GazeCloud: A Thumbnail Extraction Method using Gaze Log Data for Video Life-Log

Yoshio Ishiguro
The University of Tokyo, Japan
JSPS Research Fellow,
ishiy@acm.org

Jun Rekimoto
The University of Tokyo, Japan
Sony Computer Science Laboratories, Inc.

Abstract

We propose a method for information extraction and presentation using recorded eye gaze data, i.e., life-log video data. We call our method GazeCloud, which essentially uses gaze information for the generation of thumbnail images. One of the usages of wearable computing, personal life-logs are becoming increasingly possible. However, an aspect that needs to be addressed is information retrieval through different browsing methods. It is also well known that human memory recall is aided by effective presentation of information. Our propose method GazeCloud calculates the importance of information from gaze data that is consequently used for the generation of thumbnail images. This method performs the calculation using the eye gaze duration and hot spot information. Additionally, we construct a prototype daily-use wearable eye tracker system.

1. Introduction

Eyes do more than provide the sense of sight to a human; they can aid in communication as well. In this study, we developed GazeCloud, a method capable of presenting daily life-log data in a browsable manner, based on thumbnail images generated using gaze information (Figure 1). In the SenseCam Project [7], the approach of displaying pictures captured from daily life has proved to be a useful retrospective memory aid [5]. In this project, a wearable camera that captures images from the user's daily activities was used. Moreover, several studies have attempted to record user locations, accelerations, and data on other daily-life events [4,8,9]. However, life-log database will be very large. For example, video recording of 8 hours at a rate of 15 frames per second gives a video frame with approximately 432,000 images. An important issue is the availability of efficient viewing/browsing methods for recalling daily activity from large-scale life-log data. Therefore, we propose to use gaze information for this purpose. Eye movement is well known to be a representative of human interests [11]; corresponding research on eye movement has been ongoing for nearly the last 100 years [10,12]. We propose information extraction method by using gaze information in addition to other context information by using wearable eye-tracker (Figure 1). In this study, we explored the use of eye gaze information for developing a browsing method for pictures taken from daily life. We also constructed a noninvasive system that could be used in daily life (i.e., without causing discomfort to the wearer and other individuals around), rather than a conventional experimental system that could have high accuracy but low practical usability. Finally, we applied the proposed method and system to actual recorded data.

2. Related work

A number of works reported in the literature have addressed life-log data and its browsing methods. Notably, the SenseCam [7] is a wearable life-log device that captures digital images from daily life; it can be worn daily. Hodges et al. presented the potential memory benefits of SenseCam. This device records images but also acceleration, ambient temperate, light level, and other environmental information and a

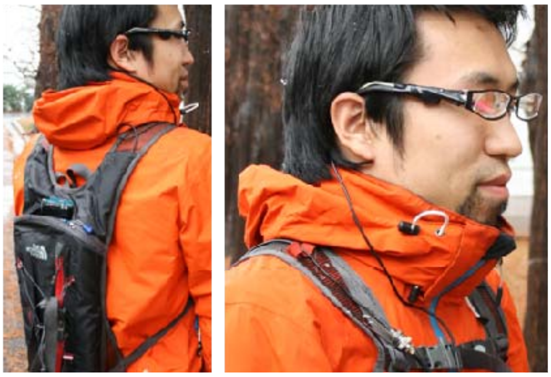

Figure 1. Wearable eye tracker capable of recording video data and gaze focus to a notebook computer for life-logging.

1550-4816/12 $26.00 © 2012 IEEE
DOI 10.1109/ISWC.2012.32

browsing method that uses these data is also proposed [14]. Another browsing tool called the “SnapTrack” [2] is used for human memory recall via a combined use of the global positioning system (GPS) and SenseCam data. Further, several researchers used image data to extract information from them [3,6]. However, we consider it difficult to estimate the user’s state by using only passive (environmental) data. For estimating a user's state, the life-log data should include not only environmental information and users’ physical movements but also users’ physiological changes. Consequently, Aizawa et al. proposed an information extraction method that uses brain waves in addition to acceleration, gyro, and other kinds of data [9].

Different life-log data visualization methods have been researched [6]. Furthermore, Goldman proposed a method for generating thumbnail images automatically, in a professional storyboard-like manner [1].

