Recognizing Faces of Moving People by Hierarchical Image-Set Matching

Masashi Nishiyama, Mayumi Yuasa, Tomoyuki Shibata, Tomokazu Wakasugi,
Tomokazu Kawahara and Osamu Yamaguchi
Corporate Research & Development, Toshiba Corporation, Japan
masashi.nishiyama@toshiba.co.jp

Abstract

This paper proposes a novel method for recognizing faces in a cluster of moving people. In this task, there are two problems caused by motion, which are occlusions, and changes in facial pose and illumination. Multiple cameras are used to acquire near-frontal faces to avoid occlusions and profile faces. The Hierarchical Image-Set Matching (HISM) creates a distribution for each individual by integrating a set of face images of the same individual acquired from the multiple cameras. By adopting a method for comparing between test and training distributions in identification, variation in pose and illumination is alleviated, and good recognition accuracy can be obtained. Experimental results using video sequences containing 349 people show that the proposed method achieves high recognition performance compared with conventional methods, which use frame-by-frame identification and a distribution obtained from a single camera.

1. Introduction

Face identification technology can be used to build a security system that is natural, non-intrusive, and easy to use [9]. Our goal is to design a security system capable of recognizing moving people in a cluster by using face images. In order to obtain high recognition performance, we have to cope with the following issues : 1. multiple individuals move in a cluster simultaneously, and 2. facial appearance varies according to pose and illumination owing to the motion.

Issue 1, the occlusion problems arise principally in the case that the system is built using a single camera. If a short individual comes following after a tall individual, a single camera sometimes cannot view the face of the short individual. In this case, a method of detecting faces and tracking 2D position in a single camera e.g. [20] is inapplicable. To avoid these problems, multiple cameras are used to acquire face images for each individual as shown in Figure 1. Multiple cameras can view each face from different positions.

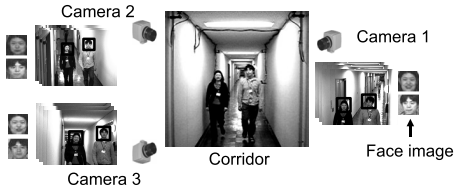

Figure 1. Concept of a security system for moving people in a cluster. Multiple cameras are used to acquire face images from moving people to avoid the occlusion problems.

Thus, although a particular camera cannot detect a face owing to occlusion, multiple cameras increase the probability of the face being detected near-frontally and without occlusion.

As solutions to issue 2, many face recognition methods have been proposed e.g. [18, 15]. In these methods, a single face image from a single camera is used for test data. A face image is represented by a vector in a feature space and is compared with a distribution of training data registered previously as shown in Figure 2(i). We refer to this as the frame-by-frame method. In the frame-by-frame method, false recognition arises frequently because test data is only a single vector and is easily influenced by variation in pose and illumination. In other methods for dealing with such variation, normalization techniques for correcting pose and illumination have been proposed e.g. [2, 21]. However, it is difficult to remove the variation completely using a single face image because of remaining ambiguity. To overcome this problem, we apply a face recognition method using a set of face images for test data instead of a single face image described in [16, 24, 1, 3]. In these methods, the variation is represented by a distribution obtained from a set of face images using a single camera. A test set is compared with the distribution of a training set as shown in Figure 2(ii). By incorporating the distribution of test data, the influence of the variation is alleviated. Thus, the probability of false recognition is decreased. We refer to this as the frame-integration method. This paper improves recognition performance of the frame-integration method using multi-

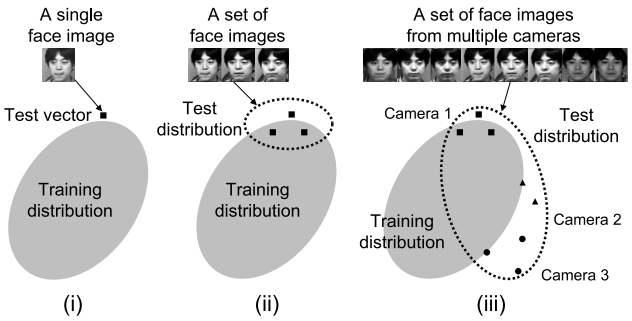

Figure 2. Comparison of three identification methods; (i)frame-by-frame using a single camera, (ii)frame-integration using a single camera, and (iii)frame-integration using multiple cameras. A set of face images has a distribution in a feature space. Similarity between test and training distributions is significantly less affected by change in facial pose and illumination than for single test vectors.

ple cameras as shown in Figure 2(iii). Multiple cameras can acquire many more face images for test data than a single camera and estimate a reliable distribution from that. Therefore, they increase the probability that a test distribution approaches a training distribution. In order to achieve the improvement, it is necessary to obtain a set of face images for each individual under multiple cameras in the situation where multiple individuals are moving in a cluster simultaneously: termed issue 1.

