Proceedings of the 2002 IEEE/RSJ
Intl. Conference on Intelligent Robots and Systems
EPFL, Lausanne, Switzerland • October 2002

Face Identification Using an Omnidirectional Image Sequence

Yu Ohara Yasushi Yagi Taro Yokoyama Masahiko Yachida

Department of Systems and Human Science,
Graduate School of Engineering Science, Osaka University

Abstract

Face is one of the most attractive information for personal identification.In this paper, we propose the face identification method from an omnidirectional image sequence. Since an omnidirectional image sensor HyperOmni Vision observes a 360-degree view around the robot, it can observe a global azimuth information of the person (face). We track the human face while the person walks around the camera. Under an assumption of smooth human motion, we identify the corresponding person from facial database.

1 Introduction

Omnidirectional vision can observe a 360-degree view around a camera in real time, and can observe a global azimuth information of persons. By using global azimuth information via omnidirectional vision, we can focus on and observe a person. Many researchers have reported regarding tracking persons and estimating their positions [1, 2, 3, 4, 5, 6, 7, 8]. The omnidirectional vision is expected to be useful for surveillance and monitoring of people. However, it is difficult to apply an omnidirectional image system to face recognition because the angular resolution of an omnidirectional image is as yet too low. However from 2000, we started a project to maximize face recognition utilizing an omnidirectional image sequence [9].

While many researchers have investigated several face recognition methods, Tark [10] most reports are based on a frontal face snapshot taken by a standard camera under controlled observational conditions. Since people naturally walk around cameras, it is not always possible to obtain a frontal face suitably posed and a convenient distance from the camera.

Recently, several research groups have proposed identification methods from a sequence of face images. For instance, a common approach uses an average of the recognition rate in an image sequence. Wang and Haung used an oblivion function for weighting the most recent images ([11]). But when people walk naturally, the information important for personal identification does not always appear in the recent face image. Furthermore, averaging never enhances the recognition rate because the image size and the angular resolution of the facial region in an omnidirectional image are a somewhat smaller than that from an ordinary camera. The approach by constrained mutual subspace, proposed by Yamaguchi ([12]), makes it possible to stably identify the person by evaluating a face image with a set of face patterns. The method assumes that facial features such as eyes and noise can be extracted precisely. But in the case of the omnidirectional vision, since a face region is too small to extract accurate facial features, such a method is not reliable for identification by omnidirectional vision.

In this paper, considering these advantages and disadvantages of omnidirectional vision, we propose a template matching based method to identify a person from sequences of face images obtained by omnidirectional vision where each face region in the image is normalized in its size and by the width of the head that can be stably estimated. As noted, a standard averaging recognition score is not suitable for low resolution and small size face images. Omnidirectional vision can obtain a face region from over a range of distances and from left to right. This means that the differences of distance and face poses in the face images are greater than those from an ordinary camera. According to our preliminary results, both the pose and distance of the face affected the magnitude of correlation. Therefore, we propose a reliable model that represents the differences of template matching results relative to distance and face pose.

2 Overview of face identification system

2.1 HyperOmni Vision

The omnidirectional image sensor HyperOmni Vision (Figure 1) can capture omnidirectional information by using a hyperboloidal mirror while it simultaneously continuously observes people as they move around an environment. These advantages are suit-

Figure 1: HyperOmni Vision

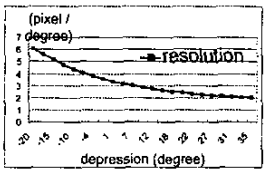


Figure 2: Angular resolution of HyperOmni Vision related to the depression

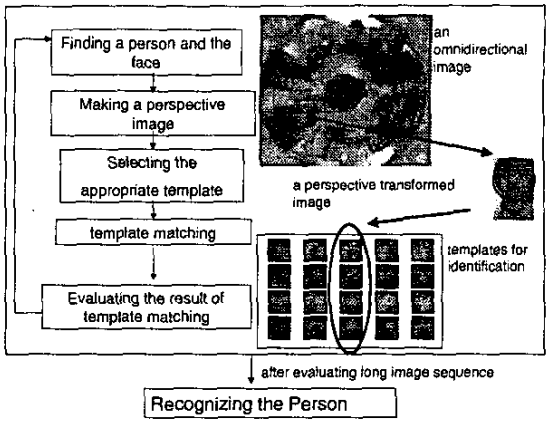


Figure 3: Face identification system

able for remote monitoring and surveillance. However, the angular resolution of HyperOmni Vision depends on the observational depression angle. Figure 2 shows the angular resolution of HyperOmni Vision as it relates to the depression. If an ordinary camera has 30 degrees of field of view and 512 pixels vertically, the angular resolution of the camera is approximately 15 pixels per degree. On the other hand, the resolution of the HyperOmni Vision is less than just 6 pixels per degree, which is about a third as low as that of standard cameras. This makes it difficult to recognize something precisely from a single image.

