Face Recognition in a Meeting Room

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

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

Abstract

In this paper, we investigate recognition of human faces in a meeting room. The major challenges of identifying human faces in this environment include low quality of input images, poor illumination, unrestricted head poses and continuously changing facial expressions and occlusion. In order to address these problems we propose a novel algorithm, Dynamic Space Warping (DSW). The basic idea of the algorithm is to combine local features under certain spatial constraints. We compare DSW with the eigenface approach on data collected from various meetings. We have tested both front and profile face images and images with two stages of occlusion. The experimental results indicate that the DSW approach outperforms the eigenface approach in both cases.

1. Introduction

While significant progress has been made with face recognition systems [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. Recently researchers have begun to work on systems to identify people from video sequences [6, 8]. 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 have developed a new face recognition algorithm. Our algorithm is capable of handling occlusions that typically appear during meetings.

The remainder of the paper is structured as follows. In Section 2 we give an overview of the meeting room environment in which we conducted our experiments. Section 3 introduces the new Dynamic Space Warping (DSW) algorithm along with the baseline Principal Component Analysis (PCA) approach to face recognition. In Section 4 we present the database of face images we 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

Face-to-face meetings usually encompass several modalities including speech, gesture, handwriting and person identification. Recognition and integration of each of these modalities is important to create an accurate record of a meeting for later reference. At the Interactive Systems Labs we are developing a multimodal meeting area [1] to continuously track, capture and integrate the important aspects of a meeting using the JANUS speech recognizer [16] and a multimodal person identification module [14]. The identity of a meeting participant is currently determined using a combination of speaker identification and color appearance identification. Our goal is to increase the robustness of the person identification system by adding face recognition.

The automatic recognition of faces constitutes a particularly difficult pattern recognition task. This is due to the substantial variations in appearance that faces undergo with changing illumination, orientation, scale and facial expressions. The possibilities of restricting this variance in our meeting room are limited since we can not restrict the meeting participants to follow specific behaviors. Therefore, the 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 limited 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. To capture high resolution pictures of a face it is necessary to closely track the person in question with a dedicated camera. With a larger number of people to be tracked and identified in a room, it becomes impossible to use a single camera per person.

- **Illumination**

  Depending on the head pose and the position of a person relative to the overhead lights, the illumination of the face changes dramatically. We can observe the full range of shade variations even though the overall lighting conditions in the room remain constant over the course of a meeting.

- **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 then results in the whole face or part of the face being obstructed by a hand, a piece of paper or other objects. Furthermore, depending on the number of cameras and their location the recognizer also has to cope with occlusion stemming from other people obstructing the field of view.

Figure 1 contains a collection of face images recorded during meetings, demonstrating these problems. Compared with the remarkable human performance in recognizing faces from pictures [2, 15] it is surprising to note that humans struggle to recognize people on low quality video if they are not familiar with the faces they are given to identify [4].

3. Face Recognition System

3.1. Local Versus Global Approaches to Face Recognition

Early computer vision systems for face recognition measured a set of geometric features in the face and compared the resulting vector with previously stored pattern (e.g. [7]). These local, feature based approaches have been superseded in recent years by global, template based algorithms. Empirical evidence suggests that algorithms based on whole face templates tend to outperform local approaches [3, 5].

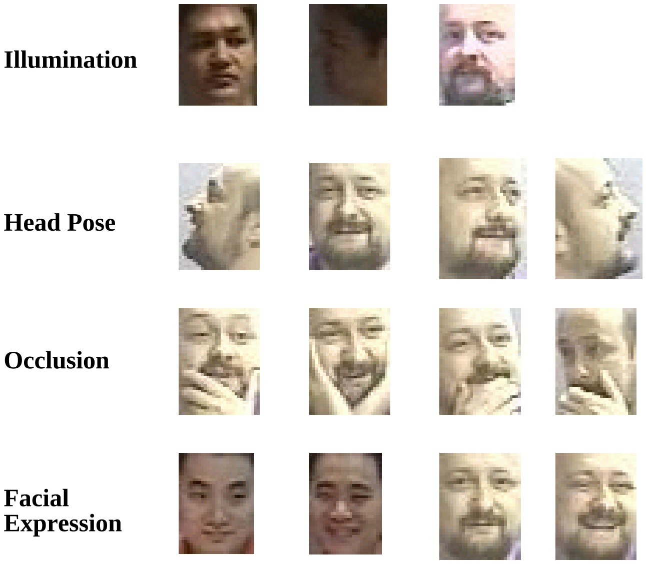

Figure 1. Enlarged face images illustrating the challenging meeting room environment.