3. Video life-log browser with gaze information

In order to obtain an efficient information-viewing method, it would be beneficial to use visualization methods such as “heat map” or “tag cloud.” The user cannot afford to spend an hour in watching a given life-log video. Factors that would contribute to quick and efficient playback are extraction of specific scenes and the data presentation method. By setting the key frame and presenting the user with appropriate context information, the user can easily access linked life-log video data. Therefore, we propose a simple data viewer and a method termed “GazeCloud.”

3.1. GazeCloud: gaze-oriented picture-tag cloud

Using tag clouds is a common practice in Web services. They can represent textual data visually, e.g., in the form of keyword metadata. The user can determine which keyword is more popular or the amount of information contained in similar tags.

Tag clouds are mainly used for the reason that they can direct attention to tags indexing the data, it is placed by alphabetical order and size is controlled by importance. Tags are normally listed alphabetically on the X-Y axis, and the text size shows their importance. In this manner, the importance of the text data is highlighted. Consequently, we apply this visualization and presentation method to life-log video data. We also use the target dwell time and the recorded date of each frame for key-frame extraction.

The tag cloud lists tags alphabetically for the X-Y mapping of the life-log video data, and the date of the frame is used for the purpose of listing thumbnail images. However, instead of acquiring frames at regular intervals, the key frame is extracted using contextual information about the importance of text.

3.2. Importance estimation

It is vital to specify the importance of each frame. This importance estimation uses some specific gaze behavior such as the gaze duration, a gazed point, the gaze angle distribution and other eye movement.

In the first approach, the importance is estimated from the standard deviation (SD) of gaze position in the defined intervals. The gaze duration information can be used to differentiate between objects glanced at and those gazed at carefully. Essentially, we define the duration as the importance ***(imp)*** where the SD of gaze point ***(x, y)*** during the defined interval $\boldsymbol{n}$ (frames) is calculated. The importance for each instant of time $\boldsymbol{t}$ is.

$$\frac{1}{imp_{(t)}} = \frac{1}{n}\sum_{i=t-n}^{t}(x_i - x_{AVE})^2 + \frac{1}{n}\sum_{i=t-n}^{t}(y_i - y_{AVE})^2$$

If the SD is over the threshold specified by the user, then the thumbnail image is adopted. Additionally, the thumbnail-image size is decided based on the difference between the threshold and the importance. The user controls the number of thumbnails by changing the threshold level.

Another approach could be to combine gaze and face information in addition to textual and other information extracted from the camera images.

4. Daily usable gaze data recording system

We constructed an eye tracker for measuring gaze direction in daily life. This system is small and lightweight, because life-log data recorded using an eye tracker that would disrupt and hinder daily-life movement and activities would be meaningless.

Figure 2 shows our constructed eye tracker. The principle of this spectacle-type (head-mounted) tracker is based on dark-pupil detection. It can record the gaze direction and the user’s view simultaneously by using

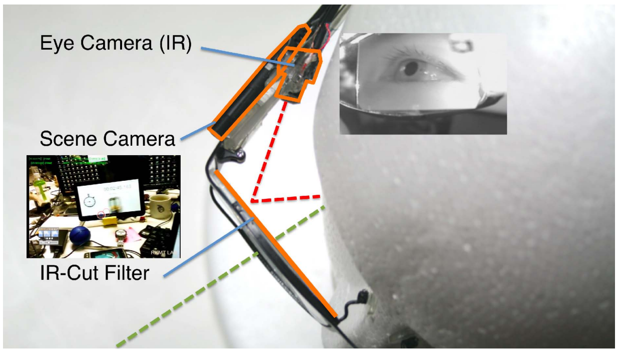

Figure 2. Our wearable eye-tracker prototype. An IR camera placed on the ear side can capture the eyeball surface that reflected the image on the filter accordingly.

