IMAGE/VIDEO INDEXING, RETRIEVAL AND SUMMARIZATION BASED ON EYE MOVEMENT

Atsuo Yoshitaka

Japan Advanced Institute of Science and Technology (JAIST), Japan, ayoshi@jaist.ac.jp

ABSTRACT. Information retrieval is one of the most fundamental functions in this era information. There is ambiguity in the scope of interest of users, regarding image/video retrieval, since an image usually contains one or more main objects in focus, as well as other objects which are considered as ‘background’. This ambiguity often reduces the accuracy of image-based retrieval such as query by image example. Gaze detection is a promising approach to implicitly detect the focus of interest in an image or in video data to improve the performance of image retrieval, filtering and video summarization. In this paper, image/video indexing, retrieval and summarization based on gaze detection are described.

Keywords: gaze detection, AoI, image retrieval, information filtering

INTRODUCTION

Rapid progress of computers has enabled us to handle large amounts of multimedia data, including images, sound, and video, even for personal use. Smart phones are no longer only telephones, but enable us to take photo/video, share images via social network services, manage our personal schedule, browse Web pages, etc. Digital broadcasting, including surface wave and satellite broadcasting, as well as content delivery via the internet have enabled us to access hundreds of channels of programs, where multiple channels can be recorded simultaneously. The type of information which plays a major part in our society has shifted from text to image/video, and the amount of information we can access is exploding year by year. Image/Video data is stored as personally managed data from a digital still camera, video camera, or personal video recorder, and is commonly used in communicating via social network services. Gigabytes or even terabytes of storage enable us to store unlimited images, from hundreds to thousands of hours of video data. Information management, including browsing, indexing, retrieving, filtering, and summarizing, is fundamental for better access to all this information, especially under the circumstances of today’s information explosion.

For issues of information management in the past, especially information retrieval, finding keywords by exact match within the textual data was a basic technique, and it was satisfactory in most cases, since information in text form is represented by specific terms with little ambiguity. Image/Video retrieval has shown us quite different issues to be taken into account, compared with text-based information retrieval, because of the diversity of interpretations and focus of interest. Ambiguity relates to the diversity of interpretations, focus of interest, and criteria of judgment. The diversity of information which can be interpreted has two aspects; one is structural and another is semantic. The structural level refers to objects which humans recognize in an image, from component parts to an entire object. The latter is classified into objective, contextual, and affective information. Objective information is information with regard to the existence of specific objects in the image, whereas the contextual information is the information which relates to the specific characteristics of the object; concentrating not on 'what it is' but on 'how it is'. Affective information is perceived from the image/video, which is explicitly represented but implicitly perceived as a result of psychological effects due to specific patterns of objects, colors, or temporal transitions of images. The type of information on which a person focuses depends on his/her interest, or the purpose of information management.

Query by image/video example is one promising technique for offering an intuitive interface from the point of view of better human-computer interaction. A query condition is specified by presenting an image/video example which contains a target object, or the motion of object(s) to be retrieved. Since it is often the case that such an example image or video data contains not only the target object, but also other irrelevant objects, it is important to detect the area of interest or object of interest of the user, in order to improve the quality of information retrieval. In other words, the major issue is "How should we detect the focus of interest or preference in image/video data in the process of data management?" This should also be considered in information filtering or summarization.

As a solution for the issue of detecting the area of interest or user's preference, gaze detection or eye movement analysis is one promising method. In this paper, several topics in gaze-detection-based information management are described: visual life logging and its retrieval, implicit social filtering based on eye movement, refinement of local features for content-based image retrieval, and video summarization based on viewer's behaviour.

HUMAN EYE MOVEMENT AND GAZE DETECTION

Human eyes are known to make typical movements, depending on an object being watched. In the case of gazing at a motionless object, the eye alternates fixation and saccade. Fixation corresponds to the state in which the eye is not moving from macro-movement point of view, whereas the saccade is a motion where the point of gaze moves instantly from one position to another. A saccade takes approximately 30msec., and a fixation usually lasts more than 300msec. During fixation, drift, tremolo, and micro-saccade are observed as micro-movements, and they are not taken into account for user interest or preference detection in this paper. In the case of staring at a moving object, smooth pursuit, i.e. smooth eye movement following the object without saccadic movement, is observed if the movement of the object does not exceed 30deg/sec., approximately.

Eye movement can be monitored with a commercial product eye tracker, however, there are a number of studies of eye tracking with mono or stereo camera. Figure 1 is a photo of a head mounted camera for gaze detection developed by our laboratory. It consists of two CCD cameras, one of which is a B/W camera with IR-LEDs to shoot an eye, and another is a kolor CCD camera for shooting user's field of view. Figure 2 is a snapshot of analysing eye movement. Two graphs correspond to vertical and horizontal distances between the previous and current gaze position, and concentration on a visual object is detected by extracting the repetition of relatively long fixations.

