# Learning Personal Specific Facial Dynamics for Face Recognition From Videos

Abdenour Hadid[1], Matti Pietikäinen[1], and Stan Z. Li[2]

[1] Machine Vision Group, P.O. Box 4500, FI-90014, University of Oulu, Finland
http://www.ee.oulu.fi/mvg

[2] Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Donglu
Beijing 100080, China

**Abstract.** In this paper, we present an effective approach for spatiotemporal face recognition from videos using an Extended set of Volume LBP (Local Binary Pattern features) and a boosting scheme. Among the key properties of our approach are: (1) the use of local Extended Volume LBP based spatiotemporal description instead of the holistic representations commonly used in previous works; (2) the selection of only personal specific facial dynamics while discarding the intra-personal temporal information; and (3) the incorporation of the contribution of each local spatiotemporal information. To the best of our knowledge, this is the first work addressing the issue of learning the personal specific facial dynamics for face recognition.
We experimented with three different publicly available video face databases (MoBo, CRIM and Honda/UCSD) and considered five benchmark methods (PCA, LDA, LBP, HMMs and ARMA) for comparison. Our extensive experimental analysis clearly assessed the excellent performance of the proposed approach, significantly outperforming the comparative methods and thus advancing the state-of-the-art

**Key words:** Facial Dynamics, Local Binary Patterns, Face Recognition, Boosting

## 1 Introduction

Psychological and neural studies [1] indicate that both fixed facial features and dynamic personal characteristics are useful for recognizing faces. However, despite the usefulness of facial dynamics, most automatic recognition systems use only the static information as it is unclear how the dynamic cue can be integrated and exploited. Thus, most research has limited the scope of the problem by applying methods developed for still images to some selected frames [2]. Only recently have researchers started to truly address the problem of face recognition from video sequences [3–9].

In [3], an approach exploiting spatiotemporal 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

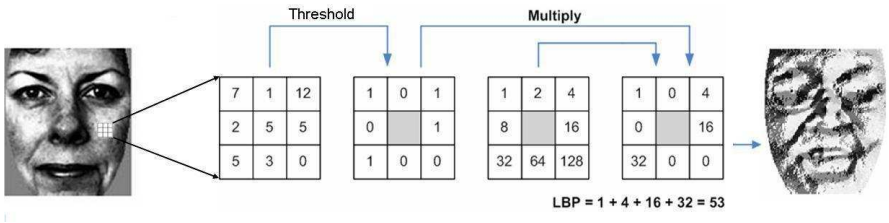


**Fig. 1.** Example of an LBP calculation

surfaces. Experimental results using 12 training sequences and the testing sequences of three subjects were reported with a recognition rate of 93.9%.

In [4], 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 reported performance enhancement over the frame to frame matching scheme. In another work, Zhou and Chellappa 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 [5].

An alternative to model the temporal structures is the use of the condensation algorithm. This algorithm has been successfully applied for tracking and recognizing multiple spatiotemporal features. Recently, it was extended to video based face recognition problems [6, 5]. More recently, the Auto-Regressive and Moving Average (ARMA) model [10] was adopted to model a moving face as a linear dynamical system and perform recognition [7].

Perhaps, the most popular approach to model temporal and spatial information is based on the Hidden Markov models (HMM) which have also been applied to face recognition from videos [8]. The idea is simple: in the training phase, an HMM is created to learn both the statistics and temporal dynamics of each individual. During the recognition process, the temporal characteristics of the face sequence are 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.