3.2. PCA Based Face Recognition

Among the numerous global face recognition algorithms introduced in recent years, the eigenface approach proposed by Turk and Pentland [12] is one of the most influential. It uses principal component analysis to linearly project the high dimensional image space to a lower dimensional feature space. Once the eigenvectors of the covariance matrix which span the feature space are determined, recognition is performed by computing the Euclidean distances between the test image and the reference images in feature space. While the eigenface approach performs well in ‘mugshot’ settings, it has difficulties handling occlusions. The algorithm encodes an input image as single point in feature space and therefore has no means to recover from the distortion induced by occlusions. This effect can be seen in Figure 8.

3.3. Dynamic Space Warping

We propose a new face recognition algorithm which tries to overcome the shortcomings of the eigenface approach. Instead of projecting the input face onto a single point in feature space we use a moving window as depicted in Figure 2 to create a sequence of points. The window passes over the face from the upper left to the lower right corner.

For each face image Γ_i out of a set of training images $\Gamma_1, \Gamma_2, \ldots, \Gamma_m$ we create a vector of subimages $\Gamma_{\nu(i)} = \Gamma_i^1 \Gamma_i^2 \ldots \Gamma_i^{n_i}$. Based on the subimages $\Gamma_i^j, i = 1, \ldots, m, j = 1, \ldots, n_i$ we perform princi-

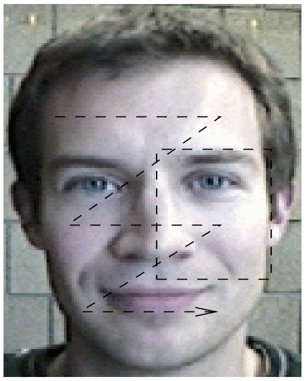

Figure 2. A sequence of subwindows over the face.

pal component analysis and project the vector $\Gamma_{\nu(i)}$ piecewise into the eigenspace. In our experiments the number of subimages is constant over all training images ($n_i = c$). The subimages Γ_i^j can be used in two ways in the PCA. Besides the obvious way of combining all subimages Γ_i^j into a single eigenspace it is also possible to build c different eigenspaces using the images $(\Gamma_1^1, \Gamma_2^1, \ldots, \Gamma_m^1), (\Gamma_1^2, \Gamma_2^2, \ldots, \Gamma_m^2), \ldots, (\Gamma_1^c, \Gamma_2^c, \ldots, \Gamma_m^c)$. However, our experiments did not show significant differences between these variants. The resulting sequence of points in feature space is stored as reference sequence for the given training image. A face image of unknown identity is compared with the stored reference sequences using dynamic programming, which makes the technique similar to dynamic time warping as used in speech recognition [9]. Due to this similarity we call it dynamic space warping. The subwindows i of the test sequence and the subwindows j of each template k define a set of grid points (i, j, k). Each grid point can be associated with a distance $d(i, j, k)$ between the respective subwindows. The algorithm now searches for the path through the grid points which provides the best match between the test pattern and a reference pattern. We define $D(i, j, k)$ as the minimum accumulated distance along any path leading to the grid point (i, j, k). With $D(1, j, k)$ initialized as follows:

$$D(1, j, k) = \Sigma_{n=1}^{j} d(1, n, k)$$

we can formulate the update rule as:

$$\begin{aligned} D(i, j, k) = d(i, j, k) + min\{&D(i-1, j, k), \\ &D(i-1, j-1, k), D(i, j-1, k)\} \end{aligned}$$

The best reference sequence is given by $min_k D(n, j, k)$ (with n being the length of the test sequence). Figure 3 depicts the different steps of the algorithm.

