of digital images available these days. The correct and sufficient annotation of billions of images is simply not feasible.

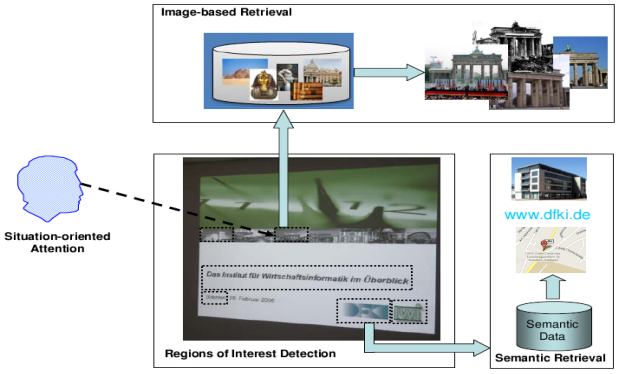

Fig. 1. Illustration of eye-tracker based region selection for the application of content based image retrieval.

Another approach to describe the image content is to extract a numerical representation of it. By computing specific statistics from the image, i.e., distribution of colors and textures, the actual image content is described in a way that similarities between these descriptions and thus the originating images can be computed. Therefore, this method is called content based image retrieval (CBIR). For CBIR, a query is made with an image showing content similar to what we look for in the image database.

However, CBIR methods do not consider which parts of an image are interesting or useful, since it depends entirely on the perspective of the viewer or user. Such a perspective can be obtained by exploiting user's attention data, for instance captured through an eye tracking device. Hence, keeping track of the user's gaze should provide all the information needed, i.e., where the user looks and for how long, to perform localized CBIR as shown in Figure 1.

In this paper, we explore how human gaze can be utilized

This work was supported by BMBF grant 01IW08002

to extraction regions of interest (ROIs) from an image and present an algorithm is proposed which is capable of automatic ROI extraction based on eye tracking data (Section 4). The so acquired ROI's are further investigated regarding their applicability to CBIR.

2. RELATED WORK

The success of CBIR systems in recent years has largely been attributed to the so-called "patch-based" methods as compared to the traditional global image descriptors. In a patch-based method, first points-of-interest in the image are detected. Then, image patches centered on those interest points are extracted and different features are computed as a descriptor of that image part. Robust methods for detection of interest points [1, 2] and extraction of their visual descriptors [3] have made it possible to reliably match images by comparing their local descriptors. Based on these methods, recognition systems can be developed that have excellent robustness against lighting and position variations, background changes, and partial occlusion [4, 5, 6]. Such approaches show impressive results in very challenging situations, and the first systems are already in commercial use [7, 8].

Patch-based methods have also been used to do part-image search for example in the *VideoGoogle* System [9]. In addition to the recognition of specific objects with patch-based methods, recognition of considerably more difficult object categories (eg "car", "tree") has also been reported [10, 11, 12].

Itti et al. [13] presented a visual attention system for the analysis of image scenes that is inspired by the early primate visual system. There, image features extracted from multiple scales are combined into a saliency map which then is presented to a neural network for the selection of attention drawing regions. We will compare our proposed algorithm against this method in Section 5. See Figure 6 for an example of the saliency-based region selection being applied. More recently, Judd et al. have bridged the eye tracking technology with the creation of computational attention models [14]. A system for object recognition and tracking in video using aspects of peripheral and foveal vision was proposed by Gould et al. [15] for the application in robotic systems. The authors use an attentive map marking unidentified objects based on a low resolution peripheral view to direct a high resolution camera for a high resolution foveal view of the region. Interactive cropping of photos based on the gaze information provided by an eye-tracker was presented by Santella et al. [16]. However, in opposition to our proposed application of eye tracking for part-image retrieval, they focus on the post-processing of photographs to increase the attractiveness to humans.

