Electronic Letters on Computer Vision and Image Analysis 5(1):1-13, 2005

An Experimental Investigation about the Integration of Facial Dynamics in Video-Based Face Recognition

Abdenour Hadid and Matti Pietikäinen

Machine Vision Group, Infotech Oulu and Department of Electrical and Information Engineering
P.O. Box 4500 FIN-90014 University of Oulu, Finland

Received 14 December 2004; accepted 1 February 2005

Abstract

Recent psychological and neural studies indicate that when people talk their changing facial expressions and head movements provide a dynamic cue for recognition. Therefore, both fixed facial features and dynamic personal characteristics are used in the human visual system (HVS) to recognize faces. However, most automatic recognition systems use only the static information as it is unclear how the dynamic cue can be integrated and exploited. The few works attempting to combine facial structure and its dynamics do not consider the relative importance of these two cues. They rather combine the two cues in an $ad\ hoc$ manner. But what is the relative importance of these two cues separately? Does combining them enhance $systematically$ the recognition performance? To date, no work has extensively studied these issues. In this article, we investigate these issues by analyzing the effects of incorporating the dynamic information in video-based automatic face recognition. We consider two factors (face sequence length and image quality) and study their effects on the performance of video-based systems that attempt to use a spatio-temporal representation instead of one based on a still image. We experiment with two different databases and consider HMM (the temporal hidden Markov model) and ARMA (the auto-regressive and moving average model) as baseline methods for the spatio-temporal representation and PCA and LDA for the image-based one. The extensive experimental results show that motion information enhances also automatic recognition but not in a systematic way as in the HVS.

Key Words: Face Recognition, Facial Dynamics, Principal Component Analysis (PCA), Hidden Markov Models (HMM), Auto-Regressive and Moving Average(ARMA), View-Based Recognition.

1 Introduction

While current face recognition systems perform well under relatively controlled environments [21, 16], they tend to suffer when variations in pose, illumination or facial expressions are present. On the other hand, the human visual system (HVS) has remarkable capabilities to recognize faces even under poor viewing conditions. Naturally, the human perception uses not only the facial structure to recognize faces but also additional cues such as color, facial motion, contextual knowledge etc.

Correspondence to: <hadid@ee.oulu.fi>

Recommended for acceptance by Fabio Roli
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

Importantly, recent psychological and neural studies show that facial movement supports the face recognition process, especially in degraded environments [15]. Inspired by these findings, researchers and developers have recently attempted to exploit facial dynamics to enhance still image based face recognition. However, most of these attempts do not exploit facial dynamics very efficiently, but apply still image based techniques to some "good" frames selected from face sequences [2]. In [5], for example, a system for face recognition from video is proposed. It is based on tracking the positions of the nose and eyes. The locations of these three points are used to decide whether the orientation of the face is suitable for face recognition. If they form an equilateral triangle, then image-based recognition is launched; otherwise the tracking continues until a "good" frame occurs. It is clear that this approach exploits only the abundance of frames in the video sequence and not the facial dynamics. By "$facial\ dynamics$" we refer to the non-rigid movement of facial features, in addition to the rigid movement of the whole face (head). Therefore, in order to more efficiently exploit the temporal information one must choose a form of spatio-temporal representation that incorporates both the facial structure and its dynamics. Some approaches to video-based face recognition using this principle include the condensation method and the method based on hidden Markov models (HMMs), which have been successfully applied to video-based face recognition [23, 13]. Recently, linear dynamical system model [19] has been also proposed and used to capture the spatio-temporal information in face sequences [1].

The few works attempting to combine facial structure and its dynamics do not consider the relative importance of these two cues. They rather combine the two cues in an $ad\ hoc$ manner. Does combining them (without considering the relative importance of each) enhances $systematically$ the recognition performance? To date, no work has extensively studied these issues. In this article, we investigate this by analyzing the effects of incorporating the dynamic information in video-based automatic face recognition. For this purpose, we analyze how the length of the face sequences and the quality of the images affect the performance of video-based face recognition. We consider the temporal HMM approach [13] and the auto-regressive and moving average model (ARMA)[19, 1] as baseline methods for the spatial-temporal representation and PCA [20] and LDA [4] for the image-based one. We perform extensive experiments using two different video databases: MoBo [6] and Honda/UCSD [10]. The considered subset from MoBo database contains 96 face sequences of 24 different subjects walking on a treadmill while the subset from Honda/UCSD database includes 20 individuals moving their heads in different combinations of 2-D and 3-D rotation, expression and speed.

The rest of this paper is structured as follows. First, we summarize the neuropsychological findings related to the importance of facial dynamics in the human visual system. Then, in Section 3, we review different methods which attempt to "$truly$" incorporate facial dynamics in the recognition process. Section 4 presents the data and the baseline methods that are used in the experiments. In order to check whether a spatio-temporal representation enhances face recognition performance, we present in Section 5 our approach of building a view-based face recognition scheme using only the static information on the video sequences. Experiments on the effects of the face sequence length and image quality are presented in Section 6. Finally, Section 7 contains discussion and concludes the paper.