The size of the moving window and the vertical and horizontal offsets are determined automatically based on the size of the input images. Empirical evidence suggests to partition the face into nine overlapping regions. In contrast to other local approaches, DSW does not require the localization of facial landmarks such as eyes, nose or mouth.

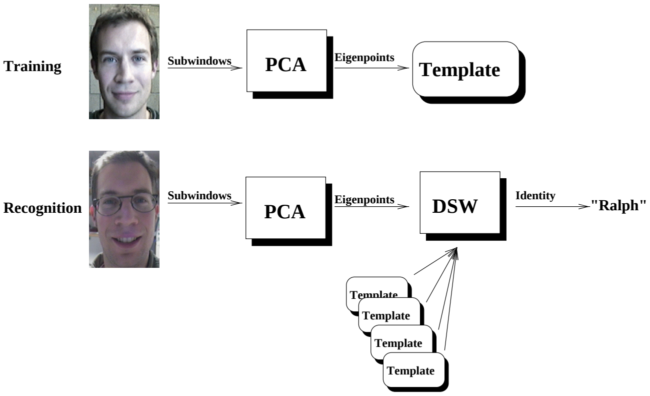

Figure 3. Processing steps of the DSW algorithm.

4. Experiments

4.1. Database

In order to evaluate our algorithm we recorded six group meetings in the meeting room. We then manually labeled face location, orientation, identity and degree of occlusion. The images in our dataset vary in size between 15x20 and 40x54. We normalize the size of the extracted images and perform a set of standard preprocessing procedures (histogram equalization, lighting correction, normalization to zero mean and unit variance). Using three different views per person (one frontal and two side views) we built models for six members of our group. The training images of all views were combined into a single eigenspace (parametric eigenspace [10]). The position of the meeting participants changed between meetings, therefore creating variance in views and illumination conditions for each face. Our database consisted of approximately 1200 pictures, averaging to about 60 pictures per model. We randomly selected training images out of the pool, built the models and tested on the remaining pictures. To assert the validity of the results we repeated this procedure and obtained the average recognition result.

4.2. Results

Figure 4 compares the recognition rates of the classical eigenface approach and DSW for varying numbers of training images. For both algorithms we evaluated two variants, termed PCA1, PCA2 and DSW1, DSW2 respectively. In the first version the pattern vectors resulting from the application of PCA or DSW on the training images for one model were averaged and only one reference vector was stored. For the second variant all vectors were retained.

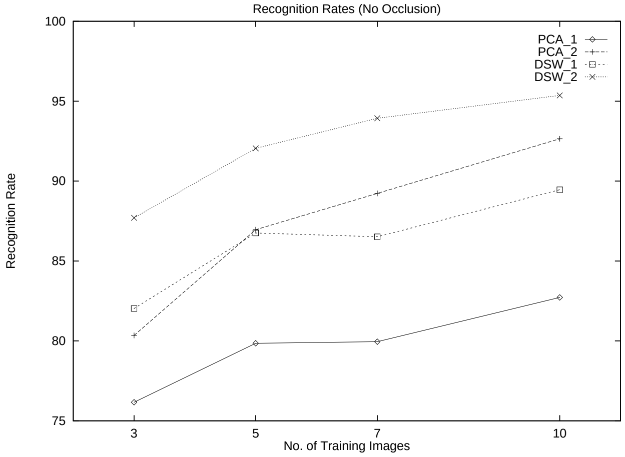

Figure 4. Recognition rates of PCA and DSW algorithms (no occlusion).

For both variants the DSW approach achieves higher recognition rates.

In addition to normal face images we also labeled approximately 150 faces with two stages of occlusions. Figure 5 depicts examples for both categories. The recognition rates obtained over those images are shown in Figures 6 and 7.

