STRUCTURING PERSONAL EXPERIENCES — ANALYZING VIEWS FROM A HEAD-MOUNTED CAMERA

Yuichi NAKAMURA † ‡ *Jun'ya OHDE* † *Yuichi OHTA* †

† Institute of Engineering Mechanics and Systems
University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573 JAPAN
‡ PRESTO, Japan Science and Technology Corporation (JST)

ABSTRACT

This paper introduces a novel method for analyzing video records which contain personal activities captured by a head-mounted camera. This aims to support the user to retrieve the most important or relevant portions from the videos. For this purpose, we use the user's behaviors which appear when he/she pays attention to something. We define two types of those behaviors, one of which is "gaze at something in a short period" and the other is "staying and continuously see something". These behaviors and the focused object can be detected by estimating camera and object motion. We describe, in this paper, the details of the method and experiments in which the method was applied to ordinary events.

1 Introduction

We often need a help for recording or memorizing our activities. Although we can usually remember impressive events, it is hard to recall things in detail, *e.g.* in which order we did something or what was there. We hope devices for augmenting our memory. Fortunately, in the near future, we will certainly get wearable hardware with enough computational power to deal with real-time image processing and large amount of videos. One of the leading works is DyPERS which gives appropriate information to the users according to what the user sees[1]. The system retrieves pre-recorded information when a pre-registered object appears in the user's view.

However, we still need considerable efforts to realize a system that we can record our activities and look up accumulated records.

One of the most important topics is data structuring, summarization, and retrieval from enormous records. Those videos taken as personal experiences can be long and redundant, and the user needs to take great pains in searching for the right portion. This disadvantage may spoil the merit of video records. Kawashima, et al., have developed experimental systems, which pick up important scenes by detecting human faces, and so on [2, 3]. This helps the user to recall the meeting or conversations with other people.

On the other hand, it is still difficult to deal with details on our experiments, for example, recording and retrieving the process of important operations. We are tackling with this problem by our structural analysis and summarization of the video data. We propose a novel method which utilizes the user's head motion for gazing a target. First, we define two types of those behaviors. Next, we propose a method for detecting those behaviors by estimating camera and object motion. Then, structuring videos based on these behaviors effectively reduces the user's efforts to recall or retrieve the information he wants.

2 Attention and Apparent Motion

2.1 Head-mounted Camera

The system needs to capture the scene around the user at anytime he/she wants. The most plausible way is to carry a camera attached to the user. We think the best position of the camera is on head, since the view from the camera can be similar to what the user sees. The user can easily recall what happened by checking the view.

As mentioned above, we need structuring and summarization of the recorded data, since they are usually long and redundant. Our system structurally analyzes the videos by detecting two kinds of scenes in which the user intentionally looks at something. Figure 1[1] shows the brief overview of our idea. By detecting scenes which would be anchors of our memory, the system enables us to browse our activity records. Thus video records can be effective tools for a variety of applications: an augmented personal memory; an instruction manual which contains teacher's view; a tool for sharing experience for the people involving the same task.

[1] Since some figures in this paper are too small to look at details, we prepared full resolution version at our web site.

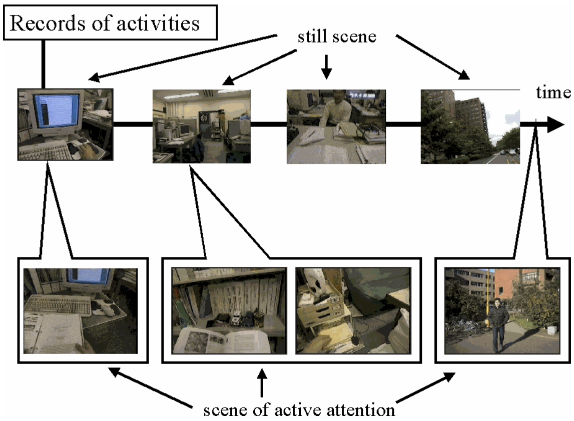

Figure 1: Overview: The upper row shows where the user was and what the user continuously saw; the lower row shows what user gazed.
(http://www.image.esys.tsukuba.ac.jp/∼yuichi/ICME/figure1.jpg)

2.2 Attention and Behaviors

When a person is paying attention to visual objects or events, head movements shown in Figure 2 occur. For these behaviors, we first define two types of attention. One is *active attention*, which means gazing and tracking something in a short period, and the other is *passive attention*, which means staying at the same place and seeing something.