2 Role of Facial Dynamics: Neuropsychological Evidence

We summarize here the main findings in psychophysics and neuroscience that have a direct relevance to research on automatic face recognition. The studies [8, 15] indicate that:

- **(i)** Both static and dynamic facial information are useful for recognition.
- **(ii)** People rely primarily on static information because facial dynamics provide less accurate identification information than static facial structure.
- **(iii)** Dynamic information contributes more to recognition under a variety of degraded viewing conditions (such as poor illumination, low image resolution, recognition from distance etc.)

(iv) Facial motion is learned more slowly than static facial structure.

(v) Facial motion contributes to recognition by facilitating the perception of the 3-D structure of the face.

(vi) Recognition of familiar faces is better when they are shown as an animated sequence than as a set of multiple frames without animation. However, for unfamiliar faces, the moving sequence does not provide more useful information than multiple static images.

How can we interpret and exploit these findings to enhance the performance of automatic face recognition systems? A possible indication from the statements in (i) and (iii) is that motion is a useful cue to enhance the performance of static image based systems. Importantly, the usefulness of the motion cue increases as the viewing conditions deteriorate (statement (iii)). Such an environment is often encountered in surveillance and access control applications. Thus, an automatic recognition system should exploit both dynamic and static information. From the evidence in (ii) and (iii), we can interpret that motion and static information do not have the same importance as the role of motion depends on a number of factors such as the familiarity with the faces (statement (vi)), the viewing conditions (statement (iii)) etc. Thus, depending on the situation, the automatic systems should bias the role of each cue rather than integrate them with fixed weights. Finally, we can see the statement in (v) as an indication of using structure from motion in the recognition process.

3 Automatic Face Recognition From Videos: An Overview

Based on the way of considering motion information, we can classify automatic face recognition evolution into three categories, as shown in Table 1.

In the first category, only one (or a few) static image(s) are available for recognition. An example of such an application is mug-shot matching, which includes the recognition of faces in driver's licenses, passports, credit cards etc. Typically, the images are of good quality and the imaging conditions are controlled. Therefore, segmenting and recognizing the face is relatively easy. The second category concerns a wider range of applications, such as access control and video surveillance, where the images are generally obtained from video sequences. The algorithms in this class, in contrast to the first one, are faced with new challenges since they generally deal with small low quality images. Nevertheless, they have an advantage from the abundance of frames in the videos. Despite the fact that both static and dynamic information are available in this category of algorithms, most research has limited the scope of the problem to the use of still image based methods to some selected frames while some other approaches have adopted 3-D construction and recognition via structure from motion or structure from shading. It is clear that both schemes do not fully exploit facial dynamics as they use mainly the spatial information contained in the video sequences. Only recently have researchers started to "$truly$" address the problem of face recognition from video sequences. These algorithms, belonging to the third category, attempt to simultaneously use the spatial and temporal information for recognizing moving faces. Comprehensive surveys on face recognition evolution, especially for the first two categories, can be found in [2, 21]. We focus the rest of this section on reviewing and discussing the third class.

Class	Input	Method	Use of motion
Class 1	Static images	Still image-based	No
Class 2	Video	Still image-based	Partially
Class 3	Video	Spatio-temporal	Yes

Table 1: Classification of face recognition algorithms according to their integration of motion information

In [12], an approach exploiting spatio-temporal information is presented. It is based on modeling face dynamics using identity surfaces. Face recognition is performed by matching the face trajectory that is constructed

from the discriminating features and pose information of the face with a set of model trajectories constructed on identity surfaces. Experimental results using 12 training sequences and the testing sequences of three subjects were reported with a recognition rate of 93.9%.

In [11], Li and Chellappa used the trajectories of tracked features to identify persons in video sequences. The features are extracted using Gabor attributes on a regular 2D grid. Using a small database of 19 individuals, the authors have reported performance enhancement over the frame to frame matching scheme. In another work, Zhou and Chellappa [22] proposed a generic framework to track and recognize faces simultaneously by adding an identification variable to the state vector in the sequential important sampling method.

An alternative way to model the temporal structures is the use of the condensation algorithm. This algorithm has been successfully applied for tracking and recognizing multiple spatio-temporal features. Recently, it has been extended to video-based face recognition problems [23, 22].

Hidden Markov models have been also applied to model temporal information and perform face recognition [13]. During the training phase, an HMM is created to learn both the statistics and temporal dynamics of each individual. During the recognition process, the temporal characteristic of the face sequence is analyzed over time by the HMM corresponding to each subject. The likelihood scores provided by the HMMs are compared. The highest score provides the identity of a face in the video sequence.

Recently, the auto-regressive and moving average (ARMA) model [19] has been adopted to model a moving face as a linear dynamical system and perform recognition [1]. Other researchers have also presented some approaches for exploiting the dynamic characteristics of contiguously moving faces in image sequences. For example, recently K. C. Lee et al. [10] have proposed an approach to video-based face recognition using probabilistic appearance manifolds.

The above works mainly aimed to propose representations which combine both shape and dynamics without taking in consideration the relative importance of these two cues. This tendency is due to the fact that both shape and dynamics contribute to face recognition. But what is the relative importance of these two cues separately? Does combining them enhance $systematically$ the recognition performance? To date, no work has extensively studied these issues.

