Multiple-Face Tracking System Using Multiple Cameras

Mitsunori Ohya*, Hitoshi Hongo†‡, Kunihito Kato*, and Kazuhiko Yamamoto*

***Gifu University, Faculty of Engineering**
1-1, Yanagido, Gifu Japan 501-1193

†Softopia Japan and JST
Gifu Prefecture Regional Intensive Research Project
4-1-7, Kagano, Ogaki, Gifu Japan 503-0006

‡Hypermedia Research Center, SANYO Electric Co.,Ltd
180 Ohmori, Anpachi-cho, Anpachi-gun, Gifu Japan 503-0195

Abstract

We propose a multi-camera system that can track multiple human faces as well as focus on the face for recognition. Our current system consists of four video cameras. Two cameras are fixed and used as a stereo system to estimate face position. The stereo camera detects faces by a standard skin color method we proposed. The distances of the faces are then estimated. To track multiple faces, we evaluated the position and size of the faces in consecutive frames. The other two cameras perform tracking of the face. Our system selects a face for recognition from among the faces by using size and motion information sequentially. If the size of the selected face is too small for recognition, the tracking cameras acquire its zoomed image. Using our system, we experimented on multiple face tracking.

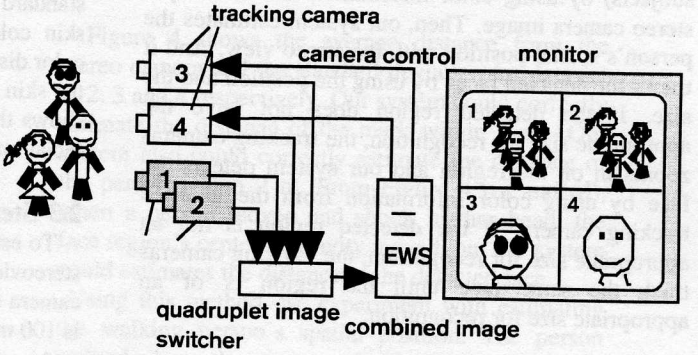
1.Introduction

In recent years, various methods based on computer vision technologies to improve man-machine interfaces and security systems have been developed[1][2][3]. The research projects achieved good results. In order to make computers interact with humans in the manner of human communication, it is important to focus on human intentions as indicated by facial expression, eye direction and so on. In practice, using a vision-based system should allow us to solve many problems such as occlusion, getting the appropriate resolution of an image to achieve a particular aim, and dealing with multiple persons.

On some face recognition systems and eye and mouth detection systems, the results depend on the image size. We already proposed the multiple camera system that can track multiple faces independent of the human's position[4]. This system had one fixed camera and two tracking cameras. We have tracked multiple face regions by using the region's center of gravity and size. We sometimes could not track the faces with this method when persons were overlapped. In this paper, we proposed a multiple camera system that can track multiple face regions by using spatial position and size. In order to estimate the person's position, our system uses stereovision. Our system can control the pan, tilt angles and zoom ratio of the cameras to acquire an appropriately sized image. In this paper, we describe our system configuration, an algorithm for detecting and tracking multiple faces. Experiments showed that our system could track multiple faces.

2. Face detection system

Our system has four video cameras and acquires a face image that has an appropriate size for recognition.



In this section, we explain the overview of our system.

Figure 1. System configuration

2.1 Overview of the face detection system

Figure 1 shows the configuration of our active face detection system. Our system uses video cameras (EVI-G20, Sony) that can be controlled in pan, tilt angles and zoom ratio by the computer. Two of the cameras are set in parallel as a stereo camera. The stereo camera detects faces and measures their respective distances from the stereo camera. The other two cameras are used for tracking and zooming in on the detected faces with the stereo camera. All of the images acquired by the stereo camera and the tracking cameras are combined by a quad switcher (YS-Q430, Sony), and the combined image is entered into the EWS (O2, SGI).

