Human–Computer Interaction — INTERACT '01
Michitake Hirose (Ed.)
Published by IOS Press, © IFIP, 2001

Unsupervised Learning of Faces for Human-Computer Interfaces

Bisser Raytchev and Hiroshi Murase

NTT CS Labs, 3-1, Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

Abstract: We propose a novel method for unsupervised face recognition from time-varying sequences of faces. 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 associations across spatio-temporal facial manifolds by two types of connecting edges depending on a local measure of similarity. Experiments with real-world data including both frontal and side-view faces were used to test the method, and encouraging results were observed.

Keywords: unsupervised incremental learning, face recognition, image sequences, human-computer interfaces

1 Introduction

Considering the fact that faces represent natural interfaces for humans, it seems inevitable that the more refined human-computer interfaces of the future will have to be provided with face recognition faculties. In spite of the extensive research conducted during the last several decades, 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, can be found in the different ways humans and computers learn:

(a) Humans learn by interacting directly with the sensory input from their environment. Category labels, like human names, are not essential for discrimination in the learning process (*unsupervised learning*). 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;

(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;

(c) Automatic face recognition is difficult because different people look more similar to each other in similar conditions (illumination, view angle, size, etc.) than same face in different conditions. 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 stream, containing the whole spectrum of variations in illumination, viewing angles, etc., which everyday life provides.

Although some researchers have already pointed out the need for incremental and unsupervised self-organization of the internal state of the learning system (Ando et al., 1999; Swets & Weng, 1999), or use of time-sequential data (Satoh, 2000), a method which takes into consideration all of these together and performs reasonably well on real-world data has not been demonstrated yet. Here we propose a new method for face recognition, *associative chaining*, inspired by observations (a)-(c) above. 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 associations (similar views) across spatio-temporal facial manifolds by two types of connecting edges depending on a local measure of similarity.

2 Learning by Associative Chains

2.1 Preprocessing of the input

The input to our system is assumed to be in the form of video sequences containing dynamic scenes of moving human subjects. The face area is automatically extracted, normalized and arranged

cally extracted, normalized and arranged into into time-segmented face-only image sequences.

2.2 Associative chaining (AC)

In the AC algorithm, to all available face image sequences are assigned "nodes" A, B,..., N, which will represent them in a graph, constructed and updated in the course of learning. 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), i.e. these nodes are considered to be linked by *consistent associations*, while "inconsistent" edges connect nodes from different categories. The "consistency" of an edge is determined by a local *consistency rule*, according to which two nodes A and B can be connected by a consistent edge with length L, only if L is not much larger than the average length of the edges of all nodes directly connected to A or B by consistent edges. An edge, which doesn't satisfy the consistency rule, is considered to be inconsistent. Any of the available face matching techniques can be used to define similarity between two faces, and here we assume that the distance between two nodes is defined as the distance between the most similar pair of faces taken from the corresponding sequences.

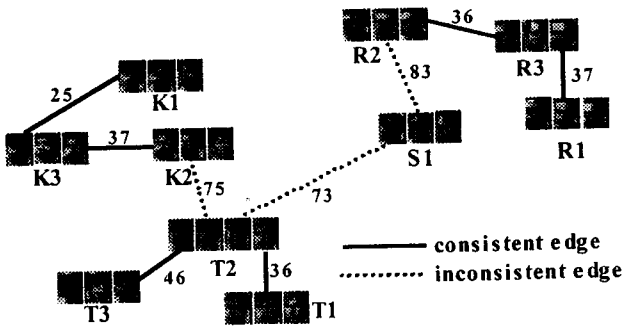

Figure 1: Part of the graph built using the learning algorithm. Edge lengths are shown near the edges.

2.3 MST formation

The algorithm starts by connecting each available node to its nearest neighbor with consistent edge, as a result of which initial chains (subclusters) of nodes are formed. These subclusters are sorted in increasing order, and the smallest subcluster is connected to the subcluster, the distance to which is minimal, by an edge with consistency determined by the consistency rule. Then their nodes are merged, forming a new subcluster. The process of sorting, connecting and merging is repeated recursively until no isolated subclusters remain. The resulting graph (Fig.1) is a minimum spanning tree (MST) with two types of edges, where all nodes connected by consistent edges are considered to belong to the same face category, while inconsistent edges separate clusters from different categories.

2.4 Incremental node addition

When a new image sequence is available from the input, a new node is added to the graph in an incremental fashion, using an algorithm for *incremental node addition*, which inserts new edges (consistency is determined by the consistency rule), or deletes existing ones if necessary, so that the resulting graph is again a MST.

3 Experimental Results

In order to evaluate the performance of the proposed method, several experiments have been done using about 370 face image sequences obtained over last 6 months from 17 different subjects. Illumination conditions were demanding and varied significantly. Recognition rate was 93.0% on data containing predominantly frontal faces (data set A), 85.9% on data containing both frontal and side view faces (data set B), and 88.6% on all available data (data set A+B).

4 Conclusion

In this paper we have proposed a novel method for unsupervised face recognition based on "chaining" of associations across spatio-temporal facial manifolds. Rather than relying on category-specific information provided by human teachers in advance (which might be biased by their limited understanding of the complex environment), the system autonomously learns the structure and underlying relations inherent in sensory input. Encouraging results were observed when the method was tested with data containing both frontal and side-view face image sequences obtained in real-world conditions.

Acknowledgment

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

References

Ando, H., Suzuki, S. & Fujita, T. (1999), Unsupervised visual learning of 3D objects using a modular network architecture, *Neural Networks*, 12, 1037-53.

Satoh, S. (2000), Comparative Evaluation of Face Sequence Matching for Content-based Video Access, *Proc. 4th Int. Conf. AFGR*, 163-8.

Swets, D. & Weng, J. (1999). Hierarchical Discriminant Analysis for Image Retrieval, *PAMI*, 21(5), 386-401.