SIMULTANEOUS TRACKING AND RECOGNITION OF HUMAN FACES FROM VIDEO

Shaohua Zhou and Rama Chellappa

Center for Automation Research
ECE Department, University of Maryland
College Park, MD 20742
{shaohua, rama}@cfar.umd.edu

ABSTRACT

This paper investigates the interaction between tracking and recognition of human faces from video under the framework proposed earlier [1, 2], where a time series model is used to resolve the uncertainties in both tracking and recognition. However, our earlier efforts employed only a simple likelihood measurement in the form of a Laplacian density to deal with appearance changes between the frames and between the observation and the gallery images, yielding poor accuracies in both tracking and recognition when confronted by pose and illumination variations. The interaction between tracking and recognition was not well understood. We address the interdependence between tracking and recognition using a series of experiments and quantify the interacting nature of tracking and recognition.

1. INTRODUCTION

Recognition of human faces from video requires resolving the uncertainties in both tracking and recognition. While conventional approaches [3] resolve both uncertainties separately, i.e. after tracking is accomplished, recognition is then applied, we have proposed [1, 2] a framework to model both uncertainties in a unified way.

However, our earlier efforts employed only a simple likelihood measurement in the form of a Laplacian density to deal with appearance changes between the frames and between the observation and the gallery images. As a consequence, this yielded unsatisfactory results in both tracking and recognition when confronted by pose and illumination variations. Also, the interaction between tracking and recognition was unclear.

In this paper, we attempt to address these issues by asking the following questions: What is the relationship between tracking and recognition in face recognition from video problem? Is a good recognition metric modeling the appearance changes between the observation and the gallery images is good for tracking? How good is our approach in terms of recognition rate compared to the still-image-based recognition system? To find the answers, we conducted a series of experiments by bringing in face modeling techniques and study the interdependence between tracking and recognition.

The rest of the paper is structured as follows. After a brief review of the literature on face modeling and recognition in Sec. 2, and a recapitulation of our previous time series mode for recognition in Sec. 3, we present our experimental results in Sec. 4, where five cases are presented to mainly address the interacting

effects between tracking and recognition. Sec. 5 concludes the paper.

2. FACE MODELING AND RECOGNITION

Statistical approaches to face modeling have been very popular since Turk and Pentland's work on eigenface in 1991 [4]. In the statistical approach, the two-dimensional appearance of face image is treated as a vector by scanning the image in lexicographical order, with the vector dimension being the number of pixels in the image. In the eigenface approach [4], all face images consists of a distinctive face subspace. This subspace is linear and spanned by the eigenvectors of the covariance matrix found using PCA. Typically we keep the number of eigenvectors much less than the true dimension of the vector space. The task of face recognition is then to find the closest matches in this face subspace. However, PCA might not be efficient in terms of recognition accuracy since the construction of the face subspace does not capture discrimination between humans. This motivates the use of LDA [5, 6] and its variants. In LDA, the linear subspace is constructed [7] in such a manner that the within-class scatter is minimized and the between-class scatter is maximized. This idea is further generalized in the approach called Bayesian face recognition [8], where intra-personal space (IPS) and extra-personal space (EPS) are used in lieu of within-class scatter and between-class scatter measures. The IPS models the variations in the appearance of the same individual and the EPS models the variations in the appearance due to a difference in the identity. Probabilistic subspace density is then fitted on each space. A Bayesian decision is taken using a *maximum a posteriori* (MAP) rule to determine the identity.

Neural-networks have also been commonly used for face recognition. In the famous EGM [9] algorithm, the face is represented as a labeled graph. The nodes of the graph are located at facial landmarks, e.g., the pupils, the tip of nose, etc. Also, each node is labeled with jets derived from responses obtained by convolving the image with a family of Gabor functions. The edge characterizes the geometric distance between two nodes. Face recognition is then formalized as a graph matching problem.