This paper proposes a novel method which identifies moving people using a set of face images generated by the *Hierarchical Image-Set Matching* (HISM). And, we describe a method for comparing sets of face images for matching accurately. HISM links together face images of the same individual from multiple cameras using three matching layers. Layer 1 generates a fragmentary set by matching face images for a single camera. Layer 2 links together fragmentary sets to a connected set between cameras. Layer 3 identifies faces using a connected set. The layers are designed to alleviate variation in pose, illumination, and the number of individuals for comparison. Figure 3 compares the variation for each layer. In Layer 1, matching of face images is subject to less influence of pose, although there is considerable variation of illumination due to motion. Although a greater number of individuals are registered for training data in Layer 3, Layers 1 and 2 compare a smaller number of candidates using time information because the number of people moving simultaneously is limited. In Layer 3, high recognition performance is obtained due to the effectiveness of the frame-integration method.

The remainder of this paper is organized as follows. First, Section 1.1 describes previous work. Next, Section 2 describes HISM, and Section 3 describes an improvement for comparing distributions from sets accurately. Then,

Variation \ Layer	1 Matching within a camera	2 Matching between cameras	3 Face identification
•Illumination	Medium	Medium	Large
•Pose	Small	Medium	Large
•Num. of individuals	Small	Small	Large
	Easy →		Difficult

Figure 3. Comparison of the influence of variation for each layer in HISM. Layer 1 is subject to less influence of pose, although there is considerable variation of illumination due to motion. Layers 1 and 2 compare a smaller number of candidates using time information.

we demonstrate the effectiveness of hierarchical matching through experiments in Section 4.

1.1. Previous Work

This section describes previous work on recognition of moving people and discusses the advantages of the proposed method. The methods described in [26, 22, 17] deal with variation in facial pose due to motion. Yang et al. [26] proposed a method for extracting frontally posed face images, Wang et al. [22] proposed a method for estimating pose using stereo cameras, and Tanaka et al. [17] proposed a method for synthesizing face images in various poses using a generic face model. Although they considered variation in pose, these methods did not consider how to recognize multiple individuals moving in a cluster simultaneously.

To recognize multiple individuals, [5, 12, 19] used a tracking algorithm based on estimating exact 3D positions of each individual using multiple cameras. These methods have an advantage in that people can move freely in the area watched by cameras. However, for estimating such positions these methods require camera calibration which is time-consuming and makes a security system expensive.

The proposed method can also recognize multiple individuals moving in a cluster, which is not dealt with in [26, 22, 17], and alleviate the influence of variation in pose and illumination by the frame-integration method. For matching face images, the proposed method uses only a pattern matching method that evaluates similarity between images and thereby does not require knowledge of 3D object positions. Thus, unlike [5, 12, 19], strict camera calibration is not required. The proposed method does not require high-frame-rate cameras, and is applicable for cameras of different frame rates. Furthermore, if occlusion arises, it is unnecessary in the proposed method to consider a state transition between detecting and tracking.

2. Hierarchical Image-Set Matching (HISM)

This section describes a recognition method for moving people in a cluster using a set of face images for each indi-

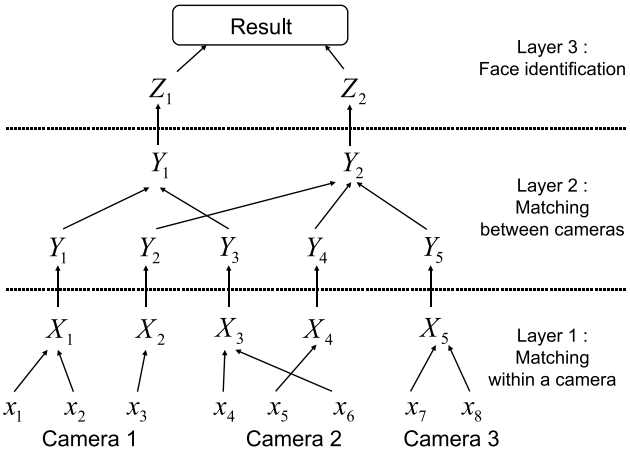

Figure 4. Example of the structure of HISM. The diagram assumes the case of two people moving in three cameras. x_i represents a face image from each camera. Firstly, x_i is matched to a fragmentary set X_i for a single camera in Layer 1 and X_i is translated to a fragmentary set Y_i of Layer 2. Secondly, Y_i is matched between cameras and matched Y_i is translated to a connected set Z that is used for identification in Layer 3.

vidual generated by matching of face images from multiple cameras. The HISM framework consists of three matching layers. Figure 4 shows an example structure in the situation where two people move under three cameras.