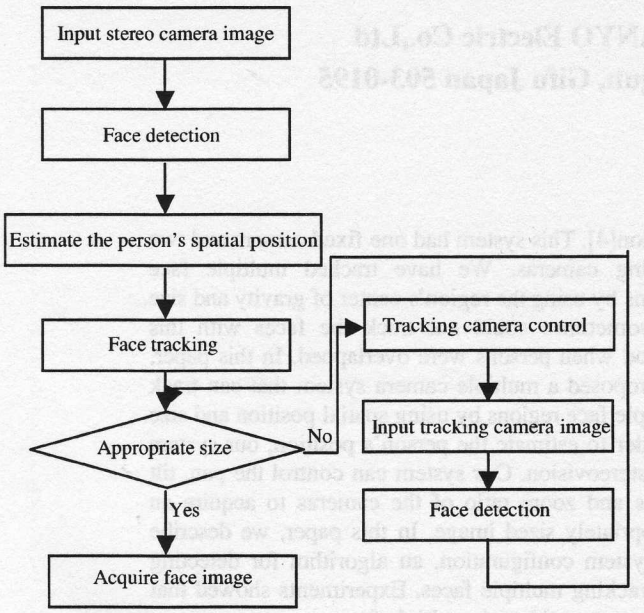
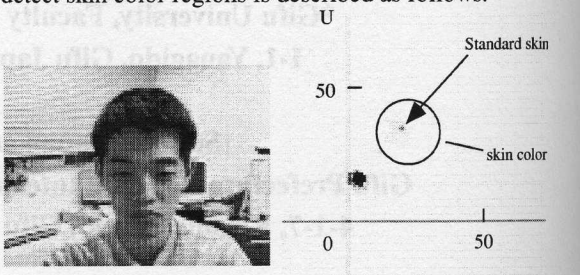


Figure 2. Flow chart of our system

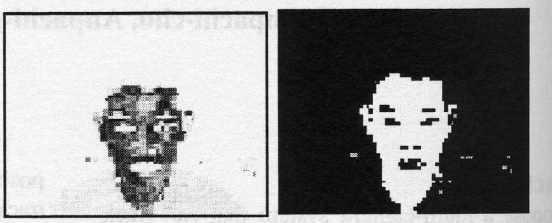
Figure 2 shows the flow chart of our system. First, our system detects the face regions (for Asian subjects) by using color information from the input stereo camera image. Then, our system estimates the person's spatial position with the stereo view. And it tracks the detected faces by using the position and the size. If the detected region does not have the appropriate size for recognition, the tracking cameras zooms in on the region and our system detects the face by using color information from the inputted tracking camera. If the detected region is not an appropriate size for recognition, the tracking cameras track the same face until the region is of an appropriate size for recognition.

2.2 Skin color detection

Various methods of face detection using color information have been developed [5][6]. Typically, such methods used the hue and chromatic values in the HSV color system. However, these methods do not work well if the face objects are not bright enough. The LUV color system consists of lightness (L) and color values (U and V). In general, the LUV color system is as capable as humans in representing the color distance between any two colors. We have already proposed a method of face detection using U, V values in the LUV color system [7]. Our method to detect skin color regions is described as follows.



(a) Original image (b) Color distribution



(c) Color distance of image (d) Detected skin color region

Figure 3. Color distribution

First, a two-dimensional UV values histogram is made from the previous frame. From the two-dimensional histogram, we determine the standard skin color that denotes the maximum number of pixels within the range of skin colors. Second, each pixel's UV value of the input image is converted to the color distance from the standard skin color. Figure 3 (a) is an original image, and Figure 3 (b) shows the standard skin color and the color distribution of the original image. Figure 3 (c) shows the color distance from the standard skin color. The level of brightness denotes the distance from the standard skin color. Dark areas show a close match to skin colors. Finally, we make a histogram of the color distance from the above results and then extract the skin region by discriminant analysis. Figure 3 (d) shows the results of the detected face region for the facial image in Figure 3 (a).

2.3 Stereo matching

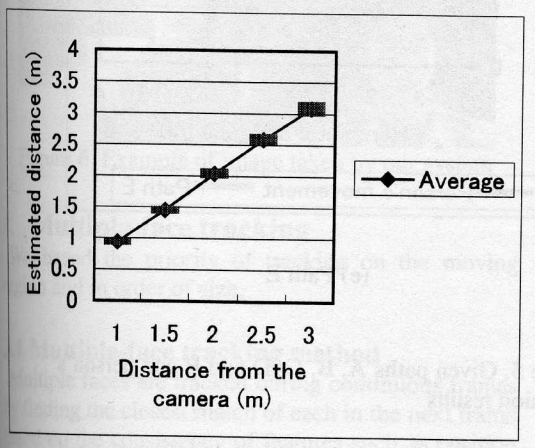
To estimate the spatial position of persons, we use stereovision. Our system consists of a parallel stereo camera setup, where the disparity of the two cameras is 100 mm.

