A Novel Wearable System for Capturing User View Images

Hirotake Yamazoe[1,2], Akira Utsumi[1],
Nobuji Tetsutani[1], and Masahiko Yachida[2]

[1] ATR Media Information Science Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun, Kyoto 619-0288, Japan
[2] Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama-cho Toyonaka-shi, Osaka 560-8531, Japan

Abstract. In this paper, we propose a body attached system to capture the experience of a person in sequence as audio/visual information. The proposed system consists of two cameras (one IR (infra-red) camera and one wide-angle color camera) and a microphone. The IR camera image is used for capturing the user's head motions. The wide-angle color camera is used for capturing frontal view images, and an image region approximately corresponding to the users' view is selected according to the estimated human head motions. The selected image and head motion data are stored in a storage device with audio data. This system can overcome the disadvantages of systems using head-mounted cameras in terms of the ease in putting on/taking off the device and its less obtrusive visual impact on third persons. Using the proposed system, we can record audio data, images in the user's view and head gestures (nodding, shaking, etc.) simultaneously. These data contain significant information for recording/analyzing human activities and can be used in wider application domains (such as a digital diary or interaction analysis). Experimental results show the effectiveness of the proposed system.

1 Introduction

The current down-sizing of computers and sensory devices will allow humans to wear these devices in a manner similar to clothes. This concept is known as 'wearable computing' [1, 2]. One major direction of wearable computing research is to smartly assist humans in daily life everywhere a user chooses to go. To achieve these capabilities, the system must recognize contextual information of human activities from sensory information. Another research direction is to record user experiences and reactions to a database for later reference or analysis. An automatic annotation mechanism should be a key issue in such research. In addition, sensing technologies are fundamental parts of both types of systems for capturing human activities and recognizing user attention/intention.

To capture human activities, various modalities are solely or jointly used. These include audio information, visual information, body motions (locations, gestures) and physiological information (e.g., heart rate, perspiration and breathing rate). Consequently, various sensory devices, such as microphones, video

N. Sebe et al. (Eds.): HCI/ECCV 2004, LNCS 3058, pp. 165–175, 2004.

cameras, gyro sensors, GPS (global positioning systems) and skin conductance sensors will be used for detecting this information. Some of these devices are already small and portable enough to wear; however, others still pose difficulties for actual use.

Here, we consider wearable camera systems. In human perception, visual information plays a significant role. Therefore, many systems have employed head-mounted cameras to record images from the user's viewpoint. Using head-mounted cameras, we can easily capture images in a similar view field to the images humans actually perceive. Here, a change of the view field (head motion) reflects a change of the user's attention (except for eye movement). Therefore, this is a very useful device for recognizing user interests.

On the other hand, head-mounted cameras are problematic for users to put on/take off. They cause fatigue to the user's head due to their weight. In addition, their large visual impact makes them difficult to use in daily life. Therefore, it is desirable for a system to have high usability and less visual impact. To achieve this, we propose a body-attached camera system that can record images corresponding to human head motions without requiring the users to wear cameras on their head. By using our system, the user's head motions can be detected in addition to capturing user view images. This is very useful for recognizing not only the orientation of the user's attentions but also the user's head gestures.

In the next section, we briefly summarize related works. Section 3 provides the system overview. Section 4 describes the image processing algorithm. In Section 5, we give experimental results for pose estimation accuracy and show examples of attention region extractions. In Section 6, we address some application domains of the proposed system and delineate future direction. Finally, we conclude this paper in Section 7.

2 Extraction of User's Attention Using Sensory Devices

Here, we briefly summarize related works regarding the recording of human activities using sensory devices.

Personal view images are useful for storing and/or recognizing the user's contextual situation. Therefore, many systems use a wearable camera to capture the user's view image[3, 4]. In these systems, the status of the attention target and surrounding environment are stored as video images in sequence. As mentioned above, visual information constitutes a major part of human experience. This leads researchers to store human experience as video images captured by head-mounted cameras [5, 6].

We proposed a system to estimate human head position and postures from multiple static cameras and head-mounted cameras [7]. In this system, we can extract interaction events among multiple persons from the estimated head motions. This is useful for annotating image data based on user interaction. In this system, however, head pose detection depends on global human positions estimated from multiple static cameras. This means that human attention cannot be extracted without static cameras.

