Unsupervised Face Recognition from Image Sequences Based on Clustering with Attraction and Repulsion

Bisser Raytchev and Hiroshi Murase
NTT Communication Science Laboratories,
3-1, Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
{bisser, murase}@eye.brl.ntt.co.jp

Abstract

We propose a new method for unsupervised face recognition from time-varying sequences of face images obtained in real-world environments. Two types of forces, attraction and repulsion, operate across the spatio-temporal facial manifolds, to autonomously organize the data without relying on any category-specific information provided in advance. Experiments with real-world data gathered over a period of several months and including both frontal and side-view faces were used to evaluate the method and encouraging results were obtained. The proposed method can be used in video surveillance systems or for content-based information retrieval.

1. Introduction

In recent years automated face recognition has attracted a lot of attention, and this seems to be motivated not only by scientific curiosity, but also by the numerous potential applications stemming from the fact that faces represent natural interfaces for humans, and face recognition is central to human communication. However, in spite of the extensive research conducted in this area during the last several decades (see [1]-[4] for surveys), face recognition still remains a domain in which humans significantly outperform computers, especially in real-time, unconstrained and unpredictable environments. Here we argue that some of the reasons for this situation, together with hints for the answers, might be found by investigating some of the discrepancies between the way humans learn faces and the way most computer-based face recognition procedures operate:

(a) Humans learn by interacting directly with the sensory input from their environment. Category labels, like human names in the case of face recognition, are not essential for discrimination in the learning process and are used just for convenience after the faces have already been learnt, based on the internal characteristics of the sensory input itself (*unsupervised learning*), rather than on any category-specific information accompanying it in a supervised manner. This is in contrast to the way most computer-based face recognition procedures operate. Computers are usually provided with input, which has been segmented and classified in advance by human teachers (*supervised learning*);

(b) Biological learning is *incremental* in nature, i.e. new categories can be learnt and added to those already in existence, without the need to "relearn" everything anew, or to represent the new categories with a restricted pre-defined set of features, either designed by humans or automatically selected to represent the available data in some optimal way;

(c) Automatic face recognition is difficult because different people's faces observed in the same conditions (illumination, view angle, size, etc.) look more similar to each other than the same person's face observed in different conditions (e.g. in frontal and side view; under extreme illumination conditions; occluded; etc.). One approach to solve this problem is to find features invariant under different conditions, but this has proven to be difficult. It might be possible that humans use a different approach – to learn from *time-sequential* input, in the form of temporally-constrained continuous sensory streams, containing the whole spectrum of variations in illumination, viewing angles and object sizes, which everyday life provides. Again, in contrast to this, computers typically are trained with few isolated samples from a large set of different face categories, taken in restricted environmental conditions.

Although some researchers have already pointed out the need for incremental and unsupervised self-organization of the internal state of the learning system ([5]-[7]; see also [8] for a relevant discussion on the differences between human and machine learning and the need for "more cognitive learning"), or use of time-sequential data [9], a method for face recognition which takes into consideration all of the concerns mentioned above and performs reasonably well on real-world data has not been demonstrated yet, to our knowledge.

In this paper we propose a new method for unsupervised face recognition from video sequences of time-varying facial images, inspired by observations (a)-(c) above. The method utilizes the higher level of sensory variation contained in the input image sequences to autonomously organ-

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

ize the data into category groups, without relying on category-specific information provided in advance. This is achieved by using the Clustering with Attraction and Repulsion (CAR) algorithm introduced below, where two types of opposing forces, attraction and repulsion, act across the spatio-temporal facial manifolds, and the partition of sample space that can be characterized by maximal structural stability is chosen as the final grouping. Several experiments, using data obtained in real-world conditions, were conducted in order to evaluate the performance of the proposed method, and encouraging results were observed. Expected areas of application of this method include visitor identification in surveillance systems, content-based face retrieval/annotation in multimedia applications, etc.

Figure 1. An example of original face image sequence (temporally subsampled) together with the corresponding normalized face-only sequence extracted from it.