2.1. Layer 1 : Matching Within a Camera

Face images are sequentially acquired from each camera. A face image x from camera index c is defined as

$$x \equiv \{\mathbf{v}, c, t\} \ , \tag{1}$$

where $\mathbf{v}$ is a feature vector; t is time when the face image is obtained. $\mathbf{v}$ is generated from a captured image after preprocessing (i)–(iv). (i) Face regions are detected from camera images by a method using the joint Haar-like features [10] and AdaBoost. (ii) 14 facial feature points, e.g. pupils, nostrils, etc., are detected from a face region using combination of a circular points detection and a pattern matching method. (iii) Facial pose is corrected to near-frontal direction by fitting facial feature points to a generic 3D shape learned from the shapes of many faces[8] . (iv) Illumination is normalized by applying filter convolution to a pose-corrected face image for extracting the ratio of albedo[13]. Figure 5 shows the flow of generating a face image. After preprocessing, $\mathbf{v}$ is obtained by raster-scanning of a face image in Figure 5(iv).

Layer 1 generates a fragmentary set by matching face images of the same individual for a single camera. A fragmentary set X_i of Layer 1 is defined as

$$X_i \equiv \{x_1, \ldots, x_m\} \ , \tag{2}$$

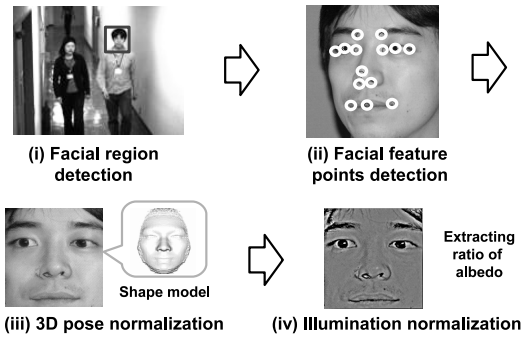

Figure 5. The flow of generation of a face image alleviating the variation.

where i is the label of a fragmentary set; m is the number of face images in X_i. The face images of X_i are acquired from the same camera. Let n_c be the number of fragmentary sets stored in Layer 1 for camera c by current time. If a new face image $x_{\rm new}$ is acquired, then $x_{\rm new}$ is added to the fragmentary set X_l of label l as

$$X_l \Leftarrow \{x_{\rm new}\} \cup X_l \ . \tag{3}$$

In order to determine l, a similarity S_i between $x_{\rm new}$ and X_i is evaluated as

$$l = \begin{cases} \underset{i}{\rm argmax}\{S_i \mid i = 1, \ldots, n_c\} & S_{\rm max} \geq \alpha_1 \\ n_c + 1 & \text{otherwise} \end{cases} \ , \tag{4}$$

where α_1 is a threshold; $S_{\rm max}$ is $\max\{S_i \mid i = 1, \ldots, n_c\}$. If $S_{\rm max}$ is less than α_1, then a new label n_c+1 is given, and a new fragmentary set is generated. The new label makes it possible to handle the situation where a new individual appears. A similarity S_i is calculated as

$$S_i = Sim1\,(x_{\rm new}, X_i) = \frac{({\mathbf{v}_{\rm new}}^T \mathbf{v}_j)^2}{1 + \gamma(t_{\rm new} - t_j)} \ . \tag{5}$$

where $\mathbf{v}_{\rm new}$ is a feature vector of $x_{\rm new}$; $\mathbf{v}_j$ is a feature vector of $x_j \in X_i$ acquired at the latest time t_j; $t_{\rm new}$ is a time of $x_{\rm new}$; γ is a constant value. The function $Sim1$ gives a large similarity to X_i obtained near the time of $x_{\rm new}$.

For transition from Layer 1 to Layer 2, X_i is translated to $Y_{\rm new}$ when a constant time β_1 passes as

$$Y_{\rm new} \Leftarrow X_i \quad \text{if} \ \ t - t_j > \beta_1 \ , \tag{6}$$

where t is current time; t_j is the time of the latest x_j added to X_i. $Y_{\rm new}$ is also defined as $\{x_1, \ldots, x_m\}$. The transition corresponds to the situation where an individual moves outside the area watched by the cameras.

False matching, in which face images of different individuals are wrongly linked together, can cause error in identification. To overcome this problem, threshold α_1 in equation (4) is important. The threshold is determined by a prior examination on a database in which all face images are labeled.

