Face-orientation Detection and Monitoring for Networked Interaction Therapy

Akira Utsumi, Shinjiro Kawato, Kenji Susami, Noriaki Kuwahara, and Kazuhiro Kuwabara
ATR Intelligent Robotics and Communication Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan

Abstract

Detecting users' attention is a key issue in our "networked interaction therapy system," which effectively attracts the attention of memory-impaired people and lightens the burden of helpers or family members. In this paper, we discuss our system for estimating users' attention with vision-based techniques. In the system, users' motions, including face positions and orientations, are robustly detected with sequential images and a scale adaptive algorithm. Users' attention level is estimated based on facial orientation and is used for controlling the contents of services of our "networked interaction therapy system." Finally, we show some preliminary experiments on attention estimation by using our system.

1 Introduction

For a computer system to efficiently support human activities, it has to recognize users' behaviors and understand their intentions. Therefore, the ability to recognize human behavior by using sensors embedded in living environments is becoming an increasingly important research endeavor.

In a human-computer interaction task, for instance, attracting and keeping the motivation of users becomes significant for extracting positive reactions. To achieve this, the system has to estimate the individual's concentration level and control the style and amount of displayed information. Since reactions vary from user to user, such control should occur dynamically.

The same situation prevails in our "networked interaction therapy system," which effectively attracts the attention of memory-impaird people and lightens the burden of helpers or family members [1]. "Networked interaction therapy" requires that the system provide remote communication with helpers and family members as well as video contents and other services. To attract the attention of users for long periods of time, the system has to detect users' behaviors and control the order and timing of provided services based on estimates of their concentration levels.

Various sensory devices can be used to detect users' behaviors. Vision-based detection of human behavior fits our needs since it does not require any special attachments [2, 3, 4, 5]. The system can remotely detect human motions, and users do not have to be aware of the sensor systems.

In this paper, we describe a method to estimate users' attention based on facial orientation detected with a vision-based face-tracking system. In the proposed method, if the user's attention leaves the target media, the system changes the contents of the displayed information or prompts the user in such a way to lure him or her back to the media. We believe that such attention estimation enables users to receive services in a manner appropriate to their interests or preferences.

In the next section, we briefly introduce our "networked interaction therapy system." Section 3 summarizes the framework of our attention estimation. Section 4 describes the process of facial detection and tracking, and section 5 shows the experimental results. In section 7, we give our conclusions.

2 Networked Interaction Therapy

Memory is frequently impaired in people with such acquired brain damage problems as encephalitis, head trauma, subarachnoid haemorrhage, dementia, cerebral vascular accidents, etc. Such people have difficulty leading normal lives due to memory impairment or higher brain dysfunction; consequently, the requirement for constant care and attention creates a heavy burden on their family members. Networked interaction therapy, a term that denotes our method for relieving the stress suffered by memory-impaired people and their family members, creates easy access to the services of networked support groups.

The main goal of networked interaction therapy is the creation of barrier-free access to the Internet to assist interactive communication between memory-impaired people, their family members, and volunteers. To provide such access, the system needs to recognize a memory-impaired person. Then, the system connects his/her terminal to another terminal, starting the process of communication. A second important issue is to support the daily activities of memory-impaired people and reduce the burden on their family. For this purpose, we aim to give networked interaction therapy the capability to automatically detect the intentions of a memory-impaired person. This provides the necessary information for guiding the individual to a more comfortable condition on behalf of his or her family members before the occurrence of such behavioral problems as wandering at night, incontinence, fits of temper, and so on. These

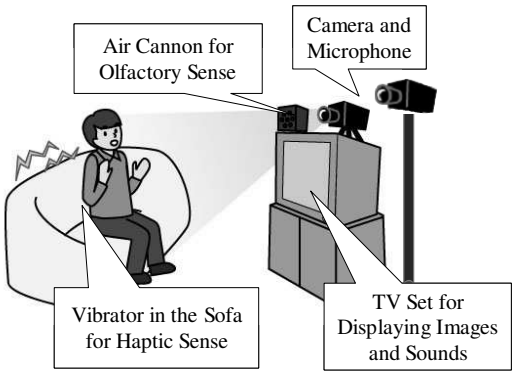

Figure 1: Networked Interaction Therapy System

anxieties are often caused by a lack of information. The system needs to detect situations where communicating helps the individual to overcome the difficulties of daily activities. Figure 1 illustrates an example of a terminal used for networked interaction therapy. Currently, we plan to provide our service by using a large screen TV and a TV-top box that controls Internet communication, cameras, a microphone, and the sensory media. We will use an air cannon and a vibrator in the sofa to provide olfactory and haptic stimulation, respectively [6].