2. Clustering with Attraction and Repulsion

The purpose of the learning algorithm introduced here is to group a set of unlabelled face image sequences, which could be pre-stored as a database (*batch mode*), or obtained in a sequential manner in the order they become available from the input device (*incremental mode*). As already mentioned, this has to be done without using any category information provided in advance, i.e. some clustering technique ([10]-[12]) has to be utilized. Our task is further complicated by the following requirements: (a) generally, the number of the categories is not known in advance and new face categories have to be accounted for in a non-destructive manner; (b) the different categories are not represented uniformly, some might be under-represented and some over-represented; (c) in sample space, the face sequences for the different face categories form complex non-linear manifolds, for which intra-class distances generally can take higher values than inter-class distances.

The above-mentioned characteristics of the problem preclude the possibility of using some of the popular clustering approaches, and this has motivated us to propose the current method. The following subsections describe the different stages of the system in more detail. Preprocessing will be briefly explained in section 2.1. Some definitions, which will be needed for the description of the learning algorithm will be given in section 2.2. Section 2.3 will introduce the batch version of the CAR algorithm, while section 2.4 deals with the incremental version and online recognition. Several experimental results will be reported in section 3, and section 4 will conclude the paper.

2.1. Preprocessing

Since the concrete implementation of this part of the system is not essential for the operation of the learning algorithm, the detailed description of this stage will be omitted. All that is required from the preprocessing is to obtain image sequences of the moving objects of interest and to guarantee that each separate image sequence corresponds to one and the same object only. Here we assume that input is provided from a video camera fixed in a constant position and continuously monitoring the scene in front of it. Subjects enter the scene, walk towards the camera and finally exit the scene. To extract face-only image sequences, a multi-resolution image pyramids are formed from the binary silhouettes of the moving subjects, and the face area is extracted after analyzing the x and y-histograms of the binary silhouettes at different resolutions. The extracted and normalized face-only image sequences (see Fig.1) are input to the next stage of the system for learning them. Alternative algorithms for face tracking/extraction may be employed, depending on the concrete task (for example, see [13],[14]).

2.2. Preliminaries

Let $S^{(a)}(i,j,t)$ and $S^{(b)}(i,j,t)$ be two face image sequences, where a and b are sequence indexes ($a, b : 1...N$), i and j are image coordinates, and t is image frame number. Let C be a non-empty set of such image sequences $\mathbf{C} = \{ S^{(a)}, S^{(b)}, \ldots, S^{(N_C)} \}$ with cardinality $\eta(\mathbf{C}) = N_{\mathbf{C}}$. Each image sequence will generally depict a complex curve in the high-dimensional image space, but assume, for simplicity, that each of the face image sequences $S^{(a)}, S^{(b)}, \ldots, S^{(k)}, \ldots$ can be represented by points $a, b, \ldots, k, \ldots$ in 2D-space, as shown in Fig. 2. Assume that the points $a, \ldots, k, \ldots$ interact with each other, i.e. each point k is being attracted or repulsed from the other members l of $\mathbf{C}$ by positive forces of attraction $\boldsymbol{A}(k, l)$, or negative forces of repulsion $\boldsymbol{R}(k, l)$ whose magnitude is a function of the distance between them, as shown in Fig. 3. The forces $\boldsymbol{D}(k, l)$ acting in the *area of doubt*, which lies between the area of attraction and the area of repulsion, can be either positive of negative, but small in magnitude, as this is an *area of doubt*, and we wouldn't like its influence to be too significant.

In Fig. 3, the examples of force $F(d)$ as a function of the distance d between two image sequences were obtained by

$$F_1(d) = C_1\left(\exp(\gamma - d)^{0.1} - \exp(d - \gamma)^{0.1}\right) \quad (1)$$

$$F_2(d) = -C_2(d - \gamma)^3 \quad (2)$$

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

$$F_3(d) = C_3\left(\exp(-\frac{d^2}{\sigma}) - \exp(-\frac{(d-2\gamma)^2}{\sigma}) \right) \quad (3)$$

$$F_4(d) = \begin{cases} A(d) = C_4(\alpha - d) & 0 \le d < \alpha \\ D^+(d) = +\delta & \alpha \le d < \gamma \\ D^-(d) = -\delta & \gamma \le d < \beta \\ R(d) = C_4(\beta - d) & d \ge \beta \end{cases} \quad (4)$$

$$\gamma = (\alpha + \beta)/2;$$

where α and β are parameters which determine the width and location of the area of doubt, C_i are positive scale constants (irrelevant to the clustering results, but introduced here just to be able to plot F_1 - F_4 in the same graph in Fig. 3, where $C_1 = 1$, $C_2 = 0.001$, $C_3 = 150$, $C_4 = 3$), and δ in (4) is a small positive constant.