Recently, eye tracking technology has got attention for its potential use in image retrieval. Oyekoya et al. [17] demonstrated navigation through a collection of images using an eye tracker as a new human-computer interface. Kienzle et al. [18] designed a biologically-inspired interest point detector by learning human eye fixations at particular points in the image. An image retrieval system using a combination of eye tracking and visual features from the image was presented in [19]. Researchers have also proposed using eye tracking as a relevance feedback mechanism for image retrieval systems [20, 21] and on applying image re-ranking based on such relevance feedback [22]. In this work, we extend the state-of-the-art in the direction of part-image search and retrieval using eye tracking technology.

3. FOVEAL VISION AND EYE-TRACKING

The fovea centralis, also known as fovea, is a small depression in the retina located at the center of the macula which provides maximal acuity for vision. For activities where detailed vision is of importance, i.e., driving, reading, and watching the foveal vision is essential. Since it only covers about 2° of the human field of vision, it is necessary to constantly adjust the gaze to observe a large object. The visual information is processed between these gaze adjustments, where the eyes fixate for about $200 - 300ms$.

The patterns of eye movement and fixation can be determined by utilizing eye tracking technology. An eye tracker collects the gaze targets based on the eye movement and generates a data stream to compute the fixations of a user. Several methods for tracking the eyes of a person exist, i.e., electrooculography, search coils, infrared purkinje method, and camera-based. The camera-based approach, which will be utilized in this paper, can be further subdivided into a *bright* and *dark pupil method*. For the bright pupil method infrared light is directed at the eye, which will be reflected by the retina. Here, the infrared light source needs to be on the same axis as camera and eye. In the images taken by the camera the pupil of the eyes will appear bright in contrast to the surrounding due to their reflection and absorption behavior. The dark pupil method is based on the absorption of visible light in the retina. A source of visible light not being on the camera-eye axis is used to illuminate the eyes of the user, which leads to a dark appearance of the pupils on the camera image. For the user study and algorithm development a Tobii 1750 desktop-mounted eye tracker is used, which applies a combination of the bright and dark pupil method. To compute the fixations from the eye tracker provided data stream, the gaze locations are analyzed. A sequence of four gaze points are considered a fixation if they are placed in an area of 30×30 pixels. To tolerate noise from the eye tracker, which can be introduced by adverse lighting conditions or eye blinking, a larger region of 50×50 pixels surrounding the previous one is examined whether a subsequent gaze point is falling into it. If this is the case, the gaze point will be assumed to belong to the previous fixation, otherwise the following gaze points are analyzed for a new fixation. Three or less gaze points falling into a 30×30 region are ignored as outliers.

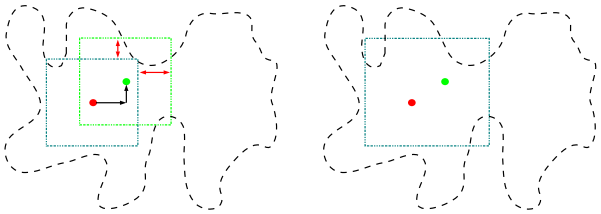

Fig. 2. Left: Cluster region extension for a new fixation (green point) by computing the distance to the centroid (red point) and between region borders. Right: New cluster after extending the cluster region.

Fig. 3. Example images from the SIVAL dataset for one object (Sprite Can) with different backgrounds.

It is obvious that the user-screen-distance affects the size of the foveal field of vision on the eye tracker screen. For instance, a user-screen-distance of $60cm$ and a field of vision of $2°$ marks a region of approximately 79 pixels diameter on the screen. For algorithmic convenience, we approximate this with the corresponding 79×79 square region.

4. FOVEAL FIXATION CLUSTERING

Our method to extract regions of interest (ROI) from an image depending on user gaze, is inspired by [23]. The algorithm iterates over all fixations of a user on a specific image. The first fixation generates a cluster of a fixed size. For every following fixation, the spatial distance to all existing clusters is calculated. If the distance to each cluster exceeds a threshold, the fixation will create a new cluster. Otherwise, the fixation is assigned to the nearest cluster. Finally, the clusters containing most fixations indicate potential ROIs.