3 Attention Estimation using Facial Orientation

The human gaze is a strong cue for estimating the target of a user's visual attention. Since visual cues play an important communicative role in our "networked interaction therapy system," we focus on facial orientation to estimate human attention. Strictly speaking, facial orientation is different from human gaze due to the lack of eye information. However, in most cases, loss of visual attention is accompanied by head movement. Therefore, we still consider facial orientation information useful for attention estimation.

Vison-based face tracking is a popular research field, and many researchers have investigated and developed face tracking systems. Many systems have employed skin color information [7, 8] and the eye and mouth regions [9] to locate human faces. However, color information is not robust against illumination changes and the appearance of eye and mouth regions can change according to the facial expression. We employed only intensity values and focus on "between-eyes" patterns for detecting face candidates to avoid these problems. After the detection process, we track face orientation and extract head motions. Some researchers have investigated head gesture recognition [10]. However, those systems still have limitations in their processing speed and motion resolution.

In this paper, we employed a fast face tracking algorithm to detect head orientations in sequence. Using the extracted information, we can estimate human behavior and control the displayed contents of our interaction therapy system. In our system, we take a two-step approach: fast extraction of a small number of candidates with light calculation duty and accurate confirmation of the candidates. In the confirmation process, the eyes are located, and the position of the tip of the nose is sought based on the eye locations. For face-orientation estimation, we use the coordinates of the eyes and the nose. In the face-tracking process, a "between-the-eyes" pattern is tracked, and this pattern's template is updated frame by frame.

In the next section, we give a detailed description of the face-tracking process.

4. Face-tracking Technology

4.1. Fast face candidate extraction

For face candidate extraction, a rectangle is scanned on the input image and segmented into six parts as shown in Figure 2. We denote the average pixel value within a segment S_i as $\bar{S}_i$. Then, when one eye and an eye brow are within S_1 and the other pair are within S_3, we can expect

$$\bar{S}_1 < \bar{S}_2 \quad and \quad \bar{S}_1 < \bar{S}_4, \tag{1}$$

$$\bar{S}_3 < \bar{S}_2 \quad and \quad \bar{S}_3 < \bar{S}_6. \tag{2}$$

A point where (1) and (2) are satisfied is a face candidate. We call this a Six-Segmented-Rectangular filter (SSR filter).

For a fast calculation of the SSR filter, we use the integral image [11]. For an image $f(x, y)$, the integral image is defined as follows.

$$ii(x, y) = \sum_{x' \le x,\ y' \le y} f(x', y') \tag{3}$$

The integral image can be computed in one pass over the original image by

$$s(x, y) = s(x - 1, y) + f(x, y), \tag{4}$$

$$ii(x, y) = ii(x, y - 1) + s(x, y), \tag{5}$$

where $s(x, y)$ is the cumulative row sum, $s(-1, y) = 0$, and $ii(x. - 1) = 0$. Using the integral image, the sum of pixels within any rectangle D defined by $(x1, y1)$, $(x2, y1)$, $(x1, y2)$,

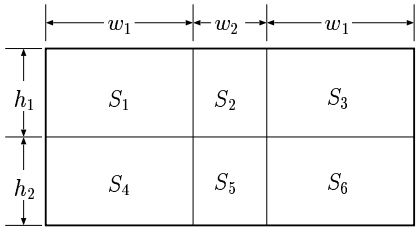

Figure 2: Six-Segmented Rectangular (SSR) Filter. Average pixel values in each segment are computed and compared to find whether they satisfy certain conditions.

and $(x2, y2)$ can be computed from only four corner values of the integral image as follows.

$$\sum_{(x,y)\in D} f(x,y) = ii(x2,y2) + ii(x1,y1) - ii(x1,y2) - ii(x2,y1). \quad (6)$$

Therefore, the computing time of an SSR filter does not depend on the filter size.

