Video-Based Face Recognition Using a Probabilistic Graphical Model

Yi-Chia Chan
Dept. of Comp. Science and Info.
Eng., National Taiwan Normal U.
No. 88, Sec. 4, Ting-Chou Rd.
Taipei 114, Taiwan, R.O.C.
yichia0427@yahoo.com.tw

Cheng-Chieh Chiang
Dept. of Info. Tech.
Takming U. of Science & Tech.
No. 56, Sec. 1, Huan-Shan Rd.
Taipei 114, Taiwan, R.O.C.
kevin@csie.ntnu.edu.tw

Kai-Ming Wang, Greg C. Lee
Dept. of Comp. Science and Info.
Eng., National Taiwan Normal U.
No. 88, Sec. 4, Ting-Chou Rd.
Taipei 114, Taiwan, R.O.C.
sibevin@yahoo.com.tw,
leeg@csie.ntnu.edu.tw

Abstract

This paper presents a probabilistic graphical model to formulate and deal with video-based face recognition. Our formulation divides the problem into two parts: one for likelihood measure and the other for transition measure. The likelihood measure can be regarded as a traditional task of face recognition within a single image, i.e., to recognize who the current observing face image is. In our work, two-dimensional linear discriminant analysis (2DLDA) is employed to judge the likelihood measure. Moreover, the transition measure estimates the probability of the change from a false recognition at the previous stage to the correct person at the current stage. Our approach for transition measure does not only consider the visual difference among persons according to the training face images but also involve prior information of the pose change in video frames. We also provide several experiments to show the efficiency of our proposed approach in this paper.

1 Introduction

Face recognition is an important and active topic in pattern recognition. That is also a key technology widely applied in computer vision. In traditional, face recognition is treated as a supervised learning, i.e., classifiers are trained by a set of prepared face images associated with persons and then new face images are recognized by use of the classifiers. Different methods of classifier learning, e.g., eigenface [12], PCA and LDA [2], or SVM [3], have been proposed to deal with the problem. There are two literature surveys for face recognition in [11] and [16].

In recent, researchers have focused on video-based face recognition that recognizes who are appeared in a video stream. In principle, video-based face recognition can be considered face recognition in a set of sequential and continuous images. However, there are, in fact, more information hidden in video frames. For example, face poses of a person could be changed in a video, and that may be helpful to improve face recognition. A face recognition method using temporal voting is proposed for image sequences in [9]. Considering in continuous video frames, visual features extracted from face images could form a manifold in high-dimensional feature space. Thus, we can convert face recognition to be a matching problem between the corresponding manifolds [1][6][13]. Another approach treats face images from video frames as 3D models and the problem is converted to a 3D model matching and recognition [4][7][10].

This work deals with video-based face recognition in a static environment such as a classroom that the members are fixed. We assume there are *K* persons in the system. In a video, these persons may be appeared with different face poses, or not appeared. Similarly, we have their face and pose images for training. Our goal is to build a model to recognize whose face in a video is.

Our basic idea is like a tracking task: to track the selection in the *K* candidates over time according to the observations of visual features in video frames. That motivates us to employ the state space model to construct a probabilistic graphical model for video-based face recognition. Our formulation divides video-based face recognition into two parts: likelihood and transition measures. The former is like a traditional task of face recognition in a single image to make a decision who the current observing face image is. The latter measures the probability of the change from a false recognition at the previous stage to the correct person at the current stage.

The rest of this paper are organized as the follows. Section 2 presents the basic state space model briefly. In Section 3, we formulate the task of video-based face recognition based on a probabilistic graphical model by revising the basic state space model. Therefore, we describe how to compute the likelihood measure using 2DLDA in Section 4 and how to measure the transition probabilities according to the training face images and the prior information of poses in Section 5. Section 6 provides several experimental results to show the efficiency of our proposed approach. In final, the conclusion and future works are drawn in Section 7.