However, two issues arise when applying this algorithm to detect the ROIs for part-image retrieval. First, regarding the relevancy of fixations, a retrospective interpretation of the data is not possible without a record of the complete viewing process. This prevents the decision about the relevancy of a fixation and if a particular fixation creates a new cluster. The second issue is related to the size and growth of the clusters. Several parameters have to be specified, i.e., starting size for clusters, spatial distance between successive fixations to form a new cluster, the amount and direction of cluster growth if a fixation falls into an existing cluster, and if the cluster size should relate to the number of fixations it contains.

Tests have shown that using the foveal region of the human field of vision provides a good base for the initialization of the cluster growth and size. This has lead to the development of the *foveal fixation clustering algorithm*, which is described in the following. For a new fixation and a fixation not falling into an existing cluster, a new cluster with the size of the foval region is created. If a new fixation is detected inside an existing cluster, first the distance of the fixation to the centroid of the cluster is computed to determine the direction for the region growth. Then, a pseudo-cluster around the new fixation is created and the distance between the borders of the pseudo and the existing cluster is computed in growth direction to find the amount of the cluster extension. In case the pseudo-cluster exceeds the existing cluster, the cluster region is extended by the computed distance value, otherwise the cluster region remains at its size. The procedure is illustrated in Figure 2. For the case of overlapping clusters and a new fixation falling into the overlapping area, the fixation will be added to the larger one of them, since its probably most important.

So far, all previously mentioned issues regarding cluster region growing are solved with this algorithm, except for the time based parameters, i.e., the duration of the fixation to indicate its relevancy. To investigate these parameters a user study has been conducted.

5. EXPERIMENTS

For all experiments presented in this paper we make use of the SIVAL dataset, which is a benchmark dataset for object-based or localized image retrieval. The SIVAL dataset has been previously used for tasks like content-based image retrieval [24, 25] and multiple instance learning [26]. This dataset is composed of twenty-five different image categories and each category contains sixty images of the same object taken from different viewpoints at varying locations (Figure 3).

To determine the fixation duration that indicates a user's interest in a region as well as to find out how long it takes to specify the region by gaze, a small user study with five participants was done. The study was carried out in a room with all interior removed except for the absolute necessary items, to minimize the distraction of the study subjects. Each of the participants was briefed about the process and the goals of the specific experiment. Then, the participants were seated in front of the eye-tracker with a user-screen-distance of $60cm$ and an example image was shown for preparation. As data, two images of each category were selected from the SIVAL dataset, forming a set of fifty images which were shown to the

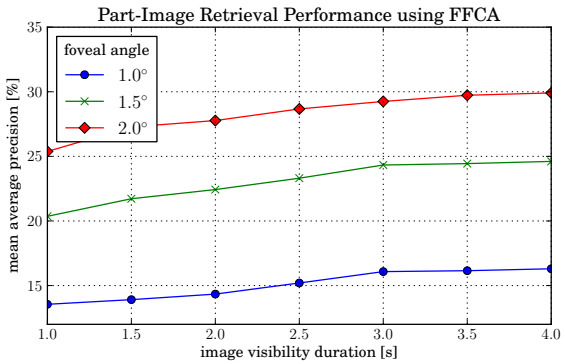

Fig. 4. Mean average precision achieved with query regions selected by the foveal fixation clustering algorithm for different image visibility durations and foveal angles.