Unfortunately, most of the methods described above use spatiotemporal representations that suffer from at least one of the following drawbacks: (1) the local information which is shown to be important to facial image analysis [11] is not well exploited with holistic methods such as HMMs; (2) while only personal specific facial dynamics are useful for discriminating between different persons, the intra-personal temporal information which is related to facial expression and emotions is also encoded and used; and (3) equal weights are given to the spatiotemporal features despite the fact that some of the features contribute to recognition more than others. To overcome these limitations, we propose an effective approach for face recognition from videos that uses local spatiotemporal features and selects only the useful facial dynamics needed for recognition. The idea consists of looking at a face sequence as a selected set of volumes (or rectangular prisms) from which we extract local histograms of Extended Volume Local

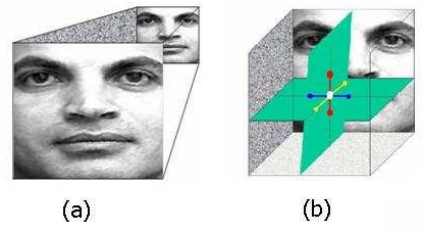


**Fig. 2.** (a): A face sequence is seen as a rectangular prism and (b): An example of 3D neighborhood of a pixel in Volume LBP

Binary Pattern (EVLBP) code occurrences. Our choice of adopting LBP (Local Binary Patterns) for spatiotemporal representation is motivated by the recent results of LBP approach [12] in facial image analysis [13] and also in dynamic texture recognition [14].

In this paper, noticing the limitations of volume LBP operator in handling the temporal information, we first extend the operator and derive a rich set of volume LBP features denoted EVLBP. Then, instead of ignoring the weight of each feature or simply concatenating the local EVLBP histograms computed at predefined locations, we propose an effective approach for automatically determining the optimal size and locations of the local rectangular prisms (volumes) from which EVLBP features should be computed. More importantly, we select only the most discriminative spatiotemporal EVLBP features for face recognition while discard the features which may hinder the recognition process. For this purpose, we use AdaBoost learning technique [15] which has shown its efficiency in feature selection task. The goal is to classify the EVLBP based spatiotemporal features into intra and extra classes, and then use only the extra-class information for recognition. To the best of our knowledge, this is the first work addressing the issue of learning personal specific facial dynamics for face recognition.

## 2 Extended Volume LBP Features (EVLBP)

The LBP texture analysis operator, introduced by Ojala *et al.* [16, 12], is defined as a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. It is a powerful means of texture description and among its properties in real-world applications are its discriminative power, computational simplicity and tolerance against monotonic gray-scale changes.

The original LBP operator forms labels for the image pixels by thresholding the $3 \times 3$ neighborhood of each pixel with the center value and considering the result as a binary number. Fig. 1 shows an example of an LBP calculation. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. Each bin (LBP code) can be regarded as a micro-texton. Local primitives which are codified by these bins include different types of curved edges, spots, flat areas etc.

The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel $(x_c, y_c)$ is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^P \tag{1}$$

where $g_c$ corresponds to the gray value of the center pixel $(x_c, y_c)$, $g_p$ refers to gray values of $P$ equally spaced pixels on a cicrle of radius $R$, and $s$ defines a thresholding function as follows:

$$s(x) = \begin{cases} 1, if\ x\ \geq\ 0; \\ 0, otherwise. \end{cases} \tag{2}$$

The occurrences of the LBP codes in the image are collected into a histogram. The classification is then performed by computing histogram similarities. For an efficient representation, facial images are first divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. In such a description, the face is represented in three different levels of locality: the LBP labels for the histogram contain information about the patterns on a pixel-level, the labels are summed over a small region to produce information on a regional level and the regional histograms are concatenated to build a global description of the face. This locality property, in addition to the computational simplicity and tolerance against illumination changes, are behind the success of LBP approach for facial image analysis [13].

The original LBP operator (and also its later extension to use neighborhoods of different sizes [12]) was defined to deal only with the spatial information. For spatiotemporal representation, Volume LBP operator (VLBP) has been recently introduced in [14]. The idea behind VLBP is very simple. It consists of looking at a face sequence as rectangular prism (or volume) and defining the neighborhood of each pixel in three dimensional space. Fig. 2 explains the principle of rectangular prism and shows an example of 3D neighborhood for Volume LBP.

There are several ways of defining the neighboring pixels in VLBP. In [14], $P$ equally spaced pixels on a circle of radius $R$ in the frame $t$, and $P + 1$ pixels in the previous and posterior neighboring frames with time interval $L$ were used. This yielded in VLBP operator denoted VLBP$_{L,P,R}$. Fig. 3 (top) illustrates an example of VLBP operator with $P$=4 and $R$=1.

