

PROBABILISTIC RECOGNITION OF HUMAN FACES FROM VIDEO

Rama Chellappa, Volker Krüger and Shaohua Zhou

Center for Automation Research
University of Maryland
College Park, MD 20742-3275

ABSTRACT

Most present face recognition approaches recognize faces based on still images. In this paper, we present a novel approach to recognize faces in video. In that scenario, the face gallery may consist of still images or may be derived from a videos. For evidence integration we use classical Bayesian propagation over time and compute the posterior distribution using sequential importance sampling. The probabilistic approach allows us to handle uncertainties in a systematic manner. Experimental results using videos collected by NIST/USF and CMU illustrate the effectiveness of this approach in both still-to-video and video-to-video scenarios with appropriate model choices.

1. INTRODUCTION

Face recognition has been studied extensively over the last years [3, 19, 15, 2]. Among the most successful ones are Principal Component Analysis (PCA) [16], Linear Discriminant Analysis (LDA) [7, 1], and Elastic Graph Matching (EGM) [18]. All these approaches consider the *still-to-still* scenario, where probe and gallery are given as mug-shots [15]. Experiments reported in [15] show that the above approaches give good results only if imaging conditions in gallery and probe set are sufficiently similar: if proper conditions are not met, e.g. due to illumination, out-of-plane rotation recognition rates drop severely.

To cope with non-optimal imaging conditions, we suggest to consider the *still-to-video* and the *video-to-video* scenarios: Here, the probe set consists of video, while the gallery consists of either mug-shots or a videos.

In the following we start discussing the *still-to-video* scenario, which we later generalize to the *video-to-video* scenario.

Denote the gallery by $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$, indexed by the identity variable $n \in \mathcal{N} = \{1, 2, \dots, N\}$.

Condensation and sequential importance sampling [10, 12, 6] are well-known approaches for probabilistic tracking where at each time instant the joint posterior distribution of a state vector describing the geometric deformation is

estimated. For our video-based approach, we enhance the probabilistic state vector with an identity variable $n \in \mathcal{N}$ of the gallery individuals. This way the identity is estimated at each time step, along with the geometric deformation vector. The probability of each identity is then given by the marginal distribution of the identity variable.

This approach can be enhanced by considering a video as a gallery for each individual. Inspired by well-known vector quantization techniques, we estimate a mixture density from the gallery video, with facial images as the mixture centers: the face region in the video is tracked and geometric distortion is compensated. The mixture model is then built from these geometrically normalized face images. The joint posterior distribution of a state vector and the identity variable is computed by marginalizing over the set of mixture centers.

Research on video-based recognition is less common than still-based recognition due to conditions generally encountered in surveillance applications[19]: poor video quality, significant illumination and pose variations, and low image resolution. Most video-based recognition systems [4] wait for good mug-shots in the video stream and therefore reduce the problem to the still-to-still scenario. This approach has severe drawbacks as it does not, above all, allow evidence integration over time. In [9, 14, 17], Radial Basis Function (RBF) networks are used for tracking and recognition. The system in [9] uses an RBF network for recognition. Since no warping is done, the RBF network has to learn the individual variations as well as possible transformations. The performance appears to vary widely, depending on the size of the training set. In [14], face tracking is based on an RBF network to provide feedback to a motion clustering process. Good tracking results were demonstrated, but person authentication results were referred to as future work. In [11], recognition of faces over time is implemented by constructing a face identity surface. The face is first warped to a frontal view, and its Kernel Discriminant Analysis (KDA) features over time form a trajectory. It is shown that the trajectory distances accumulate recognitive evidence over time.

In Section 2 we introduce our still-to-video approach and discuss its properties. In Section 3 we discuss the video-to-video based approach and we conclude the paper with an

This work was done with the support of the DARPA HumanID Project (Grant N00014-00-1-0908).

experiments in Sec. 4 and final conclusions.

2. STILL-TO-VIDEO FACE RECOGNITION

In this section we present details on how we establish the propagation model for recognition and discuss its impact on the posterior distribution of the identity variable under some weak assumptions.

