5—3

Detecting Scenes of Attention from Personal View Records
— Motion estimation improvements and cooperative use of a surveillance camera

Satoshi KUBOTA, Yuichi NAKAMURA, Yuichi OHTA
IEMS, University of Tsukuba
1-1-1 Tennodai, Tsukuba, JAPAN
{kubota, yuichi}@image.esys.tsukuba.ac.jp

Abstract

This paper introduces a novel method for analyzing video records captured by a head-mounted camera. Compared to our previous method, the new method is improved on two points. One is a new method of two-step motion estimation that adaptively uses either of the 2D affine model or the 3D rigid-body motion with central projection model. The other is a cooperative use of a wide-angled surveillance camera, which delineates the location and the situation where the user acted. Experimental results show its usability for browsing personal records.

1 Introduction

Videos captured by a head-mounted camera (hereafter abbreviated as *HMC*) are good media for recording our activities, and they are useful for recalling or sharing the experience afterwards. Videos taken as personal records, however, can be long and redundant, and a user may need considerable time for finding the information he/she requires. This disadvantage may spoil the merit of video records.

For this purpose, we previously reported that *scenes of attention* can be good indices for summarizing those videos[3]. The view from an HMC contains the central portion of the sight, and the camera's ego-motion represents the user's head motion. By estimating ego-motions and by separating object motions, we can detect typical behaviors of the user's for paying attention to something as shown in Fig. 1.

Figure 2 shows a browser that presents those scenes, and this browser is much more comprehensible than a simple arrangement of images taken at regular intervals (Fig. 3). We also reported that video summaries composed of those scenes showed good match to the summaries that were manually made by selecting important scenes from the videos[4].

This paper introduces two new approaches for the performance improvement and for the extension of the potential applications.

- A new method of two-step motion estimation that adaptively uses either of the 2D affine model or the 3D rigid-body motion with central projection model is proposed. This method is composed of two step motion estimation that reduces false detections of edges or regions with fine textures.
- A view from a wide-angled surveillance camera is cooperatively used for delineating the location where the user acted, and for clearly presenting the situation by the wide-angled view.

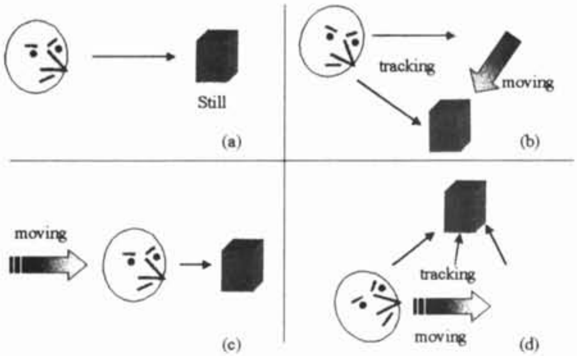


Figure 1: Head motion in paying attention

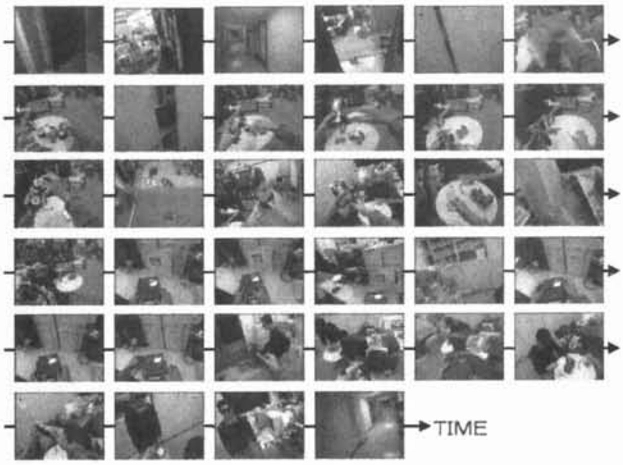


Figure 3: Frames extracted at regular intervals

In the following sections, we will briefly describe the above two approaches and their performance.

2 Detecting Scenes of Attention

We consider the following scenes of attention:

Active Attention: When *a region that moves differently from the background stays still at the same position*, the region is a good clue for the focus of attention. Figure 4 shows the motion differences when a person is gazing at a focus of attention.