4 Experimental Data and Baseline Algorithms

4.1 Experimental Data

Typically, video-based face recognition simultaneously involves three steps: segmentation, tracking and recognition of the faces. However, our goal in this paper is to analyze how to represent the faces for recognition rather than develop a full video-based face recognition system. Therefore, we focus our experiments only on the recognition phase, assuming that the faces are well segmented.

Thus, we considered two different databases: MoBo [6] and Honda/UCSD [10].The MoBo (Motion of Body) database is the most commonly used database in video-based face recognition research [22, 9, 13], although it was originally collected for the purpose of human identification from distance. The considered subset* from MoBo database contains 96 face sequences of 24 different subjects walking on a treadmill. 4 different walking situations are considered: slow walking, fast walking, incline walking and carrying a ball. Some example frames are shown in Fig.1. Each sequence consists of 300 frames. From each video sequence, we cropped the face regions, obtaining thus images of 40*40 pixels. Examples of extracted faces from a video sequence are shown in Fig.2.

The second database, Honda/UCSD, has been collected and used by K. C. Lee et al. in their work on video-based face recognition [10]. It was also used in the recent study of Aggarwal et al. [1]. The considered subset from Honda/UCSD database contains 40 video sequences of 20 different individuals (2 videos per person). During the data collection, the individuals were asked to move their face in different combinations (speed,

*Note that the original MoBo database contains 99 videos corresponding to 25 different individuals. Since there is a missing video, we considered only the 24 individuals who have 4 videos each.

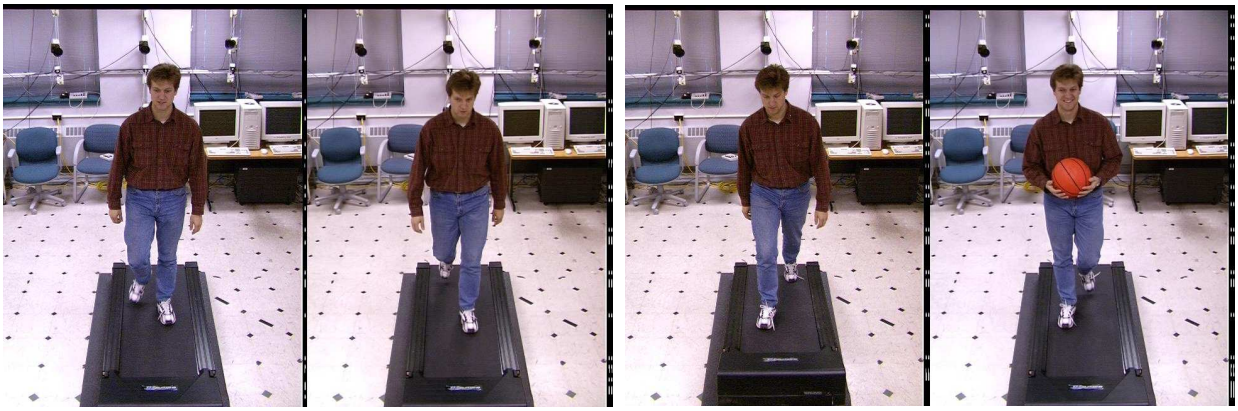

Figure 1: Example frames from MoBo database

rotation and expression). From the video sequences, we cropped the face images in the same way as we did for the MoBo database. The size of the extracted face images is 20*20 pixels.

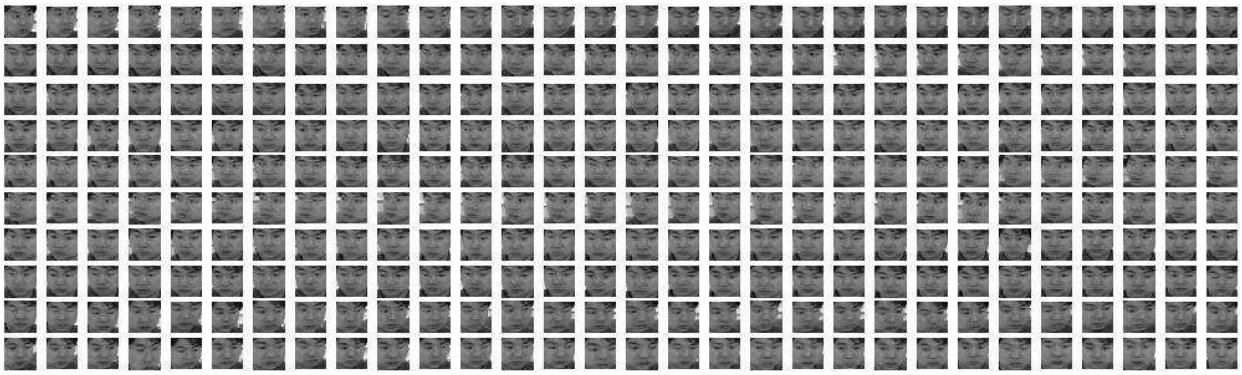

Figure 2: Examples of extracted faces from a video sequence (MoBo database)

4.2 Baseline Algorithms for Video-based Recognition

4.2.1 HMMs

The principle of using HMMs to model the facial dynamics and perform video-based face recognition is quite simple [13, 17]. Let the face database consist of video sequences of P persons. We construct a continuous hidden Markov model for each subject in the database. A continuous HMM, with N states $\{S_1, S_1, , ..., S_N\}$, is defined by a triplet $\lambda = (A, B, \pi)$, where $A = \{a_{ij}\}$ is the transition matrix, $B = \{b_i(O)\}$ are the observation probability density functions (pdf) and $\pi = \{\pi_i\}$ are the initial distributions. The model λ is built using a sequence of feature vectors, called observation sequence $O = \{o_1, o_2, ..., o_T\}$, extracted from the frames of the video sequence (T is the number of frames). Different features can be extracted and used as observation vectors (e.g. pixels values, DCT coefficients etc.). In [13], the PCA projections of the face images were considered.