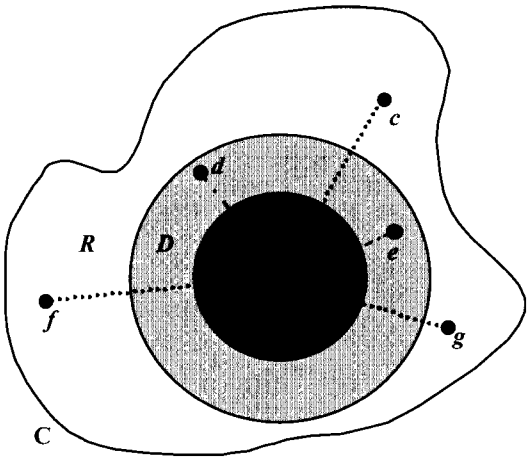

Figure 2. Forces of attraction and repulsion between image sequences. Points *a* and *b*, in the area of attraction *A* for point *k*, act with positive force on *k*; points *c*, *f*, *g* in the area of repulsion *R* act with negative force on *k*; and points *d*, *e* in the area of doubt *D* act with small positive or negative forces.

The resultant force $\tilde{F}(k \mid \mathbf{C})$ acting on each $k \in \mathbf{C}$ can be given in normalized form by

$$\tilde{F}(k \mid \mathbf{C}) = \frac{\eta(\mathbf{C})^{-1}\left(\eta_A \sum_{l \in \mathbf{C}, l \neq k} A(k,l) + \eta_R \sum_{l \in \mathbf{C}, l \neq k} R(k,l) + \eta_D \sum_{l \in \mathbf{C}, l \neq k} D(k,l) \right)}{\sum_{l \in \mathbf{C}, l \neq k} A(k,l) + \sum_{l \in \mathbf{C}, l \neq k} |R(k,l)| + \sum_{l \in \mathbf{C}, l \neq k} |D(k,l)|} \quad (5)$$

where η_A, η_R and η_D are the number of points (image sequences) in $\mathbf{C}$ from which k receives attraction, repulsion, or lie in the area of doubt for k. $\tilde{F}(k \mid \mathbf{C})$ varies in the interval [-1,...,1] and provides a measure for the extent to which k belongs to $\mathbf{C}$. At the same time, a measure for the overall stability of $\mathbf{C}$ can be provided by

$$\mathrm{Z}(\mathbf{C}) = \frac{1}{\eta(\mathbf{C})} \sum_{k \in \mathbf{C}} \tilde{F}(k \mid \mathbf{C}), \quad (6)$$

which also takes values in [-1,...,1] and the more positive its value is, the more stable the structure of $\mathbf{C}$ is considered to be. For the purposes of the clustering algorithm introduced below, a given set of points $\mathbf{C}$ is recognized as a valid cluster only if

$$F(k \mid \mathbf{C}) \ge 0 \text{ for all } k \in \mathbf{C} \quad (7)$$

i.e. when each of the nodes k receives as a whole more attraction than repulsion from the rest of the members of $\mathbf{C}$ (note that $\mathrm{Z}(\mathbf{C}) \ge 0$ is not a sufficient condition for $\mathbf{C}$ to be a valid cluster).

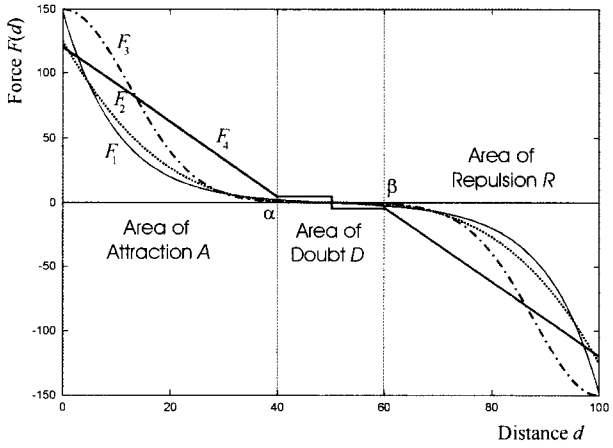

Figure 3. Several examples of forces between image sequences as a function of distance.