2 State Space Model

A state space model is based on Bayesian network to analyze dynamic systems, which estimate the states of systems changing over time from a sequence of noisy measurements [5][8]. A state space model in general contains two types of nodes at time *t*: (i) x_t for the system state and (ii) z_t for the observation measurement, whose probabilistic graphical structure is shown as Figure 1.

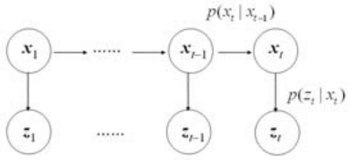

Figure 1. The probabilistic graphical structure of a state space model.

To simply express the equations, we use the notations

$X_t=\{x_1, ..., x_t\}$ and $Z_t=\{z_1, ..., z_t\}$ for all states and observations, respectively, over time t. There are two basic assumptions in the model, which can be available by use of d-separation property [8] of Bayesian Network. The first is the first-order Markov property, i.e.,

$$p(x_t \mid X_{t-1}) = p(x_t \mid x_{t-1}), \qquad (1)$$

and the second is that the observations are mutually independent:

$$p(z_t \mid X_t, Z_{t-1}) = p(z_t \mid x_t). \qquad (2)$$

3 Formulation for Video-based Face Recognition

Now we formulate the problem of face recognition using Bayesian network by extending the state space model. Assume that there are K persons in the system so that the state vector x_t indicates which person the system recognizes at time t. Also, an observation z_t means a face image which the system acquires at time t. Hence, the observation set $Z_t=\{z_1, ..., z_t\}$ collects the face images in video frames, and $X_t=\{x_1, ..., x_t\}$ shows the recognition results of face images of these observations.

However, the basic state space model shown in Figure 1 could not reach an accurate recognition while people are changing their poses in video frames. Suppose that face images of a person can be categorized into R poses denoted $H=\{h_1, ..., h_R\}$. Our approach, in the probabilistic model, is to insert additional pose nodes according to the prior analysis of face poses. The Bayesian network of our proposed model with the pose nodes is shown in Figure 2.

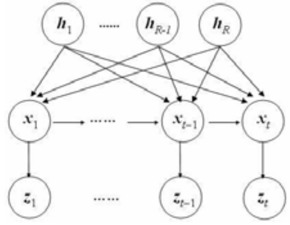

Figure 2. The probabilistic structure of the state space model for face recognition with pose nodes.

Lemma 1: Given the pose information $H=\{h_1, ..., h_R\}$ and the set of observations $Z_t=\{z_1, ..., z_t\}$ at time t for the Bayesian network in Figure 2, the posterior probability of the state x_t can be computed as:

$$p(x_t \mid Z_t, H) \propto$$
$$p(z_t \mid x_t) \int p(x_t \mid x_{t-1}, H) p(x_{t-1} \mid Z_{t-1}, H)\, dx_{t-1} \qquad (3)$$

Proof: see the appendix.

Thus, there are three factors to determine which person the state x_t is: (i) $p(z_t|x_t)$ means the likelihood measure for current observations, (ii) $p(x_t|x_{t-1}, H)$ means the transition measure based on pose information for the previous state, and (iii) $p(x_{t-1}|Z_{t-1}, H)$ is the recursive result at the previous iteration. Likelihood and transition measures are described in details in Section 4 and 5, respectively. At the beginning, moreover, an initial person should be known for the prior of the state vector. In this work, we apply 2DLDA method that is also used for likelihood measure to recognize the first face image.

In order to more simply achieve face recognition according to Eq. (3) in practice, two assumptions are held in this paper. The first is to assume that face images have been properly cropped in video frames. That can be performed by face detection. The second is to assume that poses of face images are aligned. That is to say, we define R poses for face images and each of training face images can be categorized into a pose. In our work, we apply k-means clustering to roughly divide training face images into R subsets and manually check whether face images are the same pose in the same subset.