2.2. Layer 2 : Matching Between Cameras

Layer 2 generates a connected set for each individual by matching fragmentary sets between cameras. Let Y_i represent a fragmentary set stored in Layer 2 by current time. If $Y_{\rm new}$ is obtained from Layer 1, then it is linked together to the fragmentary set Y_k of label k as

$$Y_k \Leftarrow \{Y_{\rm new}\} \cup Y_k \ . \tag{7}$$

In order to determine k, a similarity T_i between $Y_{\rm new}$ and Y_i is evaluated as

$$k = \begin{cases} \underset{i}{\rm argmax}\{T_i \mid i = 1, \ldots, n\} & T_{\max} \geq \alpha_2 \\ n+1 & \text{otherwise} \end{cases} , \tag{8}$$

where n is the number of fragmentary sets; α_2 is a threshold; $T_{\max}$ is $\max\{T_i \mid i = 1, \ldots, n\}$. To calculate a similarity T_i between $Y_{\rm new}$ and Y_i, we use the *Mutual Subspace Method* (MSM) [25]. In MSM, a distribution is represented by a subspace generated from a set of face images using principal component analysis. The basis vectors of the subspace are the eigenvectors of a correlation matrix $A = 1/m \sum_{j=1}^{m} \mathbf{v}_j \mathbf{v}_j{}^T$[14]. Let $\mathcal{Y}_{\rm new}$ be the subspace of $Y_{\rm new}$ and $\mathcal{Y}_i$ be the subspace of Y_i. The similarity T_i is defined as

$$T_i = \mathit{Sim2}(Y_{\rm new}, Y_i) \ = \cos^2\theta, \tag{9}$$

where θ is the canonical angle between $\mathcal{Y}_i$ and $\mathcal{Y}_{\rm new}$. If $\mathcal{Y}_i$ and $\mathcal{Y}_{\rm new}$ are identical, θ equals 0. Let N be the dimension of subspaces $\mathcal{Y}_i$ and $\mathcal{Y}_{\rm new}$. The similarity $\cos^2\theta$ is equal to the largest eigenvalue $\lambda_{\max}$ of a $N \times N$ matrix R using $R\mathbf{a} = \lambda\mathbf{a}$. The element r_{pr} of R is $\sum_{q=1}^{N} (\psi_p^T \phi_q)(\phi_q^T \psi_r)$ where ψ_p is the p-th basis vector of subspace $\mathcal{Y}_i$; ϕ_q is the q-th basis vector of subspace $\mathcal{Y}_{\rm new}$.

Y_i is translated to $Z_{\rm new}$ for inputting to Layer 3 when constant time β_2 passes as

$$Z_{\rm new} \Leftarrow Y_i \quad \text{if} \ \ t - t_j > \beta_2 \ , \tag{10}$$

where t is current time, and t_j is the time of the latest x_j added to Y_i.

2.3. Layer 3 : Face Identification

Layer 3 identifies the face by matching $Z_{\rm new}$ to one of the training sets using a similarity U_i between $Z_{\rm new}$ and the training set Z_i of i-th individual as

$$\mathit{Person} = \begin{cases} \underset{i}{\rm argmax}\{U_i \mid i = 1, \ldots, h\} & U_{\max} \geq \alpha_3 \\ \mathit{unknown} & \text{otherwise} \end{cases} , \tag{11}$$

where h is the number of training sets; $U_{\max}$ is $\max\{U_i \mid i = 1, \ldots, h\}$; α_3 is a threshold. To calculate U_i we use the same function *Sim2* in Layer 2. U_i is $\mathit{Sim2}(Z_{\rm new}, Z_i)$.

3. Improvement of MSM using Orthogonalization

To improve the performance of MSM, we apply an orthogonalization process to the subspaces. An orthogonalization process emphasizes the difference of the distributions between individuals reported in [7, 6, 11]. By the orthogonalization process, a similarity becomes small between fragmentary sets generated from different individuals. Thus, false matching is decreased between the sets of different individuals. For orthogonalization, this paper uses a matrix O defined by a different formulation to [7, 6, 11]. O is generated by training subspaces registered in Layer 3. Let a projection matrix $C_i = \sum_{p=1}^{m'} \psi_{ip} \psi_{ip}^T$ be the training subspace for the i-th individual from Z_i. ψ_{ip} is the p-th basis vector of the training subspace of the i-th individual and m' is the number of the basis vectors of the training subspace. O is defined as

$$O = B \Lambda^{-\frac{1}{2}} B^T \ , \tag{12}$$

where B is a matrix consisting of the eigenvectors of $C_{\rm all} = 1/h \sum_{i=1}^{h} C_i$; Λ is a diagonal matrix of the eigenvalues of $C_{\rm all}$. The basis vector $\hat{\psi}_p$ of the orthogonalized subspace is calculated as

$$\hat{\psi}_p = O\psi_p \ \ (p = 1, \ldots, m') \ , \tag{13}$$

where ψ_p is the basis vector of subspace $\mathcal{Y}_i$. Note that $\hat{\psi}_p$ is not an orthonormal basis. Then, we apply the norm normalization and Gram-Schmidt orthogonalization to $\hat{\psi}_p$. The basis vector ϕ_q of subspace $\mathcal{Y}_{\rm new}$ is also applied by O in the same way. Finally, the canonical angle between the orthogonalized subspaces of $\mathcal{Y}_i$ and $\mathcal{Y}_{\rm new}$ is calculated using MSM.