As mentioned above, skin color regions are detected from both stereo cameras. Furthermore, we perform object matching between the detected faces from the right and left cameras. If multiple candidates exist for matching, we need to resolve the combination of faces. Since parallel stereo has

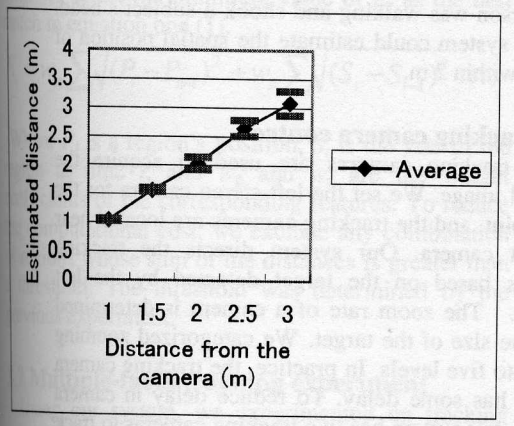
optical axes in parallel, the epipolar line becomes horizontal. Concerning the disparity of the stereo camera and the distance between humans and the system, we assume that targets cannot exchange their location between the stereo images. Therefore, we use the positions of the face's centers of gravity to match between targets as follows:

(1) Faces have the same epipolar lines
(2) Faces are the same size
(3) Faces are on a half-straight line, with no switching at right and left

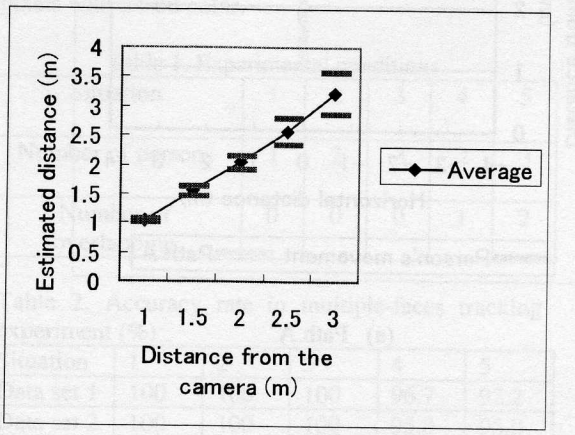
First, to examine the measurement precision of the spatial position by our system, we experimentally estimated the spatial position of a square marker and person's face. We took 50 images respectively, in which the marker was in front of the stereo camera at distances of 1, 1.5, 2, 2.5, and 3 m for data set 1. The marker was detected by subtracting the background. Next, we took 50 images respectively, in which a person stood for data set 2 in the same experimental conditions as data set 1. We recorded 50 images respectively, in which a person rolled from side to side for data set 3, and in which a person stood and shook his/her head for data set 4 in the same environment as data set 2.