For a different approach from head-mounted cameras, Starner et al. proposed the 'gesture pendant' system [8]. This system utilizes a camera that hangs from the user's neck like a pendant. The camera is used for hand gesture recognition. Healey et al. proposed the 'StartleCam' system [9]. This system can capture the user's frontal view image when the user is startled by using a body-attached camera. In these systems, the user's head motion has not been considered.

Thus, in this paper, we propose a system that does not require users to put cameras on their head by estimating human head motion using an IR (infra-red) camera that looks up. A frontal view camera can be used for extracting the attention region of users from its images in sequence. In the next section, we provide an overview of our system.

3 System Overview

Figure 1 shows the appearance of our system. Our system consists of two parts: a sensing part and a signal processing part.

The sensing part has an IR illuminator, two cameras (one wide-angle CMOS color camera and one IR camera) and a microphone. These parts are usually mounted around the middle of the upper body (as shown in Figure 1). The IR illuminator and IR camera look upwards. The IR illuminator illuminates the user's head (mainly the jowl part) and the IR camera captures the image of the illuminated parts. The wide angle color camera has higher resolution (1280×1024 pixels) than normal video cameras and captures the frontal direction image. A microphone is used for capturing audio data (human voices, environment sounds). According to the property of the CMOS device used in the frontal view camera, we can selectively retrieve a partial image with higher transfer speed. This feature is useful for extracting the user's view image.

The processing part receives signals from the sensing part and estimates head motion using the IR camera image. Then, the image area that is the closest approximately to the human view is selected from the frontal-view camera image based on the head motion estimation results. This flow is described in Figure 2.

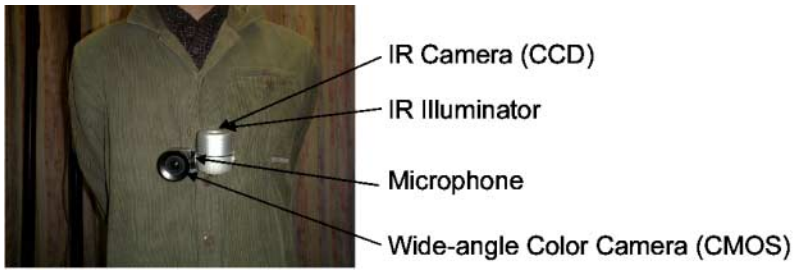

Fig. 1. System Appearance

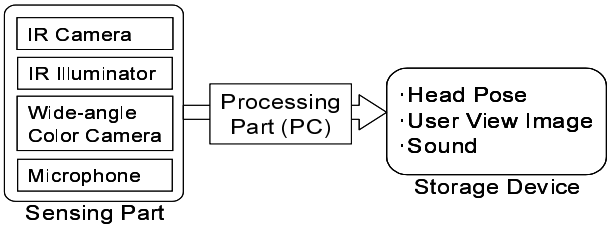

Fig. 2. Process Flow

Using the processing results, we can simultaneously record audio data, image data of the user's view and head motions (gestures).

In the next section, we describe details of the image processing algorithms.

4 Head Pose Estimation & User's View Extraction

4.1 Human Head Model

Before describing our algorithm, we will define the coordinates used in this paper. First we define body-centered coordinates X, Y and Z as shown in Figure 3. Since the direction of the user's view does not depend on rotation about the Z axis, we consider the rotations of two axes, X and Y, only. Here α and β denote the rotations about the X axis and Y axis, respectively. We assume the center of the head rotations is located near the front end of the user's neck.

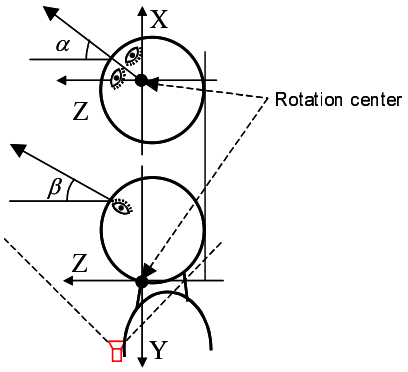

Fig. 3. Head Model

4.2 Human Region Extraction

In our system, we extract the human region (head and torso) by using the IR camera and IR illuminator. The IR camera can shoot only the part illuminated by the IR illuminator. We control the IR illuminator to illuminate only the human region (in a close range) and extract human regions by selecting larger value pixels in the IR camera images (Figure 4). The combination of IR camera and IR illuminator has been used for hand gesture recognition by Numazaki et al.[10].

