Growing Gaussian Mixture Models for Pose Invariant Face Recognition

Ralph Gross, Jie Yang, Alex Waibel
{rgross, yang+, ahw}@cs.cmu.edu

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

A major challenge for face recognition algorithms lies in the variance faces undergo while changing pose. This problem is typically addressed by building view dependent models based on face images taken from predefined head poses. However, it is impossible to determine all head poses beforehand in an unrestricted setting such as a meeting room, where people can move and interact freely. In this paper we present an approach to pose invariant face recognition. We employ Gaussian mixture models to characterize human faces and model pose variance with different numbers of mixture components. The optimal number of mixture components for each person is automatically learned from training data by growing the mixture models. The proposed algorithm is tested on real data recorded in a meeting room. The experimental results indicate that the new method outperforms standard eigenface and Gaussian mixture model approaches. Our algorithm achieved as much as 42% error reduction compared to the standard eigenface approach on the same test data.

1 Introduction

While significant progress has been made with face recognition systems in the last decade [11], the application areas are still severely limited. Most efforts concentrate on the "Face in the Crowd" problem where a probe face is matched against a potentially huge gallery of known faces. The input images are usually of high quality with controlled lighting conditions displaying faces in a restricted number of views. While the galleries contain faces of thousands of different people, individual models are usually built using only a few pictures. Only recently researchers have begun to work on systems to identify people from video sequences [7, 9]. Aside from the increased computational demands of a real-time system, this task is challenging due to the variance created by the interaction of people with each other and the surrounding environment. We are interested in the specific context of a meeting room for which we developed a novel face recognition algorithm [6]. In this work we address the problem of pose invariant face recognition. We propose a new algorithm for the controlled growing of a Gaussian Mixture Model (GMM) which uses input images labeled with only the face identity. The algorithm learns the optimal number of mixture components automatically from training data.

The remainder of the paper is structured as follows. In Section 2 we give an overview of the meeting room environment. Section 3 introduces the algorithm used to build face models of the meeting participants. Section 4 describes the database of face images collected in our meeting room and the results of our experiments. Section 5 concludes with a summary of the presented work.

2 Meeting Room Environment

At the Interactive Systems Labs we are developing a multimodal meeting area [5] to continuously track, capture and integrate the important aspects of a meeting using the JANUS speech recognizer and a multimodal person identification module [13]. The identity of a meeting participant is currently determined using speaker identification and color appearance identification. We expect that we can increase the robustness of the person identification system by adding face recognition.

The automatic recognition of faces constitutes a particularly challenging pattern recognition task. This is due to the substantial changes in appearance faces undergo with different illumination, orientation, scale and facial expressions. The possibilities of restricting this variance in our meeting room are limited since we do not want to instruct the meeting participants to follow a specific behavior. The

0-7695-0750-6/00 $10.00 © 2000 IEEE

task of performing continuous face recognition in a room with more than one person creates a number of challenges:

- **Low quality video input**
  Given a set number of cameras in fixed locations, a wide viewing angle has to be used in order to cover the whole scene. This results in relatively low resolution images of the faces.

- **Illumination**
  Depending on the head pose and the position of a person relative to the overhead lights, the illumination of the face changes dramatically.

- **Unrestricted head pose and changing facial expressions**
  Given by the dynamic nature of a meeting almost any natural head pose and facial expression can and will occur.

- **Occlusion**
  People constantly move their heads and hands during a meeting. This may result in part or the whole face being obstructed by a hand, a piece of paper or other objects at times.

Compared with the remarkable human performance in recognizing faces from pictures [14] it is surprising to note that humans struggle to identify people on low quality video if they are not familiar with the faces they are given to identify [2].

3 Growing Gaussian Mixture Models

3.1 PCA Based Face Recognition

Among the numerous face recognition algorithms introduced in recent years, the eigenface approach proposed by Turk and Pentland [12] is one of the most influential ones. A face image, if interpreted as a vector, defines a point in a high dimensional space. Different face images share a number of similarities with each other, so the points representing these images are not randomly distributed in the image space. The key idea of the recognition process is to map the face images into an appropriately chosen lower dimensional subspace and perform classification by distance computation. If we restrict ourselves to a linear dimensionality reduction, the optimal solution is provided by principal component analysis [1]. The basis functions of the lower dimensional "face space" are formed by the eigenvectors of the covariance matrix of the set of training images corresponding to the largest eigenvalues. In the context of face recognition these eigenvectors are called "eigenfaces".