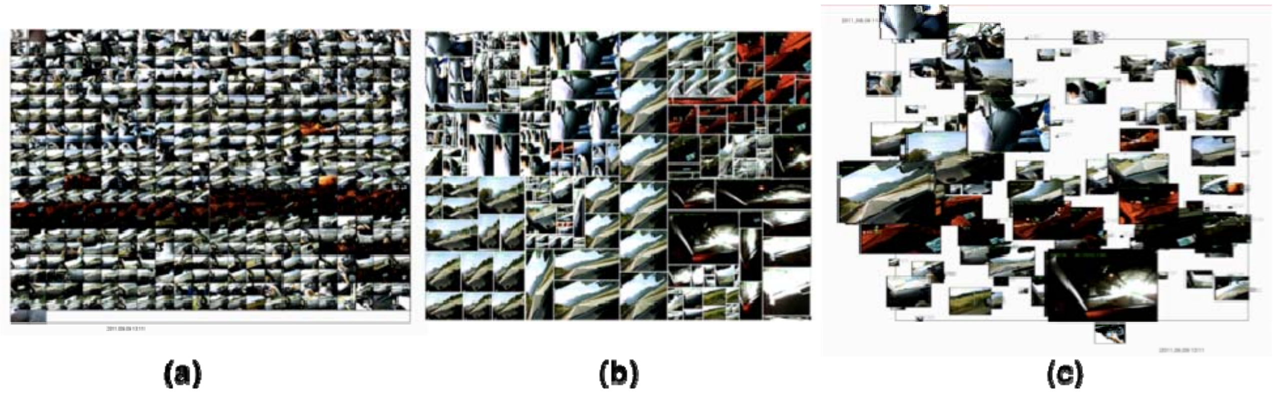

Figure 3. The result of “Time interval”, “Treemapping” and “GazeCloud” method. (a) all thumbnail are same size. (b) treemapping images show which thumbnail is larger than other without occlusion. (c) GazeCloud images show the thumbnail with importance with time line.

two cameras (see Figure 2). The first camera is an infrared (IR) one that detects dark pupils, and the second camera records the user’s view under visible light conditions. These two camera images are sent to a PC, where they are processed and saved. We used a PC with a 2.3 GHz Core i5 processor. We estimated the gaze point in real time. Life-log video data is encoded using the H.264 encryption and is stored along with textual data of the gaze information.

The system’s sensors are placed away from the eye front (or the face) in order to avoid disturbing real-life communication. We achieved this configuration by using an IR cut (reflect) filter. This filter appears similar to a clear glass panel, but it functions as a mirror during capture through an IR camera. An IR camera placed on the ear side can capture the eyeball surface that reflected the image on the filter accordingly. Thus, this system is similar to normal spectacles, but it can capture eye gaze and environmental images. It uses two small analog cameras and analog-digital converters (ADCs); the cameras have a resolution of 720 x 480 pixels. The gaze direction estimation accuracy is 1.49 degrees. The sampling rate is fixed to approximately 30 Hz, so fast eye movements such as micro-saccades cannot be measured, as they require a sampling rate of more than 500 Hz. However, the spectacle component of this system weighs less than 50 g (including two cameras, the IR cut filter, and mounting frame).

5. Examples of actual daily gaze data

Figure 3 is created from a dataset that recorded on participant for approximately one hour. This user sits on the passenger and navigated for the driver. We review our proposed presentation methods by using this dataset.

5.1 The difference between “time interval” extraction and “gaze duration” extraction

First of all, it is compare the images that are extracted thumbnail by basing on time interval or importance. Figure 3(a) show the result of time interval (10 seconds) extraction method and that method extracted 484 thumbnails. On the other hand, Figure 3 (d) and (c) show the result of importance-based extraction (fixation) method and that method extracted 441 thumbnails. The number of thumbnails has not difference between these two methods, but the extracted thumbnails are different. The thumbnails that extracted by interval, there is a various scene. On the other hand, these thumbnails are same kind of scenes. These differences show the interval extraction method did not extract thumbnails based on personal interests. But importance based extraction method automatically segment video log data into scene by weighting to time that based on user's importance.

5.2 The "Passage of Time" fidelity vs. the "Importance" fidelity

For improve visualization of life-log data, each extracted thumbnail size is changed by importance. Figure 3(b) shows the result, which applied to treemapping method [13]. The thumbnail is sorted by time stamps and these thumbnails are set out that is based on treemapping method. Each image size is changed by importance as shown as Figure 3(c). However, it is difficult to show the time exactly by using treemapping. The time course was easily understood by seeing thumbnails that is simply arranged, because these thumbnails are extracted by time interval. On the other hand, it is difficult to know the time course from thumbnails that is extracted by importance. For this reason, our proposed method "GazeCloud" allocate thumbnails by time stamps (as shown as Figure 3(c)). Additionally, the thumbnail size also is also coordinated by importance. Figure 3(c) shows the passage of time and the difference of importance at the same time.

5.3 "Time interval" vs. "GazeCloud"

We conducted an interview with two users to examine the difference between "time interval" and "GazeCloud" thumbnails by using user specific gaze log data (recorded for one hour). We conducted the interview four months (120 days) after the gaze log data recorded. The apparatus of interview is explained as following: First, participant browses "time interval" thumbnails for 10 seconds. Second, after this browsing, user talks freely about recalled information. We repeated the same approach for "GazeCloud" thumbnail browsing and recalling.