VISUAL LIFE LOGGING AND IMAGE RETRIEVAL

Mobile computers with cameras and/or sensors opened up applications called "life-log", which stores sensor data and/or photos of the surroundings where a user is located. Such data are accessed to retrieve a past record of real-world data. Focusing on image/video data, current hardware and storage capacity may be enough for recording every scene continuously which a user faced during a day. Under this scenario, indexing of image data is an important issue for better accessibility of captured data, since captured image/video is highly redundant and it is difficult to identify which objects in captured images are of the most interest or importance to the user.

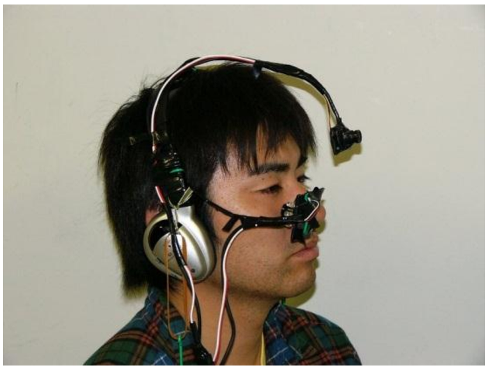


Figure 1. Head Mounted Camera

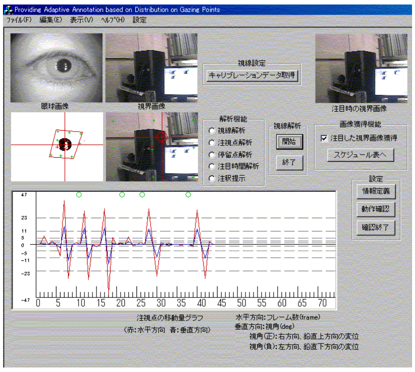


Figure 2. Displaying Eye Movement

This visual life-log system enables the capture of a user's viewed image only when he/she has watched an object with certain strength of attention. This is realized by monitoring user's eye movement, and the state of paying attention or interest is detected by measuring the duration of fixations on an object, and counting the repetitions of long fixations on the same object (Yoshitaka, 2004). When the state of attention to an object is detected, the indices of the time stamp, AoI (area of interest), saccadic pattern, and duration of fixation are stored in addition to his/her viewed image associated with this data.

It may not be easy for people to remember visual details such as the facial appearance of a person he/she met some days ago, or the appearance of goods in a showcase he/she looked at with attention. Retrieving captured viewed images with this system works to complement human's visual memory based on WYSISYR (what you see is what you retrieve). WYSISYR stands for the method of interaction in image information retrieval; images from a database which are similar to his/her current view are retrieved. His/her viewed image is treated as an image example, and retrieval is performed as a query by image example. Image retrieval using only image-related features such as kolor may result in irrelevant images, whose image features are similar, although the objects in them are of a different category. Though this issue may be resolved by bringing a local-feature-based matching for example, the computation time may not be acceptable in a mobile computing environment. In our system, similarity of saccadic patterns and similarity of duration of fixations is evaluated in addition to the similarity of basic image features in the process of retrieval.

Figure 3 is an example of WYSISYR, evaluating only kolor similarity between image example (i.e., snapshot image of user's current view) and images in visual life-log database. Snapshot images which are evaluated to be similar to the current view image are arranged from top to bottom and left to right, in the descending order of similarity measure. These results contain snapshots with irrelevant objects, such as a computer keyboard, in positions showing more similarity, because of the similarity of the grey kolor. Figure 4 shows the result of viewed image retrieval by evaluating the similarities of saccadic patterns and the duration of fixations as well as kolor similarity. As seen in the figure, most of the snapshot images, which include books of vertical text, are placed in positions of greater similarity, without showing irrelevant snapshots. Note that snapshots in the lower position are displayed, since there are only a few snapshots of vertical text.