Let the state at time t be q_t, then:

$$A = \{a_{ij} \mid a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)\} \tag{1}$$

$$B = \left\{ b_i(O) \mid b_i(O) = \sum_{m=1}^{M} C_{im} N(O, \mu_{im}, U_{im}) \right\} \tag{2}$$

$$\pi = \{\pi_i \mid \pi_i = P(q_1 = S_i)\ , 1 \leq i \leq N\} \tag{3}$$

where C_{im} is the mixture coefficient for the m^{th} mixture in state i , $N(O, \mu_{im}, U_{im})$ is a Gaussian pdf with mean vector μ_{im} and covariance matrix U_{im} and M is the number of components in the Gaussian mixture model.

During the training, a model λ_p , ($p = 1, 3, ..., P$), is built for all the subjects in the gallery. During the testing, given the gallery models $\{\lambda_1, \lambda_2, ..., \lambda_P\}$ and the sequence of the PCA feature vectors $O = \{o_1, o_2, ..., o_T\}$, the identity of the test face sequence is given by:

$$arg_p\ (max_p\ P(O|\lambda_p)) \tag{4}$$

In other terms, the likelihood scores $P(O|\lambda_p)$ provided by the HMMs are compared, and the highest score defines the identity of the test video sequence.

4.2.2 ARMA

In the ARMA (Auto-Regressive and Moving Average) framework, a moving face is represented by a linear dynamical system and described by Eqs.5 and 6:

$$x(t+1) = Ax(t) + v(t) \qquad v(t) \sim N(0, R) \tag{5}$$

$$I(t) = Cx(t) + w(t) \qquad w(t) \sim N(0, Q) \tag{6}$$

where, $I(t)$ is the appearance of the face at the time instant t, $x(t)$ is a state vector that characterizes the face dynamics, A and C are matrices representing the state and output transitions, $v(t)$ and $w(t)$ are IID sequences driven from some unknown distributions.

We build an ARMA model for each face video sequence. To describe each model, we need to estimate the parameters A, C, Q and R. Using the tools from the system identification literature, the estimatation of the ARMA model parameters is closed form and therefore easy to implement [19, 1]. While the state transition A and the output transition C are intrinsic characteristics of the model, Q and R are not significant for the purpose of recognition [19]. Therefore, we need only the matrices A and C to describe a face video sequence. Once the models are estimated, recognition can be performed by computing distances between ARMA models corresponding to probe and gallery face sequences. The gallery model which is closest to the probe model is assigned as the identity of the probe (nearest neighbor criteria).

Several distance metrics have been proposed to estimate the distance between two ARMA models [3]. Since it has been shown that the different metrics do not alter the results significantly, we adopt in our experiments the Frobenius distance (d_F^2), defined by :

$$d_F^2 = 2\sum_{i=1}^{n} sin^2\theta_i \tag{7}$$

where, θ_i are the subspace angles between the ARMA models, defined in [3].

5 Building a View-Based Face Recognition Scheme

In video-based face recognition schemes both training and test data (galleries and probes) are video sequences. The recognition consists of matching the spatio-temporal representation extracted from the probe videos to those extracted from the galleries. In order to check whether a spatio-temporal representation enhances face recognition performance, one should compare the results to those obtained using still image based techniques

under the same conditions. However, performing still-to-still face recognition when the data consists of video sequences is an ill-posed problem. Notice that the Face Recognition Vendor Test (FRVT2002) [16], which recently added video-based tests, has adopted a methodology that is more suitable for comparing performances of commercial systems (which is the main goal of the FRVT) rather than comparing video-to-video versus still-to-still face recognition.

Here we adopted a new scheme to perform static image based face recognition that exploits the abundance of face views in the videos. The approach consists of performing unsupervised learning to extract the most representative samples (or exemplars) from the raw gallery videos. Once these exemplars are extracted, we build a view-based system and use a probabilistic voting strategy to recognize the individuals in the probe videos. The probabilistic voting strategy consists of combining[†] the recognition confidences in every frame to decide on the person identity in the probe video sequence.

Thus, given a training face sequence G such as that shown in Fig. 2,

$$G = \{G_{face_1}, G_{face_2}, ..., G_{face_T}\} \tag{8}$$

we are interested in selecting the most representative samples (or exemplars)

$$E = \{e_1, e_2, ..., e_K\} \tag{9}$$

in order to consider them as models for appearance-based face recognition. The desirable samples are those that summarize the content of the face sequence G. In other words, they should capture the within-class variability due to illumination changes, poses, facial expressions and other factors.