4 Likelihood Measure Using 2DLDA

The likelihood term, denoted as $p(z_t|x_t)$, in Eq. (3) measures the possibility of the current observations given a state (i.e., a known person). That can be estimated by the similarity measure between the face image of the current observation and the training images of the given person. Thus, the computation of the likelihood measure for a face image in video frames can be regarded as a task of face recognition in the image.

In this work, we adopt 2DLDA (two-dimensional linear discriminant analysis) [15] for face recognition in a single image. 2DLDA employs IMLDA (uncorrelated image matrix-based linear discriminant analysis) [14] twice: one for horizontal and the other for vertical direction shown as Figure 3 which is taken from [15]. In principle, 2DLDA selects most discriminative features of images in vertical and horizontal directions.

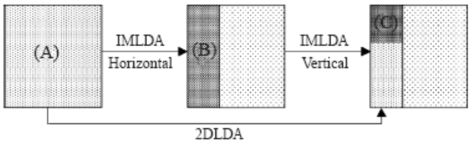

Figure 3. Illustration of 2DLDA.

Given training face images of the K persons, a dxd-dimensional subspace that can best separate these images can be trained according to 2DLDA transformation. In our work, we set d=3 and the 2DLDA plane is 9-D. Suppose that m_i is the mean of projected training images for a person M_i, i=1 to K, in the 2DLDA plane. Given a test face image z_t for the observation at time t, we can compute the distance z'_t-m_i between the projection point of z_t and the mean m_i of the person M_i in the plane. Therefore, the likelihood term $p(z_t|x_t)$ can be estimated by normalizing the distance z'_t-m_i:

$$p(z_t \mid M_i)$$
$$= (2\pi)^{-\frac{d}{2}} \mid C \mid^{-\frac{1}{2}} \exp(-\frac{1}{2}(z_t^{'} - m_i) C^{-1} (z_t^{'} - m_i)^t) \qquad (4)$$

where C is the covariance matrix of training images for M_i in the 2DLDA plane.

5 Transition Measure

The transition function, denoted as $p(x_t|x_{t-1}, H)$ in Eq. (3), measures the transitive possibility while the system makes a false recognition for the observing test images. According to the following equation

$$p(x_t \mid x_{t-1}, H) = \frac{p(H \mid x_t, x_{t-1}) p(x_t \mid x_{t-1}) p(x_{t-1})}{p(H \mid x_{t-1}) p(x_{t-1})},$$
$$= p(x_t \mid x_{t-1}) \frac{p(H \mid x_t, x_{t-1})}{p(H \mid x_{t-1})} \qquad (5)$$

we know the transition measure can be divided into two parts: one is to consider the transition $p(x_t|x_{t-1})$ without any pose information, and the other is the relationship of the pose recognitions between two successive iterations.

Regarding the first term $p(x_t|x_{t-1})$, that could be learned by use of the training images associated with persons. That is to say, we can analyze the similarity among face images of different persons to define the transition function of persons. Note that the similarity is computed in the projected 2DLDA plane. Following the notations in the previous section, let I_i be the set of the projected points of training images of a person M_i. Then, the similarity of any two persons M_i and M_j can be defined as

$$sim(M_i, M_j) = \frac{1}{|I_i|}\left(\sum_{r \in I_i}(r - m_j)(r - m_j)^t\right)^{1/2} \qquad (6)$$

and then normalized by Gaussian distribution.

Table 1. The transition probabilities of the changes of face poses. Seven poses are included: front, right, left, up, down, right-up, and left-up.

t t-1	Front	Right	Left	Up	Down	Right up	Left up
Front	0.2778	0.1667	0.1667	0.1667	0.1667	0.0278	0.0278
Right	0.24	0.4	0.04	0.08	0.08	0.04	0.12
Left	0.24	0.04	0.4	0.08	0.08	0.12	0.04
Up	0.2222	0.1111	0.1111	0.3704	0.037	0.0741	0.0741
Down	0.24	0.12	0.12	0.04	0.4	0.04	0.04
Right up	0.25	0.0417	0.125	0.0833	0.0417	0.4167	0.0417
Left up	0.25	0.125	0.0417	0.0833	0.0417	0.0417	0.04167