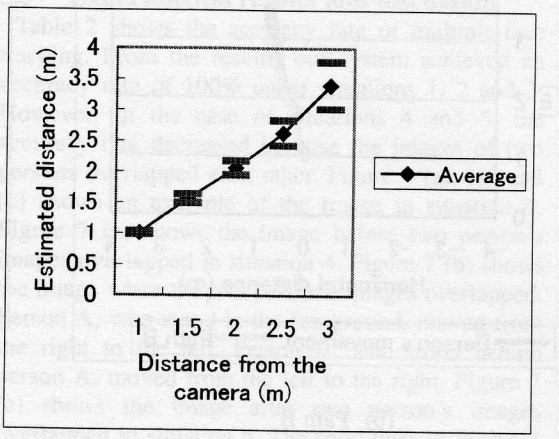
(a) Data set 1



(b) Data set 2



(c) Data set 3

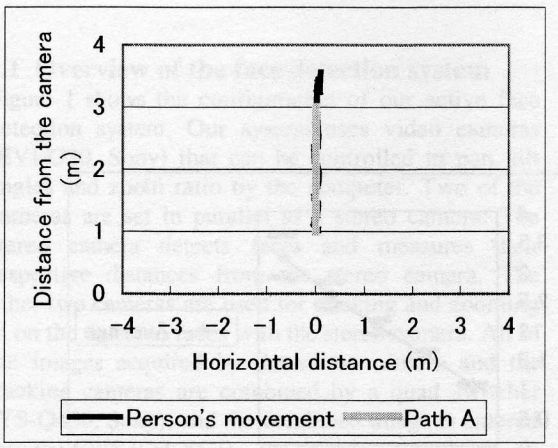


(d) Data set 4

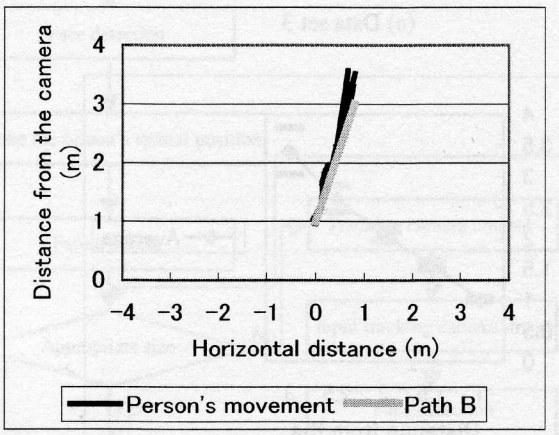
Figure 4. The average distance from the stereo camera

Figure 4 shows the average distance from the stereo camera and the standard deviation for data set 1, 2, 3 and 4 respectively. Our system could correctly estimate the distance of the mark within 3 m. Our system also could correctly estimate the distance of the person within 2 m. From Figure 4 (c) and (d), when a person moved and shook his/her head, the face region's center of gravity moved, but our system could estimates the distance of the detected face.

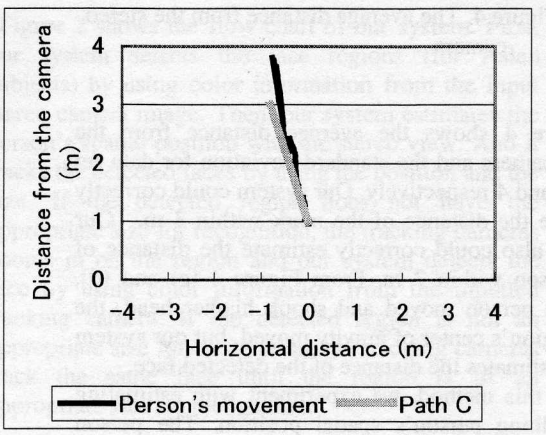
Using this method, we experiment with estimating the walking person's spatial position. The person walked along the given paths A, B, C, D and E. Figure 5 shows the given paths A, B, C, D and E and the person's movement results.



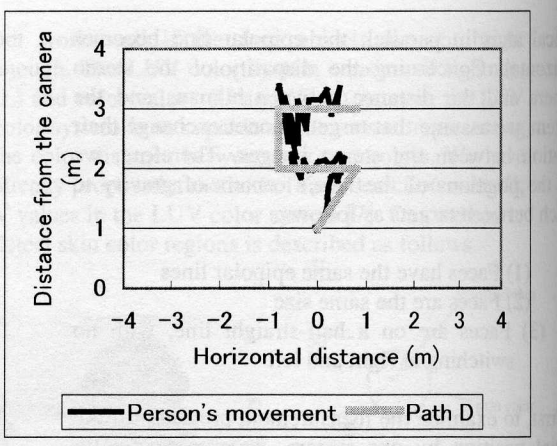
(a) Path A



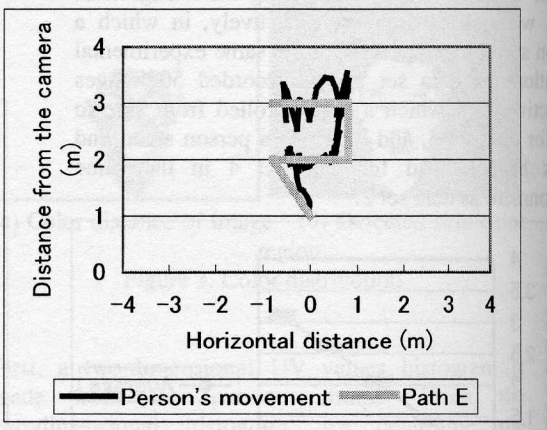
(b) Path B



(c) Path C



(c) Path D



(e) Path E

Figure 5. Given paths A, B, C, D and E and person's excursion results

In Figure 5 (d) and (e), the person's estimated spatial position varied widely at 3 m, but was accurate at 2 m.
The person was walking and shook a subject's head, but our system could estimate the spatial position of human within 2 m.