We noticed in our experiments on face recognition from videos that VLBP$_{L,P,R}$ does not encode well enough the temporal information in the face sequences since the operator considers neighboring points only from three frames and therefore the information in the frames with time variance less than L are missed out. In addition, a fixed number of neighboring points (i.e. $P$) are taken from each of the three frames, yielding in a less flexible operator with large set of neighboring points. To overcome these limitations, we introduce here an extended set of VLBP patterns by considering $P$ points in $frame\ _t$, $Q$ points in the $frames\ _{t\pm L}$ and $S$ points in the $frames\ _{t\pm 2L}$. This yields in Extended Volume LBP (EVLBP) operator that we denote by EVLBP$_{L,(P,Q,S),R}$.

By setting

$$\begin{cases} Q = P + 1 \\ S = 0 \end{cases} \tag{3}$$

EVLBP$_{L,(P,Q,S),R}$ will be equivalent to VLBP$_{L,P,R}$. Therefore, VLBP$_{L,P,R}$ can be seen as a special case of EVLBP$_{L,(P,Q,S),R}$. Fig. 3 (bottom) illustrates an example of Extended Volume LBP operator with $P$=4, $Q$=S=1 and $R$=1 (EVLBP$_{L,(4,1,1),1}$), while Fig. 3 (top) illustrates an example of VLBP$_{L,4,1}$ operator which is equivalent to EVLBP$_{L,(4,5,0),1}$.

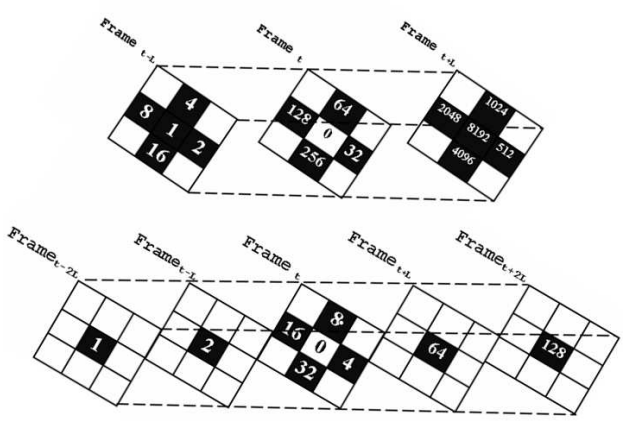


**Fig. 3.** Top: $VLBP_{L,4,1}$. Bottom: $EVLBP_{L,(4,1,1),1}$

Once the neighborhood function is defined, we divide each face sequence into several overlapping rectangular prisms of different sizes, from which we extract local histograms of EVLBP code occurrences. Then, instead of simply concatenating the local histograms into a single histogram, we use AdaBoost learning algorithm for automatically determining the optimal size and locations of the local rectangular prisms, and more importantly for selecting the most discriminative EVLBP patterns for face recognition while discarding the features which may hinder the recognition process.

## 3 Learning EVLBP Features for Face Recognition

To tackle the problem of selecting only the spatiotemporal information which is useful for recognition while discarding the information related to facial expressions and emotions, we adopt AdaBoost learning technique [15] which has shown its efficiency in feature selection tasks. The idea is to separate the facial information into intra and extra classes, and then use only the extra-class EVLBP features for recognition.

First, we segment the training face sequences into several overlapping shots of F frames each in order to increase the number of training data. Then, we consider all combinations of face sequence pairs for the intra and extra classes. From each pair

($sequence_i^1$, $sequence_i^2$), we scan both face sequences with rectangular prisms of different sizes. At each stage, we extract the EVLBP histograms from the local rectangular prisms and compute the $\chi^2$ (Chi-square) distances between the two local histograms. $\chi^2$ dissimilarity metric for comparing a target histogram $\xi$ to a model histogram $\psi$ is defined by:

$$\chi^2(\xi, \psi) = \sum_{j=0}^{l-1} \frac{(\xi_j - \psi_j)^2}{\xi_j + \psi_j}, \tag{4}$$

where $l$ is the length of feature vector used to represent the local rectangular prisms.