3.2 Pose Invariant Face Recognition

The eigenface approach works reasonably well only in 'mugshot' settings where the input space is restricted to frontal face images. An extension by Pentland et. al [10] deals with the problem of multiple head poses by building separate eigenspaces for nine different views. In the recognition stage they first determine the subspace which is most representative for the test image and then find the closest match between this image and a model in the chosen subspace.

A different approach was proposed by Graham and Allinson [4]. They built a common eigenspace from faces of all views and observed that a face which continuously changes pose between the two profile views forms a convex curve in the subspace. Using a radial basis function network they were able to exploit this fact and recognize faces in previously unseen views.

Cootes et. al [3] proposed Active Appearance Models, which combine shape and gray-level appearance. During localization a generic face model is deformed to fit the input face.

For all methods the training stage requires images of the subjects from various predefined views. In our meeting room environment the head movements of the participants are completely unrestricted. In order to avoid a lengthy registration process at the beginning of a meeting we require an automatic learning procedure which uses whatever images are available of a person. We would like to train the system with a short sequence of images from the beginning of the meeting and then update the recognizer during runtime within our multimodal people identification framework.

3.3 Growing Gaussian Mixture Models

Given a face image x and classes C_k we are interested in the probability $p(C_k|x)$, that x belongs to class C_k, where each class represents a different person. Using Bayes' rule we link the *posterior* probability $p(C_k|x)$ to the *class conditional* probability $p(x|C_k)$. Assuming equal class priors we can determine the most likely class for an image x with a maximum likelihood estimation

$$\begin{aligned} C^* &= \underset{k}{\operatorname{argmax}} p(C_k|x) \\ &\approx \underset{k}{\operatorname{argmax}} p(x|C_k) \end{aligned}$$

We model the class conditional probability $p(x|C_k)$ with a Gaussian mixture model:

$$p(x|C_k) = \sum_{j=1}^{M} p(x|j)P(j) \qquad (1)$$

where each mixture component $p(x|j)$ is a Gaussian distribution with mean μ_j and covariance matrix Σ_j:

$$p(x|j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)} \qquad (2)$$

Given a predefined number of mixtures M the models can be trained using the EM algorithm [1]. The problem lies in the determination of M. The input to the recognizer is comprised of a stream of face images captured in the meeting room. As people move freely about the room any head pose can occur. The number of different model categories and therefore the optimal choice of M is unknown. To address this problem a growing scheme for the GMM was implemented.

The process is initialized with a single mixture component for each model computed from the sample mean and covariance matrix of the training set. In order to reduce the number of parameters that have to be estimated in each step, diagonal covariance matrices $\Sigma_j = \sigma_j^2 I$ are used. As we perform principal component analysis prior to recognition we can assume that the different feature dimensions are uncorrelated.

The algorithm then proceeds in two steps. During bootstrapping the training samples are evaluated using the model for the respective class only. From a pool of samples with low probability we randomly draw a vector and use it as a seed point for a Gaussian mixture. The training set is clustered according to the previous mean and the new seed point using neural gas clustering [8], which is an extension to k-means clustering.

The model parameters are then re-estimated using the EM algorithm. After two bootstrapping iterations the algorithm trains the models discrimatively. It evaluates the training samples using all models and records those which are misclassified. From the misclassified samples the example with the highest probability is used as the starting point for a new Gaussian mixture. The same procedure of re-clustering and parameter re-estimation is then iteratively applied until a local minimum in the number of misclassified samples for each model is found. Figure 1 shows an overview over the training procedure.

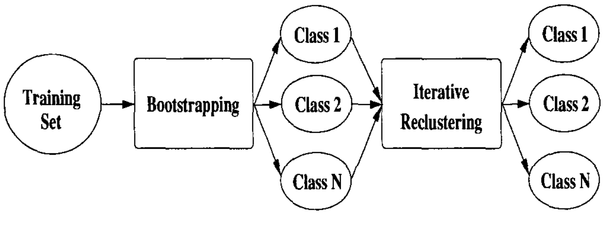

Figure 1. Overview of the training procedure.

4 Experiments

In order to evaluate our approach we recorded an internal group meeting and hand-labeled the face location of seven participants. Some examples of the images acquired are shown in Figure 2.

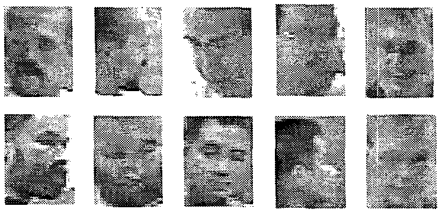

Figure 2. Examples from the dataset.