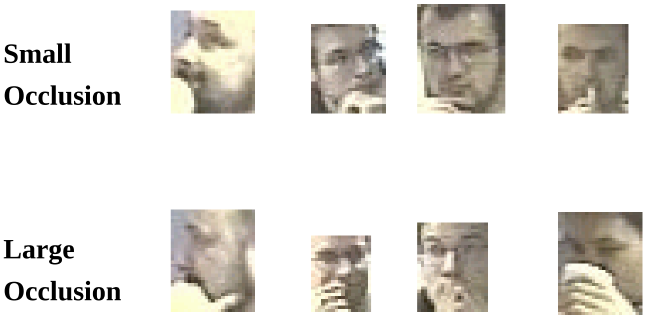

Figure 5. Examples of face images with small and large occlusions.

Again, the DSW approach clearly outperforms the standard eigenface algorithm. The results of the experiments using the first variant of PCA and DSW where only one reference vector is stored, are summarized in Table 1.

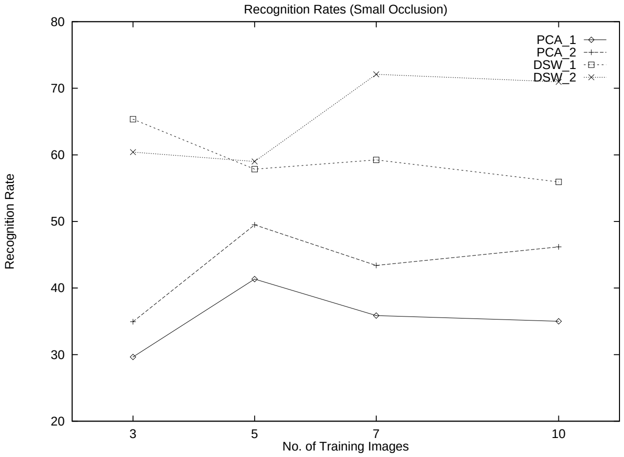

Figure 6. Recognition rates of PCA and DSW algorithms (with small occlusions).

#Train images	3	5	7	10
PCA1 w/o occl	76.2%	79.9%	80.0%	82.7%
DSW1 w/o occl	82.0%	86.8%	86.5%	89.4%
PCA1 sml occl	29.7%	41.4%	35.9%	35.0%
DSW1 sml occl	65.4%	57.9%	59.2%	55.9%
PCA1 lrg occl	25.3%	31.6%	29.0%	30.8%
DSW1 lrg occl	45.5%	49.9%	47.5%	48.6%

Table 1. Comparison of the recognition rates for PCA1 and DSW1 on different databases.

The localized approach of DSW enables the algorithm to deal better with local occlusions than standard PCA can. Figure 8 demonstrates this observation with reconstructed face images. In this procedure the original image is first projected into the eigenspace and then reconstructed using the eigenspace representation and the eigenface basis of the feature space. The figure shows original images as recorded during meetings and their counterparts reconstructed from a PCA and a DSW representation. For faces without occlusion the reconstructed images bare a strong resemblance to the originals. If parts of the face are occluded by a hand or a pen the face images reconstructed from a PCA eigenspace show strong distortions while the images obtained from DSW are still remarkably clear.

The system in its current stage is a first step towards a robust face identification system that is capable of handling real world situations that occur during meetings. Work is currently under way to integrate the face recognizer with a face tracker developed in our lab [13] and with the multimodal people ID system [14]. The face tracker is able to track multiple faces in the field of view in real time. When integrated with the face tracker we will be able to train more

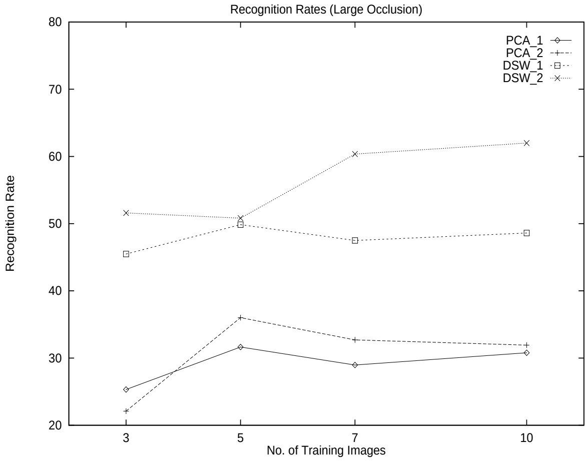

Figure 7. Recognition rates of PCA and DSW algorithms (with large occlusions).

robust models by utilizing the vast amounts of data available from many recordings of our meetings.