Active Attention: We often gaze at something and visually track it when it attracts our interest. If the target stays still, head motion will be Figure 2(a) or (c). If the person is moving, they will be Figure 2(c) or (d). This type of behavior lasts relatively short time, *e.g.* a few seconds.

Passive Attention: We often look vaguely and continuously at something around ourselves during desk works, conversations, or rests. This type of behavior does not always express a person's attention. However, this kind of scene can be a very good cue to remember where the person was. Head motions tend to be still as shown Figure 2(a), often with small movements such as nodding. The duration of those scenes is usually long, for example, 10 minutes.

We consider both of the above as important keys which effectively represent the video contents. Hereafter, we call the video frames in which those behaviors occur as *scene(s) of attention*.

2.3 Image Features

Typical patterns appear in an image sequence if the above defined behaviors occur. Head motions cause the apparent motion of the background in an image, while the region of a target is likely to stay around the image center. In most case, therefore, we have at most two important image regions which have different apparent motions.

In the case of passive attention, we have to consider Figure 2(a) with small movements or slow movements for looking around. The view does not change much, and the images taken during those periods largely overlap each other. Consequently, we can detect passive attention by detecting sections in which the total of the apparent motion vector stays small.

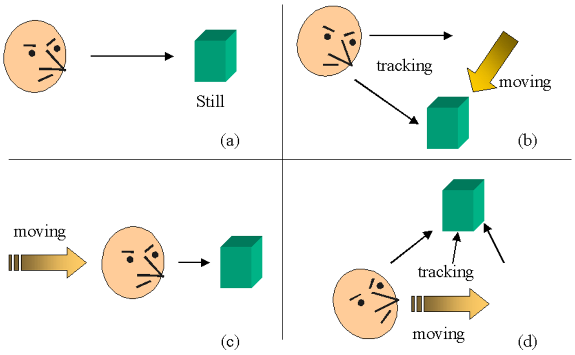

Figure 2: Head motion in paying attention

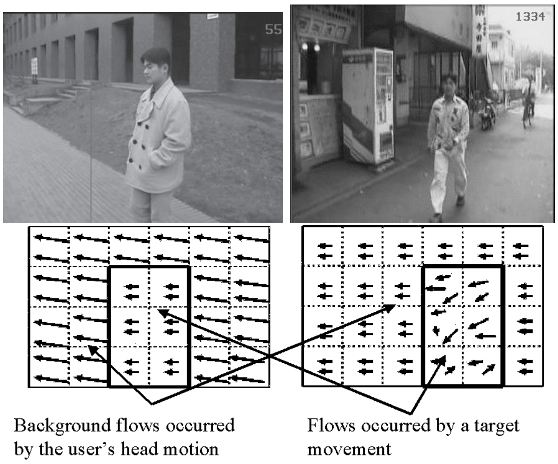
55
1334
Background flows occurred by the user's head motion
Flows occurred by a target movement

Figure 3: Apparent motion vectors on active attention

In the case of active attention, we have to consider two different apparent motions of the background and the target at which the user is gazing. Apparent motion vectors in typical situations are shown in Figure 3. When we track an object of interest, the apparent motion vectors on the object are relatively small compared to those on background as shown in the left column in Figure 3. The region stays still in the view for at least several video frames.

In addition to that, if an object is rotating or deforming, a region with complicated motion vectors appears as shown in the right column of Figure 3. If a region of this type stays in the view, we consider it drew the user's attention.

Therefore, we can detect active attention by segmenting motion vectors into two or more regions. Usually the largest region is the background, and a region which has small motion vectors is the object at which the person gazed.

3 Scene Detection

The flow of our scene detection process is shown in Figure 4.

1. Find the correspondence and motion parameters between two consecutive images, which are apart by one to several frames. We apply a computer vision technique of motion estimation based on central projection.
2. Detect still scenes which correspond to passive attention, by finding portions with small background motion and merging them.
3. Detect a target which is possibly gazed and tracked, by using the correspondence obtained by 1. If a target is detected, label the segment as a scene of active attention.

For the above process 1, we use the central projection model, the apparent motion $u(\mathbf{x})$ of a image point $\mathbf{x}$ can be calculated by using the camera translation $\mathbf{t} = (t_1, t_2, t_3)^T$ and the camera rotation $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)^T$.

$$u(\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\mathbf{A}\mathbf{t} + \mathbf{B}\omega \quad (1)$$

where,

$$\mathbf{A} = \left[\begin{array}{ccc} -f & 0 & x \\ 0 & -f & y \end{array} \right],$$

$$\mathbf{B} = \left[\begin{array}{ccc} (xy)/f & -(f^2+x^2)/f & y \\ -(f^2+x^2)/f & -(xy)/f & -x \end{array} \right]$$

f is the focal length, $Z(\mathbf{x})$ is the depth at the position $\mathbf{x}$ on the image plane.