A straightforward approach is to apply K-means directly to the data and pick up one or a few sample(s) from each cluster. In such a way, one may not find meaningful clusters especially for complex and high-dimensional data. Our approach, however, is based on two steps: first embedding the face images in a low-dimensional space in which "similar" faces are close to each other, and then applying the K-means clustering algorithm. The exemplars can be defined then as the cluster centers. Instead of using the classical manifold learning and dimensionality reduction techniques, we adopted the recently proposed LLE algorithm to represent the faces in a low-dimensional space.

In short, LLE [18] is an unsupervised learning algorithm that maps high dimensional data onto a low-dimensional, neighbor-preserving embedding space. Considering a face sequence G and organizing the faces into a matrix X (where each column vector represents a face), the LLE algorithm involves the following three steps:

1. Find the nearest neighbors of each point X_i.

2. Compute the weights W_{ij} that best reconstruct each data point from its neighbors, minimizing the cost in Eq.10

$$ReconstructError(W) = \sum_{i=1}^{length(G)} \left\| X_i - \sum_{j \in neighbors(i)} W_{ij} X_j \right\|^2 \tag{10}$$

3. Compute the embedding Y (of lower dimensionality $d << D$, where D is the dimension of the input data) best reconstructed by the weights W_{ij} minimizing the quadratic form in Eq.11 :

$$\Phi(Y) = \sum_{i}^{length(G)} \left\| Y_i - \sum_{j \in neighbors(i)} W_{ij} Y_j \right\|^2 \tag{11}$$

The aim of the two first steps of the algorithm is to preserve the local geometry of the data in the low-dimensional space, while the last step discovers the global structure by integrating information from overlapping local neighborhoods. The details of the algorithm can be found in [18]. An example of LLE embedding

[†]Note that there are many ways to combine the recognition confidences such as "sum", "product" etc. Here, we considered the sum of the confidences.

of a face sequence in 2-D is shown in Fig. 3. Once the embedding is computed, K-means is performed and the exemplars are thus defined as the cluster centers. The results of applying K-means to a face sequence in a 2-D embedded space is also shown in Fig. 3.