Actually, the sizes of SSR filters that we apply to the input images of 320×240 are 120×72, 80×48, 60×36, 40×24, and 30×18 because the facial scale in the image is not known beforehand. This series corresponds to the scaling of the scale-down images used in the tracking process mentioned below. However, this correspondence is not necessary because tracking and detection are independent processes. Nonetheless, it is convenient to monitor the process in down-scaled images. The six segment ratios of the SSR filter here are $w_1 = 4$, $w_2 = 2$, and $h_1 = h_2 = 3$, where w_1, w_2, h_1, h_2 are denoted in Figure 2.

In the detection process, larger SSR filters are applied before smaller ones. If a face is detected, a region of the face is masked so that smaller SSR filters are not applied in this region.

Eventually, points satisfying inequalities (1) and (2) emerge as clusters. Therefore, we select one point at the center of the bounding box of each cluster as a face candidate.

Since we don't use colored information except the green component of an RGB image as a gray-scale image, the algorithm is not affected by the color temperature of the lighting.

We use a motion cue to avoid false face candidates in the still background. Even if an SSR filter indicates a face candidate, when the number of pixels, whose value changes significantly from the previous frame, is less than a threshold (typically a few percent) in the SSR filter's applied area, the point is not identified as a face candidate. The same usage of motion cues appears in [12].

4.2. Confirmation by SVM

A Support Vector Machine (SVM) is applied to determine whether a candidate is a face. To minimize the effects of hair styles, beards, and other facial hair, the forehead and mouth regions are excluded from training patterns for SVM. Figure 3 shows a typical face pattern for SVM training. The pattern size is 35×21. Scale and orientation are normalized based on eye locations. The distance between the eyes is 23 pixels, and they are aligned on the 8th row. Histogram equalization is applied to the gray level.

Before feeding a candidate to the SVM, its pattern should be normalized in the same way as the training patterns. Two local minimum (i.e. dark) points are extracted from the (S_1+S_4) and (S_3+S_6) areas of the SSR filter as left and right eye candidates. Therefore, for one face candidate, a maximum of four patterns are tested by SVM. We extract two eye candidates because the darkness of eyebrows is sometimes similar to the eyes. In scaling and rotating images, we adopt the nearest pixel rule, which takes the value of the pixel nearest the calculated coordinate. The quality of the resulting image may not be perfect, but this saves processing time.

Figure 3: Typical face pattern for SVM training.

4.3. Tracking eyes

In the face-confirmation process, the eyes are located simultaneously. Once they are located, we track them. However, blinking causes drastic and rapid changes in eye patterns that even updating templates cannot follow.

To cope with this problem, we track a "between-the-eyes" template [13] instead of the eyes themselves. Its pattern is fairly stable despite changes in facial expression. It has a relatively bright part at the bridge of the nose and relatively dark parts on both sides of the eyes like wedges. This feature can be accurately located by template matching. After the detection of "between-the-eyes," the eyes are sought in very small areas because their relative positions to "between-the-eyes" are known from the previous frame. However, the success of tracking is still confirmed with SVM.

Even a "between-the-eyes" template must be updated frame-by-frame. We update this template based on the current eye positions.

To track various scales of faces with a fixed template size, we use a series of scale-down images. To save calculation time in constructing such images, we use subsampled images. A series of subsampling rates can be 2/3, 1/2, 1/3, 1/4, and 1/6. These make up an approximate $1/\sqrt{2}$ ratio series. Since, the scale-down images consist of the original pixel values, no additional calculation is required to make them. In the tracking process, an appropriate scale is selected based on the distance between the eyes.

4.4. Tip of the nose detection and tracking

Once the two eyes are located, it is easy to detect the tip of the nose because it has convex features as stated in [14] and is slightly specular. Therefore, a highlight point exists on the nose's tip. Although the precise location of the highlight depends on facial orientation and lighting direction, it invariably exists on the nose's tip.