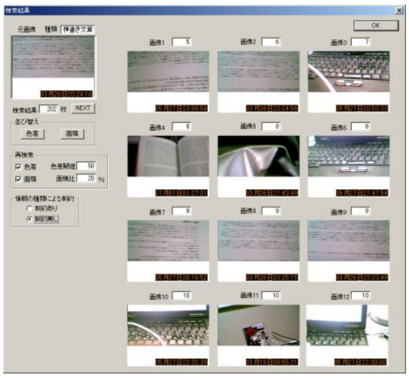


Figure 3. Retrieval by Colour Similarity

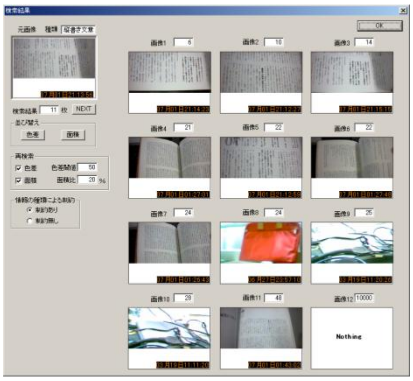


Figure 4. Retrieval by Eye Movement Pattern

SOCIAL FILTERING BASED ON IMPLICIT PREFERENCE RECOGNITION

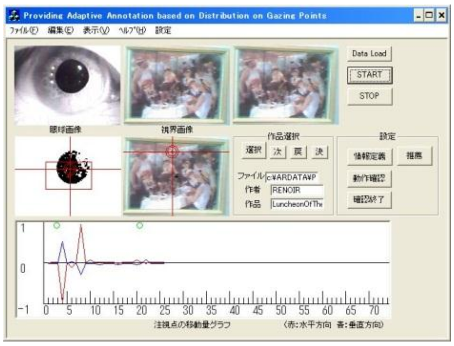
Social filtering is a technique to extract information that fits the preferences of a target user based on the similarity of his/her preferences to those of others, under the assumption that the information preferred by others is also preferred by the target person. A straightforward method to build someone's preferred model is to have him/her answer a questionnaire. However, this method requires him/her to spend time which is not for his/her primary task.

Gaze detection is one solution for this problem, since it is applicable for detecting user's preferences. It is natural behaviour for a person to spend more time watching an object of interest, and spend less time on things of less interest. For instance, assume that a person is appreciating paintings in a museum. Assume that there are a large number of paintings, too many to appreciate all of them in a day, and one would like to appreciate only the paintings which match one's preferences. One possible way to build a preference model is to detect preferences by measuring the duration of time spent in front of a painting, since someone will stay in front of a painting longer if he/she is more interested in it than in other paintings. This method enables one to build a picture-based preference, however, it is difficult to build object-based preferences because the objects of interest cannot be detected.

Figure 5 shows an intermediate process of detecting objects of interest while one is appreciating a painting. Each object in the painting is segmented into a region with annotations so as to build a preference model based on objects. In the process of information recommendation, preference models of other persons, who have already visited the museum and appreciated paintings, are extracted if they are similar to the preferred model of the target user. Finally, paintings which have not yet been seen by the user but fit his/her preferences are presented to him/her as recommendations, as shown in Figure 6 (Yoshitaka, 2007).

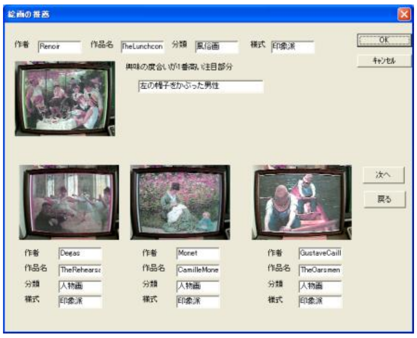
REFINEMENT OF LOCAL FEATURE SELECTION FOR OBJECT RECOGNITION

Object recognition or object categorization based on local features is one of the promising approaches, and is currently a hot topic in the area of object recognition or image retrieval. The idea of object recognition or classification by local features is to prepare small pieces of visual words which represent a category of object, and object recognition is performed by evaluating how many visual words are found which are included in a category of object.



ファイル c:¥ARDATA¥P
作者 RENOIR
作品 LuncheonOfThe

Figure 5. Detecting User Preferences[1]



作者 Renoir 作品名 TheLunchcon 分類 風俗画 様式 印象派

興味の度合いが1番高い注目部分

左の帽子をかぶった男性

作者 Degas
作品名 TheRehears
分類 人物画
様式 印象派

作者 Monet
作品名 CamilleMone
分類 人物画
様式 印象派

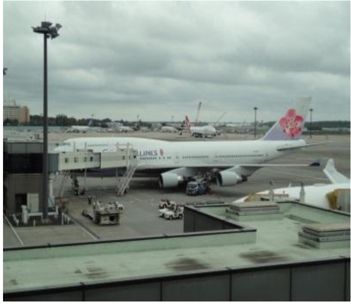
作者 GustaveCaill
作品名 TheOarsmen
分類 人物画
様式 印象派

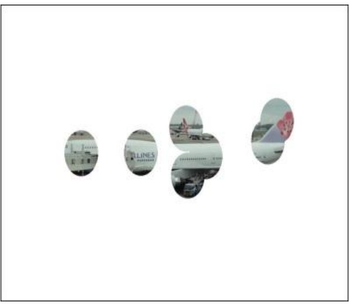
Figure 6. Presenting Recommendations[2]

When we consider query by example, by specifying an image which contains the target object, the example image may contain not only the target object which the user wishes to retrieve, but also background or irrelevant objects. The simple method of organizing a bag of visual words (Csurka, 2004) which is obtained by extracting local features from the query image regardless of whether they correspond to parts of the target object or not, will result in a bug of visual words which may reflect local features of not only the target object but also irrelevant objects or background. If such 'noise' co-occurs with target information, it may act like a part of the features of the target object. If does not, it appears that the accuracy of object detection has deteriorated.