The images in the dataset vary in size between 18x24 and 38x50 pixels. Prior to feature extraction and recognition the images are normalized for size using bilinear interpolation and processed with standard preprocessing procedures (histogram equalization, lighting correction, normalization to zero mean and unit variance). The images are then projected into a lower dimensional eigenspace using a set of generic eigenfaces. We build non-overlapping training and test sets of face images from the beginning and the end of the meeting simulating the planned use of the system.

Table 1 compares the recognition rates of the growing Gaussian mixture model with two standard algorithms based on 500 training and 200 test images for each of the seven models. The first alternative procedure implements the conventional eigenface approach where the eigenspace representations of all training images are averaged to a single model vector. The second algorithm uses a "static" Gaussian mixture model with a varying, fixed number of mixtures. Our algorithm outperforms both methods.

Figure 3 shows the recognition rates for various numbers of mixture components for the static Gaussian mixture model compared with the other algorithms. For all numbers of mixture components the discriminativly grown GMM achieves higher recognition rates than the static mixture model. Finally, in Figure 4 the recognition rates for different numbers of training images are shown. Our algorithm consistently performs best.

5 Summary

This paper presented a new algorithm for growing a Gaussian mixture model to recognize face images acquired in a real world environment. Results obtained on data collected in our meeting room demonstrate that the new algorithm outperforms traditional approaches. Future work will combine this algorithm with our previously introduced

Algorithm	Recognition Rate
Growing GMM	75.79%
GMM (11 Mixtures)	70.14%
Standard Eigenface	58.07%

Table 1. Recognition rates for the different algorithms.

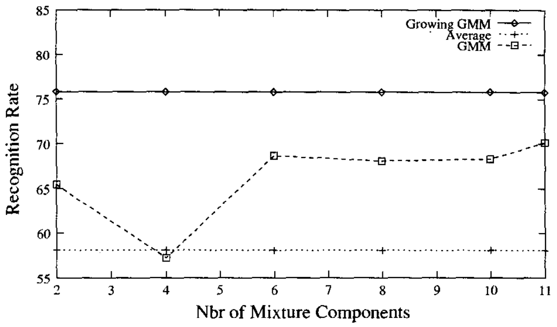

Figure 3. Recognition rates for different numbers of mixture components in a conventional Gaussian Mixture Model.

face recognition algorithm [6] and evaluate it on face images with varying degrees of occlusion, as typically seen in our meeting room environment. Work is under way to integrate both algorithms with a face tracker and our multimodal people identification system [13].

References

[1] C. M. Bishop. *Neural networks for pattern recognition.* Oxford Press, 1995.

[2] M. Burton, S. Wilson, and M. Cowan. Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science*, 1999.

[3] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, 1998.

[4] Daniel B. Graham and Nigel M. Allison. Face recognition from unfamiliar views: Subspace methods and pose dependency. In *3rd Conference on Automatic Face and Gesture Recognition*, pages 348 – 353, 1998.

[5] R. Gross, M. Bett, H. Yue, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Towards a multimodal meeting record. In *IEEE International Conference on Multimedia and Expo*, New York, July.

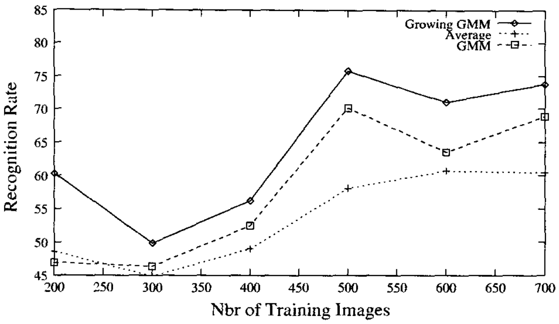

Figure 4. Recognition rates of the different algorithms for varying numbers of training images.

[6] R. Gross, J. Yang, and A. Waibel. Face recognition in a meeting room. In *Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, 2000.

[7] A. J. Howell and H. Buxton. Towards unconstrained face recognition from image sequences. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996.

[8] T. M. Martinetz, S. G. Berkovich, and K. Schulten. Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4), 1993.

[9] S. McKenna and S. Gong. Face recognition from sequences using models of identity. In *Proc. Asian Conference on Computer Vision*, 1998.

[10] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*, 1994.

[11] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *CVPR'97*, 1997.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[13] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel. Multimodal people id for a multimedia meeting browser. In *ACM Multimedia 99*, 1999.

[14] R. K. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, 1969.