All the above approaches are based on 2-D appearance and perform poorly when significant pose and illumination variations are present [10]. To completely resolve such challenges, 3-D face modeling [3] is necessary. However, building a 3-D face model is a very difficult and complicated task in the literature even though structure from motion has been studied for several decades.

3. TIME SERIES STATE SPACE MODEL FOR RECOGNITION

In this section, we briefly present the propagation model for recognition, consisting of the following three components, and define the recognition task as a statistical inference problem.

Motion equation

In its most general form, the motion model can be written as

$$\theta_t = g(\theta_{t-1}, u_t); \quad t \geq 1, \tag{1}$$

where u_t is *noise* in the motion model, whose distribution determines the motion state transition probability $p(\theta_t|\theta_{t-1})$. The function $g(.,.)$ characterizes the evolving motion and it could be a function learned offline or given a priori. One of the simplest choice is an additive function, i.e., $\theta_t = \theta_{t-1} + u_t$, which leads to a first-order Markov chain.

Choice of θ_t is application dependent. Affine motion parameters are often used when there is no significant pose variation available in the video sequence. However, if a 3-D face model is used, 3-D motion parameters should be used accordingly.

Identity equation

$$n_t = n_{t-1}; \quad t \geq 1, \tag{2}$$

assuming that the identity does not change as time proceeds.

Observation equation

By assuming that the transformed observation is a noise-corrupted version of some still template in the gallery, the observation equation can be written as

$$\mathcal{T}_{\theta_t}\{z_t\} = I_{n_t} + v_t; \quad t \geq 1, \tag{3}$$

where v_t is *observation noise* at time t, whose distribution determines the observation likelihood $p(z_t|n_t, \theta_t)$, and $\mathcal{T}_{\theta_t}\{z_t\}$ is a transformed version of the observation z_t. This transformation could be either geometric or photometric or both. However, when confronting sophisticated scenarios, this model is far from sufficient. One should seek for complicated likelihood measurement as shown in Section 5.

We assume statistical independence between all noise variables and prior knowledge on the distributions $p(\theta_0|z_0)$ and $p(n_0|z_0)$. Using the overall state vector $x_t = (n_t, \theta_t)$, Eq. (1) and (2) can be combined into one state equation (in a normal sense) which is completely described by the overall state transition probability

$$p(x_t|x_{t-1}) = p(n_t|n_{t-1})p(\theta_t|\theta_{t-1}) . \tag{4}$$

Given this model, our goal is to compute the posterior probability $p(n_t|z_{0:t})$. It is in fact a probability mass function (PMF) since n_t only takes values from $\mathcal{N} = \{1, 2, ..., N\}$, as well as a marginal probability of $p(n_t, \theta_t|z_{0:t})$, which is a mixed distribution. Therefore, the problem is reduced to computing the posterior probability.

In [1], we invoked the Condensation algorithm, a special case of Sequential Monte Carlo (SMC) methods [11], to provide numerical approximations to the posterior distribution $p(n_t, \theta_t|z_{0:t})$. In [2], we greatly improved the computational load by judiciously
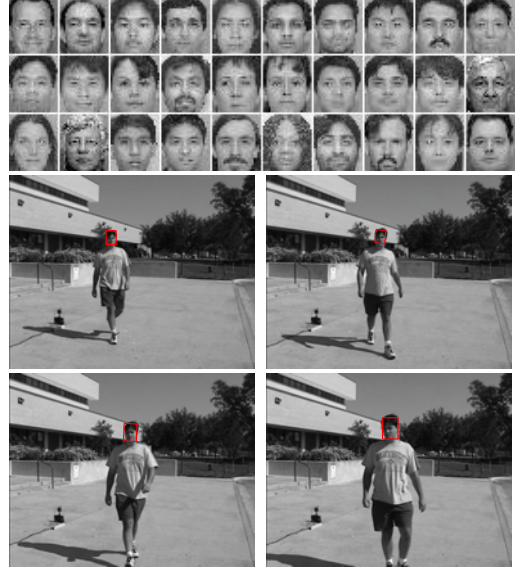
Fig. 1. Database-1. The 1st row: the face gallery with image size being 30x26. The 2nd and 3rd rows: 4 example frames in one probe video with image size being 720x480 while the actual face size ranges approximately from 20x20 in the first frame to 60x60 in the last frame. Notice the significant illumination variations between the probe and the gallery. The tracking results obtained has been illustrated by the superimposed bounding box.