5. Conclusion

We presented a new algorithm for the recognition of faces under adverse conditions and showed empirical evidence of its improved performance with respect to the standard eigenface approach. While our system is able to handle occlusions, the low quality of the input images and the changing illumination conditions, the number of views in the experiments we are reporting on is restricted. Given the low quality of the input images, we believe that building a 3D head model from the data to normalize for different view directions is not feasible. It therefore seems more promising to investigate into view tolerant algorithms which build different models for different views.

References

[1] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Multimodal meeting tracker. Under Review.

[2] E. Brown, K. Deffenbacher, and W. Sturgill. Memory for faces and the circumstances of encounter. *Journal of Applied Psychology*, 1977.

[3] R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10), 1993.

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

[5] S. Gutta, J. Huang, D. Singh, I. Shah, B. Takacs, and H. Wechsler. Benchmark studies on face recognition. In *Proceedings of International Workshop on Automatic Face and Gesture Recognition*, 1995.

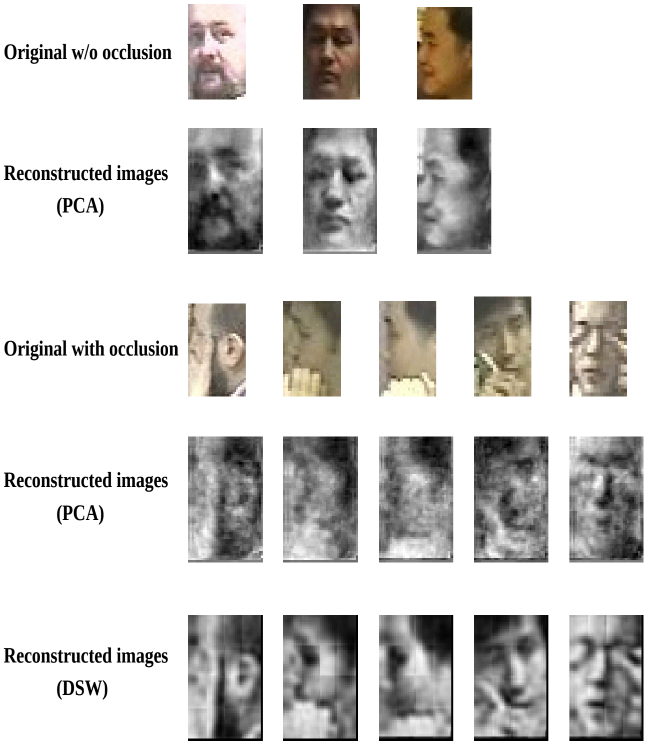
Original w/o occlusion
Reconstructed images (PCA)
Original with occlusion
Reconstructed images (PCA)
Reconstructed images (DSW)

Figure 8. Face images and their counterpart obtained through reconstruction from PCA and DSW.

[6] A. 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.

[7] T. Kanade. Picture processing by computer complex and recognition of human faces. Technical report, Dept. of Information Science, Kyoto University, 1973.

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

[9] H. Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2), 1984.

[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. Phillips, H. Moon, P. Rauss, and S. 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 and A. Waibel. A real-time face tracker. In *Proceedings of WACV'96*, 1996.

[14] J. Yang, X. Zhu, R. Gross, J. Kominek, and A. Waibel. Multimodal people ID for a multimedia meeting browser. In *Proceedings of ACM Multimedia*, 1999.

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

[16] H. Yu, M. Finke, and A. Waibel. Progress in automatic meeting transcription. In *Proceedings of the Eurospeech '99*, 1999.