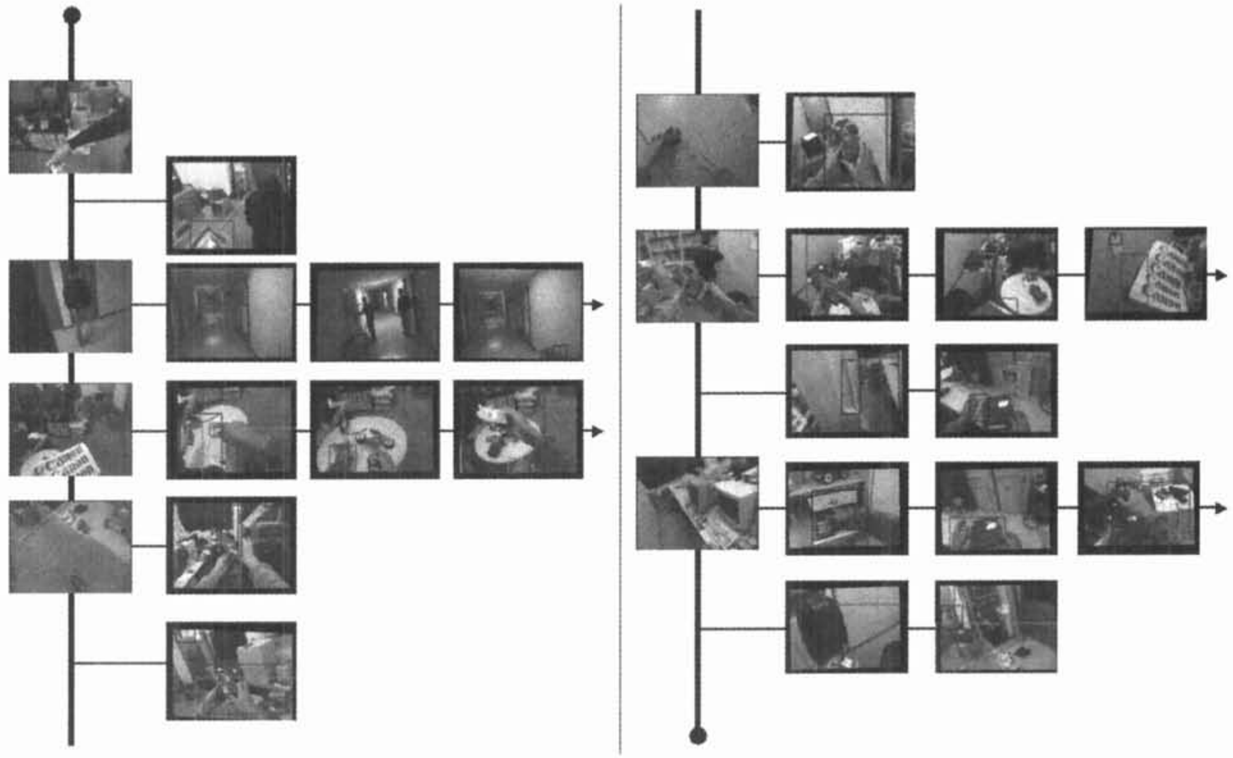


Figure 2: Detection Result : In each column, the vertical direction expresses time passing. The leftmost images in each column are the still scenes and the images on the right side are the scenes of active attention.

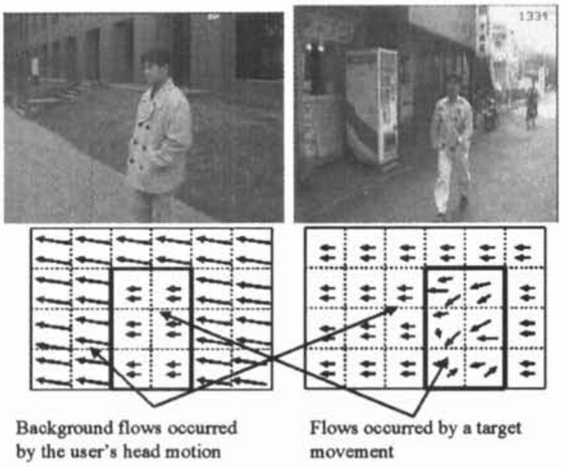

Background flows occurred by the user's head motion
Flows occurred by a target movement

Figure 4: Apparent motion vectors on active attention

Passive Attention: When *the camera motion stays small*, the user is usually seeing something continuously at the same location. This situation is also a good clue for summarization.