4. Empirical Evaluation

4.1. Recognition performance on a real-world database

To illustrate the performance of the proposed method, we have conducted experiments on recognition of moving people using a real-world database. Firstly, Section 4.1.1 describes a real-world database and the measures used in evaluation. Secondly, Section 4.1.2 demonstrates recognition performance using a set of face images from multiple cameras. Thirdly, Section 4.1.3 demonstrates the effectiveness of HISM by the simulation of multiple individuals moving in a cluster.

4.1.1 Database and Measure for Evaluation

We collected video sequences of moving people for 349 individuals. In each video sequence, a single individual was

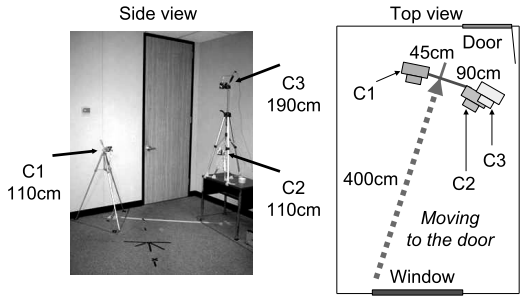

Figure 6. Camera setting to collect a real-world database for evaluation. Video sequences were collected for 349 individuals. Each individual moved along the broken line from the start position near the window to the end line near the door. Three cameras (C1, C2, C3) were used.

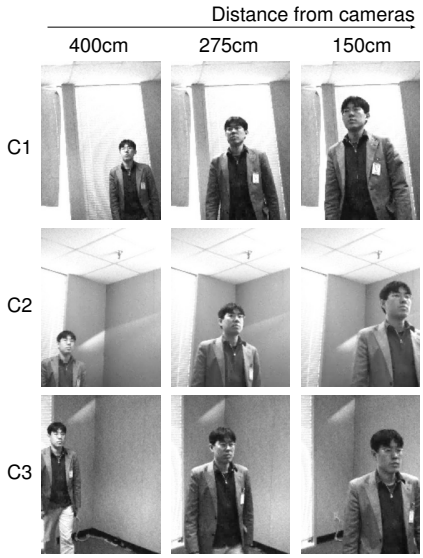

Figure 7. Sample frames of a video sequence of moving people in a real-world database in Section 4.1.1. Although camera positions are fixed, facial pose direction relatively changes from 22 degrees to 54 degrees as an individual moves from 400 cm to 150 cm from cameras.

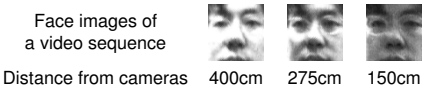

Figure 8. Samples of face images where pose is corrected in a video sequence. Illumination on a face varies relatively due to motion of subject though lighting sources are fixed.

moving. Three cameras (C1, C2, C3) were set as shown in Figure 6. Each individual moved along the broken line from the start position near the window to the end line near the door. They continued to look straight ahead while moving. For each individual, two video sequences were collected.

Table 1. Recognition performance in terms of RR and EER. *The number of moving people in a cluster simultaneously is only one*. (A) is frame-by-frame method using a single camera [23]. (B) is frame-integration method using a single camera [25]. (C) and (D) are frame-integration method using multiple cameras.

Condition	Frame rate	RR	EER
(A)	15fps	80.3%	8.7%
(B)	15fps	93.7%	5.6%
(C)	15fps	98.6%	2.0%
(D)	5fps	97.4%	2.3%

One is for training data and the other is for test data. The resolution of video sequences for each camera is 768×1024 pixels, and the frame rate is 15fps. A video sequence was captured for intervals of 4 seconds. Figure 7 shows sample frames of the video sequences and Figure 8 shows samples of face images acquired from C1. Pose and illumination are corrected for an extracted face image, scaled to 64×64 pixels.

For evaluation we use the recognition rate (RR) and the equal error rate (EER). RR is the probability that a self-similarity from self-training data has the highest value compared with other similarities from other training data. EER is the probability that the false acceptance rate (FAR) equals the false rejection rate (FRR) when α_3 of Equation (11) is used.

4.1.2 Evaluation of identification by a connected set from multiple cameras.

This experiment demonstrates recognition performance of the frame-integration method using multiple cameras. We assume that all face images are matched correctly, which is realized by limiting the number of moving people in a cluster simultaneously to one. The evaluation of automatic matching of face images is described in Section 4.1.3.

In order to compare the performance of identification for a single moving individual, we conducted experiments under three conditions:

(A). frame-by-frame method using a single camera (The frame rate of a camera is 15 fps) [23],

(B). frame-integration method using a single camera (The frame rate of a camera is 15 fps) [25],

(C). frame-integration method using multiple cameras (The frame rate of each camera is 15 fps).

