Structuring Personal Activity Records based on Attention — Analyzing Videos from 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 by visual information processing. 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 can record our activities and provides an appropriate memory.

One of the most important topics is data structuring, summarization, and retrieval from enormous video records. 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.

For this purpose, we propose a new method for structural analysis and summarization of the video data. First, we define two types of behaviors that occurs when the user pays attention. Next, we describe a method for detecting those behaviors by separating the camera motions and object motions. Then, we describe that structuring videos based on these behaviors effectively reduces the user's efforts to recall or retrieve the information he/she wants.

2 Attention and Apparent Motion

2.1 Views from Head Mounted Camera

The system needs to capture the views around the user at anytime he wants. One of the best locations of a camera is on the user's 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.

To deal with videos taken from a head mounted camera, however, we have to solve the following problems.

- Views can be shaky. The user may sometimes feel pains in watching those videos.
- Videos are usually long and redundant. It requires considerable time to look through them.

For this purpose, we propose a new method for structurally summarizing those videos:

- It picks up scenes that the user tends to remember, and that can be anchors of his/her memory.
- It presents a comprehensible overview and to enable quick access to the video contents by providing the above scenes to the user.

It first applies motion estimation to an image sequence, and detects two kinds of scene in which the user intentionally looks at something. Figure 1 shows the brief overview of our idea. By presenting those scenes, the system enables us to browse our activity records.

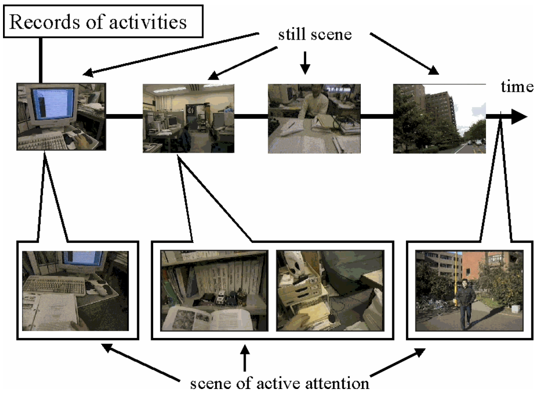

Figure 1. Overview: The upper row shows where the user was and what the user continuously saw; the lower row shows what user gazed.

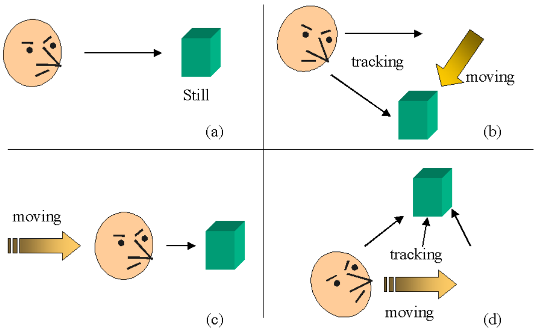

Figure 2. Head motion in paying attention

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.

Active Attention: We often gaze at something and track it when it attracts our interest. If the target stays still, head motion will be Figure 2(a) or (b). 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 in 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*.

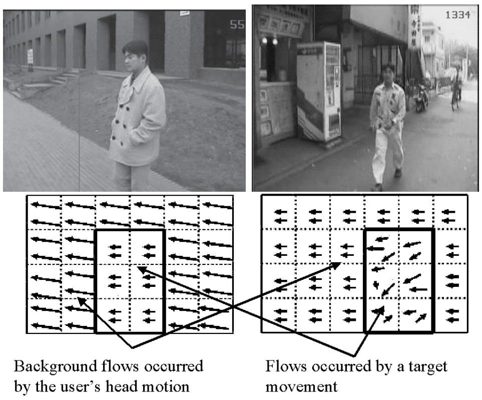
55
1334
Background flows occurred by the user's head motion
Flows occurred by a target movement

Figure 3. Apparent motion vectors on active attention

2.3 Image Features

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.

background motion: Apparent motions caused by the camera motion, *e.g.*, the head motion.

target motion: We assume that the biggest region which has motions different from that of the background draws the user's attention. If the region is staying almost at the same position of our view, it has high possibility of being the target at which the user is gazing.