For the former one, our method detects the regions by the following steps as shown in Fig. 5:

- A pair of images are taken from the video data. They may be consecutive or several frames apart.
- The apparent motion estimation is applied to the image pair. For this purpose, we previously used the 3D rigid-body motion model with central projection, which is briefly presented in Appendix A.

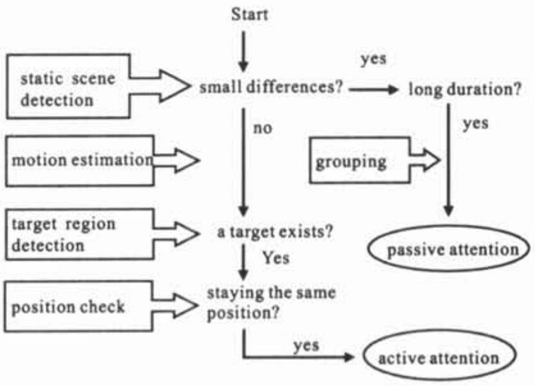


Figure 5: Flow of scene detection

- After the ego-motion is estimated, one image is transformed so that the viewing position and the camera orientation are equal to those of the other image. The difference between the two images are evaluated, and the regions that moved differently from the background is detected.

Although not a few methods with simpler models have been proposed for video mosaicing(for example, [5]), most of them assume conditions which do not hold for our application. Indoor objects can be close to an HMC, and the depth range may widely vary in the view. Thus, our method has advantages for dealing with the shaky videos taken from an HMC.

This method, however, has the following drawbacks:

- This model estimates the depth by assuming disparities. This computation is not appropriate when the camera translation is small.

- Since it is extremely difficult to estimate accurate motions, *e.g.*, less than one pixel error for all over an image, edges and fine textures are often detected as the differences between an image pair.

To cope with these problems, we improved our method on the following points:

- Our new method uses 2D affine model and it switches between the 2D model and the 3D model adaptively. The 2D model is briefly described in Appendix B.
- For finding differences between an image pair, the new method has two steps of motion compensation. By applying the second motion estimation to the candidate region detected at the first motion estimation step, most of false detections such as edges are eliminated.

The overview of the new process flow is shown in Fig. 6. The 2D model is first applied for every pair, and the 3D model is used if the apparent motion is larger than a pre-determined threshold.

For detecting differences, we first gather statistics on the accuracy of the apparent motion estimation for synthesized images. Through the experiment, the average displacement error is around 3 to several pixels, which varies according to the position in a image. Thus, we found that one-step motion estimation is not enough for comparing two images with a considerable ego-motion. The second step of our method is to find matches around the candidate regions detected by the first step. This step is a simple template matching process with the tolerance around the average errors of the first step. Additional explanation of this process is presented in Appendix C.

By this new method we obtain better results as shown in Fig. 7.

- Because of the improved motion estimation and the two-step difference detection, regions as shown in Fig. 7(a) are not detected, and regions as shown in (b) are well detected.
- Since the 2D motion estimation requires less computational time, the total computation time is reduced to almost the half of the previous one. It currently requires around 5 times as long as the video length (on PentiumIII 933MHz), and we can expect that it will be computed online and real-time in the near future.

The result for a ten-minute video is already shown in Fig. 2. Compared with the summaries in Fig. 3, redundant portions are shortened, and important events in each location are well chosen by the proposed method.