Regarding the second term $p(H|x_t, x_{t-1})/p(H|x_{t-1})$, it is difficult to induce a simple closed-form. In this work, instead of, that is approximated to the possibility of the change of the face poses in successive iterations t and t-1. For example, the front pose (0°) should be more easily change to a small-degree pose (e.g., 30°) than a large-degree pose (e.g., 60°). Hence, two tasks are applied to measure the term $p(H|x_t, x_{t-1})/p(H|x_{t-1})$. The first is to recognize which pose the current observing face image is. We also employ 2DLDA to make face pose classifiers like face classifiers in Section 4. The second is to define prior probabilities of the change between any two face poses. According to our experiences of face pose changing in video, we summarize the counting of pose changes and normalize them to be the prior probabilities shown as Table 1. Note that the matrix is not symmetric. For example, a pose change from right to front is of course more possible than that from front to right.

6 Experimental Results

In the experiments, the Honda/UCSD Video Database [6][17] is adopted for our training and test dataset. We arbitrarily select 10 persons from the dataset for our experiments. Figure 4 illustrates their photos and names. In these video frames, people move their head with the seven face poses listed in Table 1.

The basic experiment is to evaluate the precision of face recognition in video frames. That is computed by (the number of video frames correctly recognized)/(the total number of video frames). Table 2 lists the details of precisions for each person and their averages according to Eq. (3). We employ 2DLDA and PCA to face or pose recognition described in Section 4 and 5. For example, "PCA+2DLDA" means that PCA is used for likelihood measure in Section 4 and 2DLDA is used for pose recognition in Section 5. In additional, we perform our method (2DLDA+2DLDA) with and without transition for comparison. Table 2 clearly presents that our approach makes a significant improvement, either in using 2DLDA for likelihood measure or in applying transition measure as well as pose information.

Figure 4. Photos and names of the ten persons used for the experiments.

Table 2. The detailed and average precisions using our proposed approach with different classifiers in the likelihood and transition measures.

Name	2DLDA+ 2DLDA+ with transition	2DLDA+ 2DLDA+ without transition	PCA+ 2DLDA+ with transition	PCA+ PCA+ with transition
Behzad	0.91906	0.8225	0.86684	0.84334
Danny	0.82289	0.7847	0.68119	0.65938
Fuji	0.65862	0.6621	0.65862	0.67931
James	0.88179	0.77	0.96166	0.96166
Jeff	0.73801	0.7168	0.64425	0.60708
Joey	0.66567	0.5701	0.43582	0.49851
Ming	0.76607	0.7404	0.46787	0.49356
Rakesh	0.98201	0.9537	0.94087	0.94858
Wei	0.77236	0.7175	0.68293	0.66666
Yokoyama	0.85953	0.7559	0.69565	0.66889
AVG	**0.806601**	**0.74937**	**0.70357**	**0.70269**

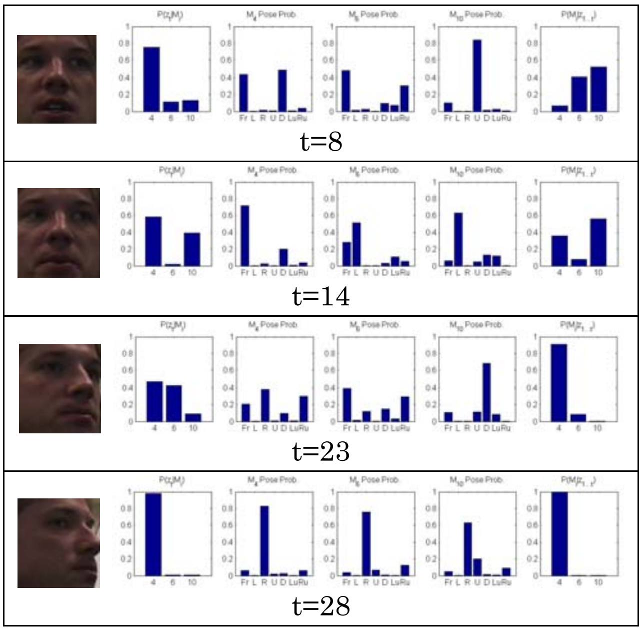

Figure 5. Illustration of the recognition process over time.