2.2 Assumptions

The initial condition is that nobody is in the room and the lighting condition is constant. A person comes into the room and, looking forward naturally, passes through beside the HyperOmni Vision, which is mounted at a height of 1.6 meters.. This means that the change of the face appearance in the image sequence is only caused by the face direction relative to the HyperOmni Vision and that it changes continuously.

3 A process sequence of the face identification system

Figure 3 shows a sequence of face recognition process. Our face recognition system consists of two parts. The first part is a human and head tracking process. The second is the head pose estimation and personal identification. In this section we introduce the outline of the human and head tracking process. The second part of the system is described in the next section.

First, the subtracted region is made by point-to-point subtractions between the stationary background image and the input image. After smoothing the subtracted region, to find the human region, radial lines in the image are transformed into 2-D polar coordinates(r, θ) and projected onto the θ axis to get a 1-D projection. The azimuth of the person is estimated by calculating the peak of the 1-D projection. After extraction of the human region, the width of the head in the transformed perspective image is estimated. To estimate the width of the head, we first extract the top of the head. The top of head is extracted by projecting the extracted human region to the r-axis to get the 1-D projection as shown in Figure 4. Then we define the top of head as the position where the magnitude of the 1-D projection is a five parts of the maximum magnitude of the 1-D projection. This threshold has been determined by our experience. in a similar matter, we extract the neck, and then the width of the face is defined by the following equation.

$$s = y((r_{top} + r_{bottom})/2) \quad (1)$$

Figure 5 shows the sequence of extracted faces.

4 Face identification from an image sequence of HyperOmni Vision

The proposed method identifies a person by using the characteristic that the appearance of a face changes smoothly. First, under the smoothness constraints of a change of a face, we estimate the face poses in

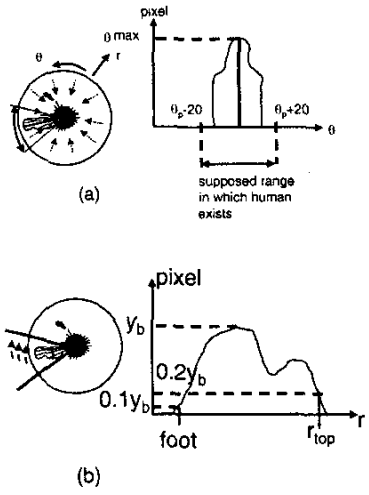


Figure 4: Human and face region detection

the image sequence. Actually the input face image is matched with template face images in which the pose of each template changes smoothly. Once the face pose is estimated, we calculate the recognition rate against all persons in the facial database. Continuously calculating the recognition rate against all persons, the system adds up the score of the person whose recognition rate is the best in that frame. Then the first person whose score is over a certain threshold is identified as the corresponding person.

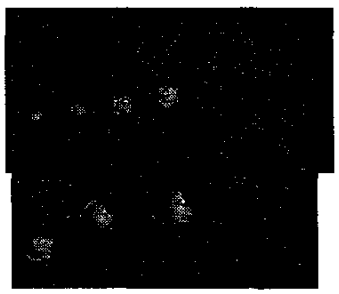


Figure 5: Result of face region extraction

4.1 A sequence of template images for each person

Before recognition, there is a need to make a facial database. In our system, the database contains sequences of face images that change appearance smoothly for each person. Templates for each person consists of 20 images that change face pose forms -60 degrees to 60 degrees by 6 degrees. (Note that a frontal face is the center of an image sequence in a pose that is approximately 0 degrees) An example of a sequence of template images is shown in Figure 6 . The size of a template image is 45 × 45 pixels square.