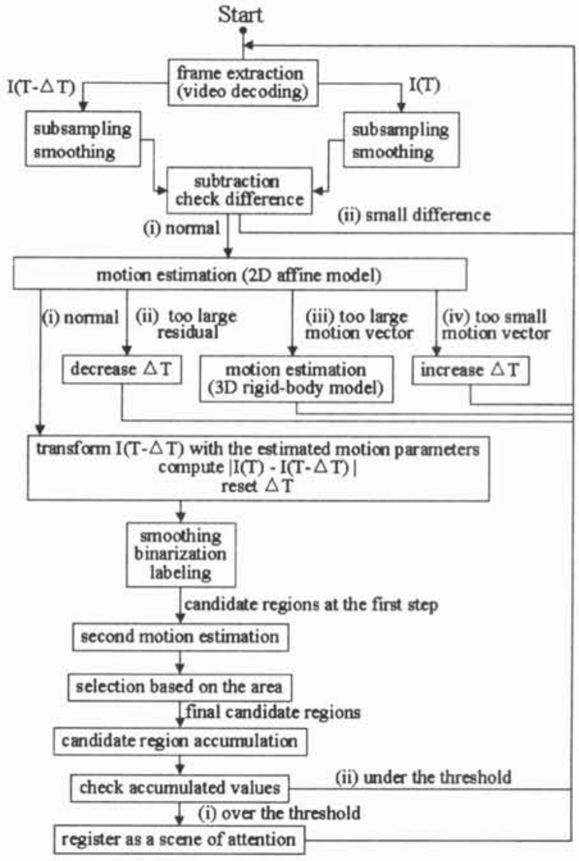
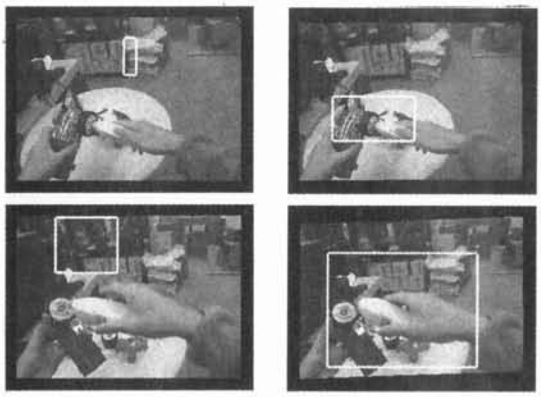

Start
frame extraction (video decoding)
I(T-ΔT)
I(T)
subsampling smoothing
subsampling smoothing
subtraction check difference
(i) normal
(ii) small difference
motion estimation (2D affine model)
(i) normal
(ii) too large residual
(iii) too large motion vector
(iv) too small motion vector
decrease ΔT
motion estimation (3D rigid-body model)
increase ΔT
transform I(T-ΔT) with the estimated motion parameters
compute |I(T) - I(T-ΔT)|
reset ΔT
smoothing
binarization
labeling
candidate regions at the first step
second motion estimation
selection based on the area
final candidate regions
candidate region accumulation
check accumulated values
(ii) under the threshold
(i) over the threshold
register as a scene of attention

Figure 6: Process flow of the new method



(a) previous results (b) improved results

Figure 7: Example of improved results

3 Cooperative Use of a Surveillance Camera

An HMC potentially misses the important events out of its view field, and it is sometimes difficult to recognize the user's location. The cooperative use of a surveillance camera greatly reduces those difficulties. With a wide-angled view as shown in Fig. 8, the system can use the user's location for indexing the videos, and the locus of the tracking is a good clue for recalling or explaining the activities.

By applying the following operations to the videos captured by a surveillance camera, we can obtain the user's location and its locus for the above purpose:

- Detection of a moving region and tracking.
- Detection of a standstill point, which is usually a location a person stops and does something.
- Linking of a standstill point to the scene of attention captured by an HMC.

For detecting and tracking a human, we use the subtraction from the background image that is always up-

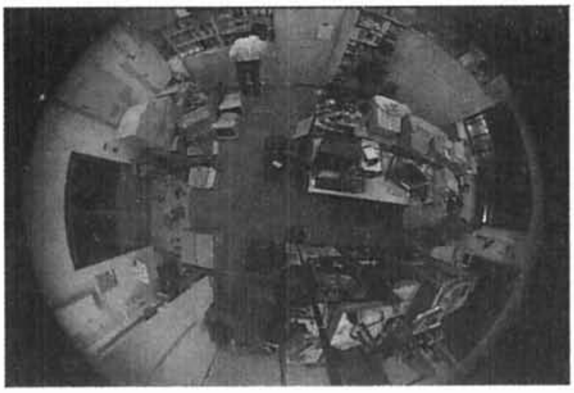


Figure 8: Example image captured by the surveillance camera

dated. Except sleeping persons, a human moves more or less in a short period of time, and this makes easier to detect humans and to update the background image. The details are skipped because of the length limit.