Next, let us discuss the convergence process with the likelihood and the transition measure over time. Figure 5 illustrates an example of face recognition at time 8, 14, 23, 28. Note that the person of the example is James with index “4” in plots. His face poses changed from front to left in this example. The observing person is identified incorrectly as “yokoyama” in initial (*t*=1), but he is recognized correctly in final (*t*=28). There are five plots at each row. This example only displays the probability values for three persons for simplicity. The first plot shows the likelihood measure of the current observation according to Eq. (4). The second to fourth plots display the probabilities of face poses for different persons given the observation. The last plot shows the final probability of persons given the observing face image. In general, it is difficult to avoid false decision either for face or pose recognition. However, our method makes a possibility converging to the correct decision by aggregating the recognitions in likelihood and transition measures such as illustrated in the last two iterations.

7 Conclusion and Future Works

This paper proposes a probabilistic graphical model to deal with face recognition in video frames. Our approach, consisting of two measures: likelihood and transition, does not only perform a basic face recognition in a single video frame but also consider the change of poses over time. We employ 2DLDA to recognize faces for likelihood measure and poses for transition measure. In the future, our plan is to extend the dataset for experiments to achieve more significant results. We also plan to design an incremental learning algorithm with our probabilistic model. That could improve face or pose recognition in likelihood and transition measure and make our proposed model more robust.

Acknowledgement

This work was supported by National Science Council, Taiwan, under Grant No. NSC 97-2511-S-003-007-MY3, NSC 97-2631-S-003-003, and NSC 97-2221-E-147-005.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, “Face Recognition with Image Sets Using Manifold Density Divergence,” in Proceedings of CVPR, 2005.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,” IEEE Trans. on PAMI, Vol. 19, No. 1, pp. 711− 720, 1997.

[3] B. Heisele, P. Ho, and T. Poggio, “Face recognition with support vector machines: Global versus component-based approach,” in Proceedings of ICCV, 2001.

[4] Y.-X. Hu, D.L. Jiang, S.-C. Yan, L. Zhang, and H.-J. Zhang, “Automatic 3D reconstruction for face recognition,” in Proceedings of International Conference on Automatic Face and Gesture Recognition, 2004.

[5] Z. Ghahramani, “An Introduction to Hidden Markov Models and Bayesian Networks”, International Journal of Pattern Recognition and Artificial Intelligence, 15(1): 9-42, 2001.

[6] K.C. Lee and J. Ho and M.H. Yang and D. Kriegman, “Visual Tracking and Recognition Using Probabilistic Appearance Manifolds,” Computer Vision and Image Understanding, Vol. 99, No. 3, pp. 303-331, 2005.

[7] X.-G. Lu, A.K. Jain, and D. Colbry, “Matching 2.5D face scans to 3D models,”, IEEE Trans. on PAMI, Vol. 28, No. 1, pp. 31-43, 2006.

[8] K. P. Murphy, “Dynamic Bayesian Networks: Representation, Inference and Learning”, U. C. Berkeley, PhD. Thesis, 2002.

[9] G. Shakhnarovich, J. W. Fisher, and T. Darrell, “Face recognition from long-term observations,” in Proceedings of ECCV, pp. 851-865, 2002.

[10] F. B. ter Haar and R. C. Veltkamp, “3D Face Model Fitting for Recognition,” in Proceedings of ECCV, 2008.

