From Stills to Video: Face Recognition Using a Probabilistic Approach

Yongbin Zhang and Aleix M. Martínez
Department of Electrical and Computer Engineering
The Ohio State University
{zhangyo, aleix}@ece.osu.edu

Abstract

While it is generally possible to do recognition from video sequences, the training process is usually done over static images. This is due to the fact that, in many applications (e.g., homeland security), one does not have large video sequences which can be used for training. For example, law enforcement agencies generally have a frontal and a profile view of wanted individuals, but do not usually keep video sequences in file. Nonetheless, in these applications, it is still possible to analyze the information of video sequences for subsequent recognition tasks. This paper presents a probabilistic algorithm that learns from small sets of static images and then recognizes faces from video sequences. The proposed algorithm is robust to partial occlusions, different orientations and expression changes and does not require of precise face localizations. Our preliminary results with a small database show that the proposed method is more robust to such changes than static-to-static recognition of faces.

1 Introduction

As computers become more ubiquitous and technology more accessible to the end user, face recognition systems are expected to play an increasing role in our society. A recent change that has influence the way we study object recognition and other computer vision problems is the improvement in quality and reduction in price of video cameras. This has drawn interests in the design of systems that can recognize objects (such as faces) from video sequences rather that from static images. A good example of this trend, it is obviously this workshop.

Although much progress has been achieved in the recognition of faces from static images, the literature on face recognition from video is still relatively small [24, 18, 20, 8, 1, 11]. One reason for this unbalance was due to the low accessibility of high-quality video cameras. The second reason is, obviously, algorithmical. While it is generally difficult to successfully do feature extraction from static images, this process has proven even more challenging in dynamic sequences [7, 13, 12].

The low accessibility of video cameras has also shaped the way law enforcement agencies and other institutions recorded biometric data such as face images. Current databases are mostly based on static images and only few have some (if any) video sequences.

Therefore, we believe that it is necessary to define algorithms that can learn from static images yet be able to do recognition from video sequences.

Our method builds on previous work defined for the recognition of faces in a static-to-static recognition scenario and proposes extensions to deal with pose and facial expression variations that are not present in the training set.

Our first step is to extend the approach of [14] to work with video sequences rather than stills. This we will show in Section 2. In Section 3, we will present a way to make the approach of Section 2 more robust to expression changes. And, in Section 4, we will extend the method to be robust to small pose variations. Section 5 presents the experimental results where we compare our method to several approaches which only use static images. We conclude in Section 6.

1.1 Related work

Some recent papers have addressed some of the problems posed by the recognition of human faces from video sequences. In [10], the authors use RBF (Radial Basis Function) networks to tackle the view-varying problem in image sequence. Wechsler *et al.* [21] define an automatic video-based person authentication system where RBFs are used to identify subjects. Li and Chellappa [11] use the parameters obtained from facial feature tracking to construct a recognition system based on feature motion. In their algorithm, tracking is formulated as a Markov Chain Monto Carlo problem. Hidden Markov models are used in [13] to learn the appearance of faces over a set of frames for static-to-static or video-to-video recognition of faces. In [12], the authors propose to learn view-based appearance manifolds from video. Robustness is increased with the help of a transition probability matrix defined between adjacent views. And, Zhou and Chellappa [25] define an improved probabilistic tracking and recognition system which can also be

used for recognition tasks.

The systems summarized above did not simultaneously tackle the problems we will address in this paper: occlusions, expression changes, pose variation and localization errors. Moreover, most of them were designed to work as video-to-video recognition systems

(i.e., where both the training and testing data are video sequences).

2 From Static-to-Video Analysis

In addition to the alignment errors introduced by inaccuracy in face detection (or face tracking), recognition from video sequences is made difficult by variations in expression, pose and partial self-occlusions.

2.1 Training

Any face recognition system should be robust to localization errors. To address this problem, we have previously proposed [14] to first learn the localization error of the localization algorithm and, then, find the subspace (within our feature-space) which represents all images under all possible errors of localization for each of the training images. In this paper, we model this subspace as a mixture of Gaussians. More formally, let $\{\mathbf{I}_1, \ldots, \mathbf{I}_n\}$ be the set of training images, where n is the number of samples. Since we assume that the localization error of our localization algorithm is known, we can now synthetically generate the set of all images under all possible errors of localization; i.e., $\hat{\mathbf{I}}_i = \{\mathbf{I}_{i,1}, \ldots, \mathbf{I}_{i,r}\}$, where $\mathbf{I}_{i,j}$ is one of the images obtained when we crop the face from the original image with one of the r possible localizations given by our localization algorithm and r is the maximum number of images we can generate with the known (average) localization error [14]. We then estimate the subspace that represents each of the image sets $\hat{\mathbf{I}}_i$ as a mixture of Gaussians. This mixture of Gaussians is learned using the EM algorithm [6].

In order to be robust to self-occlusions, we divide the face into six local areas (as shown in Fig. 1). Each sample image $\mathbf{I}_j$ will generate r possible images (to account for the localization error) for each of the six subimages; i.e., $\{\mathbf{I}_{i,1,1}, \ldots, \mathbf{I}_{i,1,r}, \ldots \mathbf{I}_{i,6,r}\}$, where $\mathbf{I}_{i,j,k}$ is the k^{th} image of the j^{th} subimage that accounts for the localization error of the i^{th} sample.

We then estimate the localization error subspace of every subimage by means of a mixture model of H Gaussians. Formally, let $\{\mu_{1,1,1}, \ldots, \mu_{1,6,H}, \ldots, \mu_{n,6,H}\}$ be the H means of the H Gaussians of each of the subareas of each of the sample images and $\{\Sigma_{1,1,1}, \ldots, \Sigma_{1,6,H}, \ldots, \Sigma_{n,6,H}\}$ the corresponding covariance matrices. Fig. 1 shows the subspace representing one of the subarea of a set of sample images (with $H = 3$).

Finally, we compute the eigen-representation of each of the six local areas; i.e., $\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_6\}$ where $\mathbf{\Phi}_j$ is the projection matrix of the j^{th} subimage of the face whose columns are the p eigenvectors associated to the largest p eigenvalues of the eigen-decomposition problem

$$\left(\sum_{i=1}^{n}\sum_{k=1}^{H}\Sigma_{i,j,k}\right)\mathbf{V} = \mathbf{V}\Lambda\,. \tag{1}$$

Each sample image is now represented as a mixture of Gaussians in each of these eigenspaces. Formally, $\{\hat{\mu}_{1,1,1}, \ldots, \hat{\mu}_{1,6,H}, \ldots, \hat{\mu}_{n,6,H}\}$ and $\{\hat{\Sigma}_{1,1,1}, \ldots, \hat{\Sigma}_{1,6,H}, \ldots, \hat{\Sigma}_{n,6,H}\}$ where $\hat{\mu}_{i,j,k} = \mathbf{\Phi}_k\mu_{i,j,k}$ and $\hat{\Sigma}_{i,j,k} = \mathbf{\Phi}_k\Sigma_{i,j,k}$ respectively.

2.2 Testing from video sequences

Given a test sequence, $\mathbf{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_m\}$ (where m is the total number of images in the sequence), of an unknown person, we work as follows. We divide each image into the six local areas defined above, $\{\mathbf{T}_{i,1}, \ldots, \mathbf{T}_{i,6}\}$, and calculate the probability of each subimage to belong to each of the learned classes,

$$P(\mathbf{T}_{i,j}, c) = \sum_{h=1}^{H}\left(\hat{\mathbf{T}}_{i,j} - \hat{\mu}_{c,j,h}\right)\hat{\Sigma}_{c,j,h}\left(\hat{\mathbf{T}}_{i,j} - \hat{\mu}_{c,j,h}\right)^T, \tag{2}$$

where $P(\mathbf{T}_{i,j}, c)$ is the probability of subimage $\mathbf{T}_{i,j}$ to belong to class c and $\hat{\mathbf{T}}_{i,j} = \mathbf{\Phi}_j\mathbf{T}_{i,j}$. Then add the result of each local area together to yield the probability of each image $\mathbf{I}_i$ to belong to class c:

$$P(\mathbf{T}_i, c) = \sum_{j=1}^{6} P(\mathbf{T}_{i,j}, c)\,. \tag{3}$$

To have a real probability, the above equation should take values between zero and one. We can compute the probability (with a value between 0 and 1) of correct classification according to class c as

$$Q_{i,c} = \frac{P(\mathbf{T}_i, c)}{\sum_{a=1}^{C} P(\mathbf{T}_i, a)}, \tag{4}$$

where C is the total number of classes (i.e. people).

Our initial algorithm will use as many images (say, q) out of the total m as necessary such that $Q_c = 1/q\sum_{i=1}^{q} Q_{i,c}$ is larger than a pre-specified threshold T for a given c (and, obviously, $q \leq m$). Naturally, the most convenient way of choosing T is by means of an evaluation set. When the number of people in our database is large, we find it convenient to use different thresholds for each person.

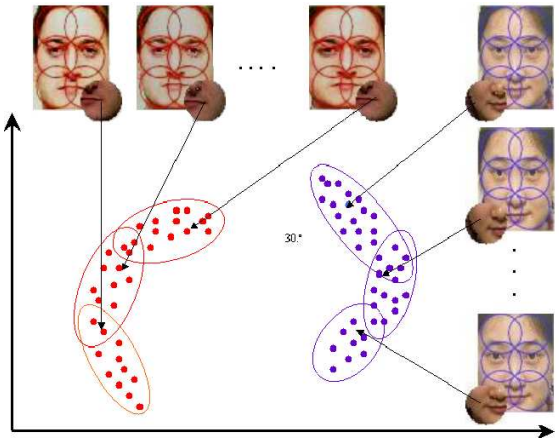

Figure 1: **We first generate all cropped images according to the localization error of our localization algorithm. Then, each face image is divided into six local areas. The localization error is finally estimated using a mixture of Gaussians.**

3 Dealing with Facial Expression Changes

As mentioned above, the facial expressions in the training images will generally diverge from some of the expressions in some of the images of our sequence. Ideally, we would like to use those images of the sequence that have a similar expression to those in the training set and discard the rest – because it is known that the recognition rate decreases when images with different expressions are used [15, 14, 23].

The first problem to be addressed is, therefore, to estimate the difference in expression between the testing and each of the images used for training. A common way to do this is by means of the optical flow approach [4, 3, 22, 14]. More formally, we write $\mathbf{O}(\mathbf{T}_i, \mathbf{I}_j)$ to represent the optical flow vector that describes the motion between the expression in $\mathbf{T}_i$ and that of $\mathbf{I}_j$, where here $\mathbf{I}_j$ is the j^{th} sample image. Since faces have many motion discontinuities, a robust algorithm (such as Black and Anandan's [5]) is generally preferred [15].

Once we have calculated $\mathbf{O}(\mathbf{T}_i, \mathbf{I}_j)$, we can estimate the "usefulness" of each image in our video sequence as $W_{i,j} = O_{max} - \|\mathbf{O}(\mathbf{T}_i, \mathbf{I}_j)\|$, where $O_{max} = max_{(\mathbf{T}_i, \mathbf{I}_j)} \|\mathbf{O}(\mathbf{T}_i, \mathbf{I}_j)\|$. Small values of $W_{i,j}$ mean that the expressions in each of the two images ($\mathbf{T}_i$ and $\mathbf{I}_j$) is very different. Large values imply similar expressions in both images.

We can now use these values to weight each of the probabilities in Eq. (3). To do this, we will need to calculate the usefulness of each of the local areas, which is given by

$$W_{i,j,k} = O_{max} - \|\mathbf{O}(\mathbf{T}_i, \mathbf{I}_{j,k})\| \tag{5}$$

where $\mathbf{I}_{j,k}$ is the k^{th} local area of image j. We then normalize these values appropriately,

$$\widehat{W}_{i,j,k} = \frac{W_{i,j,k}}{\sum_{j=1}^{n} W_{i,j,k}}, \tag{6}$$

and redefine Eq. (3) as

$$P(\mathbf{T}_i, c) = \sum_{k=1}^{6} \widehat{W}_{i,c,k} P(\mathbf{T}_{i,k}, c)\,. \tag{7}$$

The reader may have noted that since the motion field between each pair of training and testing images is now known (i.e., pixel correspondences are available), we could have morphed one of the faces to equal the shape of the other. We have indeed experimented with this idea and observed that the results were comparable to those obtained with the approach described above but with a higher computational cost.

4 Dealing with Pose Variations

When the face changes orientation with respect to a fix camera, the brightness patterns in the image are distorted. This problem is exacerbated in appearance-based algorithms, because reconstructing the object would require to recover both, the shape and radiometric information lost [2, 4].

Similarly to what we did in our previous section, we can now select those images of the video sequence which have a similar pose to those images in the training set and discard the rest. Following this idea, our algorithm works as follows. We first estimate the pose (with respect to the camera) of the training and testing images. Then, we weight each local area according to the pose similarity between each pair of training and testing images. Identical pose in both will require a weight of one. Pose differences of over 90 degrees a zero weight.

4.1 Pose estimation using view-based eigenspaces

The view-based approach advanced by Pentland and colleagues [17] has been extensively used for object recognition under varying pose as well as for pose estimation [16, 19]. To do pose estimation, we first need to learn the subspace (e.g. eigenspace) that represents the images of all faces as viewed from a given orientation α.[1] Formally, let Σ_α be the sample covariance matrix of a set

[1] In this communication, we have only considered rotations about the y-axis, but our formulation can be extended to deal with the more general (although rare) case where all three angles are considered.

of sample images with faces at orientation α. Then, the eigenspace representing those face images at orientation α is given by the projection matrix $\mathbf{\Phi}_\alpha$ whose column vectors are the eigenvectors associated to the p largest eigenvalues of $\Sigma_\alpha \mathbf{V} = \mathbf{V}\Lambda$. Therefore, if we have sample images belonging to l different orientations, we will obtain l eigenspaces; i.e., $\{\mathbf{\Phi}_{\alpha_1}, \ldots, \mathbf{\Phi}_{\alpha_l}\}$.

We can now estimate the pose of any new image, $\mathbf{T}$, of a face as follows. First, we calculate the distance from the vector $\mathbf{T}$ to each of the subspaces. This can be estimated by the sum of the Euclidean distance from $\mathbf{T}$ to each of our eigenspaces and the Mahalanobis distance from the projected vector within each of these eigenspaces to their means. We will denote these distances as $\{D_{\alpha_1}, \ldots, D_{\alpha_l}\}$. Next, we select the d smallest distances (with $d \leq l$) and estimate the pose of $\mathbf{T}$ as

$$\theta_{\mathbf{T}} = \sum_{i=1}^{d} \frac{D_{\alpha_{(d-i+1)}}}{\sum_{j=1}^{d} D_{\alpha_j}} \alpha_i \tag{8}$$

where here α_i corresponds to the pose of the face images represented by the i^{th} closest eigenspace $\mathbf{\Phi}_{\alpha_i}$.

4.2 Weighting local areas

Once the pose of each pair of training and testing images has been computed, we can calculate the weights for each of the six local areas of the face. As we argued above, the local weights, $W'_{i,j,k}$ (where i is the testing image $\mathbf{T}_i$ and k is the k^{th} local area of the j^{th} training image $\mathbf{I}_j$), should be one when the orientation of both images is identical and zero when the difference is above 90 degrees. I.e., $W'_{i,j,k} = 1$ when $\theta_{\mathbf{T}_i} = \theta_{\mathbf{I}_{j,k}}$ and $W'_{i,j,k} = 0$ when $|\theta_{\mathbf{T}_i} - \theta_{\mathbf{I}_{j,k}}| \geq 90^o$. Any other difference in between will be approximated using a linear function as follows:

$$W'_{i,j,k} = 1 - \frac{g\left(|\theta_{\mathbf{T}_i} - \theta_{\mathbf{I}_{j,k}}|\right)}{90}, \tag{9}$$

where

$$g(\theta) = \begin{cases} \theta \leq 90 & : \quad \theta \\ \theta > 90 & : \quad 90. \end{cases}$$

We finally redefine Eq. (7) to yield our final probability equation

$$P(\mathbf{T}_i, c) = \sum_{k=1}^{6} W'_{i,c,k} \widehat{W}_{i,c,k} P(\mathbf{T}_{i,k}, c)\,. \tag{10}$$

5 Experimental Results

To test the approach described above, we have collected a dataset of static images which we used for training, and a dataset of video sequences that we used for testing. The training set consists of three neutral face images per person. An example of the training set for one of the subjects is shown in Fig. 2. The testing set includes three sequences. The first and second are video sequences of nearly frontal view faces with random talking; see Fig. 3. The third sequence corresponds to faces with orientations raging from (roughly) -50^o to $+50^o$ and with some facial expression changes, Fig. 4. Our current database consists of twenty two people.

Figure 2: **The three training images for one of the subjects in our database.**

For our experimental results, we have used the algorithm of Heisel *et al.* [9] to automatically localize the position of the eyes, mouth and nose of each face. Once these facial features have been localized, using the differences between the x and y coordinates of the two eyes, the original image is rotated until obtaining a frontal view face where both eyes have the same y value; i.e., $atan(\|y_1 - y_2\|/\|x_1 - x_2\|)$, where $(x1, y1)$ and $(x2, y2)$ are the right and left eye image coordinates. A contour algorithm is used to search for the lateral boundaries of the face (i.e., left and right). The top and bottom of the face are assigned as a function of the above features and the face is then warped to a final standard face of 120 by 170 pixels. Fig. 5 shows the warping results for the images shown in Fig. 3.

The localization algorithm of the algorithm described in [9] was found to have an average localization error of about 16 by 16 pixels. In our experiments we also set $H = 3$ and $d = 3$.

In Fig. 6 we show the results of our algorithm on the first two video sequences which include expression changes. We

Figure 3: **A few frames of one of the video sequences with random talking.**

Figure 4: **Some frames of a video sequence with faces at varying pose.**

have compared our results to those obtained with global PCA (as defined in [20]) and global LDA (as defined in [1]) where only one frame from the sequence is used. To obtained these results, we calculated the recognition rate of every frame of each sequence and then compute the average recognition rate. These results are labelled G-PCA and G-LDA in the figure.

In addition, we have also compared our results to a local PCA and a local LDA approach. In these two cases, our local weighted approach defined in this paper was used to obtain the results. Thus, these methods should be robust to localization errors, partial occlusions and variations in pose. However, these two algorithms will only use a single image. The average recognition rates obtained with these two methods are labelled L-PCA and L-LDA. These results show the improvement achieved when using our probabilistic approach in a video sequence rather than using a single frame; i.e., video versus static.

Of course, we could have also manually localized the facial features to see how the system works when the localization error is reduced to a minimum. In this case, our algorithm (using video sequences) obtained a recognition rate of **100%** while the local PCA and LDA approaches (which only use a single frame) do not do better than 92%. As before, the global PCA and LDA approaches felt far behind with results below 75%.

Figure 5: **Example of warped faces.**

In Fig. 7 we show the results of both, our method and our global and local implementations of PCA and LDA, on the video sequences with faces at varying pose.

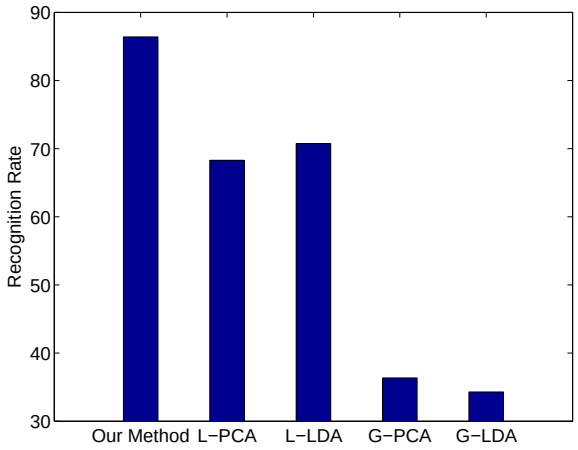

Figure 6: **Shown here are the recognition rates obtained with our probabilistic algorithm and two implementations of PCA and LDA – one local, one global.**

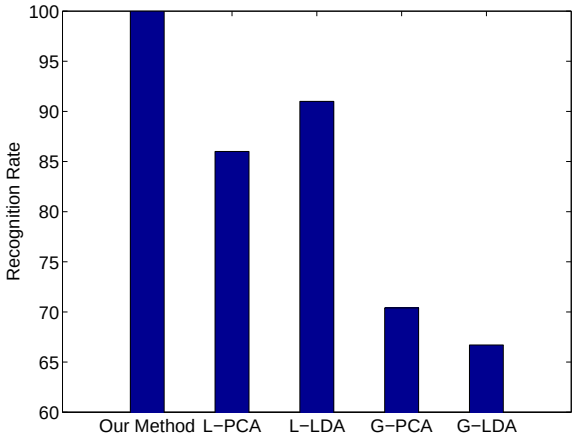

Figure 7: **Recognition rates obtained on the sequences with faces at different pose.**

6 Conclusion

In this paper we have shown derivations for an algorithm that can learn to recognize faces from just a few training images. While the training process uses static images, the recognition task is performed over video sequences. Our results show that higher recognition rates are obtained when we use video sequences rather than statics – even when the algorithm using static images and that using video sequences address the same problems with exactly the same techniques.

Acknowledgments

We would like to thank the people of the CBCL lab at MIT for sending us the code of their face localization algorithm used in our experiments. This research was partially supported by NIH.

References

[1] P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recogniton using class specific linear projection", IEEE Trans. PAMI 19(7):711-720,1997.

[2] A. Georghiades, P.N. Belhumeur and D.J. Kriegman, "From Few to Many: Generative Models of Object Recognition", IEEE Trans. PAMI 23(6):643-660, 2001.

[3] M.S. Bartlett, J.C. Hager, P. Ekman and T.J. Sejnowski, "Measuring spatial expressions by computer image analysis", *Psychophysiology* 36: 253-263, 1999.

[4] D. Beymer and T. Poggio, "Face Recognition from one Example View", In Proc. ICCV, 1995.

[5] M.J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piece-wise smooth flow fields," Computer Vision and Image Understanding 63(1): 75-104,1996.

[6] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likehood from Incomplete Data via the EM algorithm", Journal of Royal Statistical Society 30(1): 1:38, 1977.

[7] R.O. Duda, P.E. Hart and D.G. Stork, "Pattern Classification (*second edition*)", Addison Wesley 2002.

[8] K. Etemad and R. Chellapa, "Discriminant analysis for recognition of human face images", Journal of Optical Socieaty of American A 14(8): 1724-1733, 1997.

[9] B. Heisel, T. Sere, M. Pontil, T. Vetter, T. Poggio, "Categorization by Learning and Combining object parts", In Proc. NIPS, 2001.

[10] A.J. Howell and H. Buxton, "Towards Unconstrained Face Recognition from Image Sequences", In Proc. IEEE Conf. on Automatic Face and Gesture Recognition, pp. 224-229, 1996.

[11] H. Li and R. Chellappa, "Face Verification through tracking facial feature", JOSA-A, 8(12), 2001.

[12] K. Lee, J. Ho, M. Yang and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds", In Proc. IEEE CVPR, Vol. I, pp. 313-320, 2003.

[13] A.M. Martinez, "Face Image Retrieval Using HMMs", In Proc. IEEE Workshop on Content-Based Access of Images and Video Libraries, 1999.

[14] A.M. Martinez, "Recognizing Imprecisely Localized, Partially Occuded and Expression Variant Faces from a Single Sample per Class," IEEE Trans. PAMI 24(6):748-763, 2002.

[15] A.M. Martinez, "Matching Expression Variant Faces," Vision Research 43(9):1047-1060, 2003.

[16] H. Murase and S. K. Nayar, "Visual Learning and Recognition of 3-D Object from Appearance", International Journal of Computer Vision, 14:5-24, 1995.

[17] A. Pentland, B. Moghaddam and T. Starner, "View-based and modular eigenspaces for face recognition", In Proc. IEEE CVPR, 1994.

[18] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," Journal of the Optical Society of America A 4 519–524,1987.

[19] S. Srinivasan and K. L. Boyer, "Head Pose Estimation using View Based Eigenspaces", In Proc. International Conference on Pattern Recognition, pp. 302-305, 2002.

[20] M. Turk and A. Pentland, "Eigenfaces for recogntion", J. of Cognitive Neuroscience 3(1):71-86,1991.

[21] H. Wechsler, V. Kakkad, J. Huang, S. Gutta and V. Chen, "Automatic Video-based Person Authentication Using the RBF Network", In Proc. International Conference on Audio and Video-based Biometric Person Authentication, pp. 117-183, 1997.

[22] Y. Yacoob and L. Davis Computing Spatio-Temporal Representation of Human faces. *In Proc. CVPR*, pp. 70-75, 1994.

[23] Y. Yacoob and L. Davis, "Smiling Faces are Better for Face Recognition," In Proc. International Conference Face Recognition and Gesture Analysis, pp. 59-64, 2002.

[24] W.Y. Zhao, R. Chellappa, A. Rosenfeld and J.P. Phillips, "Face Recognition: A Literature Survey", ACM Computing Surveys, 2003.

[25] S. Zhou, V. Krueger, and R. Chellappa. "Probabilistic recognition of human faces from video", Computer Vision and Image Understanding 91:214-245, 2003.