A set of points $\mathbf{C}$ which doesn't satisfy (7) can be modified into a valid cluster by removing from it those points l^* for which $\tilde{F}(l^* \mid \mathbf{C}) < 0$, to get a new set $\mathbf{C}^* = \mathbf{C} - \{l^*\}$. The order in which points are being removed from $\mathbf{C}$ is important. The point removal starts with the point

$$l^* = \arg\min_{l \in \mathbf{C}} \{\tilde{F}(l \mid \mathbf{C})\} \quad (8)$$

and after it is removed, the resultant forces $\tilde{F}^*(k \mid \mathbf{C}^*)$ for all $k \in \mathbf{C}^*$ are recalculated using (9)-(13) below, after which again the point with the most negative resultant force (if

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

such one exists) is removed and the forces acting on the remaining points are recalculated. This procedure is repeated until either no more points with negative resultant forces exist, or $\mathbf{C}^*$ is a singleton.

The updated resultant force for any $k \in \mathbf{C}^*$ can be compactly represented in a vector form using the following notation:

$$\tilde{F}^*(k \mid \mathbf{C}^*) = \frac{\frac{1}{\eta - 1}\left(\mathbf{n} - \frac{\mathbf{B}^*(\mathbf{C})}{\left\|\mathbf{B}^*(\mathbf{C})\right\|}\right)\left(\mathbf{B}(\mathbf{C}) - \tilde{\mathbf{B}}(\mathbf{C})\right)}{\mathbf{e}^{\mathrm{T}}\left(\hat{\mathbf{B}}(\mathbf{C}) - \mathbf{B}^*(\mathbf{C})\right)} \tag{9}$$

where

$$\mathbf{B}(\mathbf{C}) = \begin{pmatrix} \sum_{l \in \mathbf{C}, l \neq k} A(k,l) \\ \sum_{l \in \mathbf{C}, l \neq k} R(k,l) \\ \sum_{l \in \mathbf{C}, l \neq k} D(k,l) \end{pmatrix}, \quad \hat{\mathbf{B}}(\mathbf{C}) = \begin{pmatrix} \sum_{l \in \mathbf{C}, l \neq k} A(k,l) \\ \sum_{l \in \mathbf{C}, l \neq k} \left|R(k,l)\right| \\ \sum_{l \in \mathbf{C}, l \neq k} \left|D(k,l)\right| \end{pmatrix} \tag{10}$$

$$\mathbf{B}^*(\mathbf{C}) = \begin{pmatrix} A(k,l^*) \\ \left|R(k,l^*)\right| \\ \left|D(k,l^*)\right| \end{pmatrix}, \quad \tilde{\mathbf{B}}(\mathbf{C}) = \begin{pmatrix} A(k,l^*) \\ R(k,l^*) \\ D(k,l^*) \end{pmatrix} \tag{11}$$

$$\mathbf{n} = \left(\eta_A \;\; \eta_R \;\; \eta_D\right)^{\mathrm{T}}, \quad \mathbf{e} = \left(1 \;\; 1 \;\; 1\right)^{\mathrm{T}}. \tag{12}$$

In order to calculate the forces in (1)-(12), it is necessary to adopt a suitable measure for the distance between two image sequences. Different ways to do this are conceivable, but for simplicity and out of computational considerations, we define the distance between two face sequences $S^{(a)}(i,j,t)$ and $S^{(b)}(i,j,t)$ as

$$\begin{aligned} d\{a,b\} &= \min_{x,y} dist\left\langle S^{(a)}(i,j,x), S^{(b)}(i,j,y)\right\rangle \\ &= \min_{x,y} \sum_{i,j} T_\delta\left\{\left|S^{(a)}(i,j,x) - S^{(b)}(i,j,y)\right|\right\} \end{aligned} \tag{13}$$

In (13), $T_\delta\{\cdot\}$ is a threshold function with suitable threshold parameter δ. More elaborate face distance measures than the one defined above might be used, if processing time is not a problem. Having defined the distance between any two image sequences, the nature of the force acting between them, together with its magnitude, can be calculated as a function of the distance, e.g. using one of the functions plotted in Fig. 3. The only parameters that have to be set are α and β, which determine the width and location of the doubt area.