2.4 Tracking camera control

The tracking cameras are used to acquire the zoomed image. We set the left stereo camera for the base point, and the tracking cameras are located near the left camera. Our system directs the tracking cameras based on the target detected by the left camera. The zoom rate of a camera is determined from the size of the target. We categorized zooming rates into five levels. In practice, the tracking camera system has some delay. To reduce delay in camera control, our system has two tracking cameras to track independent targets at the same time. In addition, to

collect the images of targets efficiently, our system controls the tracking cameras as follows:

(1) Decide the priority of tracking for the detected targets by motion and by target size.
(2) Select the camera whose angle is aligned most closely with the target.
(3) Direct the selected camera to the target.

In this work, we based the priority of tracking on the moving region and according to size. The detected target images are acquired in order. After acquiring the target image, its priority is set to the lowest level. Figure 6 shows an example image in which our system detected two faces and zoomed in on them.

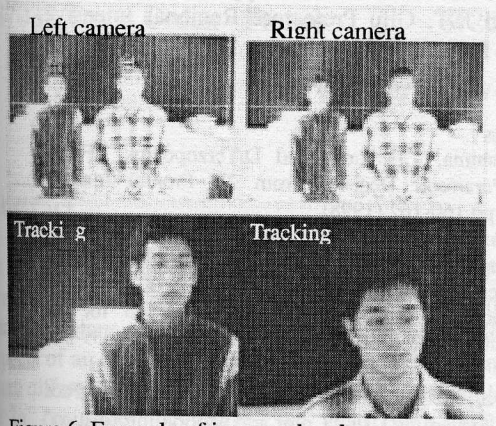


Figure 6. Example of image taken by our system

3. Multiple-face tracking

We based the priority of tracking on the moving region and in order of size.

3.1 Multiple-face tracking method

Multiple faces are tracked during continuous frames by finding the closest match of each in the next frame based on the consistency of features such as position and visual features. If multiple candidates exist for matching, all combinations of the faces are tested. Then we select the minimum value of E_t as the best match in equation one (1).

$$E_t = w_p \sum \sqrt{(P_t - P_{t-1})^2} + w_s \sum \sqrt{(S_t - S_{t-1})^2} \quad (1)$$

Where P_t is a region's position, S_t is the size of the region at time t, and w_p and w_s are the weight coefficients of the corresponding features. To reduce the computational cost, we exclude any combination of regions whose sum of the distances is greater than a threshold. The threshold was determined by the previous experiment.

3.2 Multiple-faces tracking experiment

Using our system, we experimented on tracking multiple faces. As shown in Table 1, we took image data under five situations, each having a different numbers of persons and incidents of overlapping by humans. We gathered three data sets with different subjects in each situation. Each data set consists of 60 frames. The sampling rate was 5 frames/sec. Each captured image is digitized to a matrix of 640 x 480 pixels with 24-bit color.

Table 1. Experimental conditions

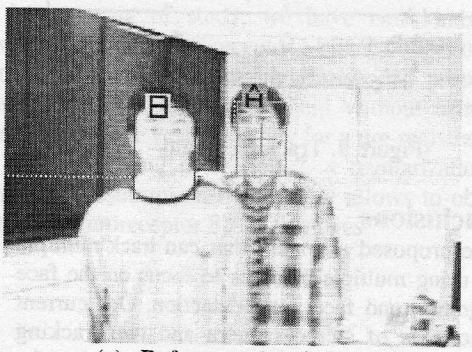
Situation	1	2	3	4	5
Number of persons	1	2	3	2	3
Number of overlapping	0	0	0	1	2

Table 2. Accuracy rate in multiple-faces tracking experiment (%)

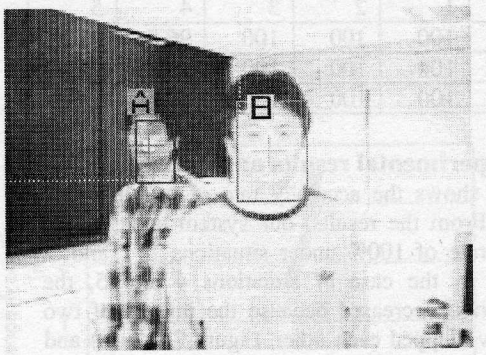
Situation	1	2	3	4	5
Data set 1	100	100	100	96.7	92.2
Data set 2	100	100	100	95.8	95.0
Data set 3	100	100	100	95.8	95.0