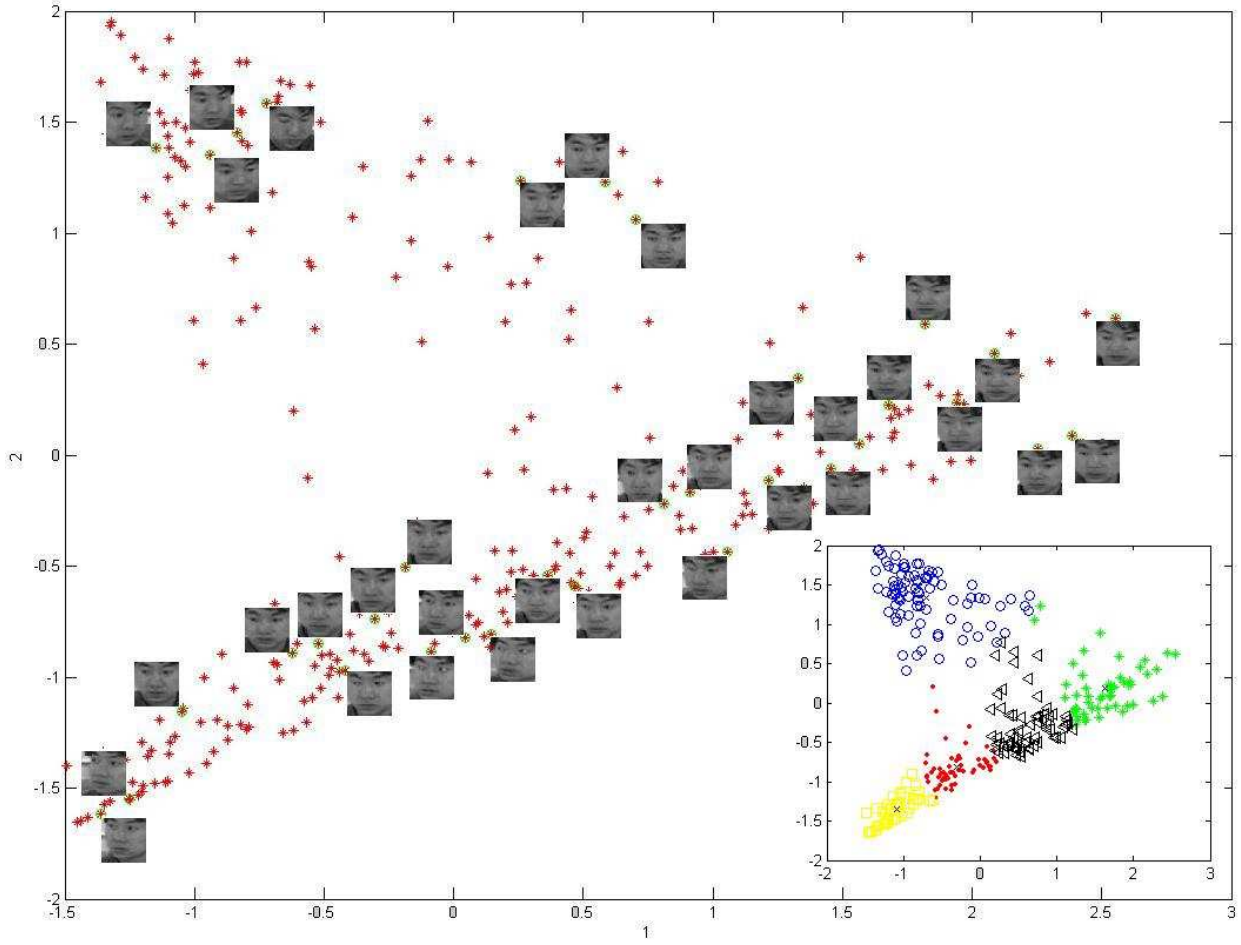

Figure 3: An example of embedding the face sequence in Figure 2 in a 2-D space using LLE. Although the intrinsic dimensionality of the faces is higher than two, LLE provides quite a good embedding. The result of applying the K-means is also shown

Once the set of exemplars are extracted from each video sequence, we use them as training samples for appearance-based recognition. To determine the identity of the probe video B, we use a probabilistic voting strategy over all frames in B. The probabilistic voting strategy consists of combining the recognition confidences in every frame to decide on the person identity in probe video B. In contrast, a majority voting scheme consists of identifying the face in every frame and then performing a majority voting to decide on the identity of the person in the sequence. Another alternative is to perform recognition only on some selected good frames. In [14], it is argued that the probabilistic voting strategy performs better than other alternatives.

Thus, we applied the proposed approach and extracted $K = 5$ exemplars from each training video and built a view-based scheme using PCA and LDA as baseline methods for still image based face recognition (more details on extracting the face models and building the view-based recognition system can be found in our recent work [7]). Since the MoBo database contains videos in 4 different situations, we considered one situation for training and the others for testing. We report the average recognition rates for the 4 combinations: 1 training situation/3 testing situations. The second database (Honda/UCSD) contains 40 videos of 20 individuals (2

videos per subject). We considered one video for training and the other for testing. The two first rows in Table 2 summarize the recognition rates on both databases.

6 Effect of Face Sequence Length and Image Quality on Recognition

Our first goal is to analyze the effects of the face sequence length on both spatio-temporal and image-based representations. For this purpose we have considered the temporal HMM [13] approach and the ARMA model [1] as baseline methods for spatio-temporal analysis while PCA [20] and LDA [4] are used as baselines for the still image based analysis. As explained in the previous section, we adopted the locally linear embedding approach to extract the exemplars and the probabilistic voting strategy for recognizing the faces in the still image based scenarios. For the temporal HMM, we used 30 eigenvectors for dimensionality reduction and a 16-state fully connected HMM (See Section 4.2.1). We summarize in Table 2 the performance of the spatio-temporal representation (HMM and ARMA) and their static image based counterpart (PCA and LDA). We noticed, as shown in the table, that the four methods performed quite well but the spatio-temporal representations outperformed the PCA and LDA methods on both databases. It is early to make any conclusion from the present results since it is not clear whether the good performances of the HMMs and ARMA are due to the combination of facial structure and its dynamics or due to their different modeling of the facial structure.

	MoBo	**Honda-UCSD**
PCA	87.1 %	89.6 %
LDA	90.8 %	86.5 %
HMM	92.3 %	91.2 %
ARMA	93.4 %	90.9 %

Table 2: Recognition rates using all probe frames (MoBo and Honda/UCSD databases)

In the above experiments, we considered all frames of the probe videos (i.e. 300 frames for both databases). However, in a real application, a subject may appear in front of a camera only for a short duration while some other subjects may stay longer. Therefore, the length of the face sequence can be as small as a few frames or as long as hundreds or thousands of frames. To analyze the effect of face sequence length on recognition performance, we conducted a set of experiments where we used only a portion (L among M frames) of the probe videos for testing. Thus, for a given probe video B, we extracted $S = 10$ sub-sequences of length L as follows:

$$B = \{B_{face_1}, B_{face_2}, ..., B_{face_M}\} \tag{12}$$

$$Set_L = \{\{B_{frame_i}, B_{frame_{i+1}}, ..., B_{frame_{L+i-1}}\}\}, \tag{13}$$

where $i = Random(1, (M - L + 1))$ and $M = 300$.

Therefore, we extracted $S = 10$ sub-videos of length L from each probe video. We performed extensive experiments with different values of L and the results are shown in Figures 4 and 5. The results indicate that for short sequences, the performance of the HMM-based system deteriorates while the image-based systems (PCA and LDA) are less sensitive to this factor. This can be explained by the fact that short sequences do not contain enough dynamic information to discriminate between the individuals. Another possible explanation might be also that the HMMs need sequences which are long enough in order to be trained [17]. Analyzing the performance of the ARMA approach, we noticed also better results for longer face sequences. The ARMA method performed better than the HMM approach especially for short face sequences.