4.2 Weighting table for estimating head pose and normalization of face size

By calculating the cross-correlation between an input image and a template image, the pose of the input image is estimated as being a high correlation pose of the template image. By doing that for each frame, we can track the face pose while a person walks through a room beside the camera.

When the head pose around the vertical axis enlarges, even if a person walks while keeping looking forward, a small tilting motion of the head is amplified into a large appearance change of the head in the image. In such cases, it becomes difficult to get a high correlation between the input image and the templates, and it is difficult to track the correct pose. To make for some stability of the correlation value against the difference of a head pose, we have defined a weighting table for estimating head poses. To make a weighting table for each person, the first step is to compute the correlations between adjacent images in a sequence of template images of each person. We define the weight of a face pose as the inverse value of the average of these correlations Figure 7 (a) is the calculated weighting table. The abscissa axis is the pose of the head. This shows that the change of appearance is small around a frontal face, and is sensitive in case of a slanted face.

Based on the width of the face, the face in a transformed perspective image is normalized. Actually, the width of every face image is normalized to 90 pixels. The template size in the normalized image is then 45 × 45 pixels.

4.3 The estimation of face pose

Let $K(r^p, t)$ be a correlated value between an image r^p that is a template image of person p and the input image at frame t. And $V(r^p, t)$ is defined as following equation.

$$V(r^p, t) = K(r^p, t) \times w(r) \quad (2)$$

Figure 6: Example of template images

w(r) is a weight value to a pose of template image r. Before face recognition, we select persons whose faces are different types from the database. These data are used for face pose tracking. These are called representative templates. First, we match the input image with representative templates. Against each representative person, the pose with the highest correlation is selected. Next, we vote on selected poses. The first ranked pose $r_0^\star$ and second ranked pose $r_0^{\star\star}$ are selected as candidate poses of the first frame. $Y(p,t)$ is defined in the following equation.

$$Y(p,t) = V(r_{\max}, t) \tag{3}$$

Note that $r_{\max}$ means $r_t^\star$ or $r_t^{\star\star}$ at frame t.

Each candidate selects the best scored adjacent pose on the next frame. Finally, two trajectories of the face pose are extracted. Although this operation does not have a good reliability, as do dynamic programming and beam search algorithms, the computational cost of our tracker is low.

4.4 Identification process at each frame

Initially, we have to select the person whose evaluation value is the highest, and is higher than the threshold D. If no evaluation value of any person is higher than the threshold D, such a frame is not used for recognition. An evaluation value of a frame that is higher than threshold D is used for recognition. Let define recognition value $Q(p,t)$ as the following equation.

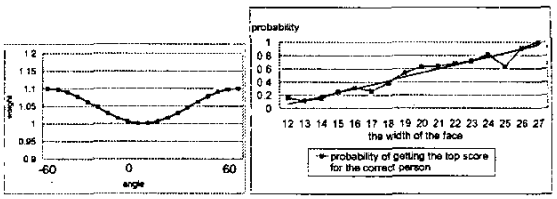
- In the case of a person p who got the highest evaluation value

$$Q(p,t) = d(t)Y(p,t) \tag{4}$$

- In the case of a person p 's evaluation value is not the highest

$$Q(p,t) = 0 \tag{5}$$

Here, $d(t)$ is the weight for the distance of a person, and is determined by the distance between a person and the sensor. When 15 people walk straight toward the sensor, the probability of getting the top score is shown in fig 7 (b). As showed in fig 7 (b), the probabirity is propotion to the width of a face in



(a) (b)

Figure 7: (a)Weighting for face pose, (b)Probability of getting the top score by correct person at each face size

omnidirectional image. That why we define $d(t)$ is defined as a following equation.

$$d(t) = b \cdot s(t) \tag{6}$$

The estimated slope b is 0.05.

4.5 Identification of a person considering time sequence

It can happen that because of observational noise an wrong person gets a higher correlation value than the correct one. A person, who in fact does not look like someone, can look like that person if we observe them from a certain position and in a particular pose. However, it can be expected that an incorrect person will not always get a high correlation score. Therefore, we can define the recognition rate $R(p)$ of a person p as in equation 7 $R(p)$ is an average of the recognition values at each frame.

$$R(p) = 1/F \sum_{i=0}^{t} Q(p,i) \tag{7}$$

Here, F is the sum of the frames that are useful for recognition, meaning that someone has got a higher evaluation value than threshold D. Then, in our system, this evaluation is done for both $r_t^\star$ and $r_t^{\star\star}$, and the final answer of the identification is the person who got the higher score for recognition rate $R(p)$ than the threshold in either of the pose sequences which start $r_0^\star$ or $r_0^{\star\star}$, when F is bigger than the threshold. But if following two situations occur, we believe that no one can be recognized.