After the human region extraction, we determine a boundary line between head and torso regions using human region histograms. Here, we determine the center of head rotation to be the mid-point of the boundary line between the head and torso regions. Then, the nose top position in the head region is selected by maximizing the distance between a point and the center of the head rotation (Figure 5). Figure 6 shows some examples of the described process. Here, $\times$ denote estimated positions of nose top points and lines show the estimated head-torso boundaries.

4.3 Head Pose Estimation

In this section, we describe our algorithm for estimating head poses from IR images. Figure 7 describes the relationships among image features. Here, we represent the nose top point and the head rotation center in the IR images as $\boldsymbol{P}_n(=[u_n, v_n])$ and $\boldsymbol{P}_c(=[u_c, v_c])$, respectively. The u axis corresponds to the estimated boundary line between head and torso (Figure 7). v is the perpendicular axis to the u axis.

Fig. 4. Human Region Extraction

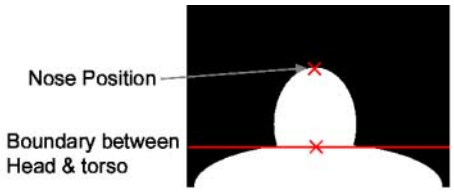

Fig. 5. Feature Detection

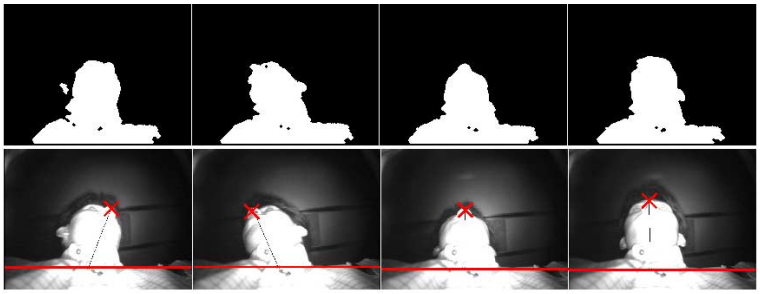

Fig. 6. Human Region Extraction (top: Extracted human regions, bottom: IR images and extraction results)

We assume the head rotation center is located at the front end of the user's neck. Then, head rotation angles α and β can be calculated as follows.

$$\alpha = \tan^{-1}\left(\frac{u_n}{v_n}\right), \tag{1}$$

$$\beta = \cos^{-1}\left(\frac{\sqrt{u_n^2 + v_n^2}}{k}\right) - \cos^{-1}\left(\frac{\sqrt{(u_n^{(0)})^2 + (v_n^{(0)})^2}}{k}\right), \tag{2}$$

where k is a constant value determined by head size and intrinsic parameters of the IR camera, and $\boldsymbol{P}_n^{(0)}$ is the value of $\boldsymbol{P}_n$ when $\alpha = 0$ and $\beta = 0$.

In practice, the relative pose between the camera system and a human body can be changed due to self-rotation (fluctuation) of the camera system. This makes the estimation values of the real pose α and β shifted. We estimate the offset values ($\Delta\alpha$ and $\Delta\beta$) and compensate the self-rotation. $\Delta\alpha$ is determined as the tilt angle of the boundary line (Figure 7). $\Delta\beta$ can be calculated as follows.

$$\Delta\beta = \tan^{-1}\left(\frac{v_c^{(0)} - v_c}{f}\right), \tag{3}$$

where $\boldsymbol{P}_c^{(0)}$ is the value of $\boldsymbol{P}_c$ when $\Delta\alpha = 0$ and $\Delta\beta = 0$. f is the focal length of the IR camera.

4.4 Extraction of User's View

Using the results of the head pose estimation, we can extract an image region corresponding to the user's view from frontal camera images.