We denote the intensity $I(\mathbf{x}, T)$ at point $\mathbf{x}$ at time T. If the above camera motion and rotation occurred during $[T - \delta t, T]$, the following relationship ideally holds.

$$I(\mathbf{x}, T) = I(\mathbf{x} - \mathbf{u}, T - \delta t)$$

Thus we can expect to get the motion parameters by minimizing the following error E.

$$E = \sum_{x,y} \{I(\mathbf{x}, T) - I(\mathbf{x} - \mathbf{u}, T - \delta t)\}^2 \quad (2)$$

We use a new method for applying the above calculation to shaky videos from a head-mounted camera, though the above idea is based on Bergen's method [4]. Technical details are skipped because of the space. They will be reported in the near future[5].

4 Experiments

We applied our method to two 12 minute videos recorded during office work. The videos contain the following stories: (i) the user went to his supervisor's room, and had a meeting; (ii) the user came back to his lab and did some assembling for a video projector. We assume that (i) and (ii) are consecutive events in his daily life.

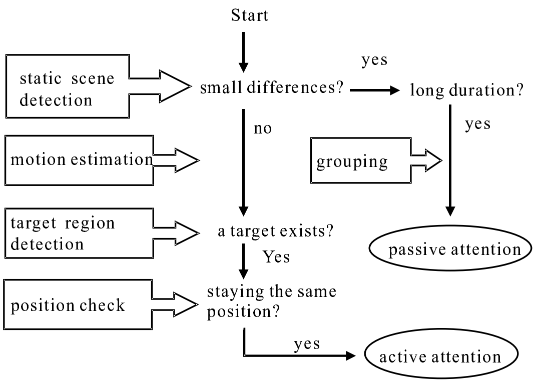

Figure 4: Flow of scene detection

Table 1: Time required for recalling the details of the operations: Person A through C used our system, and person D through F used a VCR.

Person	A	B	C	D	E	F
Time(sec)	220	240	330	420	450	450

Scene Detection Results

The detected scenes are shown in Figure 5. The result shows our motion estimation and scene detection method works well for videos on ordinary activities, though we have not precisely evaluated yet.

In each column, the vertical direction expresses the pseudo time axis. The left column and the right column are consecutive. The leftmost images in each column are detected scenes of passive attention (still scenes) and the images on the right side are scenes of active attention.

For each scene of passive attention, the representative frame is the frame at the middle of the duration, and it is shown with the size proportional to the scene duration. Scenes of active attention are connected to the corresponding scene of passive attention. If it has no corresponding scene, it is directly connected to the vertical line. The rectangle in each scene expresses a candidate of the target to which paid attention.

In the left column, since the user stayed still, only several scenes are detected for (i). On the other hand, for the subtask (ii), especially in the right column, many scenes of active attention are detected. That shows the subtasks require rather complicated procedures.

The detection result includes not a few false positives, in this case, 15 scenes of active attention out of 47 detected scenes. We need further investigation to eliminate false detection, though this is not a serious problem.

For comparison, Figure 6 shows the same number of frames (47 frames) picked up with constant interval. We can see many redundant images, while enough images are not detected for the complicated procedures.

User Study

To evaluate the effectiveness of our scheme, we had a preliminary experiment.