Figure 4 shows the nose-tip search area relative to the location of the eyes. The brightest point in this area is a candidate for the tip of the nose. If the distance from this point to the eyes are identical, we assume it is the tip and start tracking.

For such tracking, we also use updating template matching. A small rectangular pattern centered on the tip is saved as a

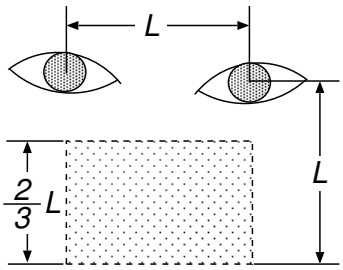

Figure 4: Nose tip search area relative to eyes.

template for the next frame. In the current frame, the best match with the template is searched around the tip position in the previous frame. Then the tip is reregistered to the brightest point in a very small region around the matching point, updating the nose-tip template. If it leaves the region shown in Figure 4, we assume that the tip is lost and resume detection.

4.5. Face orientation estimation

For an accurate estimation of facial orientation, we need a stereo system to measure the 3-D positions of a face's feature points. However, here we adopt a single camera approach for simplicity to roughly determine the direction in which the the subject is looking.

In general, when we take a frontal face image, the distances from the tip of the nose to the each of the two eyes are nearly equal. Therefore, a vertical line passing the tip of the nose that is perpendicular to a horizontal line connecting the eyes (base line) crosses the base line at the middle of the distance between the eyes. When the face turns right or left, this crossing point moves right or left in correspondence to the facial orientation. Therefore, we measure the facial orientation by using a ratio of the distance from the crossing point to the eyes. (Figure 5)

We cannot expect a highly accurate measurement because face profiles differ from person to person. However, a single camera approach has the advantage of long shot images with appropriate zooming capabilities.

When the coordinates of the eyes and the tip of the nose are $(x_1, y_1), (x_2, y_2), and (x_3, y_3)$, expressing

$$\begin{aligned} x_2 - x_1 &= x_{21} \\ &and \\ y_2 - y_1 &= y_{21}, \end{aligned} \tag{7}$$

the x coordinate of the crossing point is

$$x_t = \frac{y_{21}(y_{21}x_1 - x_{21}y_1) + x_{21}(x_{21}x_3 + y_{21}y_3)}{(x_{21}^2 + y_{21}^2)}. \tag{8}$$

Take

$$\begin{aligned} p &= x_t - x_1 \\ q &= x_2 - x_t \end{aligned}$$

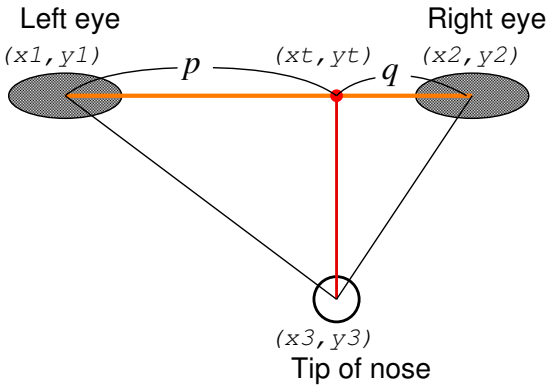

Figure 5: Estimation of facial orientation

and

$$r = \frac{q}{p+q} - 0.5, \tag{9}$$

here, r is the amount of the facial rotation, and the sign of r means the direction of facial orientation (right or left).

Figure 6 shows an example of facial orientation. Graphics are overlayed on the facial image, indicating detected eyes and the tip of the nose. The detected facial orientation is indicated by the heavy line in the narrow region along the top of the image.