Using estimated angles α and β, we can determine the 2D point in a frontal camera image that corresponds to the center of the user's view as follows.

$$x_\alpha = f \cdot \tan(\alpha + \Delta\alpha), \quad y_\beta = f \cdot \tan(\beta + \Delta\beta). \tag{4}$$

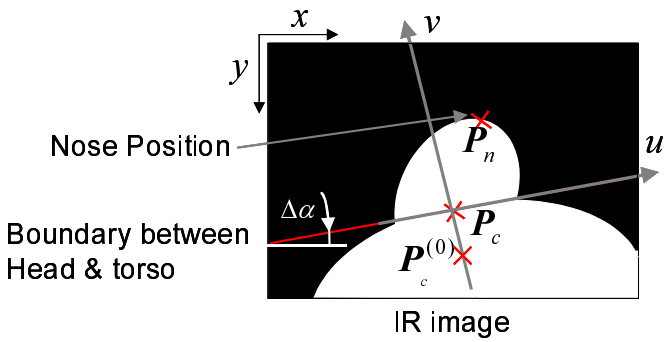

Fig. 7. Head Pose Estimation

Strictly speaking, as the frontal camera is mounted at a different position from human eyes, there is a gap between the extracted view and a real user view. This gap becomes larger when the observed objects are located closer to human eyes. It is hard to solve this problem completely; however, we can reduce the impact of the problem by introducing a non-linear mapping between the head rotation and view selection. This is left for future work.

5 Experiments

To confirm the effectiveness of the proposed system, we performed the following experiments.

First, we evaluated the accuracy of the nose position extraction. We performed nose position extraction with our system and compared the results with the manually selected nose positions. Figure 8 shows the results (Here, solid lines denote the positions estimated with our system and dashed lines denote the ground truth (by manual selection)). As can be seen, nose top points are properly located with our system. The estimation errors here are less than 10 pixels.

Next, we performed head pose estimation using our system. Figure 9 shows the results. Here, solid lines correspond to the trajectory of the estimated gaze points. In the sequence, a subject who wore our system mainly looked at three objects in the scene (a clock, a calendar and a chair).

Based on the duration of a stational head pose, we can extract the image regions that the user is interested in. Figure 10 shows the attention regions automatically extracted from the frontal camera images. In this example, the regions related to the three objects above are properly selected.

As can be seen, our system can successfully detect the changes of user attentions.

Figure 11 shows the user activities as a time sequence obtained in this experiment (from the above captured user's view images, head poses and audio

information). By using this information, we can extract the user's head gestures (nodding, shaking, etc.), and the moment when the user is talking.