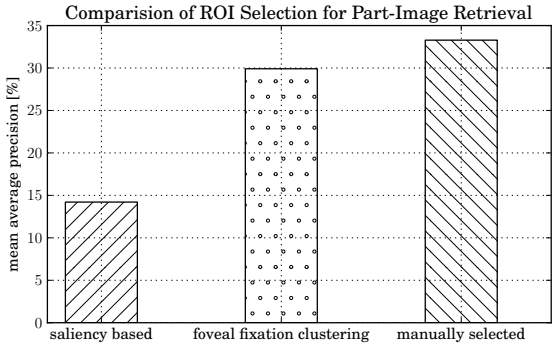

Fig. 5. Part-image retrieval performance based on query regions selected by the proposed foveal fixation clustering algorithm, saliency-based visual attention model, and manually selected ROI's.

subjects with a intermediate white screen for the relaxation of their eyes. The participants were instructed what the regions of interest in the upcoming images are, i.e. apple, sprite can, or tennis ball as well as to proceed to the next image whenever they think the region has been sufficiently marked by their gaze. Based on the outcome of this study, the threshold for the fixation duration that indicates the users interest in a region was determined at an average of $377ms$ and the duration for specifying a ROI in an image was at $3837ms$.

To evaluate the developed algorithm and analyze the effect of setting different parameter values, an experiment involving 20 participants was performed. Each image was shown to the users for four seconds. The participants were asked to look at a specific object in the image. The data resulting from this experiment was then used to form several query sets with image duration times of $1.0, 1.5, ..., 4s$ and foveal angles of $1°$, $1.5°$, and $2°$ to construct ROIs. These query sets were then utilized to evaluate the efficiency of the foveal fixation clustering algorithm by performing a part-image retrieval. We use a simple SIFT feature matching [6] based image retrieval engine, which ranks the database images based on the number of features matching to those extracted from the query image ROIs. The retrieval performance based on these queries was measured by mean average precision (MAP).

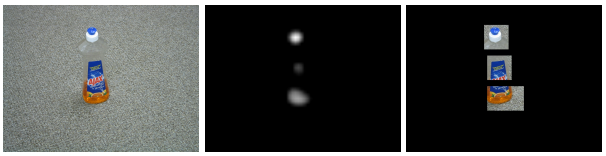

Fig. 6. Example for saliency-based region selection. Original image (l), saliency map (c), selected regions of interest (r).

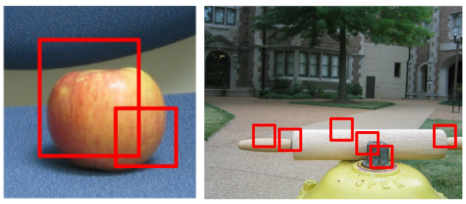

Fig. 7. Examples for regions selected by applying foveal fixation clustering.

Figure 4 shows the performance values for the investigated foveal angles in relation to the presentation time of the image. As can be seen, the large foveal angle of $2°$ provides the best retrieval results. Furthermore, it can be inferred that the optimal duration for the selection of the ROI is between $3.5s$ and $4s$, with a small performance increase between these timing values.

The last experiment compares the achievable performance in part-image retrieval using regions selected by the proposed foveal fixation clustering algorithm (Figure 7) against the selection of the saliency-based visual attention model [13] (Figure 6) and manually labeled ROIs marking the closest bounding box. The results of this experiment (Figure 5) show that the proposed algorithm performs nearly as good for selecting query regions in images as using manually labeled ones. Thus, it can be assumed that the previously found parameters, i.e., the foveal angle, the relevant fixation time, and the image presentation time allow the foveal fixation clustering algorithm to specify similarly well selected query regions. On the other hand, regions identified by applying the saliency-based attention model give a much lower performance compared to our proposed approach.

6. CONCLUSION

In this paper we presented a novel approach for clustering eye-tracker based fixations to select query regions for the purpose of part-image retrieval. Based on two user stud-

ies involving 5 and 20 participants respectively, parameters for the proposed foveal fixation clustering algorithm were determined and verified. Furthermore, it was shown that query regions selected by this algorithm provide a similar performance (29.9%) compared to manually selected regions (33.3%) for the application of part-image retrieval and outperform a saliency-based method for ROI selection in images.

7. REFERENCES

[1] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.