utilizing the discrete nature of the identity variable. We [2] also theoretically justified the evolving behavior of the recognition density $p(n_t|z_{0:t})$ under one weak assumption.

4. EXPERIMENTAL RESULTS

In this section we describe the still-to-video scenarios used in our experiments and their practical model choices, followed by a discussion of experimental results.

In Database-1, we have video sequences with subjects walking in a slant path towards the camera. There are 30 subjects, each having one face template. There are one face gallery and one probe set. The face gallery is shown in Fig. 1. The probe contains 30 video sequences, one for each subject. Fig. 1 gives some example frames extracted from one probe video. As far as imaging conditions are concerned, the gallery is very different from the probe, especially in lighting. This is similar to the 'FC' test protocol of the FERET test [10]. These images/videos were collected, as part of the HumanID project, by National Institute of Standards and Technology and University of South Florida researchers.

4.1. Model Choices

We consider affine transformation. Specifically, the motion is characterized by $\theta = (a_1, a_2, a_3, a_4, t_x, t_y)$ where $\{a_1, a_2, a_3, a_4\}$ are deformation parameters and $\{t_x, t_y\}$ are 2-D translation parameters. It is a reasonable approximation since there is no significant out-of-plane motion as the subjects walk towards the camera. Regarding the photometric transformation, only zero-mean-unit-variance operator is performed to partially compensate for con-

trast variations. The complete transformation $\mathcal{T}_\theta\{z\}$ is processed as follows: affine transform z using $\{a_1, a_2, a_3, a_4\}$, crop out the interested region at position $\{t_x, t_y\}$ with the same size as the still template in the gallery, and perform zero-mean-unit-variance operation.

Prior distribution $p(\theta_0|z_0)$ is assumed to be Gaussian, whose mean comes from the initial detector and whose covariance matrix is manually specified.

A time-invariant first-order Markov Gaussian model with constant velocity is used for modeling motion transition. Given the scenario that the subject is walking towards the camera, the scale increases with time. However, under perspective projection, this increase is no longer linear, causing the constant-velocity model to be not optimal. However, experimental results show that as long as the samples of θ can cover the motion, this model is sufficient.

4.2. Results on Database-1

Case 1: Tracking and Recognition using Laplacian Density

We first investigate the performance when the likelihood measurement is simply set as a 'truncated' Laplacian:

$$p_1(z_t|n_t, \theta_t) = LAP(\|\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}\|; \sigma_1, \tau_1) \quad (5)$$

where, $\|.\|$ is sum of absolute distance, σ_1 and λ_1 are manually specified, and

$$LAP(x; \sigma, \tau) = \begin{cases} \sigma^{-1}\exp(-x/\sigma) & \text{if } x \leq \tau\sigma \\ \sigma^{-1}\exp(-\tau) & \text{otherwise} \end{cases} \quad (6)$$

Gaussian distribution is widely used as a noise model, accounting for sensor noise, digitization noise, etc. However, given the observation equation: $v_t = \mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}$, the dominant part of v_t becomes the high-frequency residual if θ_t is not proper, and it is well known that the high-frequency residual of natural images is more Laplacian-like. The 'truncated' Laplacian is used to give a 'surviving' chance for samples to accommodate abrupt motion changes.