2.3. Clustering by Attraction and Repulsion

Initially given are N unlabeled face sequences from L categories, and the objective is to group them into clusters without using any category-specific information provided in advance (L, the number of different people is also unknown).

Step 1. MERGING

For the purpose of clustering, each of the face sequences $S^{(a)}$, $S^{(b)}$,..., $S^{(k)}$, ... is represented by a point a, b, ..., k, ..., each point initially forming a separate set, so that we have the singletons $\mathbf{C}_a = \{a\}$, $\mathbf{C}_b = \{b\}$, ..., $\mathbf{C}_k = \{k\}$,..., . For all k (k: 1...N), merge set $\mathbf{C}_k$ with set $\mathbf{C}_l$ for which

$$\begin{aligned} \mu_{kl} &= \mathrm{Z}(\mathbf{C}_k \cup \mathbf{C}_l) \times \Omega(\mathbf{C}_k \cup \mathbf{C}_l) \\ &= \max_i \{\mathrm{Z}(\mathbf{C}_k \cup \mathbf{C}_i) \times \Omega(\mathbf{C}_k \cup \mathbf{C}_i)\} \\ &\qquad \left(\Omega(\mathbf{C}) = \mathbf{e}^{\mathrm{T}}\mathbf{B}(\mathbf{C})\right) \end{aligned} \tag{14}$$

where $\mathbf{C}_i$ must satisfy the following conditions:

$$\tilde{F}(x \mid \mathbf{C}_k \cup \mathbf{C}_i) \geq 0 \text{ for all } x \in \mathbf{C}_k \cup \mathbf{C}_i \tag{15}$$

$$\exists A(x,y) > 0, \quad x \in \mathbf{C}_k, \; y \in \mathbf{C}_i \tag{16}$$

$$\eta(\mathbf{C}_k) \leq \eta(\mathbf{C}_i) \tag{17}$$

so that $\mathbf{C}_k$ is not changed if no $\mathbf{C}_i$ satisfying (15)-(17) exists. In (14), using solely $\mathrm{Z}(\cdot)$ to calculate the merge factor μ_{kl} would favor the formation of stable (even if small) cluster structures, which might lead to over-fragmentation. We observed better results when both $\mathrm{Z}(\cdot)$ and $\Omega(\cdot)$ were combined, favoring the formation of larger (even if not so stable) cluster structures. Condition (16) is necessary to guarantee that there exists force of attraction between the two sets candidates for a merge (this is not necessarily satisfied if (15) is satisfied).

Step 2. SPLITTING

If M sets are obtained after the merging step, for each set $\mathbf{C}_j$ (j: 1 ,..., M) check whether it forms a valid cluster satisfying (7), i.e. whether each member of that cluster receives more attraction than repulsion from the other members of the same cluster. The sequence-member l^* with lowest value of $\tilde{F}(l^* \mid \mathbf{C}_j) < 0$ (if such l^* exists), is removed from $\mathbf{C}_j$ to form a new singleton $\mathbf{C}_{M+1}$, which will participate (i.e. will be recycled) in the following merge step. After the removal of the most negative member, the resultant forces for the remaining sequences are updated using (9), again the most negative member (if existing) is removed to form the singleton $\mathbf{C}_{M+2}$ and the above procedure is repeated until

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

either no more negative members remain in $\mathbf{C}_j$ or it is a singleton.

The merging and splitting steps above are repeated until their execution doesn't lead to the formation of any new sets, or stopped after a pre-determined number of cycles (less than 7 cycles were enough for the algorithm to converge to unchanging set structures for all experiments described in the next section). The grouping of the face sequences into sets obtained after the algorithm is stopped, determines the final clustering result. Additionally, the resultant force acting on each member of a certain final set provides (if necessary) a measure in the range [0,...,1] for its membership in that set.

2.4. Incremental learning and online recognition

The batch version of the algorithm introduced in the previous section can be easily modified to operate in sequential mode. In the incremental version, each new face sequence available from the preprocessor is treated as a new singleton, which initiates a succession of new merge/split cycles as explained in 2.3, as a result of which it is either merged to some of the already existing sets or remains a singleton.