[2] F. Fraundorfer and H. Bischof, "Evaluation of Local Detectors on Non-Planar Scenes," in *28th OAGM/AAPR Workshop*, June 2004, pp. 125–132.

[3] K. Mikolajczyk, "A Performance Evaluation of Local Descriptors," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, June 2003, pp. 257–263.

[4] P. Roth, "Survey of Appearance-based Methods for Object Recognition," Tech. Rep. ICG-TR-01/08, Computer Graphics & Vision, TU Graz, 2008.

[5] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, *Toward Category-Level Object Recognition*, Springer-Verlag New York, Inc., 2007.

[6] D. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Int. Conf. Computer Vision*, September 1999, pp. 1150–1157.

[7] "Evolution Robotics VIPR Technology," February 2009.

[8] "kooaba Mobile Visual Search: point - snap - find," February 2009.

[9] J. Sivic and A. Zisserman, "Video Google: Efficient Visual Search of Videos," in *Toward Category-Level Object Recognition*. 2006, pp. 127–144, Springer-Verlag New York, Inc.

[10] B. Leibe and B. Schiele, "Robust Object Detection by Interleaving Categorization and Segmentation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, 2007.

[11] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, June 2003, pp. 264–271.

[12] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," October 2008.

[13] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[14] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," in *ICCV*. 2009, pp. 2106–2113, IEEE.

[15] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Meissner, G. Bradski, P. Baumstarck, S. Chung, and A.Y. Ng, "Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video," in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007.

[16] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based Interaction for Semi-Automatic Photo Cropping," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 771–780.

[17] O. K. Oyekoya and F. W. M. Stentiford, "Eye Tracking as a New Interface for Image Retrieval," *BT Technology Journal*, vol. 22, no. 3, pp. 161–169, 2004.

[18] W. Kienzle, F.A. Wichmann, B. Scholkopf, and M.O. Franz, "Learning an Interest Operator from Human Eye Movements," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 24–24.

[19] Z. Liang, H. Fu, Y. Zhang, Z. Chi, and D. D. Feng, "Content-based Image Retrieval Using a Combination of Visual Features and Eye Tracking Data," in *2010 Symposium on Eye-Tracking Research and Applications*, Mar. 2010, pp. 41–44.

[20] Y. Zhang, H. Fu, Z. Liang, Z. Chi, and D. D. Feng, "Eye Movement as an Interaction Mechanism for Relevance Feedback in a Content-Based Image Retrieval System," in *2010 Symposium on Eye-Tracking Research and Applications*, Mar. 2010, pp. 37–40.

[21] A. Faro, D. Giordano, C. Pino, and C. Spampinato, "Visual Attention for Implicit Relevance Feedback in a Content Based Image Retrieval," in *2010 Symposium on Eye-Tracking Research and Applications*, Mar. 2010, pp. 73–76.

[22] D. Hardoon and K. Pasupa, "Image Ranking with Implicit Feedback from Eye Movements," in *2010 Symposium on Eye-Tracking Research and Applications*, Mar. 2010, pp. 291–298.

[23] H. Katti, R. Subramanian, M. Kankanhalli, N. Sebe, T. Chua, and K. Ramakrishnan, "Making Computers Look the Way We Look: Exploiting Visual Attention for Image Understanding," in *Proceedings of the International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 667–670, ACM.

[24] R. Rahmani, S.A. Goldman, Hui Zhang, S.R. Cholleti, and J.E. Fritts, "Localized content-based image retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1902–1912, 2008.

[25] R. Vieux, J. Benois-Pineau, and J.-P. Domenger, "Content Based Image Retrieval Using Bag-Of-Regions," in *Proceedings of the 18th International Conference on Advances in Multimedia Modeling*, Berlin, Heidelberg, 2012, MMM'12, pp. 507–517, Springer-Verlag.

[26] R. Rahmani and S.A. Goldman, "MISSL: Multiple-Instance Semi-Supervised Learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 705–712.