Thus, for each pair of face sequences, we obtain a feature vector $X_i$ whose elements are $\chi^2$ distances. Let us denote $Y_i \in \{+1, -1\}$ the class label of $X_i$ where $Y_i = +1$ if the pair ($sequence_i^1$, $sequence_i^2$) defines an extra-class pair (*i.e.* the two sequences are from different persons) and $Y_i = -1$ otherwise. This results in a set of training samples $\{(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)\}$. Algorithm 1. summarizes our procedure of constructing the training data.

Inputs: Given a set of face sequences $\{Sequence\}$
**forall** *combinations of pairs* $(Sequence_i^1, Sequence_i^2)$ **do**
  Set $Y_i = +1$ for extra-class pairs;
  Set $Y_i = -1$ for intra-class pairs;
  **forall** *locations and sizes of local prisms* **do**
    – Extract local EVLBP$_{L,(P,Q,S),R}$ histograms with different parameters;
    – Compute $\chi^2$ distances between corresponding local histograms in the given pair of sequences;
    – Collect the $\chi^2$ distances in a feature vector X
  **end**
**end**
Outputs: $\{(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)\}$ ;

**Algorithm 1**: The construction of the training samples for feature selection using AdaBoost

Given the constructed training sets, we then apply the basic AdaBoost learning algorithm [15] in order to $(i)$ select a subset of rectangular prisms from which EVLBP features should be computed, and $(ii)$ learn and determine the weights of these selected features.

Once the rectangle prisms are selected and their weights are determined, we perform the recognition of a given probe video sequence by extracting local histograms of EVLBP patterns from the selected prisms and then applying nearest neighbor classification using weighted $\chi^2$ distance:

$$\chi^2_\alpha(\xi, \psi) = \sum_{t=0}^{T-1} \sum_{i=0}^{l_t-1} \alpha_t \frac{(\xi_{i,t} - \psi_{i,t})^2}{\xi_{i,t} + \psi_{i,t}} \tag{5}$$

where $T$ is the number of selected local prisms; $\alpha_t$ are the weighting coefficients resulted from AdaBoost learning, and $l_t$ the lengths of the feature vectors used to represent local rectangular prisms.

## 4 Experimental Analysis

### 4.1 Benchmark Methods

For comparison, we implemented five different algorithms including Hidden Markov models (HMMs) [8] and Auto-Regressive and Moving Average (ARMA) models [7] as benchmark methods for spatiotemporal representations, and PCA, LDA and LBP [13] for still image based ones. In the following, we briefly describe the implementation of these benchmark methods.

a) ***HMMs***

The principle of using HMMs to model the facial dynamics and perform video-based face recognition is quite simple [8, 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 state conditional probability density functions (pdf) and $\pi = \{\pi_i\}$ are the initial distributions. The model $\lambda$ is built using a sequence of feature vectors, called observation sequence $O = \{o_1, o_2, ..., o_l\}$, extracted from the frames of the video sequence ($l$ is the number of frames). Different features can be extracted and used as observation vectors (e.g. pixel values, DCT coefficients etc.). In [8], the PCA projections of the face images were considered. Here in our experiments, we implemented a similar approach using 30 eigenvectors for dimensionality reduction and 16-state fully connected HMM.

During our training, using the Baum-Welch procedure [17], a model $\lambda_p$ , ($p = 1, 2, ..., P$), 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 $O = \{o_1, o_2, ..., o_l\}$, the identity of the test face sequence is given by:

$$\underset{p}{argmax}\ P(O|\lambda_p) \tag{6}$$

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.

b) ***ARMA***

In the ARMA framework, a moving face is represented by a linear dynamical system and described by Eqs. 7 & 8:

$$x(t+1) = Ax(t) + v(t) \qquad v(t) \sim N(0, R) \tag{7}$$

$$I(t) = Cx(t) + w(t) \qquad w(t) \sim N(0, Q) \tag{8}$$

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 estimation of the ARMA model parameters is closed-form and therefore easy to implement [10, 7]. 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 [10]. 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 [18]. Since it has been shown that the different metrics do not alter the results significantly, we adopted in our experiments the Frobenius distance ($d_F^2$), defined by :

$$d_F^2 = 2\sum_{i=1}^{n} sin^2\theta_i(\lambda_j, \lambda_k) \tag{9}$$

where, $\theta_i$ are the subspace angles between the ARMA models $\lambda_j$ and $\lambda_k$, defined in [18].

c) ***PCA, LDA and LBP***

For comparison, we also considered still image based methods such as PCA, LDA and LBP. However, in video-based face recognition schemes both training and test data (galleries and probes) are video sequences. Therefore, performing still-to-still face recognition when the data consists of video sequences is an ill-posed problem (i.e. which frame from the test sequence to compare to which frame in the reference sequence?). Here, we adopt a scheme proposed in [19] 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 a set of K most representative samples (or exemplars) from the raw gallery videos (K=3 in our experiments). Once these exemplars are extracted, we build a view-based system and use a probabilistic voting strategy to recognize the individuals in the probe video sequences.

### 4.2 Experimental Data

For experimental analysis, we considered three different publicly available video face databases (MoBo [20], Honda/UCSD [9] and CRIM [21]) in order to ensure an extensive evaluation of our proposed approach and the benchmark methods against changes caused by different factors including face image resolution, illumination variations, head movements, facial expressions and the size of the database.

The first database, MoBo (Motion of Body), is the most commonly used in video-based face recognition research [5, 22, 8], although it was originally collected for the purpose of human identification from distance. The considered subset from MoBo

**Fig. 4.** Examples of cropped facial images from MoBo video database

**Fig. 5.** Examples of facial images from CRIM video database

database contains 96 face sequences of 24 different subjects walking on a treadmill. Some example images are shown in Fig. 4. Each sequence consists of 300 frames. From each video sequence, we automatically detected and rescaled the faces, obtaining images of $40\times40$ pixels.

The second database, Honda/UCSD, has been collected and used by Lee *et al.* in their work on video-based face recognition [9]. It was also used in the recent study of Aggarwal et al. [7]. 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, 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 resulted facial images is $20\times20$ pixels.

In order to experiment with a large amount of facial dynamics, resulted for example from the movements of the facial features when the individuals are talking, we considered a third video database called CRIM. This is large set of 591 face sequences showing 20 persons reading broadcast news for a total of about 5 hours. The database is originally collected for audio-visual recognition. There are between 23 and 47 video sequences for each individual. Some cropped images are shown in Fig. 5. The size of the extracted face images is $130\times150$ pixels.

### 4.3 Experimental Results and Analysis

From each of the three video databases (MoBo, USCD/HONDA and CRIM), we randomly selected half of the face sequences of each subject for training while the other half was used for testing. In addition, given the limited number of training samples in MoBo and Honda/UCSD databases, we also segmented the face sequences into several overlapping shots in order to increase the number of training samples. In all our experiments, we considered the average recognition rates of 100 random permutations.

First, we applied PCA, LDA, LBP, HMMs and ARMA to the test sequences in the three databases. The performances of these methods are shown in Tables 1-3. From

the results on MoBo database (Table 1), we notice that all the methods perform quite well and the spatiotemporal based methods (*i.e.* HMMs and ARMA) are slightly better that the static image based methods (PCA, LBP and LDA). The better performance of the spatiotemporal methods is in agreement with the neuropsychological evidence [1] stating that facial dynamics are useful for recognition. From these results we can also see that the benefit of the spatiotemporal approach is not very significant. Perhaps, in MoBo database, this is due to the few amount of facial dynamics which are mainly limited to the rigid movements of the head.

However, the results on Honda/UCSD database (Table 2) show that the low-image resolution ($20 \times 20$ pixels) affects all these five methods and that image based ones are more affected. This is also in agreement with the neuropsychological findings that indicate that facial movement contributes more to the recognition under degraded viewing conditions.

Surprisingly, the results on CRIM database (Table 3) show that HMM and ARMA approaches gave worse results than those of PCA, LDA, and LBP based methods. While one may not expect worse performances using spatiotemporal representations, the obtained results attest that PCA, LDA and LBP based representations might perform better. This means that combining face structure and its dynamics in an *ad hoc* manner does not systematically enhance the recognition performance.

From the experiments, we also noticed that the basic LBP approach [13] performed quite well and outperformed PCA and LDA in all our tests. This confirms the validity of LBP based descriptions in face analysis. A bibliography of LBP-related research can be found at $http://www.ee.oulu.fi/research/imag/texture/lbp/bibliography/$.

| Method | Recognition rate |
|---|---|
| PCA | 87.1% |
| LDA | 90.8% |
| LBP [13] | 91.3% |
| HMM [8] | 92.3% |
| ARMA [7] | 93.4% |

**Table 1.** Comparative recognition results of 5 benchmark methods on MoBo database

| Method | Recognition rate |
|---|---|
| PCA | 69.6% |
| LDA | 74.5% |
| LBP [13] | 79.6% |
| HMM [8] | 84.2% |
| ARMA [7] | 84.9% |

**Table 2.** Comparative recognition results of 5 benchmark methods on Honda/UCSD database

| Method | Recognition rate |
|---|---|
| PCA | 89.7% |
| LDA | 91.5% |
| LBP [13] | 93.0% |
| HMM [8] | 85.4% |
| ARMA [7] | 80.0% |

**Table 3.** Comparative recognition results of 5 benchmark methods on CRIM database

We also experimented with Volume LBP spatiotemporal approach which has been successfully applied to dynamic texture analysis in [14]. We divided each face sequence into several overlapping local rectangular prisms of fixed sizes. Then, we extracted the VLBP based spatiotemporal representation using different VLBP operator parameters. For recognition, we adopted the $\chi^2$ distance. Using such an approach, we obtained best recognition rates of 90.3%, 78.3% and 88.7% with $\text{VLBP}_{2,4,1}$, $\text{VLBP}_{1,4,1}$ and $\text{VLBP}_{1,4,1}$ on MoBo, Honda/UCSD and CRIM databases, respectively. Surprisingly, these results are worse than those obtained using still image LBP based approach which yielded in recognition rates of 91.3% (versus 90.3%), 79.6% (versus 78.3%) and 93.0% (versus 88.7%) on MoBo, Honda/UCSD and CRIM databases, respectively. This supports our earlier conclusion indicating that using spatiotemporal representations do not systematically enhance the recognition performances. The most significant performance degradations of VLBP approach are noticed on CRIM database which contains the largest amount of facial dynamics. This indicates that some of these facial dynamics are not useful for recognition. In other terms, this means that some part of the temporal information is useful for recognition while another part may also hinder the recognition. Obviously, the useful part is that defining the extra-personal characteristics while the non-useful part concerns the intra-class information such as facial expressions and emotions. For recognition, one should then select only the extra-personal characteristics.

To verify this hypothesis, we considered our proposed approach which consists of using AdaBoost for learning and selecting only the most discriminative spatiotemporal features. First, we tested AdaBoost with VLBP features and obtained recognition rates of 96.5%, 89.1% and 94.4% on MoBo, Honda/UCSD and CRIM databases, respectively. As shown in Tables 4-6, performing feature selection yields in significant performance enhancement on all these three databases. This validates our hypothesis that only some part of the temporal information is useful for recognition while another part may hinder the recognition process.

Then, we experimented with the proposed extended set of VLBP features (EVLBP) introduced in Section 2 and used AdaBoost for learning the most discriminative spatiotemporal EVLBP features. As expected, this enhanced further the performances, yielding in excellent recognition rates of 97.9%, 96.0% and 98.5% on MoBo, Honda /UCSD and CRIM databases, respectively. This additional performance enhancement explains the benefit of enriching the VLBP feature set by deriving EVLBP and shows the limitations of $\text{VLBP}_{L,P,R}$ operator which does not encode well enough the temporal information in the face sequences since the operator considers neighboring points

only from three frames and therefore the information in the frames with time variance less than L are missed out.

Notice that the obtained results significantly outperform those of all benchmarks methods (PCA, LDA, LBP, HMM and ARMA) on the three databases (comparison between Tables 1-3 and Table 4-6). To our knowledge, this is also the best performance on these databases. Perhaps, these excellent results can be explained by the followings: $(i)$ the spatiotemporal representation using extended volume LBP features, in contrast to the HMM based approach, is very efficient as it codifies the local and global facial dynamics and structure; and more importantly $(ii)$ the temporal information extracted by the extended volume LBP features consisted of both intra and extra personal information (facial expression and identity). Therefore, there was need for performing feature selection. In addition, the selected EVLBP spatiotemporal features were assigned different weights reflecting their contributions to recognition, while this was not the case in other methods.

| Method | Recognition rate |
|---|---|
| VLBP [14] | 90.3% |
| VLBP+AdaBoost | 96.5% |
| EVLBP+AdaBoost | **97.9%** |

**Table 4.** Recognition results of VLBP, VLBP with AdaBoost and EVLBP with AdaBoost on MoBo database

| Method | Recognition rate |
|---|---|
| VLBP [14] | 78.3% |
| VLBP+AdaBoost | 89.1% |
| EVLBP+AdaBoost | **96.0%** |

**Table 5.** Recognition results of VLBP, VLBP with AdaBoost and EVLBP with AdaBoost on Honda/UCSD database

| Method | Recognition rate |
|---|---|
| VLBP [14] | 88.7% |
| VLBP+AdaBoost | 94.4% |
| EVLBP+AdaBoost | **98.5%** |

**Table 6.** Recognition results of VLBP, VLBP with AdaBoost and EVLBP with AdaBoost on CRIM database

**Fig. 6.** Examples of the four first selected rectangular prisms from which EVLBP spatiotemporal features are extracted on CRIM face sequences

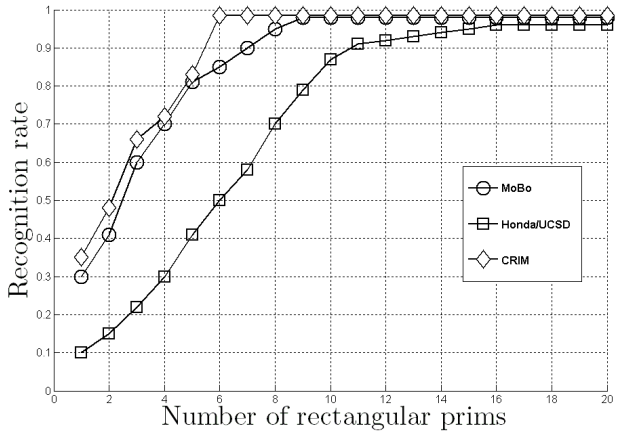


**Fig. 7.** The recognition rates function of the number of selected regions with AdaBoost from which EVLBP features are extracted

Analyzing the selected local regions (the rectangular prisms) from which the EVLBP features were collected, we noticed that the dynamics of the whole face and the eye area are more important than that of the mouth region for identity recognition. This is quite surprising in the sense that one can expect from the mouth region to play the most important role as it is the most non-rigid region of the face when an individual is talking. Probably, mouth region does play an important role but for facial expression recognition. Fig. 6 shows examples of the most discriminative spatiotemporal regions returned by AdaBoost for CRIM face sequences and from which EVLBP spatiotemporal features are extracted. Notice that these four first selected features are extracted from global an local regions. This supports the results of other researchers indicating that both global and local features are useful for recognition. From how many selected regions the EVLBP features are computed? Fig. 7 shows the recognition results as a function of the number of regions selected by AdaBoost. The best results are obtained with 9, 16 and 6 regions on MoBo, Honda/UCSD and CRIM databases, respectively. Using additional regions did not enhance the recognition performance.

Table 7 summarizes the obtained results using the different methods (PCA, LDA, LBP, HMM, ARMA, VLBP and EVLBP) on the three databases (MoBo, Honda/UCSD and CRIM).

| Method | Results on MoBo | Results on Honda/UCSD | Results on CRIM |
|---|---|---|---|
| PCA | 87.1% | 69.9% | 89.7% |
| LDA | 90.8% | 74.5% | 91.5% |
| LBP [13] | 91.3% | 79.6% | 93.0% |
| HMM [8] | 92.3% | 84.2% | 85.4% |
| ARMA [7] | 93.4% | 84.9% | 80.0% |
| VLBP [14] | 90.3% | 78.3% | 88.7% |
| VLBP+AdaBoost | 96.5% | 89.1% | 94.4% |
| EVLBP+AdaBoost | **97.9%** | **96.0%** | **98.5%** |

**Table 7.** Summary of the obtained results using the different methods on the three databases

## 5 Conclusion

The few works attempting to use spatiotemporal representations for face recognition from videos ignore the fact that some of the facial information may also hinder the recognition process. Indeed, while one may not expect worse results using spatiotemporal representations instead of still image based ones, our results showed that still image based methods can perform better than spatiotemporal based ones. This suggests that the existing spatiotemporal representations have not yet shown their full potential and need further investigation.

From this observation, we presented a novel approach for spatiotemporal face recognition with excellent results. The efficiency of the proposed approach can be explained by the local nature of the spatiotemporal EVLBP based description, combined with the use of boosting for selecting only the personal specific information related to identity while discarding the information which is related to facial expression and emotions.

## Acknowledgment

The partial financial support of the National Agency for Technology and Innovation (Tekes) is gratefully acknowledged.

## References

1. O'Toole, A.J., Roark, D.A., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Science **6** (2002) 261–266
2. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys **34(4)** (2003) 399–458
3. Li, Y.: Dynamic Face Models: Construction and Applications. PhD thesis, Queen Mary, University of London (2001)
4. Li, B., Chellappa, R.: Face verification through tracking facial features. Journal of the Optical Society of America **18** (2001) 2969–2981
5. Zhou, S., Chellappa, R.: Probabilistic human recognition from video. In: European Conf. on Computer Vision. (May 2002) 681–697

6. Zhou, S., Krueger, V., Chellappa, R.: Face recognition from video: A condensation approach. In: IEEE Int. Conf. on Automatic Face and Gesture Recognition. (May 2002) 221–228
7. Aggarwal, G., Chowdhury, A.R., Chellappa, R.: A system identification approach for video-based face recognition. In: 17th International Conference on Pattern Recognition. Volume 4. (August 2004) 175– 178
8. Liu, X., Chen, T.: Video-based face recognition using adaptive hidden markov models. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition. (June 2003) 340–345
9. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition. (June 2003) 313–320
10. Soatto, S., Doretto, G., Wu, Y.: Dynamic textures. In: International Conference on Computer Vision. Volume 2., Vancouver, BC, Canada (July 2001) 439–446
11. Heisele, B., Ho, P., Wu, J., Poggio, T.: Face recognition: Component based versus global approaches. Computer Vision and Image Understanding **91**(1-2) (2003) 6–21
12. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 971–987
13. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12) (2006) 2037–2041
14. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **29(6)** (2007) 915–928
15. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1) (1997) 119–139
16. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition **29** (1996) 51–59
17. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE. Volume 77(2). (1989) 257– 286
18. Cock, K., Moor, B.D.: Subspace angles between ARMA models. Systems and Control Letters **46**(4) (2002) 265–270
19. Hadid, A., Pietikäinen, M.: Selecting models from videos for appearance-based face recognition. In: 17th International Conference on Pattern Recognition. Volume 1. (August 2004) 304–308
20. Gross, R., Shi, J.: The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University (June 2001)
21. CRIM: http://www.crim.ca/
22. Krueger, V., Zhou, S.: Exemplar-based face recognition from video. In: European Conf. on Computer Vision. (May 2002) 732–746