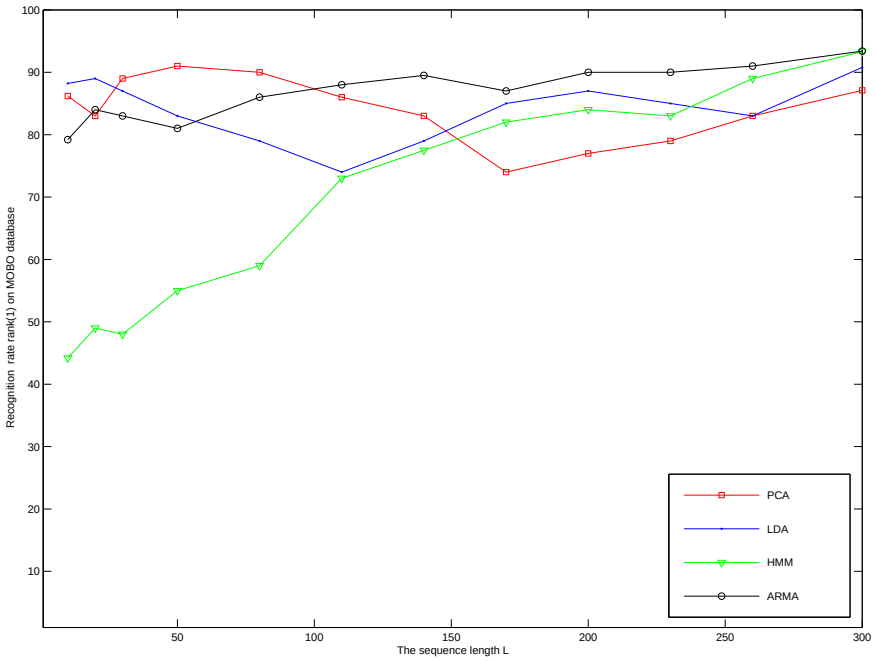

Figure 4: Recognition rates for different face sequence lengths on MoBo database

This means that, when the shape and dynamics cues are combined without consideration of their relative importance, a joint spatio-temporal representation is not $always$ efficient in case of short face sequences. The HMM-based system did not perform as well as its PCA and LDA-based counterparts on both databases. A similar conclusion for the ARMA-based system on the MoBo database can be made. This is an interesting and important result since one might $always$ expect better performance using the joint representation ($which\ is\ not\ actually\ the\ case$). However, as we increase the length of the face sequence, the superiority of the HMM and ARMA approaches becomes clear. The recognition rates on MoBo database increased from 79.2% to 93.4% for the ARMA method and from 44.2% to 92.3% for the HMM-based approach. This confirms the evidence that facial dynamics support face recognition. On the Honda/UCSD database, the recognition rates increased from 80.2% to 90.9% for the ARMA method and from 69.2% to 91.2% for the HMM-based approach.

Additionally, we performed a set of experiments to check how image resolution affects the recognition rates. We downsampled each face image in the MoBo database from 40*40 to 20*20 and then to 10*10 pixels. We noticed that recognition rates decrease for all the methods (see Table 3). However, it seems that the HMM and ARMA are least affected by image quality.

Resolution	**40*40**	**20*20**	**10*10**
PCA	87.1 %	81.3 %	60.6 %
LDA	90.8 %	79.5 %	56.5 %
HMM	92.3 %	85.2 %	71.2 %
ARMA	93.4 %	84.1 %	74.2 %

Table 3: Recognition rates for different face image resolutions using the MoBo database

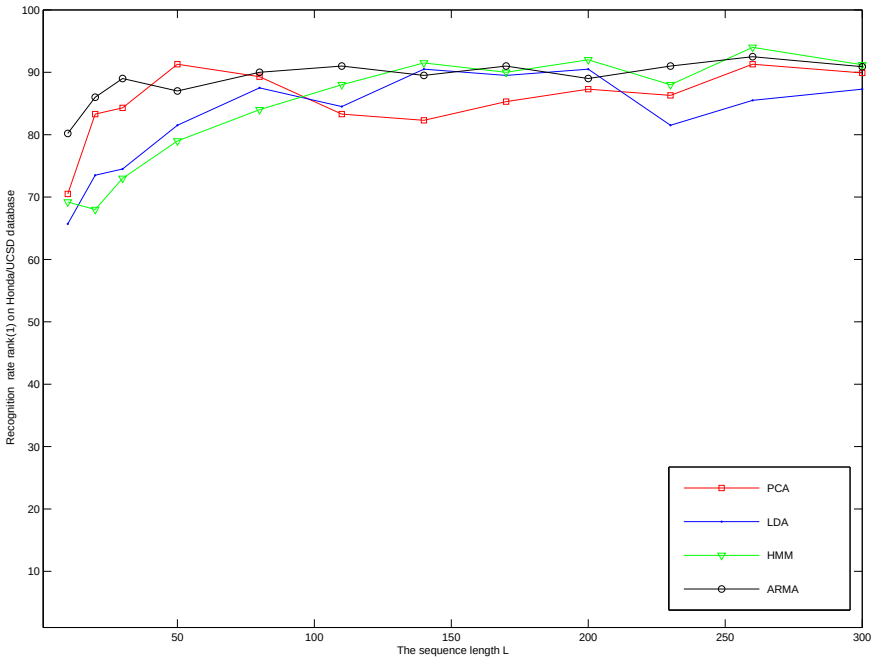

Figure 5: Recognition rates for different face sequence lengths on Honda/UCSD database

7 Discussion

Psychological and neural studies indicate that dynamic personal characteristics support and do not hinder face recognition in the human visual system (HVS). However, it is unclear how this dynamic cue is integrated and combined with the static facial information. In this work, we considered the automatic face recognition and analyzed the effects of incorporating this dynamic information. We considered two factors (face sequence length and image quality) and studied their effects on the performance of video-based systems that attempt to use a spatio-temporal representation instead of one based on a still image.