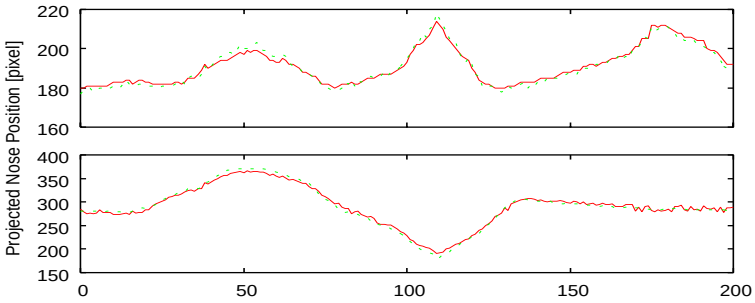

Fig. 8. Nose Position Estimation Accuracy

Fig. 9. Estimated Gaze Trajectory

Fig. 10. Extracted User Attention Area

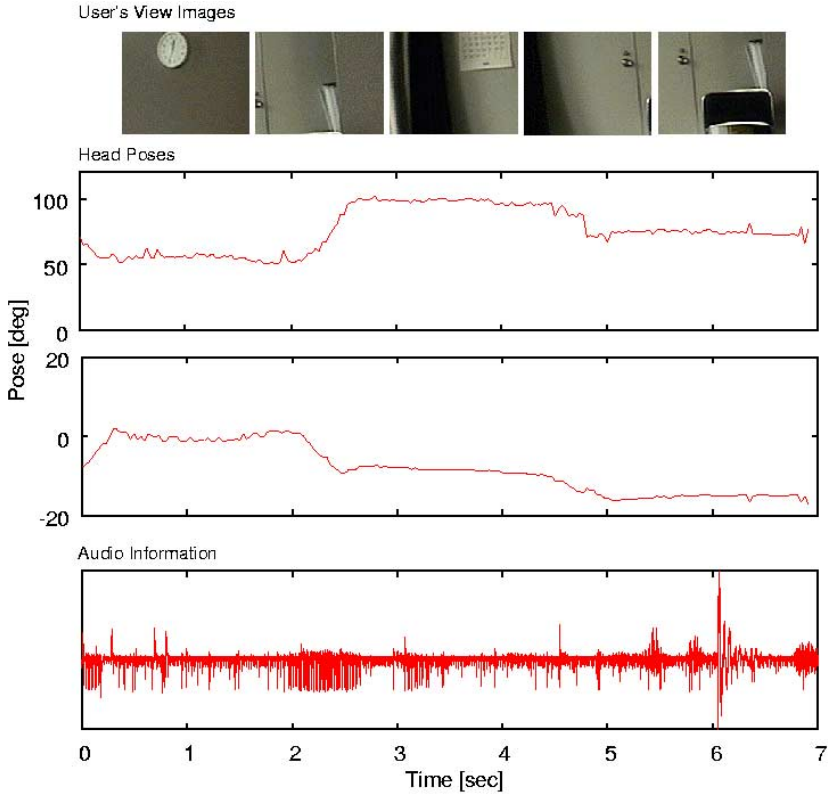

Fig. 11. Captured User's Activities

6 Discussion and Future Direction

In our system, as the user wears the cameras on his/her chest, the user suffers less fatigue and the visual impact to third persons is less than with head-mounted cameras. The system is easy to put on and take off, and is suitable for capturing human daily activities.

Stored data contains useful information for summarizing user experience. Using the head motion data and temporal changes of the image itself, we can extract the user's attention. In experiments (Section 5), no object recognition is considered and just the time length is used to extract the user's attention. Therefore, the system fails to extract attention information if the user is moving. To overcome this problem, we plan to evaluate the similarity of images. Some color histogram-based matching will be considered for this purpose.

Using the proposed system, image sequences approximately in the user's view and head motion information can be stored with audio data. Head gestures can be recognized using head motion information. These data are analyzed for

detecting predefined interactive events, extracting user interests, etc. In addition, head gestures (nodding, shaking, etc.) can be recognized from head motion data. This property of the system is useful for interaction analysis [7].

The prototype system is still insufficient in terms of its size and weight. However, our system has a simple structure and further miniaturization is not difficult. Future versions will become much more compact. Our system lacks global positioning capability though some relative user motions can be estimated from frontal camera images. Position information is helpful for enhancing the quality of video annotation and interaction analysis. We plan to integrate location sensors such as gyro sensors and GPS with our system for that purpose.

7 Conclusion

We proposed a wearable system to capture audio and visual information corresponding to user experience. Using our system, audio information, head motions and images in the user's view are easily recorded in sequence. The system is easy to put on and take off and has less visual impact to third persons. These properties are desirable for capturing human daily activities. We confirmed the effectiveness of our system through experiments.

Future work includes improvement of the head-motion tracking algorithm, head gesture recognition and miniaturization of the entire system. We will also address the analysis of human activities based on user location and content of captured images.

This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

[1] Lamming, M., Flynn, M.: Forget-me-not intimate computing in support of human memory. Technical Report EPC-1994-103, RXRC Cambridge Laboratory (1994) 165

[2] Mann, S.: Wearable computing: A first step toward personal imaging. Computer **30** (1999) 25–32 165

[3] Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computers. In: Proc. of Intl. Symp. on Wearable Computers. (1998) 50–57 166

[4] Clarkson, B., Mase, K., Pentland, A.: Recognizing user context via wearable sensors. In: Proc. of Intl. Symp. on Wearable Computers. (2000) 69–75 166

[5] Sumi, Y., Matsuguchi, T., Ito, S., Fels, S., Mase, K.: Collaborative capturing of interactions by multiple sensors. In: Ubicomp 2003. (2003) 193–194 166

[6] Aizawa, K., Shiina, M., Ishijima, K.: Can we handle life-ling video? In: Int. Conf. Media Futures. (2001) 239–242 166

[7] Yamazoe, H., Utsumi, A., Tetsutani, N., Yachida, M.: Vision-based human motion tracking using head-mounted cameras and fixed cameras for interaction analysis. In: Proc. of Asian Conference on Computer Vision 2004. (2004) 682–687 166, 174

[8] Starner, T., Auxier, J., Ashbrook, D., Gandy, M.: Gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In: Proc. of Intl. Symp. on Wearable Computers. (2000) 87–94 167

[9] Healey, J., Picard, R. W.: Startlecam: A cybernetic wearable camera. In: Proc. of Intl. Symp. on Wearable Computers. (1998) 42–49 167

[10] Numazaki, S., Morishita, A., Umeki, N., Ishikawa, M., Doi, M.: A kinetic and 3d image input device. In: Proc of CHI'98. (1998) 237–238 169