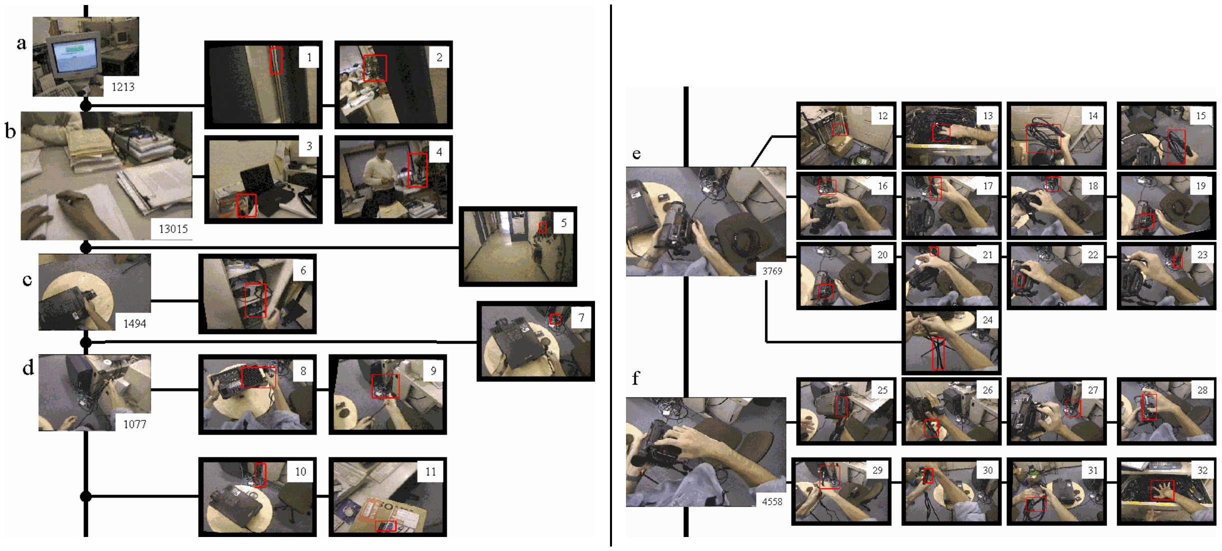

Figure 5: Detection Result : In each column, the vertical direction expresses time passing. The leftmost images in each column are the still scenes and the images on the right side are the scenes of active attention. Each rectangle in the images expresses a candidate of the target to which paid attention. (http://www.image.esys.tsukuba.ac.jp/∼yuichi/ICME/figure5-1.jpg, figure5-3.jpg)

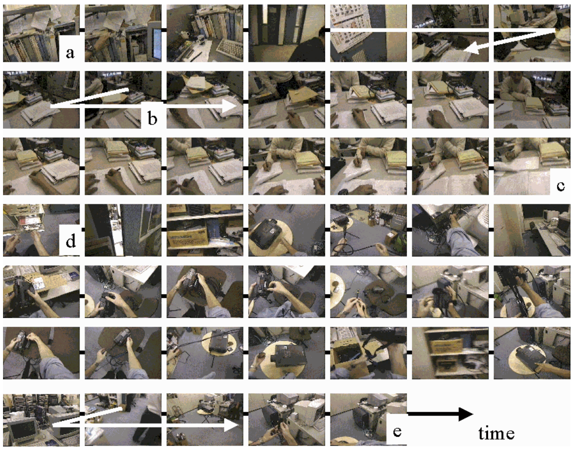

Figure 6: For comparison, the same number of frames (48 images) are picked up with constant interval. (http://[the same as figure5]/figure6.jpg)

- We asked six users to do the same task according to our memo.
- Several days after, we asked them some questions which require to recall the details of the task.
- We asked one group (person A through C) to use our GUI with the above detection result, and the other group (person D through F) to use a VCR.

The result is shown in Table 1. Our system shortened the time needed for answering the questions at the rate of 1.5 – 2. Although this is a naive experiment, this shows the potential of our scheme. We are now continuing to further experiments.

5 Summary

In this paper, we presented the overview of our video structuring scheme. We first showed how the user's attention can be estimated from videos taken by head-mounted cameras. Then, we described the method for detecting scenes of attention by motion estimation between frames. Although our system is at the first step toward archiving our experiments, the preliminary experiments showed the good potential.

One of the most important area for future work is evaluation. Other experiments for verifying our scheme is currently ongoing. One is, for example, comparison between out system outputs and the user's summary made by hand. Those will be reported in the near future.

6 References

[1] Jebara,T., Schiele,B., Oliver,N., Pentland,A., "DyPERS: Dynamic Personal Enhanced Reality System", MIT Media Laboratory, Perceptual Computing Technical Report ♯463

[2] Yoshikawa,T., et al., "Selective Scene Recording System Based on Observation of Behavior" (in Japanese), IEICE, SIG-PRMU95-97, 1995

[3] Iijima, T., et.al., "Human Image Extraction from Video Recordings of Daily Life for Mental Retrace" (in Japanese), IEICE, SIG-PRMU97-196, 1998

[4] Bergen,J., Anandan,P., and Hanna,K.: "Hierarchical model-based motion estimation" Proc. ECCV, pp.237-252, 1997.

[5] Nakamura, Y., Ohde, J., Ohta, Y.: "Personal Activity Records based on Attention" Proc. 15th ICPR (to appear), September, 2000.