(D). frame-integration method using multiple cameras (The frame rate of each camera is 5 fps).

In (A) the *Subspace Method* (SM)[23] was used. In SM, a face image from each camera is a vector for test. Note that

O of Section 3 was applied to a test vector and a training subspace to improve recognition performance. In (B) MSM and a single camera[25] were used. A set of face images of each camera is generated by matching of Layer 1. In (C) and (D) MSM and multiple cameras were used. The frame rate of each camera in (D) is one-third compared with that in (C). O is applied to subspaces calculated from sets in (B), (C), and (D). The dimension of $\mathbf{v}$ is $32 \times 32 = 1024$ by down-sampling of a corrected face image in Layer 1. The dimension of $\mathbf{v}$ is $16 \times 16 = 256$ by down-sampling the face image again in Layer 2. For identification of Layer 3, the dimension of a training subspace generated from a training set Z_i is 4. A training set for each individual is generated by linking together face images of C1, C2, and C3 manually. In (B), (C), and (D), if the number of face images in a set is less than 4, the set is not considered for identification. In this case, a test subspace cannot be generated from a moving individual, and then a similarity is equal to zero for FRR and RR.

Table 1 shows the evaluation result in terms of RR and EER for each experimental condition. In (A) and (B) RR and EER are average values calculated from each single camera. We can see that the conditions (C) and (D) are superior to the other conditions. In particular, (D) obtains good performance under the low-frame-rate cameras compared with (A) and (B). From this result, we believe that a set of face images from multiple cameras has significant effectiveness for identification.

Next, we demonstrate the effectiveness of an orthogonalization process described in Section 3. We evaluated in the same condition (C) except using a matrix O. In this condition, RR is 97.7% and EER is 2.9%. RR and EER in Table 1 (C) are superior to a method without an orthogonalization process.

4.1.3 Evaluation of identification using HISM.

This experiment considers the situation where multiple individuals are moving side by side in a cluster simultaneously. Recognition performance is evaluated by changing the combination of individuals in a cluster. This experiment assumes that near-frontal faces are obtained in the case that there are no occlusions because of the use of so many cameras. This experiment uses the real-world database introduced in Section 4.1.1. Using this database, we can simulate to evaluate recognition performance with the large number of individuals.

Firstly, M individuals were randomly selected from 349 individuals in a real-world database. Secondly, a face image for each selected individual was input to Layer 1 at initial time. After 66 milliseconds the next face image for each individual was input. M individuals were selected 200 times. The parameters were set as $\alpha_1 = 0.4, \alpha_2 = 0.4, \beta_1 = 1 \text{ second}, \beta_2 = 3 \text{ seconds}$. The parameters were experimentally determined to gain a probability of false matching less than 0.1%. If false matching occurs, then we calculated only FAR for mismatch as the correct correspondence for identification is not defined.

To compare recognition performance for multiple individuals moving simultaneously, we used five experimental conditions:

(i). frame-integration method using multiple cameras and HISM *(Proposed method)*,

(ii). frame-integration method using a single camera[25],

(iii). frame-by-frame method using a single camera[23],

(iv). frame-integration method using multiple cameras and hierarchical clustering[4],

(v). frame-integration method using multiple cameras and manual matching *(Ground truth, Ideal performance)*.

(i), (ii) and (iv) linked together face images automatically and (v) linked them together manually. (i) and (v) used a connected set, (ii) used a fragmentary set, (iii) used a single face image, and (iv) used a set $\hat{X}$ of face images generated by hierarchical clustering[4]. We applied hierarchical clustering to all face images of three cameras of selected M individuals. For the clustering, we used a distance measure $d_{\min}(\hat{X}_a, \hat{X}_b) = \min(1 - \mathbf{v}_i^T \mathbf{v}_j | \mathbf{v}_i \in x_i \in \hat{X}_a, \mathbf{v}_j \in x_j \in \hat{X}_b)$. The clustering was repeated until the minimum distance was less than a threshold. The experimental conditions of (iii) and (v) are equal to the conditions in Table 1(A) and (C) except the number of moving people M.

Figure 9 shows recognition performance in terms of RR and EER versus the number of M. RR and EER of (i) are superior to those of (ii), (iii), and (iv). However, (i) is inferior to the ideal performance of (v). This appears to be because face images of the same individual are occasionally not linked together. To improve recognition performance we need to develop a method of matching more accurately. EER of (ii) and (iv) is decreased compared with (iii). We consider that, as a result of degradation, it frequently happens that a test subspace for identification cannot be generated since the number of face images in a set is less than 4. This degradation does not arise in (i). From the results we conclude that HISM is effective for recognition of moving people.