[11] A. S. Tolba, A.H. El-Baz, and A.A. El-Harby, “Face Recognition: A Literature Review”, International Journal of Signal Processing, Vol. 2, No. 1, pp. 88-103, 2005.

[12] M. Turk and A. Pentland, “Eigenfaces for recognition”, Journal of Cognitive Neuroscience, Vol. 3, pp. 72-86, 1991.

[13] R.-P. Wang, S.-G. Shan, X.-L. Chen, and W. Gao, “Manifold-Manifold Distance with application to face recognition based on image set,” in Proceedings of CVPR, 2008.

[14] J. Yang, J.-Y. Yang, A.F. Frangi, and D. Zhang, “Uncorrelated projection discriminant analysis and its application to face image feature extraction,” International Journal of Pattern Recognition and Artificial Intelligence, Vol. 17, No. 8, pp. 1325–1347, 2003.

[15] J. Yang, D. Zhang X. Yong, and J. Yang, “Two-dimensional Discriminant Transform for Face Recognition,” Pattern Recognition, Vol. 38, No. 7, July 2005.

[16] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, “Face recognition: a literature survey”, ACM Computing Surveys, 35(4): 399-458, 2003.

[17] The Honda/UCSD Video Database, http://vision.ucsd.edu/~leekc/HondaUCSDVideoDatabase/HondaUCSD.html.

Appendix

Proof of Lemma 1

According to the two assumptions in Eq. (1) and (2), and using d-separation property [8] of Bayesian network for Figure 2, we could have the following four properties of conditional independence:

$$p(x_t \mid X_{t-1}, H) = p(x_t \mid x_{t-1}, H)\,, \tag{7}$$

$$p(z_t \mid x_t, Z_{t-1}, H) = p(z_t \mid x_t)\,, \tag{8}$$

$$p(x_t \mid X_{t-1}, Z_{t-1}) = p(x_t \mid X_{t-1})\,, \tag{9}$$

$$p(H \mid X_t, Z_t) = p(H \mid X_t)\,. \tag{10}$$

Then, we can induce the following equations

$$\begin{aligned}
& p(x_t \mid Z_t, H) \propto \int p(X_t, Z_t, H)\,dX_{t-1} \\
& = \int p(H \mid X_t, Z_t)\, p(X_t, Z_t)\,dX_{t-1} = \int p(H \mid X_t)\, p(X_t, Z_t)\,dX_{t-1} \\
& = \int p(H \mid X_t)\, p(z_t \mid x_t)\, p(x_t \mid X_{t-1})\, p(X_{t-1}, Z_{t-1})\,dX_{t-1} \\
& = \int \frac{p(H, X_t)}{p(X_t)}\, p(z_t \mid x_t)\, p(x_t \mid X_{t-1})\, p(X_{t-1}, Z_{t-1})\,dX_{t-1} \\
& = \int \frac{p(H, x_t \mid X_{t-1})}{p(x \mid X_{t-1})}\, p(z_t \mid x_t)\, p(x_t \mid X_{t-1})\, p(X_{t-1}, Z_{t-1})\,dX_{t-1} \\
& = p(z_t \mid x_t) \int p(x_t \mid X_{t-1}, H)\, p(H \mid X_{t-1})\, p(X_{t-1}, Z_{t-1})\,dX_{t-1} \\
& = p(z_t \mid x_t) \int p(x_t \mid X_{t-1}, H)\, p(H \mid X_{t-1}, Z_{t-1})\, p(X_{t-1}, Z_{t-1})\,dX_{t-1} \\
& = p(z_t \mid x_t) \int p(x_t \mid X_{t-1}, H)\, p(H, X_{t-1}, Z_{t-1})\,dX_{t-1} \\
& \propto p(z_t \mid x_t) \int p(x_t \mid x_{t-1}, H)\, p(x_{t-1} \mid Z_{t-1}, H)\,dx_{t-1}
\end{aligned}$$

and the proof is done.