UNSUPERVISED FACE RECOGNITION FROM IMAGE SEQUENCES

B. Raytchev and H. Murase

NTT Communication Science Laboratories, 3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan

ABSTRACT

We propose a novel method for unsupervised face recognition from time-varying sequences of face images obtained in real-world environments. The method utilizes the higher level of sensory variation contained in the input image sequences to autonomously organize the data in an incrementally built graph structure, without relying on category-specific information provided in advance. This is achieved by "chaining" together similar views across the spatio-temporal facial manifolds by two types of connecting edges depending on a local measure of similarity. Experiments with real-world data gathered over a period of several months and including both frontal and side-view faces from 17 different subjects were used to test the method, achieving correct self-organization rate of 88.6%. The proposed method can be used in video surveillance systems or for content-based information retrieval.

1. INTRODUCTION

Although in recent years much attention has been drawn to and extensive research conducted in the area of automated face recognition, it still remains a domain in which humans significantly outperform computers, especially in real-time, unconstrained and unpredictable environments. A closer look at the problem reveals many discrepancies between the way humans learn faces and the way most computer-based face recognition procedures operate. For example, computer algorithms typically use *few isolated* samples from a large set of different face categories, taken in restricted environmental conditions, while humans utilize a multitude of *time-sequential* samples (in the form of a continuous input sensory stream) which are learnt in natural conditions including the whole spectrum of variations in illumination, viewing angles and object sizes which everyday life provides. Another important difference between human and machine learning is that while computers are provided with input which has been collected, segmented and classified in advance by human teachers (i.e. *supervised learning* is predominantly utilized), humans themselves learn by interacting directly with the sensory input from their environment (*unsupervised learning*). 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, rather than on any category-specific information accompanying it in a supervised manner. Also, 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 predefined set of features. Although the need for incremental learning, unsupervised self-organization of the internal state of the learning system and use of time-sequential data has already been pointed out by some researchers ([1]-[3]), a method for face recognition which takes into consideration all of these together and performs reasonably well on real-world data has not been demonstrated yet, to our knowledge.

Here we propose a novel method for unsupervised face recognition from time-varying sequences of face images. An incremental self-organizing learning process is implemented, which does not rely on class-specific information provided in advance. Several experiments, using data obtained in real-world conditions, were conducted in order to test the method. Expected areas of application of this method include visitor identification in surveillance systems, content-based face retrieval/annotation in multimedia applications, etc.

2. DESCRIPTION OF THE SYSTEM

The system proposed here operates in three stages. In the first *preprocessing stage,* from video sequences containing dynamic scenes of moving human subjects, the face area is automatically extracted, normalized and provided in the form of time-segmented face-only sequences to the next stage. In the second, *learning* stage, the algorithm introduced in section 2.2 is used to organize the face sequences into different category groups, without utilizing any category information provided in advance. *Recognition*, the last stage of our system, is performed by considering the result of attempting to assign a test input sequence to one of the category groups obtained during the learning stage.

0-7803-6725-1/01/$10.00 ©2001 IEEE

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

2.1. Preprocessing stage

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 for an example) 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 [4]).

2.2. Learning and recognition

Let $F^{(a)}(i,j,t)$ and $F^{(b)}(i,j,t)$ be two face image sequences, where a and b are sequence indexes ($a, b : 1 \ldots N$), i and j are image coordinates, and t is image frame number. For all available face image sequences $F^{(n)}$ compute the proximity matrix $\boldsymbol{P}$, whose elements $p\{a, b\}$ give the minimal distance between $F^{(a)}$ and $F^{(b)}$, i.e.

$$p\{a,\ b\} = \min_{x,y} dist\left(F^{(a)}(i,j,x),\ F^{(b)}(i,j,y)\right) = \min_{x,y} \sum_{i,j} T_d\{|\ F^{(a)}(i,j,x) - F^{(b)}(i,j,y)\ |\} \quad (1)$$

so that each face sequence is represented by a row in the symmetric $\boldsymbol{P}$. In (1), $T_d(x)$ is a threshold function with suitable threshold parameter d. More elaborate face distance measures than the one defined above might be used, if processing time is not a problem.

To all face image sequences $F^{(a)}$, $F^{(b)}$,..., $F^{(N)}$ are assigned "nodes" A, B,..., N, which will represent them in a graph, constructed and updated by the learning algorithm. In order to group the different face image sequences without using any category information provided in advance, two types of edges are used in the algorithm: "consistent" edges are used to connect nodes which belong to the same category (same subject), and "inconsistent" edges are used to discriminate between nodes belonging to different categories. The "consistency" of an edge is determined by the following consistency rules.