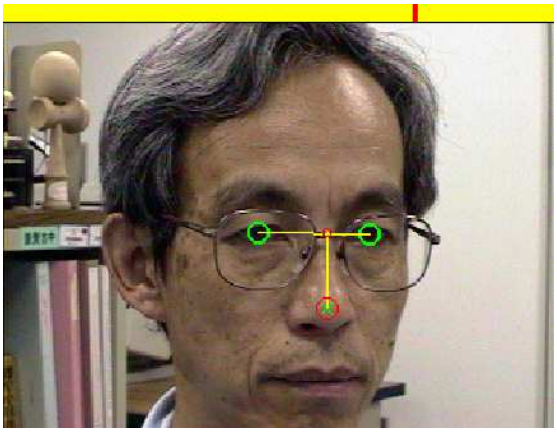

Figure 6: Example of facial orientation estimation

5 Experiments

We conducted the following experiments to confirm the feasibility of our proposed method.

First, we applied our facial orientation detection to a sequence of head rotation images. Figure 7 shows the result of face-orientation detection. Here, a user rotated his head hori-

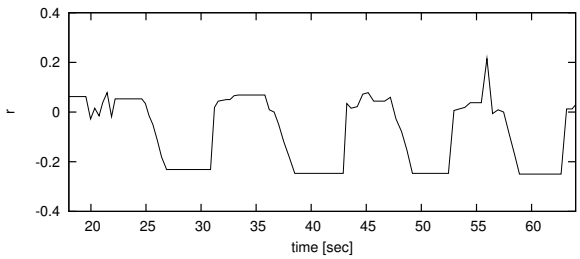

Figure 7: Orientation Detection

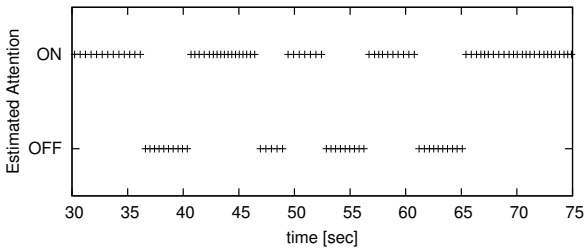

Figure 8: Results of Attention Estimation

zontally about every 10 seconds. As can be seen, our system properly detected the head rotation.

Figure 8 denotes the results of attention estimation. Here, the subject mainly watched our TV system and sometimes looked down at a magazine on his lap. Figure 10 shows a part of the image sequence (Figure 9 shows typical sample images of two states: "attention on (looking TV set)" and "attention off (looking other directions)"). As can be seen, in this case, our system accurately detected the loss of attention away from the TV system.

According to the detection results, our system gives audio feedback to the user. In the experiment, the system output a prerecorded message, "terebi miyouyo," which asks the user to watch TV.

In Figure 10, the subject took his eyes off the TV system to read. Three seconds later, the system recognized the change of attention and output the voice data (the 6th image in Figure 10). In this case, the user returned to watching the TV system as a response to the voice (Figure 10 right).

If a user gets bored by TV, it is hard to force him to watch. In that case, we have to detect this state and switch the contents. We are investigating a 'boring-state' detection method using sequential patterns of body and face motions.

6 Future Directions: Contents Control for Networked Interaction Therapy

Figure 11 shows a process diagram of our networked interaction therapy system. User motions are observed by cameras and other sensory systems. The location of the person in the room

Figure 9: Two states: "attention on (looking TV set)" and "attention off (looking other objects)"

is detected by using an IR camera and IR illumination patterns that give the 3-D position of the user. Then another camera that mainly observes the user's facial region detects the position and orientation of the face as described above.

If the system recognizes a loss of visual attention to the presented media, the system switches contents and/or gives audio feedback to recover the user's attention.

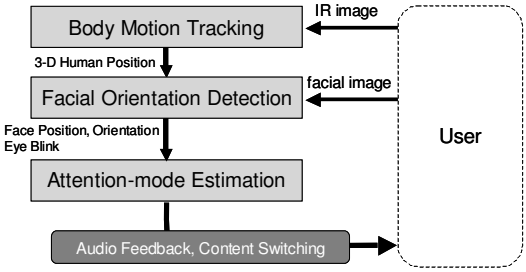

Figure 11: Attention Estimation using Vision-Based Tracking

We have implemented a preliminary system to test the video content switching based on the estimated states of visual attention. In the system, several video streams can be stored in the system and displayed to users on demand. The video content is automatically switched to others when the face orientation is turned away from a TV-set. Figure 12 shows the results of this content switching for a sample sequence. As can be seen, the system controls the video contents according to human be-