One participant browses his recorded data with two different thumbnails methods (Figure 3(a) and Figure 3(b)). After browsing interval-based thumbnails (Figure 3(a)) a participant said, "I was traveling, sitting in the front passenger seat, and the car drove through a tunnel, but I can not remember who was driving the car." Then he browsed Figure 3(c), and said, "Traveling, sitting front passenger seat and (*Driver name*) drive car, and (*passenger name*) was sitting in the backseat. We talked about some interesting stories."

Our qualitative study on the interview session show the time interval method is able to present the passage of time, and outline of life events. Where's, GazeCloud method is able to present object of interest. Thus, this participant could recall whom he traveled with, and the details of conversation topics in car. Other participant also denotes the same tendency.

6. Conclusion

We proposed "GazeCloud," a browser for life-log video data based on gaze information. Additionally, we constructed a prototype of a daily-use wearable eye tracker. The life-log video data and gaze data were obtained by recording for a few hours using our proposed method. However, collection of data over a longer duration (e.g., one month or one year) is required to perform a detailed evaluation of our proposed method. Information (thumbnail images) extraction will be more accurate if we can combine not only gaze information but also acceleration, location, and other sensor information. In the future, we intend to combine longer-duration life-log data with gaze and other contextual data using our proposed method.

In this study, we focused on only personal life-log data. However, we are also interested in relating gaze information to a second individual's life-log data. We also want to explore the feasibility of human memory enhancement with data obtained by recording for a longer duration and browsing by using our proposed method.

7. ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows, 21-8596, 2009.

8. References

[1] Goldman, D. B., Curless, B. C., Salesin, D., and Seitz, S. M., Schematic Storyboarding for Video Visualization and Editing. Proc. of SIGGRAPH 2006, Vol. 25, No. 3, 862-871.

[2] Kalnikaite, V., and Sellen, A., Whittaker, S., Kirk, D., Now let me see where i was: understanding how lifelogs mediate memory. Proc. of CHI '10. ACM, New York, NY, USA, 2045-2054.

[3] Doherty A. R., and Smeaton, A. F., Automatically Segmenting LifeLog Data into Events. Proc. of WIAMIS '08. IEEE, Washington, DC, USA, 20-23.

[4] Wang, Z., Hoffman, M. D., Cook, P. R., and Li, K., VFerret: content-based similarity search tool for continuous archived video. Proc. of CARPE '06. ACM, New York, NY, USA, 19-26.

[5] Vemuri, S., and Bender, W., Next-Generation Personal Memory Aids. BT Technology Journal 22, 4, 125-138, 2004.

[6] Byrne, D., Kelliher, A., and Jones,G. J. F., Life editing: third-party perspectives on lifelog content. Proc. of CHI '11. ACM, New York, NY, USA, 1501-1510.

[7] Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K., SenseCam: A retrospective memory aid. Proc of Ubicomp '06. Springer-Verlag, Berlin, 177-193.

[8] Vemuri, S., Schmandt, C., Bender, W., Tellex, S., and Lassey, B., An audio-based personal memory aid. Proc of Ubicomp '04. Springer-Verlag, Berlin, 400-417

[9] Aizawa, K., Tancharoen, D., Kawasaki, S., and Yamasaki, T., Efficient retrieval of life log based on context and content. Proc. of CARPE'04. ACM, New York, NY, USA, 22-31.

[10] Jacob, R. J. K., Eye movement-based human-computer interaction techniques: Toward non- command interfaces. In Advances in Human- Computer Interaction, Ablex Publishing Co, (1993), 151-190.

[11] Conde, S. M., and Macknik, S. L., Windows on the mind. Scientific American, 297(2):56–63, 2007.

[12] Findlay, J. M., and Gilchrist, I. D., Active Vision: The Psychology of Looking and Seeing. Oxford University Press, 2003.

[13] Johnson, B., and Shneiderman, B., Tree-Maps: a space-filling approach to the visualization of hierarchical information structures. Proc of VIS '91, IEEE Computer Society Press, Los Alamitos, CA, USA, 284-291.

[14] Lee, Hyowon and Smeaton, Alan F. and O'Connor, Noel E. and Jones, Gareth J.F. and Blighe, Michael and Byrne, Daragh and Doherty, Aiden R. and Gurrin, Cathal. *Constructing a SenseCam visual diary as a media process.* Multimedia Systems Journal, 14 (6). pp. 341-349.