3.3 Experimental results and discussion

Table 2 shows the accuracy rate of multiple-face tracking. From the results, our system achieved an accuracy rate of 100% under situations 1, 2 and 3. However, in the case of situations 4 and 5, the accuracy rate decreased because the images of two persons overlapped each other. Figure 7 (a), (b) and (c) shows an example of the image in situation 4. Figure 7 (a) shows the image before two person's images overlapped in situation 4. Figure 7 (b) shows the image when the two person's images overlapped. Person A, who stood in the foreground, moved from the right to the left. Person B, who stood behind person A, moved from the left to the right. Figure 7 (c) shows the image after two person's images overlapped in situation 4. The solid lines in Figure 8 denote the tracking results of each person for the data from Figure 7. While occlusions occurred in frames 42 to 45, person A disappeared without a line of trace. Our system tracked the detected target as person B. After the occlusions stopped, our system could track both persons correctly. Tracking a person hidden by occlusions will be our next task.


(a) Before overlapping

(b) Two persons overlapping



(c) After overlapping

Figure 7. Example images in situation 4

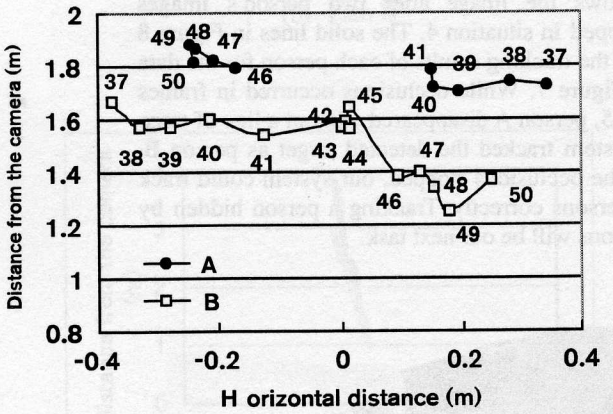


Figure 8. Tracking results

4. Conclusions

We have proposed a system that can track multiple faces by using multiple cameras to focus on the face for recognition and face parts detection. Our current system consists of stereo camera and two tracking cameras. The stereo camera detects faces by the standard skin color method and estimates the person's spatial position. The tracking cameras zoom in on images to obtain the appropriate size of faces.

Experiments showed that our system could estimate the person's spatial position by using the standard skin color method and track multiple faces by using the position and the size of regions.

The current system takes approximately 0.5 sec for one frame to complete all processes, including stereo matching and two tracking cameras control. We consider that it is easy to improve the speed by using hardware to convert RGB to LUV color and process Gaussian filters. Our next tasks will be to improve the accuracy rate of human's position and to estimate face orientation and eye direction.

Acknowledgments

A part of this research was supported by Softopia Japan and JST, Gifu Prefecture Regional Intensive Research Project.

References

[1]S.Morishima, T.Ishikawa and D.Terzopoulos: "Facial Muscle Parameter Decision from 2D Frontal Image", ICPR'98, pp.160-162 (1998)

[2] Jin Liu: "Determination of Point of Fixation in a Head-Fixed Coordinate System", ICPR'98, pp.501-504 (1998)

[3]T.Shigemura, M.Murayama, H.Hongo, K.Kato and K.Yamamoto: "Estimating the Face Direction for the Human Interface", Proc. Vision Interface'98, pp.339-345 (1998)

[4]M.Ohya, H.Hongo, K.Kato and K.Yamamoto: "Face Detection System by Using Color and Motion Information" ACCV2000 proc 1 pp.717-722 (2000)

[5] K.Sobottka and I.Pitas: "Extraction of Facial Regions and Features Using Color and Shape Information", IAPR 96, Vol.3, pp.421-425 (1996)

[6] Gang Xu and Takeo Sugimoto: "A Software-Based System for Realtime Face Detection and Tracking Using Pan-Tilt-Zoom Controllable Camera", ICPR'98 , pp.1194-1197 (1998)

[7]H.Hongo, A.Murata and K.Yamamoto: "Consumer Products User Interface Using Face and Eye Orientation", IEEE, International Symposium on Consumer Electronics ISCE'97, pp.87-90 (1997)