Table 1 shows that the recognition rate is very poor, only 13% are correctly identified using top match. The main reason is that the 'truncated' Laplacian density is far from sufficient to capture the appearance difference between the probe and the gallery, thereby indicating a need for a different appearance modeling. Nevertheless, the tracking accuracy [1] is reasonable with 83% successfully tracked because we are using multiple face templates in the gallery to track the specific face in the probe video. After all, faces in both the gallery and the probe belong to the same class of human face and it seems that the appearance change is captured by the class model.

Case 2: Pure Tracking using Laplacian Density

In Case 2, we measure the appearance change within the probe video as well as the noise in the background. To this end, we introduce a dummy template T_0, a cut version in the first frame of the video. Define the observation likelihood for tracking as

$$q(z_t|\theta_t) = LAP(\|\mathcal{T}_{\theta_t}\{z_t\} - T_0\|; \sigma_2, \tau_2), \quad (7)$$

where σ_2 and τ_2 are set manually. The other setting, such as motion parameter and model, is the same as in Case 1. We still can run the CONDENSATION algorithm to perform pure tracking.

Table 1 shows that 87% are successfully tracked by this simple tracking model, which implies that the appearance within the video remains similar.

Accuracy	Case 1	Case 2	Case 3	Case 4	Case 5
Tracking	83%	87%	93%	100%	NA
Recognition (top 1)	13%	NA	83%	93%	57%
Recognition (top 3)	43%	NA	97%	100%	83%

Table 1. Performances of algorithms when applied to Database-1.

Case 3: Tracking and Recognition using Probabilistic Subspace Density

As mentioned in Case 1, we need a new appearance model to improve the recognition accuracy. As mentioned in Section 2, various approaches have been proposed in the literature. We decided to use the approach suggested by Moghaddam et. al. [8] due to its computational efficiency and high recognition accuracy. However, in our implementation, we model only intra-personal variations instead of both intra/extra-personal variations for simplicity.

We need at least two facial images for one identity to construct the intra-personal space (IPS). Apart from the available gallery, we crop out the second image from the video ensuring no overlap with the frames actually used in probe videos. Fig. 2 (top row) shows a list of such images. Compare with Fig. 1 to see how the illumination varies between the gallery and the probe.

We then fit a probabilistic subspace density [12] on top of the IPS. It proceeds as follows: a regular PCA is performed for the IPS. Suppose the eigensystem for the IPS is $\{(\lambda_i, e_i\}_{i=1}^d$, where d is the number of pixels and $\lambda_1 \geq ... \geq \lambda_d$. Only top s principal components corresponding to top s eigenvalues are then kept while the residual components are considered as isotropic. We refer the reader to the original paper [12] for full details. Fig. 2 (middle row) show the eigenvectors for the IPS. The density is written as follows:

$$PS(x) = \{\frac{exp(-\frac{1}{2}\sum_{i=1}^s \frac{y_i^2}{\lambda_i})}{(2\pi)^{s/2}\prod_{i=1}^s \lambda_i^{1/2}}\}\{\frac{exp(-\frac{\epsilon^2}{2\rho})}{(2\pi\rho)^{(d-s)/2}}\}, \quad (8)$$

where $y_i = e_i^T x$ for $i = 1, ..., s$ is the ith principal component of x, $\epsilon^2 = \|x\|^2 - \sum_{i=1}^s y_i^2$ is the reconstruction error, and $\rho = (\sum_{i=s+1}^d \lambda_i)/(d-q)$. It is easy to write the likelihood as follows:

$$p_2(z_t|n_t, \theta_t) = PS(\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}). \quad (9)$$

Table 1 lists the performance by using this new likelihood measurement. It turns out that the performance is significantly better than in Case 1, with 93% tracked successfully and 83% recognized within top 1 match. If we consider the top 3 matches, 97% are correctly identified.