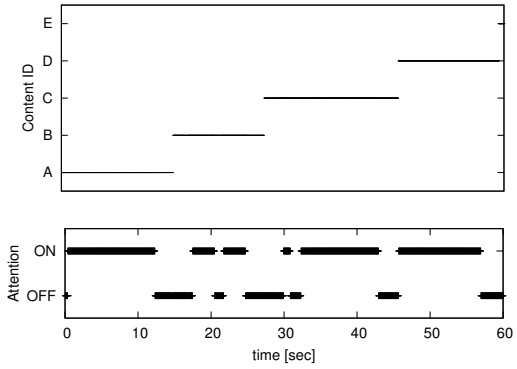

Figure 12: Content Switching Based on Face Orientation

Figure 10: Sample Sequence Data

haviors. However, in the current system, the rule of content switching is very simple and has many limitations. We have to focus on refining the switching rule to attract the user for a longer period of time. Other display modalities such as audio and olfactory output should also be considered.

7 Conclusions

We described vision-based techniques for estimating human attention based on facial orientation using an SSR filter. We employed SVM to model human faces. Facial orientation is determined by using the geometrical relation among the eyes and the tip of the nose. We applied these techniques to estimate the attention of users. In experiments, we controlled the timing of audio triggers to attract a user's attention to the presented video contents. Such capabilities are very promising for our "networked interaction therapy system," which effectively attracts the attention of memory-impaired people and lightens the burden on helpers or family members.

Future works include enhancing the face and body tracking system and carrying out studies involving a large number of participants.

This research was supported in part by the National Institute of Information and Communications Technology.

References

[1] N. Kuwahara, K. Kuwabara, A. Utsumi, K. Yasuda, and N. Tetsutani. Networked interaction therapy: Relieving stress in memory-impaired people and their family members. In *Proc. of IEEE Engineering in Medicine and Biology Society*, 2004. to be appeared.

[2] A. Utsumi and J. Ohya. Multiple-camera-based human tracking using non-synchronous observations. In *Proceedings of Fourth Asian Conference on Computer Vision*, pages 1034–1039, 2000.

[3] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Proc. of Computer Vision and Pattern Recognition*, pages 73–80, 1996.

[4] Q. Cai and J. K. Aggarwal. Tracking human motion using multiple cameras. In *Proceedings of 13th International Conference on Pattern Recognition*, pages 68–72, 1996.

[5] J. Segen and S. Pingali. A camera-based system for tracking people in real time. In *Proceedings of 13th International Conference on Pattern Recognition*, pages 63–67, 1996.

[6] Y. Yanagida, S. Kawato, H. Noma, N. Tetsutani, and A. Tomono. A nose-tracked personal olfactory display. In *SIGGRAPH 2003 Sketches & Applications*, 2003.

[7] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel. A real-time face tracker. In *Proc. 3rd IEEE Workshop on Application of Computer Vision*, pages 142–147, 1996.

[8] J. C. Terrillon, M. David, and S. Akamatsu. Automatic detection of human faces in natural scene image by use of skin color model and invariant moment. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 112–117, 1998.

[9] J. Heinzmann and A. Zelinsky. 3-d facial pose and gaze point estimation using a robust real time tracking paradigm. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 142–147, 1998.

[10] H. Wu, T. Shioyama, and H. Kobayashi. Spotting recognition of head gestures from color image series. In *Proc. of ICPR 98*, pages 1:83–85, 1998.

[11] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of CVPR 2001*, pages 1:511–518, 2001.

[12] H.-X. Zhao and Y.-S. Huang. Real-time multiple-person tracking system. In *Proc. of ICPR 2002*, pages 2:897–900, 2002.

[13] S. Kawato and N. Tetsutani. Real-time detection of between-the-eyes with a circle frequency filter. In *Proc. of ACCV 2002*, pages II:442–447, 2002.

[14] D. O. Gorodnichy. On importance of nose for face tracking. In *Proc. of IEEE 5th Int. Conf. on Automatic Face and Gesture Recognition*, pages 188–193, 2002.