Figure 9 shows an example of the results obtained by the above method. Figure 9(a) shows the loci of the users' movements, and they show which portions in the scene were frequently accessed. For example, the image shows that the cabinet on the upper side of the image were accessed three or four times. Each gray mark on the image represents a scene of attention was detected at its location. Each of the icons on the right hand side is a clickable icon that represents a scene of attention, and it holds movie clips captured through an HMC. By reviewing videos through its icon, we can efficiently browse the activity records. If necessary, by playing-back the video from the surveillance camera, we can check what happened around or behind the user. Thus, the cooperative use of a wide-angled surveillance camera greatly increase the usability for browsing personal records.

4 Summary

This paper introduced the two new approaches that improve the performance of indexing and summarizing videos from a head-mount camera. The results show that we can obtain a comprehensible summarization of a video, and it enables efficient browsing of the contents. For future works, we need to evaluate our method, and we also need to apply the method to real problems.

References

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

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

[3] Y.Nakamura, J. Ohde, Y.Ohta, "Structuring Personal Activity Records based on Attention — Analyzing Videos from Head-mounted Camera", Proc. 15th International Conference on Pattern Recognition, Track4, pp.220–223

[4] J.Ohde, Y.Nakamura, Y.Ohta, "Structuring Personal Activity Records —Evaluation of Scene Detection and Summarization (in Japanese)", Proc. MIRU2000, pp.I-499-504

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

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

Appendix

A 3D Rigid-body Motion Estimation

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 $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)^T$.

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

where

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

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

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.

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

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

$$E = \sum_{x,y} \{I(\mathbf{x}, T) - I(\mathbf{x} - \mathbf{u}, T - \delta t)\}^2 \qquad (3)$$

The process is as follows:

1. By dyadic down-sampling, for example, 1/2, 1/4, and 1/8, multi-resolution images are created.

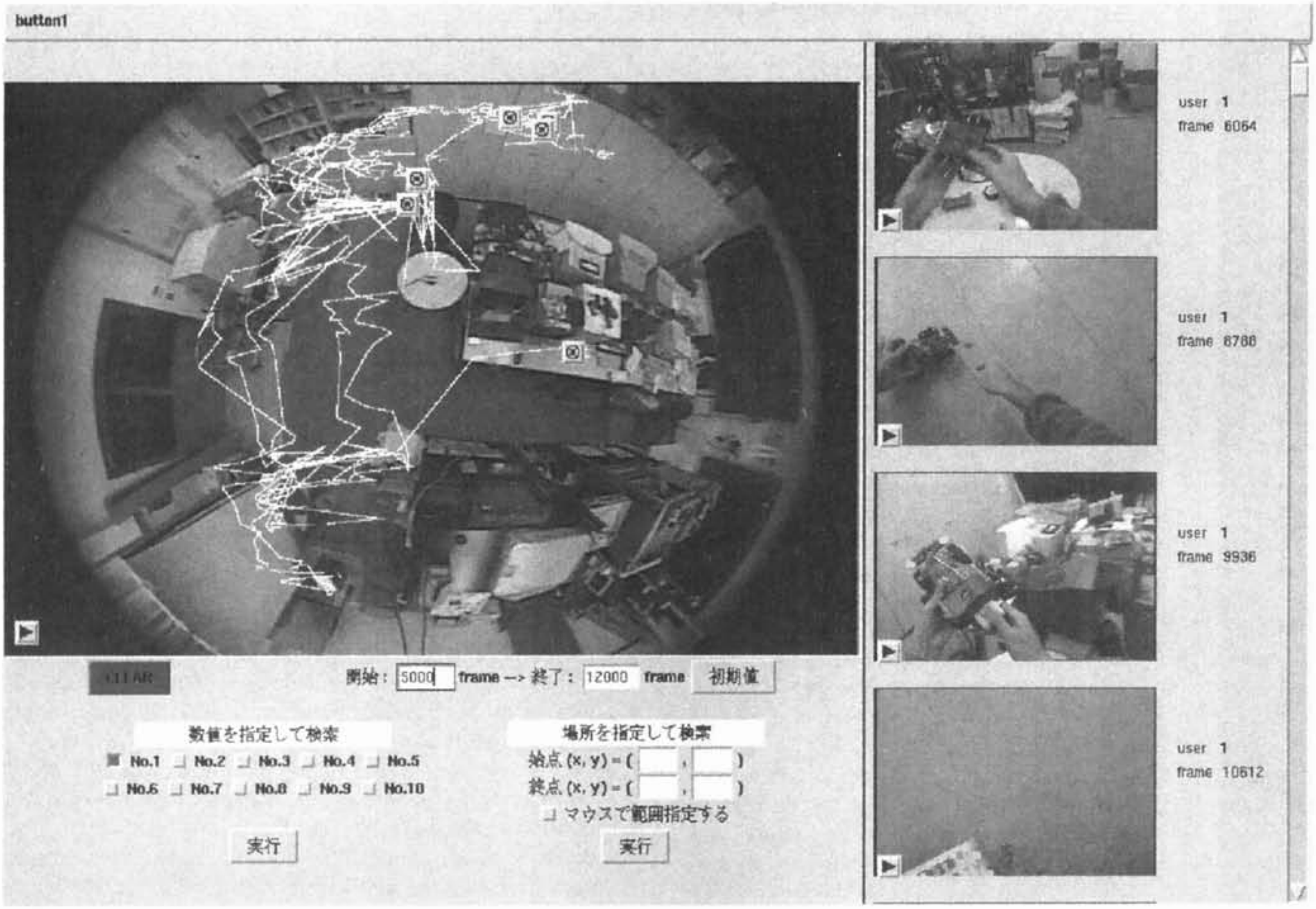
2. The initial motion parameters are given to the system. For the most coarse image, the motion parameters obtained for the previous frame are given[1]. For finer images, the parameters obtained by the calculation for more coarse images are given.

3. The error defined in Equation 3 is minimized by the Levenberg-Marquardt method.

4. The above operations are applied for all resolutions throughout the video.

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

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



開始：5000 frame --> 終了：12000 frame

数値を指定して検索

☒ No.1 □ No.2 □ No.3 □ No.4 □ No.5
□ No.6 □ No.7 □ No.8 □ No.9 □ No.10

場所を指定して検索

始点 (x, y) = (,)
終点 (x, y) = (,)
□ マウスで範囲指定する

user 1
frame 6064

user 1
frame 8788

user 1
frame 9936

user 1
frame 10612

Figure 9: Personal activity browsers with the surveillance camera view

B 2D Affine Motion Estimation

The apparent motion $u(\mathbf{x})$ of a image point $\mathbf{x}$ is calculated by using the linear transformation.

$$u(\mathbf{x}) = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

The actual calculation is similar to the 3D model. The system minimizes error E in Equation 3 by the Levenberg-Marquardt method.

C Detecting Scenes of Active Attention

Active attention of the user is detected by separating ego-motion, *i.e.* apparent motion caused 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 are equal to those of the current image. By using the motion parameters, image $I_{T-\delta T}$ at the frame $T - \delta T$ is transformed to the view $I^T_{T-\delta T}$ at time T.

2. The difference between the image I_T and $I^T_{T-\delta t}$ are evaluated after smoothing is applied to each image.

$$d(x, y) = |I_T(x, y) - I^T_{T-\delta t}(x, y)|$$

   if $d(x, y)$ is larger than a pre-determined threshold, the pixel (x, y) is considered as a candidate pixel.

3. Candidate regions are grouped by labeling the candidate pixels. For each region R_i detected by this process, the second motion estimation is done by simple template matching by using normalized correlation. The range for the template matching is determined by the average displacement error of the first step of the motion estimation, which varies according to the position in a image.

4. If the maximum correlation value is lower than a pre-determined threshold, the region is considered as a final candidate region for the focus of attention.

5. Then, the score for each pixel 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 the threshold (th_e), we consider the pixel is composing the target of attention.