Case 4: Tracking and Recognition using Combined Density

In Case 2, we have studied appearance changes within a video sequence. In Case 3, we have studied the appearance change between the gallery and the probe. In Case 4, we attempt to take

---

[1] We manually inspect the tracking results by imposing the MMSE motion estimate on the final frame as shown in Fig. 1 and determine if tracking is successful or not for this sequence. This is done for all sequences and tracking accuracy is defined as the ratio of the number of sequences successfully tracked to the total number of all sequences.

Fig. 2. Database-1. Top row: the second facial images for training probabilistic density. Middle row: top 10 eigenvectors for the IPS. Bottom row: the facial images cropped out from the largest frontal view.

advantage of both cases by introducing a combined likelihood defined as follows:

$$p_3(z_t|n_t, \theta_t) = p_2(z_t|n_t, \theta_t)q(z_t|\theta_t) \tag{10}$$

Again, all other setting is the same as in Case 1. We now obtain the best performance so far: no tracking error, 93% are correctly recognized as the first match, and no error in recognition when top 3 matches are considered.

Case 5: Still-to-still Face Recognition

To make a comparison, we also performed an experiment on still-to-still face recognition. We selected the probe video frames with the best frontal face view (i.e. biggest frontal view) and cropped out the facial region by normalizing with respect to the eye coordinates manually specified. This collection of images is shown in Fig. 2 (bottom row) and it is fed as probes into a still-to-still face recognition system with the learned probabilistic subspace as in Case 3. It turns out that the recognition result is 57% correct for the top one match, and 83% for the top 3 matches. The cumulative match curves for Case 1 and Cases 3-5 are presented in Fig. 3. Clearly, Case 4 is the best among all. We also implemented the original algorithm by Moghaddam et. al. [12], i.e., both intra/extra-personal variations are considered, the recognition rate is similar to that obtained in Case 5. One reason for the superiority of still-to-video approach is that we essentially compute the recognition score based on all video frames and, in each frame, all kinds of transformed versions of the face part corresponding to the sample configurations that are considered, while still-to-still approach derives its decision based some isolated shots.

5. CONCLUSION

We have studied the interaction between tracking and recognition in a framework proposed earlier. It has been found that (i) an effective recognition metric is effective for tracking as well, (ii) combining both contribution from tracking and recognition yields best
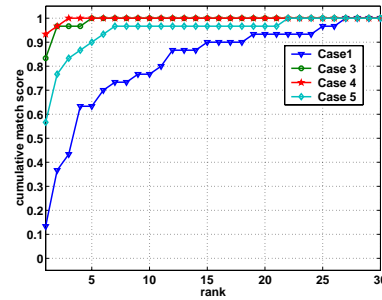
Fig. 3. Cumulative match score curves for Database-1.

performance in both tracking and recognition; and (iii) the still-to-video approach is superior to the still-to-still counterpart in terms of recognition rate.

6. REFERENCES

[1] Shaohua Zhou, V. Krueger, and R. Chellappa, "Face recognition from video: A condensation approach," *Proc. of the 5th International Conference on Face and Gesture Recognition*, 2002.

[2] Shaohua Zhou and R. Chellappa, "Probabilistic human recognition from video," *Proceedings of European Conference on Computer Vision*, 2002.

[3] T. Jebara and A. Pentland, "Parameterized structure from motion for 3d adaptive feedback tracking of faces," *Proc. of CVPR*, pp. 144 –150, 1997.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 72–86, 1991.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, pp. 711–720, 1997.

[6] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of America A*, pp. 1724–1733, 1997.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2001.

[8] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian modeling of facial similarity," *Advances in Neural Information Processing Systems*, vol. 11, pp. –, 1999.

[9] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Computers*, vol. 42, no. 3, pp. 300–311, 1993.

[10] P. J. Philipps, H. Moon, S. Rivzi, and P. Ross, "The feret evaluation methodology fro face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, pp. 1090–1104, 2000.

[11] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[12] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. PAMI*, vol. 24, pp. 780–788, 2002.