- When the useless frames for recognition continue for a long time

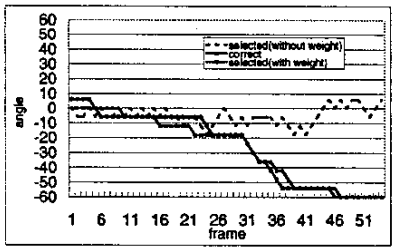


Figure 8: Example of face pose estimation

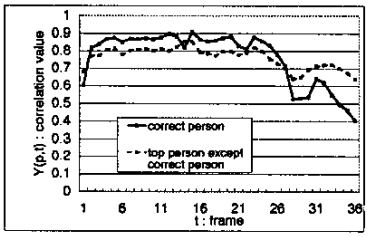


Figure 9: Example of evaluation value (person B)

- When the face pose continuously selects 60 degrees or -60 degrees.

5 Experiments

5.1 Making template images

Templates are made from an image sequence in which a person rotates their head from left to right at distance of 30 cm from the HyperOmni Vision. We used image sequences of fifteen persons. There were 20 template images for each person, and each image differed by 6 degrees. Appearances changed from -60 degrees to 60 degrees Face pose estimation and personal identification start from the distance at which we can recognize the face in the omnidirectional images. With our HyperOmni Vision, we can recognize a face when the size of the template region is approximately 10×10 pixels in the omnidirectional images. In such a case a person whose height is 1.7 m (an average height for Japanese) is 1.0m from the sensor. To evaluate our proposed method, subjects walked five different courses, The number of subjects was thirteen. For of walking pattern 1, each subject walked once, while for the others each walked twice. In the case of pattern 1, the size of the face for the templates changed from about 10×10 pixels to 60×60 pixels in the omnidirectional images.

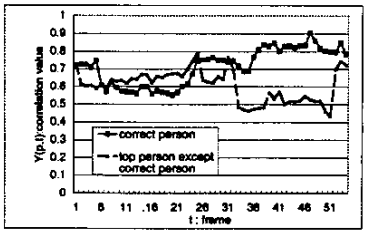


Figure 10: Example of evaluation value (person A)

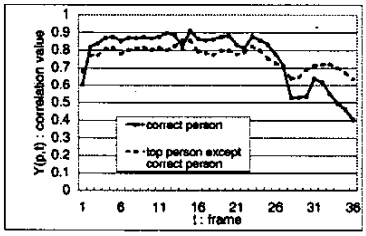


Figure 11: Example of evaluation value (person B)

The sampling rate was 30 frames per second for both templates and the input images.

5.2 Result of face pose estimates

Figure 8 (b) shows the comparative results of face pose estimations. The correct one is defined manually looking at images. As shown in the figure, in cases without weighting, only the templates around the frontal faces are selected, On the other hand, using the weighting table, our method can select almost exact poses. In our experiment, using a template sequence in which the pose of face changes smoothly, we got 90.9 % correct pose estimation.

5.3 Result of identifying a person

In Figure 11 (a), when a person is not near the sensor, the evaluation value of the correct person (person B) is high, but when a person comes close to the sensor, the evaluation value of person B becomes low. Because when the appearance was different from the frontal one, even by just a small rotation of the head, the appearance of the face changed and the evaluation value became low. Initially, we discussed the evaluation and recognition values for time sequences of pattern 1. Figure 12 (a) and Figure 11 (a) are the

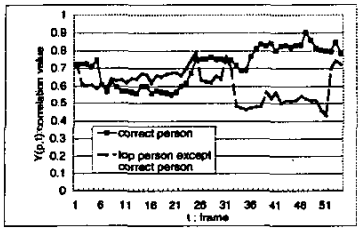


Figure 12: *Example of evaluation value (person A)*

results of evaluation values when subjects A and B walk.

The threshold D for the evaluation value $Y(p,t)$ is 0.75 (Taken as appropriate from our experiences). In Figures. 12 (a) and 11 (a) In the beginning of Figure 12 (a), when a person walks distant from the sensor, another person gets a high evaluation value. When a person approaches the sensor, the evaluation value of the correct person (person A) increased. Figures 12 (b) and 11 (b) are results of the recognition rates of Figures 12 (a) and 11 (a), respectively. In both Figures 12 (b) and 11 (b), the recognition rate of the correct person became high from the accumulating recognition rate. After evaluating along the frames, the correct person's recognition rate remained stable and high. Thus we identified the correct person. Our proposed method makes it possible to identify the correct person in 85 % of people, and over many walking patterns. Mis-identification rate and non-identification rate are 7.5 %, and 7.5 %, respectively. Thus with our system we can get a 77.3 % identification rate.

6 Discussion and Conclusion

In this paper, we proposed a method for personal identification utilizing the omnidirectional image sensor, HyperOmni Vision. Usually, to recognize a person from a single face shot in HyperOmni Vision is difficult because of its low resolution. But by our method, observing and evaluating an image sequence in which a face pose changes smoothly, and by summing the evaluation results of each frame, we could identify the correct person. By observing the person for a long time, the size and pose of the face in the omnidirectional images changed significantly, but considering the weighting table related the size and pose of a face, we could use results combined from these different conditions. Our basic idea does not limit the method of correlation. It is acceptable that PCA and ICA will give a higher recognition rate. The important point here is the weighting table and idea of accumulating. Although the current system supposes that there is only one person in the images, HyperOmni Vision is also good for observing many people at the same time. Our face identification method did not restrict the number of persons. We are presently planning to connect a real-time multiple person tracking system.

References

[1] Nayer.S.K. and Boult.T.E: Omnidirectional VSAM System, *Proc. of DARPA Image Understanding Workshop*, pp. 55–61 (1997).

[2] Onoue.Y, Yokoya.N, Yamazawa.K and Takemura.H: Visual Surveillance and Monitering System Using an Omnidirectional Video Camera, *Proc. of 14th IAPR Int. Conf. on Pattern Recognition*, pp. 588–592 (1998).

[3] R.Miki, N.Yokoya and andH.Takemura, K.: A Real-time Surveillance and Monitering System using Multiple Omnidirectional Video Cameras, *Proc. of ACCV*, pp. 528–534 (2000).

[4] Ng.K.C, Ishiguro.H, Trivedi.M and Sogo.T: Monitering Dynamucally Changing Environments by Ubiquitous Vision System, *Proc. Workshop on Visual Surveillance*, pp. 67–73 (1999).

[5] T.Sogo and H.Ishiguro: Real-time Target Localization and Tracking by N-Ocular Stereo, *Workshop on Omnidirectional Vision*, pp. 153–160 (2000).

[6] I.Kopilovic, B.Vagvolgyi and T.Sziranyi: Application of Panoramic Annular Lens for M87otion Analysis Tasks, Surveillance and Smoke Detection, *Proc. of ICPR*, pp. 714–717 (2000).

[7] T.Nishimura, H.Yabe and R.Oka: Indexing of Human motion at Meeting Room by Analyzing Time-varying Images of Omni directional Camera, *Proc. of the Fourth Asian Conference on Computer Vision*, pp. 1–5 (2000).

[8] R.Stiefelhagen, J.Yang and A.Weibel: Simulatanous Tracking of Head Poses in a Panoramic View, *Proc. of ICPR*, pp. 726–729 (2000).

[9] T.Yokoyama, Y.Ohara, Y.Yagi and M.Yachida: Face Recognition from an Omnidirectional image sequence, *IPSJ SIG Notes CVIM - 125* , pp. 119–123 (Jan. 2001).

[10] M.Turk and A.Pentland: Eigenfaces for recognition, *Cognitive Neurocience*, Vol. 3, No. 1, pp. 71–86 (1991).

[11] Weng.J, Evans.C.H and Hwang.W.S: An Incremental lerning methodfor face recognition under continuousvideo stream, *proc. IEEE conf. onAutomatic Face and Gesture Recognition*, pp. 251–256 (2000).

[12] Yamaguchi.O, Fukui.K and Maeda.K: Face recognition using temporal image sequence, *Proc. of the International Conf. on Automatic Face And Gesture Recognition*, pp. 318–323 (1998).