In most of the experiments the HMM- and ARMA-based approaches outperformed their PCA and LDA-based counterparts. This is in agreement with the evidence discussed in Section 2, which state that facial dynamics are useful for recognition. However, for short sequences, HMM gave poor results. This is probably due to the fact that HMMs need quite long sequences in order to be trained [17]. However, we noticed also that the ARMA-based system performed better with longer face sequences than with shorter ones. For short sequences, the ARMA approach gave worse results than its PCA and LDA counterparts on the MoBo database. Importantly, one may not expect worse results using spatio-temporal representations. However, the obtained results attest that PCA- and LDA-based representations might perform better in such cases. This means that the spatio-temporal representations did not succeed in discovering the importance of the spatial cue over its temporal counterpart. This leads us to the conclusion that combining face structure and its dynamics in an $ad\ hoc$ manner (i.e. without considering the relative importance of each cue) does not $systematically$ enhance the recognition performance.

The experiments also showed that image quality affects both representations but the image-based methods are more affected. Therefore, in cases of face recognition applications with low-quality images, a spatio-temporal representation is more suitable. Again, this is in agreement with the neuropsychological findings that indicate that facial movement contributes more to the recognition under degraded viewing conditions.

If we refer to the evidence discussed in Section 2, we notice that the role of facial dynamics depends on several factors such as the familiarity of the face, the viewing conditions, etc. Accordingly, the human visual system adapts the contribution of the facial dynamics. However, in automatic face recognition, the contribution of this cue is integrated in the joint spatio-temporal representation and generally not adapted (not biased) to the given situation. For instance, in our experiments, joint representation did not increase the contribution of the facial dynamics for low-resolution images and long face sequences and did not decrease this contribution for higher image resolution and shorter face sequences. This suggests that the existing spatial-temporal representations have not yet shown their full potential and need further investigation.

In our experiments, the benefit of using joint representation is noticeable but not very significant. This is due to the fact that the facial movement in both databases is almost limited to the rigid motion of the head. However, one may expect more benefit when the persons are also making non-rigid movements with their facial features (such as when the subjects are talking).

Acknowledgment

We would like to thank Dr. R. Gross for providing us the MoBo database and Dr. K-C Lee for the Honda/UCSD database. This research was sponsored by the Academy of Finland and the Finnish Graduate School in Electronics, Telecommunications and Automation (GETA). Finally, we gratefully acknowledge the comments of the anonymous reviewers.

References

[1] G. Aggarwal, A. R. Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. In *Proc. the 17th International Conference on Pattern Recognition*, volume 1, pages 175–178, August 2004.

[2] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. In *Proc. IEEE*, volume 83(5), pages 705–740, May 1995.

[3] K. Cock and B. Moor. Subspace angles and distances between arma models. In *Proc. of the Fourteenth International Symposium on Mathematical Theory of Networks and Systems*, June 2000.

[4] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America*, 14:1724–1733, 1997.

[5] D. O. Gorodnichy. On importance of nose for face tracking. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 181–186, May 2002.

[6] R. Gross and J. Shi. The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.

[7] A. Hadid and M. Pietikäinen. Selecting models from videos for appearance-based face recognition. In *Proc. the 17th International Conference on Pattern Recognition*, volume 1, pages 304–308, August 2004.

[8] B. Knight and A. Johnston. The role of movement in face recognition. *Visual Cognition*, 4:265–274, 1997.

[9] V. Krueger and S. Zhou. Exemplar-based face recognition from video. In *Proc. European Conf. on Computer Vision*, pages 732–746, May 2002.

[10] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 313–320, 2003.

[11] B. Li and R. Chellappa. Face verification through tracking facial features. *Journal of the Optical Society of America*, 18:2969–2981, 2001.

[12] Y. Li. *Dynamic Face Models: Construction and Applications*. PhD thesis, Queen Mary, University of London, 2001.

[13] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 340–345, 2003.

[14] S. McKenna and S. Gong. Recognizing moving faces. In H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 578–588. SpringerVerlag, Berlin, 1998.

[15] A. J. O'Toole, D. A. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science*, 6:261–266, 2002.

[16] P. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. Face recognition vendor test 2002 results. Technical report, 2003.

[17] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2), pages 257– 286, 1989.

[18] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[19] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 439–446, Vancouver, BC, Canada, July 2001.

[20] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[21] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 34(4):399–458, Dec. 2003.

[22] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *Proc. European Conf. on Computer Vision*, pages 681–697, May 2002.

[23] S. Zhou, V. Krueger, and R. Chellappa. Face recognition from video: A condensation approach. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 221–228, May 2002.