One solution for this issue is to separate the target object from irrelevant objects or background before organizing a bag of visual words. A saliency map (Itti, 1998) simulates conspicuity of visual information based on the stimulus processing model of the human vision system. We can use the saliency map of an example image to extract the area of interest with kolor-based segmentation.

Visual conspicuity, however, does not always correspond to the object in a person's focus. Filtering visual words by gaze detection is another possible solution. Gaze point directly corresponds to the position where the user focused on an example image, and the gaze points which correspond to exploring eye scans can be excluded by gaze duration filtering.





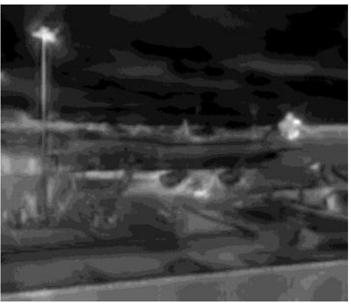


Figure 7. Original Photo **Figure 8. Saliency Map** **Figure 9 Gazed Regions**

Figure 7 is an example of an airplane photo. Conspicuous region obtained by applying saliency map is shown in Figure 8. Figure 9 shows that the regions gazed at applying the

[1] Painting in the figure: Luncheon of the Boating Party(Renoir)

[2] Painting in the figure: Luncheon of the Boating Party(Renoir), left to right on the bottom:The Rehearsal(Degas), Camille Monet and a Child in the Artist's Garden in Argenteuil(Monet), The Oarsmen (Gustave Caillebotte)

range of central vision. By comparing these pictures, saliency map with segmentation does not always correspond to the actual area of interest, and therefore, gaze-based visual word refinement outperforms saliency map based refinement.

VIDEO SUMMARIZATION BASED ON VIEWER'S BEHAVIOR

Video summarization is one technique for efficient access to information in the era of information overload. There is no algorithm which produces an ideal summary regardless of the category of video program or viewer's interest, since audio/visual features that correspond to significant segments differs depending on the category of contents. Significant segments which are commonly recognized by people can be extracted based on film grammar (Arijon, 1976). However, adjusting personal interest or preference is another issue to be solved.

Personal video taken by a non-professional user is a good example in which the film grammar does not work well in detecting significant segments. In addition, significant segments of such video may depend on each individual, since the contents depend on personal experience or sensitivity. Significant segments of video data are detected by monitoring eye gaze while a viewer watches the video. The pattern of the occurrence of typical fixation, as well as explicit video operations such as rewinding, fast-forwarding, and pausing, are referred to for measuring the degree of significance (Yoshitaka, 2012).

CONCLUSION

In this paper, several aspects of utilizing gaze detection for better accessibility of image/video data are described. Gaze detection is a promising method to measure user's attention, interest, or preference implicitly. It is applicable to a wide range of visual information management such as retrieval, filtering, and summarization. Saliency map can be considered as a substitute for gaze detection, however, visual saliency does not always follow one's interest or preference. Gaze detection is one of the key strategies in the area of human computer interaction to improve accuracy and/or quality of image/video data retrieval, information filtering, and video summarization.

ACKNOWLEDGMENTS

The author would like to express appreciation to the members of the image laboratory at JAIST for their valuable discussion, and to Ms. M. A. Mooradian for her help in revising English.

REFERENCES

Arijon, D. (1976) Grammar of the film language. *Silman-James Press*.

Csurka, G., Dance, C.R., Fan, L., Bray, C. (2004) Visual categorization with bags of keypoints. *Proc. of European Conference on Computer Vision*, 1-22.

Itti, L., Koch, C., Niebur, E. (1998) A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11), 1254-1259.

Yoshitaka, A., Hori, Y., Seki, H. (2004) Digital Reminder: Real World-Oriented Database System. *EURASIP Journal on Applied Signal Processing*, 2004 (11), 1663-1671.

Yoshitaka, A., Wakiyama, K., Hirashima, T., (2007) Recommendation of Visual Information by Gaze-Based Implicit Preference Acquisition. *Proc. of 13th Multimedia Modeling Conference*, 126-137.

Yoshitaka, A., Sawada, K., (2012) Personalized Video Summarization based on Behavior of Viewer. *Proc. IEEE Eighth International Conference on Signal Image Technology and Internet Based Systems*, 661-667.