The same strategy can be used for online recognition or verification – each of the test sequences is treated as a new singleton, which initiates a succession of new merge/split cycles, and the category of the test sample is determined to be the same as the one of the cluster to which it is finally merged. The normalized resultant force acting on the test sample provides also a quantitative measure (in the range [0,...,1]) showing how reliable the decision is. In case the test sample is not merged to any of the existing clusters, it is rejected as a face which has not been learnt yet. Thus, in principle there is no explicit distinction between learning and recognition in our system.

3. Experiments

In order to evaluate the performance of the proposed method, several experiments have been conducted using over 500 face image sequences obtained during the last several months from 33 different subjects. A typical example of the experimental setting can be seen in Fig.1, and several time-subsampled face sequences for different people, together with time stamp labels obtained from the preprocessor, can be seen on Fig. 4. Illumination conditions were very demanding and varied significantly with the time of the day during which the samples were taken. The video sequences' length varied between 30-300 frames, depending on the speed at which the subjects walked in front of the camera, in the range between slow walking with occasional stops, and running. For each one of 17 of the subjects were gathered between 10 and 50 sequences, while less than 3 (typically only 1) sequences were available for the remaining 16 subjects (these were called "rare visitors").

Two different data sets were used in the following experiments:

(a) **Data set A**: in this data set, the subjects were just walking forward toward the camera. Predominantly frontal faces were included in this data set, with a few side-view faces at the end of the sequences, when the subjects passed beside the camera. Sequences T1, F1, K1, K3, R1, R3 on Fig. 4 are representative for the data included in this set;

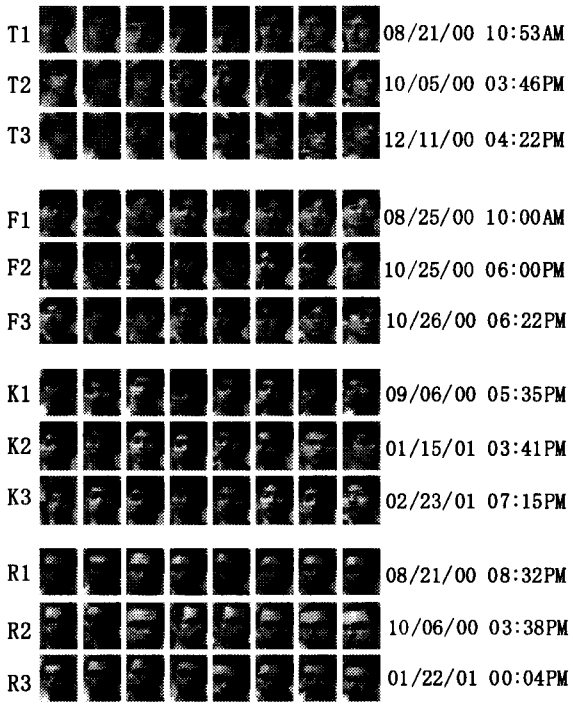

Figure 4. An example of several time subsampled face sequences with category labels (to be obtained by the algorithm) shown to the left, and the time stamp labels available from the preprocessor, shown to the right.

(b) **Data set B**: in this data set, the subjects were told to look to the left and right, up and down, as they moved towards the camera. Both frontal and side-view faces were represented in this data set. Sequences T2, T3, F2, F3, K2, R2 on Fig. 4 are representative for the data included in this set.

Samples with and without glasses were included for all subjects (except for the "rare visitors"), and hairstyles changed with time. Resolution of the original images was 320x240 pixels, and 18x22 pixels for the normalized face-only images. Near real-time processing was achieved on a SGI O2 workstation with R12000 (300MHz) processor.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

The following formula was used for calculating the recognition (self-organization) rate R:

$$R = (1.0 - \frac{E_{AB} + E_O}{N}) \times 100\%, \quad (18)$$

where N is the total number of sequences to be grouped, E_{AB} is the number of sequences which are mistakenly grouped into the cluster for certain category A, although in reality they come from category B, and E_O is the number of samples gathered in clusters in which no single category occupies more than 50% of the nodes inside them. The following 3 experiments were conducted, with results given in Table 1. In all experiments data from all 33 subjects were used.