Consistency rules : Two nodes A and B can be connected by a consistent edge with length L, only if one of the following conditions is satisfied (which one is used depends on the current stage of the learning algorithm):

R0: $L < \min(C \times L_1,\ C \times L_2)$ (2)

R1: $L < \min(C \times L_3,\ C \times L_4)$ (3)

where L_1 is the length of the consistent edge between node A and its *nearest* neighbor (if it exists), L_2 - the length of the consistent edge between node B and its *nearest* neighbor, L_3 - the length of the consistent edge between node A and its *furthest* neighbor, L_4 - the length of the consistent edge between node B and its *furthest* neighbor, and C is a constant (C=1.4 has been used throughout). If the relevant consistency rule above is not satisfied, the edge between A and B is considered to be inconsistent.

First, an initial set of "subclusters" is formed by connecting with consistent edge each node A to a node B for which B= $\arg\min_X p\{A, X\}$. After the initial connections are performed for all nodes, each node will be connected to its nearest neighbor, and depending on the data in $\boldsymbol{P}$, initial chains (subclusters) of nodes will be formed. The subclusters are not connected between each other yet. Since initially all connections are "consistent", it is possible for some nodes to be connected to members of different category, simply because they happen to be their nearest neighbors, however big the distance between them might be. This necessitates the "consistency" of the initial connections to be checked using the more restricting *R0* in (2), and all edges which do not satisfy it are changed to inconsistent. After that, the subclusters are sorted in increasing order of the number of nodes in them, and the subcluster containing the least number of nodes is connected to a subcluster, the distance to which is minimal, by an edge whose consistency is determined by *R1* in (3). After the two clusters are connected, either with consistent or inconsistent edge, a new subcluster is formed by merging their nodes. The process of sorting and connecting subclusters is repeated recursively until all subclusters are merged into one big cluster. In the graph obtained from the execution of the procedure above, all nodes connected by consistent edges are considered to belong to the same face category, while inconsistent edges separate clusters be-

longing to different categories. Fig.2 shows an example of the resulting graph using some real data.

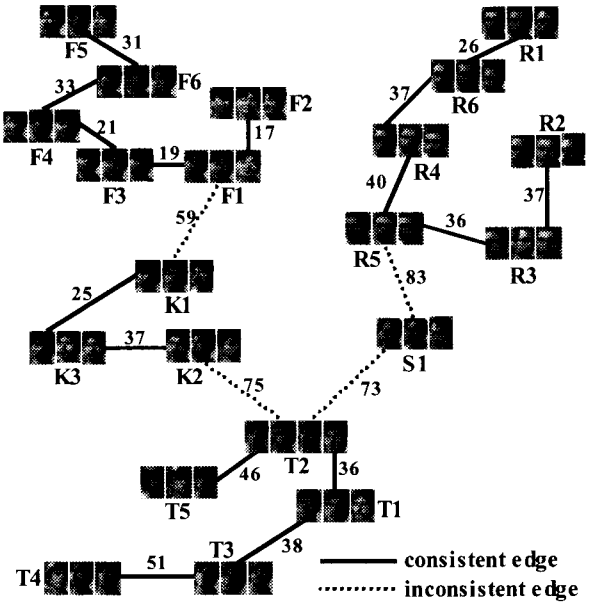

Fig. 2. An example of the graph obtained as a result of the learning algorithm. Different letters are used for nodes corresponding to face sequences from different categories. Edge lengths are shown near the edges.

When a new image sequence is available, its corresponding node is added to the graph in an incremental fashion, using the algorithm for incremental node addition below. It is assumed that the current internal state of the system is represented by the relations between N nodes in the graph, to which the newly available (N+1)st node has to be added (see Fig. 3).

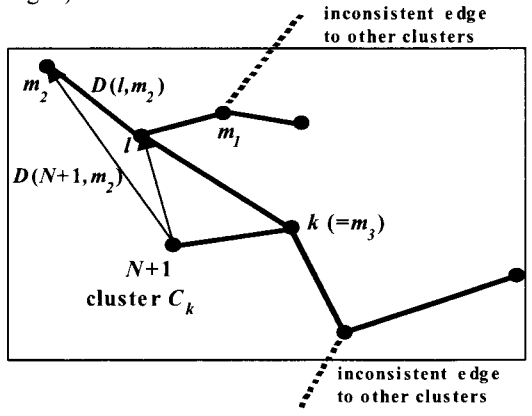

Fig. 3. Addition of a new node (steps 1-3)

Algorithm for incremental node addition

STEP 1 In the proximity matrix $\boldsymbol{P}$ compute the (N+1)st row corresponding to the new node.