2.1. Tracking and Recognition in the Bayesian Framework

In the following, I will denote an image of raw pixel data. A transformed version of image I is denoted by $T_\theta\{I\}$, where θ is the geometric deformation vector. The output of transformation $T_\theta\{I\}$ is an inner face region from image I , normalized w.r.t. the allowed set of geometric deformations Θ . In our experiments, θ is an affine deformation vector. We will denote the gallery by $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$, indexed by an *identity variable* $n \in \mathcal{N} = \{1, 2, \dots, N\}$. A video image at time step t will be denoted by Z_t .

By defining $X_t = (\theta_t, n_t) \in \mathcal{A} \times \mathcal{N}$ to be the state variable, we want to find X_t at each timestep t such that

$$p(X_t|Z_1, \dots, Z_t) \quad (1)$$

is maximal. Using the classical Bayesian propagation over time we get

$$\begin{aligned} p(X_t|Z_1, Z_2, \dots, Z_t) &\equiv p_t(\alpha_t, n_t) \\ &= \sum_{n_{t-1}} \int_{\alpha_{t-1}} p(Z_t|\alpha_t, n_t) p(\alpha_t, n_t|\alpha_{t-1}, n_{t-1})_{t-1} \\ &\quad p(\alpha_{t-1}, n_{t-1}) . \end{aligned} \quad (2)$$

The dynamic model we define as

$$p(X_t|X_{t-1}) = p(\theta_t|\theta_{t-1}) . \quad (3)$$

Marginalizing the posterior over the possible transformations $\theta \in \mathcal{A}$ we get a probability mass function for the identity:

$$p(n_t|Z_1, \dots, Z_t) = \int_{\theta_t} p(\theta_t, n_t|Z_1, \dots, Z_t) . \quad (4)$$

Maximizing (4) leads to the desired identity. To find a numerical solution for (1), we use a particle method [10, 12, 6].

2.2. Computation of the Observation Likelihood

For both, tracking and recognition, the observation likelihood for a hypothesis $X = (\theta, n)$ is given as

$$\begin{aligned} p(Z|X) &\equiv p(Z|\theta, n) \\ &\propto z \exp \left[-\frac{1}{2\sigma^2} d(I_n, T_\theta\{Z\}) \right] . \end{aligned} \quad (5)$$

Here, the input image Z is transformed according to the proposed transformation θ and then compared with the proposed gallery image I_n . Note, that d may be any distance function, e.g. one that has been well tested in face recognition applications.

2.3. The Posterior Probability of Identity Variable

To measure the evolving uncertainty remaining in the identity variable as observations accumulate, we use the notion of the conditional entropy $H(n_t|z_{0:t})$ [5]. The conditional entropy captures the evolving uncertainty of the identity variable given observations $z_{0:t}$. Since knowledge of $p(z_{0:t})$ is needed to compute $H(n_t|z_{0:t})$ we assume it degenerates to an impulse around the actual observations $\tilde{z}_{0:t}$ i.e., $p(z_{0:t}) = \delta(z_{0:t} - \tilde{z}_{0:t})$. We can therefore compute

$$H(n_t|z_{0:t}) = - \sum_{n_t \in \mathcal{N}} p(n_t|\tilde{z}_{0:t}) \log_2 p(n_t|\tilde{z}_{0:t}) . \quad (6)$$

3. VIDEO-TO-VIDEO FACE RECOGNITION

3.1. Exemplar-based Learning

In order to realize video-to-video based recognition, we generate for each individual a mixture model from his/her gallery video. Denote the gallery as $\mathcal{V} = \{V_1, V_2, \dots, V_N\}$.