Experiment 1 Only data from data set A were used where predominantly frontal faces were included.

Experiment 2 Only data from data set B were used, i.e. both frontal and side-view face images were included.

Experiment 3 Both data sets A and B (all data available until now) were used.

Data set	Sequences	E_{AB}	E_O	R (%)
A	277	5	3	97.1
B	275	12	38	81.8
A+B	536	17	40	89.3

Table 1. Experimental results

4. Conclusion

In this paper we have proposed a novel method for unsupervised face recognition from video sequences of time-varying face images obtained over an extended period of time in real-world conditions. The learning process implemented by the method does not rely on category-specific information provided by human teachers in advance, but rather lets the system find out by itself the structure and underlying relations inherent in the sensory input. The proposed method provides the following important advantages: (a) it allows all stages of the resulting system to be completely automated, avoiding the need for manual segmentation and labeling of the input stream, which might be biased by our limited understanding of the complex real-world environment. Moreover, manual segmentation and labeling of the input stream might be impractical and sometimes impossible, e.g. in on-line video surveillance systems; (b) this permits to train the system with a sufficient quantity of input data, providing the higher level of sensory variation necessary for such a challenging task as the one attempted here; (c) both frontal and side view faces can be learnt/recognized by the method; (d) the proposed method has a natural incremental implementation, allowing for "non-destructive" learning, which may be important in online systems dealing with large databases.

Results from several experiments using both frontal and side-view face sequences obtained under demanding illumination conditions were reported here, achieving recognition rate of 89.3% for the data set obtained until now. Although the preliminary results are encouraging (having in mind the difficulty of the task and the bottleneck of the features/distance measures used), additional tests with much larger data sets have to be done in order to obtain further insights about the limitations and possibilities of the present method.

Acknowledgment

The authors are grateful to Dr. K. Ishii and Dr. N. Hagita of NTT CS Laboratories for their help and encouragement.

References

[1] A. Samal and P. A. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, *Pattern Recognition* 25, pp. 65-77, 1992.

[2] R. Chellapa, C. L. Wilson, and S. Sirohey, Human and machine recognition of faces: a survey, *Proc. IEEE* 83, pp.705-740, 1995.

[3] M. A. Grudin, On internal representation in face recognition systems, *Pattern Recognition* 33, pp.1161-1177, 2000.

[4] H. Wechsler, P. J. Philips, V. Bruce, F. F. Soulie, and T. S. Huang, (eds.), *Face Recognition: From Theory to Applications*, Springer-Verlag, 1998.

[5] H. Ando, S. Suzuki and T. Fujita, Unsupervised visual learning of three-dimensional objects using a modular network architecture, *Neural Networks*, vol. 12, pp.1037-1053, 1999.

[6] J. J. Weng and W. S. Hwang, Toward Automation of Learning: The State Self-Organization Problem for a Face Recognizer, *Proc. 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pp.384-389, 1998.

[7] D. L. Swets and J. Weng, Hierarchical Discriminant Analysis for Image Retrieval, *IEEE Trans. PAMI*, 21(5), pp.386-401, 1999.

[8] S. M. Omohundro, Best-First Model Merging for Dynamic Learning and Recognition, in Moody, J. E., Hanson, S. J., and Lippmann, R. P., (eds.), *Advances in Neural Information Processing Systems 4*, pp. 958-965, San Mateo, CA: Morgan Kaufmann Publishers, 1992.

[9] S. Satoh, Comparative Evaluation of Face Sequence Matching for Content-based Video Access, *Proc. 4th Int. Conf. on Automatic Face and Gesture Recognition*, pp.163-168, 2000.

[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[11] B. S. Everitt, *Cluster Analysis*, Wiley, 1993.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley, 2001.

[13] H. A. Rowley, S. Baluja and T. Kanade, "Neural network based face detection," *IEEE Trans. Pattern Analysis abd Machine Intelligence*, 20(1): pp. 23-38, 1998.

[14] T. Darrell, G. Gordon, J. Woodfill and M. Harville, A Virtual Mirror Interface Using Real-Time Robust Face Tracking, *Proc. 3rd Int. Conf. on Automatic Face and Gesture Recognition*,pp.616-621,1998.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE