A System Identification Approach for Video-based Face Recognition

First Author Second Author

Abstract

The paper poses video-to-video face recognition as a dynamical system identification and classification problem. Video-to-video means that both gallery and probe consists of videos. We model a moving face as a linear dynamical system whose appearance changes with pose. An autoregressive and moving average (ARMA) model is used to represent such a system. The choice of ARMA model is based on its ability to take care of the change in appearance while modeling the dynamics of pose, expression etc. Recognition is performed using the concept of subspace angles to compute distances between probe and gallery video sequences. The results obtained are very promising given the extent of pose, expression and illumination variation in the video data used for experiments.

1. Introduction

Humans make use of face as an important cue for identifying people. This makes automatic face recognition very crucial from the point of view of a wide range of commercial and law enforcement applications. Although significant work has been done [18], the current systems are still not close to the human perceptual system. Traditionally, face recognition research has been limited to recognizing faces from still images. Most of these approaches discount the inherent 3-D structure of the face and therefore are very susceptible to pose changes. One way to overcome this is to generate 3-D models using multiple still images or video and then use them while testing any probe image. Even if the resolution of the images/video is high (which is usually not the case), the face model generated by the known techniques is usually far from perfect which makes this approach often not practical for face recognition.

Recently, methods based on multiple images/video sequences that do not involve creating an explicit 3-D model have been suggested. Such an approach is supported by many psychophysics works like [4], where authors argue that a 3-D object is represented as a set of 2-D images (instead of a 3-D model) in our brains. Leaving out the algorithms based on simple voting, most of these methods make use of either the natural variability in a face (due to variation in pose or expression) or the information present in the temporal variation of face. In [3], Biuk $et\ al$ recognize a face from a sequence of rotating head images by computing the Euclidean distances between trajectories formed by face sequences in PCA feature space. The Mutual Subspace Method (MSM) proposed in [17], considers the angle between input and reference subspaces formed by the principal components of the image sequences (not necessarily ordered) as the measure of similarity. This approach discounts the inherent temporal coherence present in a face sequence that might be crucial for recognition. In [12], face recognition is cast as a statistical hypothesis testing problem, where a set of images is classified using the Kullback-Leibler divergence between the estimated density (assumed to be Gaussian) of the probe set and that of gallery sets. This method is based on the underlying assumption that face recognition can be performed by matching distributions. However, two such distributions for the same subject might look very different depending on the range of poses and expressions covered by the two sets. Moreover, this approach is sensitive to illumination changes. In [9], Liu $et\ al$ learn temporal statistics of a face from a video using adaptive Hidden Markov Models to perform video-based face recognition. In [16], kernel principal angles, applied on the original image space and a feature space, are used as the measure of similarity between two video sequences. Zhou $et\ al$[19] propose a tracking-and-recognition approach by resolving uncertainties in tracking and recognition simultaneously in a probabilistic framework. Lee $et\ al$ [7], in their recent work, represent each person by a low-dimensional appearance manifold, approximated by piecewise linear subspaces. They present a maximum a posteriori formulation for recognizing faces in test video sequences by integrating the likelihood that the input image comes from a particular pose manifold and the transition probability to this manifold from the previous frame. Among the methods mentioned, Lee $et\ al$ [7] method seems to be the one most capable of handling large 2-D and 3-D rotations.

Although many previous methods make use of temporal information present in face videos to improve recognition, there has been no attempt to model a moving face as a dynamical system. Our work can be seen as an attempt to

explore this. In this paper, we present a method for modeling a moving face as a linear dynamical system to perform recognition. Each frame of a video is, therefore, assumed to be the output of the dynamical system particular to the subject. Our work follows [6] and [13], where Soatto *et al* used a very similar idea to characterize dynamic textures. In [2], they use the same approach for recognizing different types of human gait. As in [13], we also use a first order ARMA model. The difference is that here we try to capture the varying appearance (due to pose and expression variation) and dynamics of face using this framework. Once the models are estimated, recognition is performed by computing distances between ARMA models corresponding to probe face and gallery faces. We use several distance metrics which make use of subspace angles between the ARMA models.

The rest of the paper is organized as follows: First, we give an intuition for the proposed approach in Section 2 . This is followed by the details of our approach in Section 3. Section 4 describes various distance metrics used for comparing the generated ARMA models to estimate the degree of similarity between two face videos. We present details of our experimental results and their significance in Section 5. Section 6 concludes the paper.

2. Motivation

Suppose we want to model a point constrained to move along a fixed line. The position of the point at any time instant is guided by its position at the previous time instant. The point has an attribute, say color, that varies with time depending on the position of the point. In this framework, color of the point is the only thing that is visible to the outside world. Modeling such a phenomenon essentially requires two mappings viz.,

$$Position_{t+1} = \phi(Position_t) \tag{1}$$

$$Color_t = \psi(Position_t) \tag{2}$$

where the subscript denotes time instant. Given a sequence of observations (colors), if we can estimate ϕ and ψ, we are done. This is quite similar to the case of face videos if we think of the pose of the face as the position of the point and the 2-D appearance of the face as the color of the point. The dependence of the appearance on the pose is analogous to that of the color on the position. The degree of goodness of such a model is limited by the choice of the forms of the mappings ϕ and ψ and the accuracy of their estimation. In general, these mappings can be arbitrarily complex but methods to estimate them are not known. In our work, we get promising results by assuming them to be linear.

3. Framework for modeling

In this section, we develop a mathematical formulation that helps us in estimating the unknown parameters of the model, we use, to characterize a moving face sequence.

If the mappings ϕ and ψ are some linear operators, (1) and (2) can be written as:

$$x(t+1) = Ax(t) + v(t) \tag{3}$$

$$I(t) = Cx(t) \tag{4}$$

where, $I(t)$ is appearance of the face at time instant t, $x(t)$ is a state vector that characterizes the pose of the face, A and C are matrices representing the linear mappings and $v(t)$ is an IID (independent and identically distributed) realization from some unknown density $q(.)$, that takes care of the implicit assumption that the dynamical system is driven by an IID process.

Suppose at each time instant t, we can measure only a noisy version of $I(t)$ i.e., $y(t) = I(t) + w(t)$ where $w(t)$ is an IID sequence drawn from a known distribution. This leads to a first order auto-regressive, moving average (ARMA) model as follows:

$$x(t+1) = Ax(t) + v(t) \tag{5}$$

$$y(t) = Cx(t) + w(t) \tag{6}$$

This formulation has similarities with the pioneering work by Ali [1], where he addresses the problem of estimation and prediction for stationary spatial-temporal processes. He too uses a simultaneous linear model to represent spatial-temporal processes.

At this stage, we use the closed-form solution as described in [13], where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$, $v(t) \sim \mathcal{N}(0, Q)$ and $w(t) \sim \mathcal{N}(0, R)$. This makes our model a linear dynamical system driven by zero-mean Gaussian noise. Given a video sequence (i.e., a sequence of observation vectors $y(1), \ldots, y(\tau)$), we need to estimate the parameters A, C, Q and R to model the face in the video.

3.1 Closed-form solution to estimate the parameters

Let $Y^\tau = [y(1), \ldots, y(\tau)] \in \mathbb{R}^{m \times \tau}$ with $\tau > n$, then for $\{t = 1 \ldots \tau\}$, (6) can be written as

$$Y^\tau = CX^\tau + W^\tau; \quad C \in \mathbb{R}^{m \times n} \tag{7}$$

where X and W are defined in a manner similar to Y. If singular value decomposition (SVD) of Y^τ is $Y^\tau = U\Sigma V^T$, where Σ is a diagonal matrix, $U \in \mathbb{R}^{m \times n}$, $U^T U = I$, $V \in \mathbb{R}^{\tau \times n}$ and $V^T V = I$, then

$$\hat{C}(\tau) = U \tag{8}$$

$$\hat{X}(\tau) = \Sigma V^{\tau} \tag{9}$$

$$\hat{A}(\tau) = \Sigma V^T D_1 V (V^T D_2 V)^{-1} \Sigma^{-1} \tag{10}$$

where $D_1 = \begin{pmatrix} 0 & 0 \\ I_{\tau-1} & 0 \end{pmatrix}$ and $D_2 = \begin{pmatrix} I_{\tau-1} & 0 \\ 0 & 0 \end{pmatrix}$, and

$$\hat{Q}(\tau) = \frac{1}{\tau} \sum_{i=1}^{\tau} \hat{v}(i) \hat{v}^T(i) \tag{11}$$

where $\hat{v}(t) = \hat{x}(t+1) - \hat{A}(\tau)\hat{x}(t)$, give a closed-form solution (suboptimal in the sense of Frobenius).

4 Framework for Recognition

Given gallery and probe face videos, the model parameters (as explained in previous section) for each one of them are estimated. The gallery model, which is *closest* to the probe model, is assigned as the identity of the probe. We here discuss the metrics used to measure this degree of similarity.

Computing the L_2-norm of the difference between corresponding model matrices as a measure of distance will not suffice as it implicitly ignores the underlying geometry of the subspaces which is non-Euclidean. We make use of subspace angles between ARMA models for this cause. We follow the mathematical formulation given in [5] to compute these angles. The subspace angles are defined as the principal angles between the column spaces generated by the observability matrices of the two matrices extended with the observability matrices of the corresponding inverse models. Principal angles between two subspaces are the angles between their principal directions.

Cock *et al*, in [5], convert the ARMA model as represented in (5) and (6) into a forward innovation model:

$$\hat{x}_{t+1} = A\hat{x}_t + Ke_t \tag{12}$$

$$y_t = C\hat{x}_t + e_t \tag{13}$$

where $K \in \mathbb{R}^n$ is the Kalman gain as described in [11]. The problem of computing the subspace angles between the two models can be transformed into an eigenvalue problem involving the system parameters of forward and inverse innovation models.

In order to estimate the distance between two models, we need certain distance measures based on the computed subspace angles. There are several distance metrics based on subspace angles between ARMA models. The first one is due to Martin [10] and can mathematically be written as:

$$d_M(M1, M2)^2 = ln \prod_{i=1}^{n} \frac{1}{cos^2\theta_i} \tag{14}$$

where M_1 and M_2 are two ARMA models and θ_i's are the subspace angles between them. Other distance measures include gap and Frobenius norm based distances defined as follows :

$$d_g(M1, M2) = sin\theta_{max} \quad and \tag{15}$$

$$d_f(M_1, M_2)^2 = 2 \sum_{i=1}^{n} sin^2\theta_i \tag{16}$$

There is another distance described in [15] which is the largest principal angle between the two models. In our experiments, all these metrics give similar recognition performance.

5 Experiments, Results and Discussion

We conducted face recognition experiments using the proposed framework on two datasets. The first one is same as the one used by Li *et al* in [8]. It has face videos for 16 subjects with 2 sequences per subject. In these sequences, the subjects arbitrarily move their heads and change their expressions. The illumination conditions for 2 sequences of each subject were quite different. For each subject, one sequence was put in the gallery while the other formed a probe. A few example images from this dataset are shown in Figure 1.

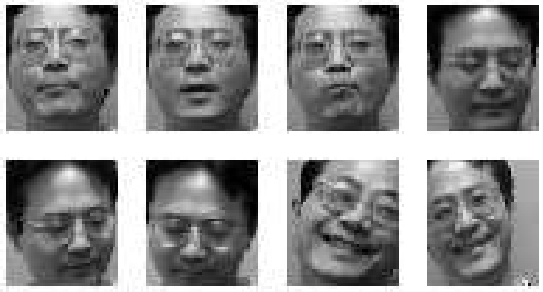

Figure 1. Few cropped faces from a video sequence in the first dataset.

The second dataset (obtained from UCSD/Honda) is the one used by Lee *et al* in [7]. With this dataset, we had a gallery of size 15 and probe containing 30 video sequences. In each video, subject moves his/her face in an arbitrary sequence of 2-D and 3-D rotations while changing facial expression and speed. There is even partial occlusion in a few frames of several video sequences. The illumination conditions vary significantly among the various sequences. Although both the datasets used are small, we consider them good tests for our algorithm because of the extreme pose and expression variations and varying illumination as is evident from Figures 1 and 2.

Figure 2. Few cropped faces from a video sequence in the UCSD/Honda dataset.

Our experiment broadly consists of three steps: preprocessing, model estimation and recognition. The preprocessing step involves cropping out the face from each frame of the video sequence. We use a variant of KL tracker [14] to track the nose tip location and an edge-based rough pose estimator. The nose tip location gives an idea about the location of the face while the pose information helps in getting the expanse of the face image relative to the nose. Figures 1 and 2 show few of the images cropped using this automatic method. Model estimation involves estimating A, C and K for each face sequence using the closed form solution explained in Section 3 while recognition involves computing the principal angles between probe and gallery models and using them to compute the distances between the models.

With both the data sets, we got recognition performance of more than 90% ($15/16$ for the first dataset and $27/30$ for the second). These numbers are very promising given the extent of pose and expression variations in the video sequences. The results reported in [7] are on per-frame basis and are not directly comparable even though one of the datasets used is the same.

6. Conclusion

We presented a structured approach to the problem of video-based face recognition. In particular, we dealt with the problem of recognizing faces when both gallery and probe consists of face videos. In our framework, a moving face is represented as a linear dynamical system whose appearance changes with time. Subspace angles based distance metrics are used to get the measure of similarity between ARMA models representing moving face sequences. The experiments conducted show that the system performs well even in case of extreme 2-D and 3-D pose variations, expression changes and ordinary illumination conditions.

References

[1] M. M. Ali. Analysis of stationary spatial-temporal processes: Estimation and prediction. *Biometrika*, 66:513–518, 1979.

[2] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2001.

[3] Z. Biuk and S. Loncaric. Face recognition from multi-pose image sequence. In *Proc. of 2nd International Symposium on Image and Signal Processing and Analysis*, Pula, Croatia, 2001.

[4] H. Bulthoff, S. Edelman, and M. Tarr. How are three-dimensional objects represented in the brain? *MIT AI Memo #1479*.

[5] K. D. Cock and D. B. Moor. Subspace angles and distances between ARMA models. In *Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000.

[6] G. Doretto and S. Soatto. Editable dynamic textures. In *ACM SIGGRAPH Sketches and Applications*, 2002.

[7] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-base face recognition using probabilistic appearance manifolds. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2003.

[8] B. Li and R. Chellappa. Face verification through tracking facial features. In *Journal of the Optical Society of America*, 2001.

[9] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2003.

[10] R. J. Martin. A metric for ARMA processes. *IEEE Transactions on Signal Processing*, 48:1164–1170, 2000.

[11] P. V. Overschee and B. D. Moor. Subspace algorithms for the stochastic identification problems. *Automatica*, 29:649–660, 1993.

[12] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. of European Conference on Computer Vision*, Kopenhagen, Denmark, 2002.

[13] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *Proc. of Intl. Conf. on Computer Vision*, 2001.

[14] C. Tomasi and T. Kanade. Detection and tracking of point features. *Carnegie Mellon University Technical Report CMU-CS-91-132*, April 1991.

[15] A. Weinstein. Almost invariant submanifolds for compact group actions. *Berkeley CPAM Preprint Series*, 1999.

[16] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. In *Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, 2003.

[17] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.

[18] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

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