STEP 2 Find the nearest neighbor, node k, of the new node, and using the consistency rule *R1* connect the two nodes with a consistent/inconsistent edge. In case the edge between them is inconsistent jump to step 4.

STEP 3 For all nodes l belonging to the same cluster as node k, compare the distances $D(l, m_i)$ between l and nodes m_i connected directly to l by a consistent edge, and the distances $D(N+1, m_i)$ between the new node and nodes m_i (see Fig. 3). If l and node N+1 are on the same side of the edge between m_i and l, and $D(l, m_i) > D(N+1, m_i)$ for some m_i, delete the edge between m_i and l, and insert a consistent edge between the new node and m_i. If m_i and node N+1 are on the same side of the edge between m_i and l, and $D(l, m_i) > D(l, N+1)$ for some m_i, delete the edge between m_i and l, and insert a consistent edge between the new node and l.

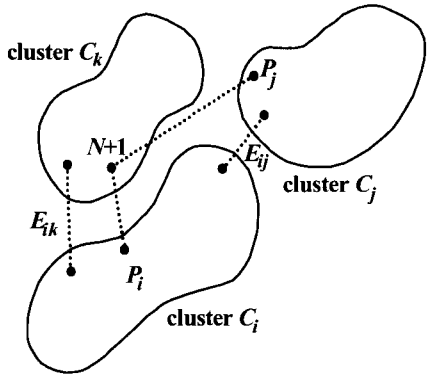

Fig. 4. Addition of a new node (step 4)

STEP 4 Let C_k be the cluster to which the new node belongs (see Fig. 4), and C_i be a cluster connected by inconsistent edges E_{ik} and E_{ij} to C_k and C_j (which may or may not be connected to C_k). Using the same logic as in step 3, the distances $D(N+1, P_i)$ and $D(N+1, P_j)$ between the new node and all nodes P_i in C_i and P_j in C_j are compared to E_{ik} and E_{ij}, and the necessary edge deletions/insertions performed (consistency is determined by *R1*), so that the resulting graph remains a minimal spanning tree in the inconsistent edges.

Recognition is performed in the same way as the node addition explained above, and the category of the test sample is determined to be the same as the one of the cluster to which it is connected by a consistent edge. In case the test sample is connected by an inconsistent edge, 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.

The learning algorithm explained above starts with some data gathered in advance, which are processed in an off-line "batch" manner, while subsequent additions of new data are done incrementally in an on-line manner. If the learning has to be on-line and incremental from the very beginning, the only change which has to be made is to connect the first two available nodes with a consistent edge (however far they might be from each other) and further incoming input data are added sequentially, node by node, using the node-addition algorithm above.

3. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed method, several experiments have been conducted using over 300 face image sequences obtained over the last 6 months from 17 different subjects. A typical example of the experimental setting can be seen in Fig.1. 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 100-600 frames, depending on the speed with which the subjects walked in front of the camera, in the range between slow walking with occasional stops, and running. Between 7 and 40 sequences were gathered for each subject. Two data sets were used: in the first (data set A), the subjects were just walking forward toward the camera, while in data set B the subjects were told to look to the left and right, up and down, as they moved towards the camera. Samples with and without glasses were included for all subjects, 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 SGI O2 workstation with R12000 (300MHz) processor. The following formula was used for calculating the recognition (self-organization) rate R:

$$R = (1.0 - \frac{E_{AB} + E_O}{N}) \times 100\%, \quad (4)$$

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 17 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	200	11	3	93.0
B	177	6	19	85.9
A+B	377	9	34	88.6

Table 1. Experimental results

4. CONCLUSION AND FURTHER WORK

In this paper we have proposed a novel method for unsupervised face recognition from a long video sequence of time-varying face images obtained over an extended period of time. The incremental learning process implemented by the method does not rely on category-specific information provided by human teachers in advance (which might be biased by their limited understanding of the complex real-world environment), but rather lets the system find out by itself the structure and underlying relations inherent in the sensory input. All stages of the system are completely automated, which makes it possible to train it 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. Results from several experiments using both frontal and side-view face sequences obtained under demanding illumination conditions were reported here, achieving recognition rate of 88.6% for the data set obtained until now. Although the preliminary results are encouraging (having in mind the difficulty of the task), 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] Ando, H., 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.

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

[3] Satoh, S., "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.

[4] Rowley, H. A., S. Baluja and T. Kanade, "Neural network based face detection," *IEEE Trans. PAMI*, 20(1): pp. 23-38, 1998.