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 the background motion.

In the case of active attention, 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. 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 the above types stays in the view, we can assume it draws the user's attention. Although this is not always true, we anyway tend to remember the biggest target moving in front of us. Thus the scene or the object could be a good cue to recall our activities.

3 Scene Detection

The flow of our scene detection process is as follows:

1. Find the correspondence and motion parameters between two consecutive images, which are apart by one to several frames. We apply a motion estimation method based on the central projection model.

2. Detect still scenes which correspond to passive attention, by finding portions with small background motion and by 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.

3.1 Motion Detection

First, two images are taken from video data. They are the images apart from each other by one to several frames. We applied a motion estimation method based on the central projection model. Although not a few methods with simpler models have been proposed for video mosaicing (for example, [3]), most of them assume conditions which do not hold in our environment. Indoor objects can be close to the camera, and the depth range in the view widely varies.

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

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

where,

$$\mathbf{A} = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} (xy)/f & -(f^2+x^2)/f & y \\ -(f^2+x^2)/f & -(xy)/f & -x \end{bmatrix}$$

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.

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

To make this calculation possible, we assume the depth is uniform within each small block, *e.g.*, a block of 5x5 pixels.

We need 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]. Note that the objective of this calculation is the correspondence between two images, and the above central projection model is used mainly as the constraints. In this sense, the accuracy of the obtain motion parameters, especially for the depth, is not important if the correspondence is correctly obtained.

The process is as follows:

1. By dyadic down-sampling, for example, 1/2, 1/4, and 1/8, multi-resolution images are created.
2. The initial motion parameters are given to the system. For the most coarse image, the motion parameters obtained for the previous frame are given[1]. For finer images, the parameters obtained by the calculation for more coarse images are given.
3. The error defined in equation 2 is minimized by the Levenberg-Marquardt method.
4. The above operations are applied for all resolutions throughout the video.

The above method does not work well when the camera motion is small, since the method needs to determine the depth. To prevent this, we first check the apparent motion by simply checking the differences of the two images. The two images are blurred, and one of these is subtracted from the other. If the sum of the differences is smaller than the threshold, we think that the camera motion is too small to calculate the above motion parameters.

In this case, we just skip the motion estimation step and the images are gathered into a group with no motions (hereafter abbreviated as *no-motion-group*).

3.2 Still Scene Detection

A still scene is detected by combining the above obtained no-motion-groups. For this purpose, we check the apparent motions of the image center between two still no-motion-groups. If the total motion is smaller than the predetermined threshold, we regard those groups belong to the same scene. Thus we merged them, and label the group as a scene of passive attention.

3.3 Active Attention and Target

Active attention of the user is detected by separating egomotion, *i.e.*, apparent motion by the camera movement, and object motions.

1. The image at the previous frame is transformed so that the viewing position and the camera orientation is equal to those of the current image. By using the motion parameters, $(\boldsymbol{t}, \boldsymbol{\omega}, Z(\mathbf{x})$, transform the image $I_{T-\delta t}$ at the frame $T - \delta t$ to the view $I^T_{T-\delta t}$ at time T.
2. The similarity between the image I_T and $I^T_{T-\delta t}$ are evaluated. We consider a window around each pixel, *e.g.*, a window of 5x5 pixels. Correlation of the two windows from different images at the same position is calculated.
3. The candidate region is detected. The image plane is divided into small blocks B_k, and the number N_k of pixels which correlation value is smaller than the threshold th_d is counted in each block. If N_k is larger than the threshold th_c, the block is labeled as a candidate region. If two or more such candidate blocks are touching each

[1] For the first (initial) frame of a sequence, the initial motion parameters are all set to zero, *i.e.*, no motion.

other, they are merged into one region. Then, the largest region is detected as a candidate at the current frame.

4. The score $P_k(t)$ for each candidate block at time t is determined.

$$P_k(t) = \begin{cases} P_k(t-1) + p & \text{if candidate} \\ P_k(t-1) - q & \text{otherwise} \end{cases}$$

where p is the score obtained from one frame, and q is the forgetting factor. At any time when the score is greater than threshold th_e, we consider the block is the target of attention.

4 Experiments

We applied our method to several videos, any of which is around 10 minutes in length.

Apparent motion is estimated at every four frames. The results are usually satisfactory for our purpose. Since the accuracy of motion estimation throughout the sequences is hard to measure, we evaluated the results by the number serious errors. If the number of the pixels which have no correspondence pixels in the other image exceed 30% of the total number of pixels, we regard the case as a serious error. In our experiments, the rate of those serious errors is less than 1%, though the rate is slightly different among videos.

Here we show one example in detail. The video is 12 minute (22,000 frames) in length, and recorded during cooking in the user's home. The detected scenes are shown in Figure 4. In each column, the vertical direction expresses the pseudo time axis. The leftmost images 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 frames are the frames at the first or the last of the duration, and their sizes are determined by the scene duration. Seven scenes are reasonably grouped except that Scene3 and Scene4 are separated. Motion estimation failed at some frames between them because of the rapid head movement which caused blurred images.

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 red rectangle in each scene expresses a candidate of the target to which paid attention.

The detected scenes are satisfactory for the summarization of the video. Most of the detected targets are the objects at which the user gazed. The detection result includes not a few false positives, in this case, 4 scenes (d, e, g, and o) out of 23 detected scenes. We need further investigation to eliminate false detection, though this is not a serious problem.

5 Conclusion

In this paper, we briefly 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 experiments are simple, our method showed enough potential for realizing augmented memory. This research is still at the beginning stage. We need further investigation such as improvement toward real-time processing, evaluation, combination with other image analyses, and so on.

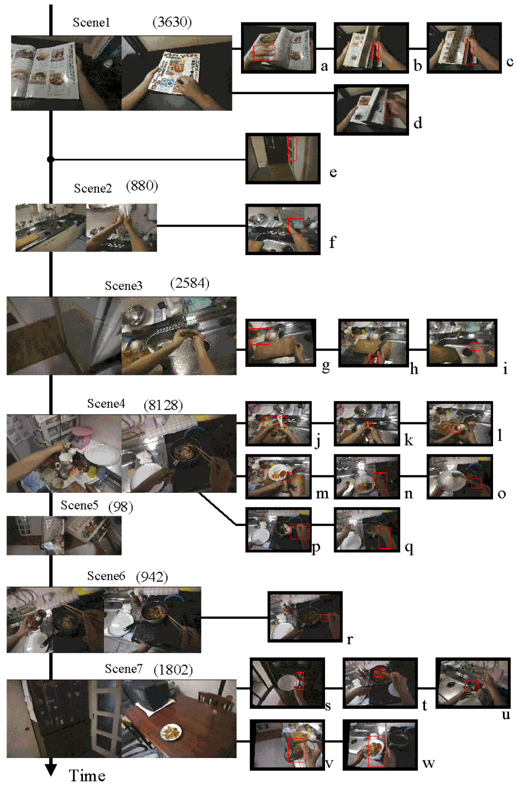

Figure 4. Detection Result : In each column, the vertical direction expresses time passing. The leftmost images are the still scenes and the images on the right side are the scenes of active attention. Each rectangle in the images expresses a candidate of the target to which paid attention.

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] Kawashima, T., et.al., "Situation-based Selective Video-Recording System for Memory Aid", Proc. ICIP, III, 835-838, 1996

[3] Szeliski,R. and Shum,H.: "Creating Full View Panoramic Image Mosaics and Environment Maps", Proc. SIGRAPH, pp.251-258, 1997.

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