4.2. Generation of a connected set under occlusion

This section demonstrates the effectiveness of the proposed method for a video sequence including occlusion. Occlusion frequently causes false matching in the case of conventional tracking methods. For collecting a video sequence, two cameras were used in a corridor, with one being

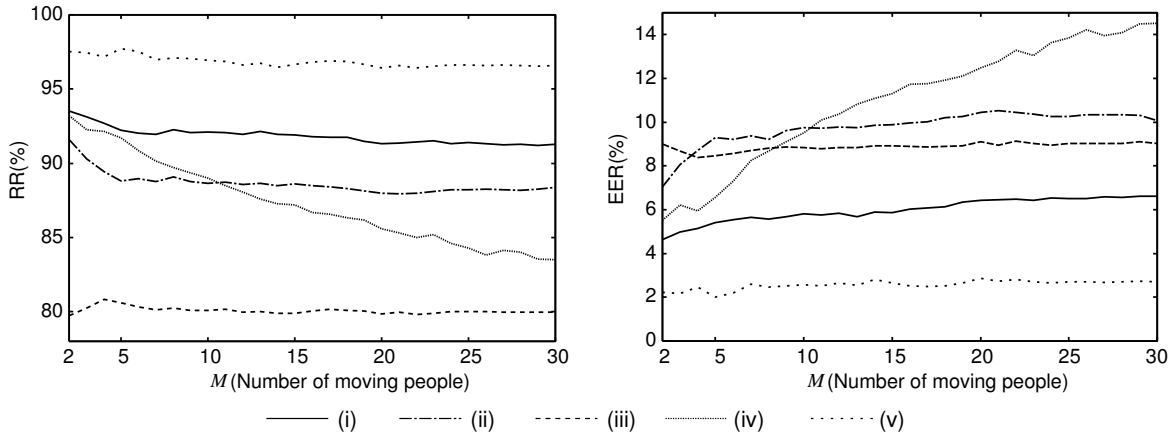

Figure 9. Recognition performance in terms of RR(%) and EER (%) with increasing the number of moving people M. *The performance is evaluated by simulating that multiple individuals are moving in a cluster simultaneously.* (i) frame-integration method using multiple cameras and HISM *(Proposed method)*, (ii) frame-integration method using a single camera[25], (iii) frame-by-frame method using a single camera[23], (iv) frame-integration method using multiple cameras and hierarchical clustering[4], (v) frame-integration method using multiple cameras and manual matching *(Ground truth, Ideal performance)*. Our method (i) yields better performance than the conventional methods (ii) to (iv).

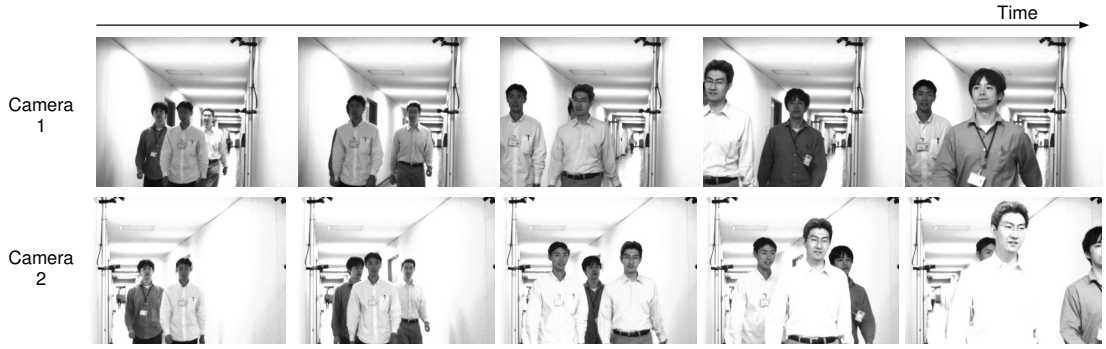

Figure 10. Examples of frames in a video sequence of moving people with occlusion in Section 4.2. The pair of upper and lower images are acquired at the same frame from different cameras.

placed on each wall. The video sequence includes three individuals moving in a cluster simultaneously and where occlusion arises in each camera how one individual moves in front of other individuals. Figure 10 shows sample frames from the video sequence. The width of the passage is 160 cm and the height of the cameras from the floor is 110 cm. We set the frame rate of each camera to 7.5fps and the resolution to 1024×768 pixels. 68 face images were acquired from the video sequence. Figure 11 shows a connected set generated by HISM for each individual. In the result, the number of face images in a set for individual 1 is 20, for individual 2 is 19 , and for individual 3 is 24. Even though an individual disappeared and reappeared, matching of face images was successful and a connected set for each individual was generated correctly. From this result, we conclude that our method can match face images even when occlusion arises.

5. Conclusion

This paper presented a method for recognition of moving people in which each individual was identified by the frame-integration method using multiple cameras. The proposed method matches face images of the same individual to generate a set of face images. The matching process is hierarchically divided into sub-processes. Each sub-process can stably match face images by limiting the variation which causes false matching. The experiment demonstrated the effectiveness of the proposed method on a real-world database of 349 moving people.

In future work, we intend to evaluate recognition performance by not only the simulation described in Section 4.1.3 but also in the case of actual video sequences in which occlusion is addressed. Also, we intend to develop an automatic method of determining parameters and a method of determining the relative, but not exact, positions of multiple

Figure 11. Connected set for each individual generated from a video sequence of Figure 10. Face images of the same individual are matched correctly.

cameras for high recognition performance.

References

[1] O. Arandjelovic and R. Cipolla. Face recognition from image sets using robust kernel resistor-average distance. *The First IEEE Workshop on Face Processing in Video*, 2004.

[2] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063 – 1074, 2003.

[3] R. Chellappa, V. Kruger, and S. Zhou. Probabilistic recognition of human faces from video. *The IEEE International Conference on Image Processing*, I:41 – 44, 2002.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification second edition. *John Wiley & Sons*, 2000.

[5] T. Kato, Y. Mukaigawa, and T. Shakunaga. Cooperative distributed registration for robust face recognition. *Systems and Computers in Japan*, 33(14):91 – 100, 2002.

[6] T. K. Kim, J. Kittler, and R. Cipolla. Incremental learning of locally orthogonal subspaces for set-based object recognition. *Proceeding British Machine Vision Conference*, 2006.

[7] J. Kittler. The subspace approach to pattern recognition. *in Progress in cybernetics and systems research.(Edited by R. Trappl, G. J. Klir, and L. Ricciardi), Hemisphere Publ. Co.*, page 92, 1978.

[8] T. Kozakaya and O. Yamaguchi. Face recognition by projection-based 3d normalization and shading subspace orthogonalization. *7th International Conference Automatic Face and Gesture Recognition*, pages 163 – 168, 2006.

[9] S. Z. Li and A. K. Jain. Handbook of face recognition. *Springer*, 2005.

[10] T. Mita, T. Kaneko, and O. Hori. Joint haar-like features for face detection. *Tenth IEEE International Conference on Computer Vision*, pages 1619 – 1626, 2005.

[11] K. Nagao and M. Sohma. Weak orthogonalization of face and perturbation for recognition. *IEEE Proceeding Conference on Computer Vision and Pattern Recognition*, pages 845 – 852, 1998.

[12] A. Nakazawa, H. Kato, S. Hiura, and S. Inokuchi. Tracking multiple people using distributed vision systems. *International conference on Robotics & Automation*, pages 2974 – 2981, 2002.

[13] M. Nishiyama and O. Yamaguchi. Face recognition using the classified appearance-based quotient image. *7th International Conference Automatic Face and Gesture Recognition*, pages 49 – 54, 2006.

[14] E. Oja. Subspace methods of pattern recognition. *Research Studies Press*, 1983.

[15] J. H. P.N. Belhumeur and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711 – 720, 1997.

[16] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. *7th European Conference on Computer Vision*, pages 851 – 868, 2002.

[17] H. Tanaka, I. Kitahara, H. Saito, H. Murase, K. Kogure, and N. Hagita. Dynamically visual learning for people identification with sparsely distributed cameras. *14th Scandinavian Conference on Image Analysis*, pages 130 – 140, 2005.

[18] M. Turk and A. Pentland. Face recognition using eigenfaces. *IEEE Proceeding Conference on Computer Vision and Pattern Recognition*, pages 586 – 591, 1991.

[19] N. Ukita and T. Matsuyama. Real-time cooperative multi-target tracking by communicating active vision agents. *Computer Vision and Image Understanding*, 97(2):137 – 179, 2005.

[20] R. C. Verma, C. Schmid, and K. Mikolajczyk. Face detection and tracking in a video by propagating detection probabilitie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1215 – 1228, 2003.

[21] H. Wang, S. Z. Li, and Y. Wang. Generalized quotient image. *IEEE Proceeding Conference on Computer Vision and Pattern Recognition*, 2:498 – 505, 2004.

[22] J. G. Wang, R. Venkateswarlu, and E. T. Lim. Face tracking and recognition from stereo sequence. *4th International Conference on Audio- and Video-based Biometric Person Authentication*, pages 145 – 153, 2003.

[23] S. Watanabe and N. Pakvasa. Subspace method of pattern recognition. *International Joint Conference on Pattern Recognition*, pages 25 – 32, 1973.

[24] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, pages 913–931, 2003.

[25] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *Third International Conference on Automatic Face and Gesture Recognition*, pages 318 – 323, 1998.

[26] Z. Yang, H. AI, B. Wu, S. Lao, and L. Cai. Face pose estimation and its application in video shot selection. *International Conference on Pattern Recognition 2004*, pages 322 – 325, 2004.