We find a set of mixture centers (*exemplars*) that minimize the expected distance between the given set of video images $V = \{Z_1, Z_2, \dots, Z_T\}$ and a set of mixture centers (*exemplars*) $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$:

$$E \{d(\mathcal{Z}, \mathcal{C})\} . \quad (7)$$

In other words, we search for a set of exemplars $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$ such that

$$p(Z_t) = \sum_{c \in \mathcal{C}} \int_{\Theta} p(Z_t|\theta, c) p(\theta|c) p(c) d\theta \quad (8)$$

is maximal for all t . Here, $p(Z_t|\theta, c)$ is the observation equation, given as

$$\begin{aligned} p(Z_t|x) &\equiv p(Z_t|\theta, c) \\ &\propto \exp \left[-\frac{1}{2\sigma^2} d(T_\theta\{z_t\}, c) \right] . \end{aligned} \quad (9)$$

Inspired by the probabilistic interpretation of the RBF neural network [13], we use an online technique to learn the exemplars: At each time step t , $p(Z_t|\theta, c)$ of Eq. (8) is maximized. If $p(Z_t|\theta, c) < \rho$ for some ρ (which depends on the choice of d) then $T_\theta\{Z_t\}$ is added to the set of exemplars.

3.1.1. Exemplars as Mixture Centers

To take into account the exemplars $\mathcal{C}^n = \{c_1^n, c_2^n, \dots, c_{K_n}^n\}$ for individual n , we refine the likelihood computation of Eq. (5) as follows:

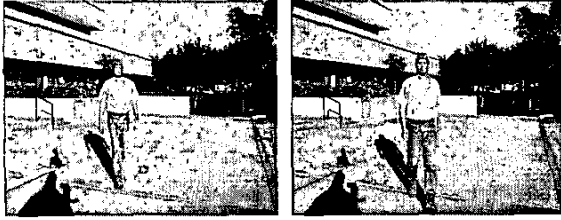


Fig. 1 Example frames in one probe video in Database-1. The image size is 720x480 while the actual face size ranges from approximately 20x20 in the first frame to 60x60 in the last frame.

$$\begin{aligned}
 p(Z_t|x) &\equiv p(Z_t|n, \theta) \\
 &\propto \sum_{c \in C^n} p(Z_t|\theta, c) p^n(c) \quad (10)
 \end{aligned}$$

$$\propto \sum_{c \in C^n} \exp \left[-\frac{1}{2\sigma^2} d(\mathcal{I}_\theta\{Z_t\}, c) \right] \mu_c^n \quad (11)$$

The *exemplars* in C^n are used as the mixture center of a joint distribution and $p^n(c) = \mu_c^n$ is the prior for mixture center c of individual n_t .

4. EXPERIMENTS

We have carried out an extensive set of experiments for the still-to-video and the video-to-video experiment.

To test the still-to-video scenario we have used a video data set provided by NIST and the University of South Florida as part of the Human ID project. The data set contains an outdoor video as well as an indoor mug-shot for 30 different individuals. In the videos, the individuals approach the camera, similar to a real surveillance situation. Fig. 1 shows some example frames in one probe video, where the tracked face is superimposed on the image using a bounding box.

Fig. 2 presents the plot of the posterior probability $p_t(n_t|Z_{0:t})$, and the conditional entropy $H(n_t|z_{0:t})$, all against t .

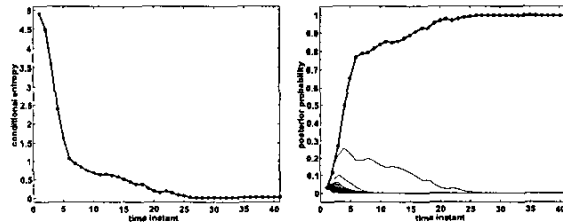


Fig. 2 Conditional entropy $H(n_t|z_{0:t})$ (left) and the posterior probability $p_t(n_t)$ (right) against time instant. Both are obtained using the proposed algorithm.

Recognition rate within top 1 match	50%
Recognition rate within top 3 matches	77%

Table 1

Performances of the still-to-video approach. The probeset consists of outdoor videos, the gallery set consists of indoor mug-shots.

To test the video-to-video scenario we have used the "MoBo" database [8], provided by the Carnegie-Mellon University. The video sequences show 25 different individuals walking on a treadmill so that they move their heads naturally. Different walking styles have been simulated to assure a variety of conditions that are likely to appear in real life: *walking slowly, walking fast, inclining, and carrying an object*. Example images from the videos (*slowWalk*) are shown in Fig. 3. Therefore, four videos per person are available. In the experiments we used one or two of the video types as gallery videos for training, while the remaining ones were used as probes for testing. From each gallery video, a first

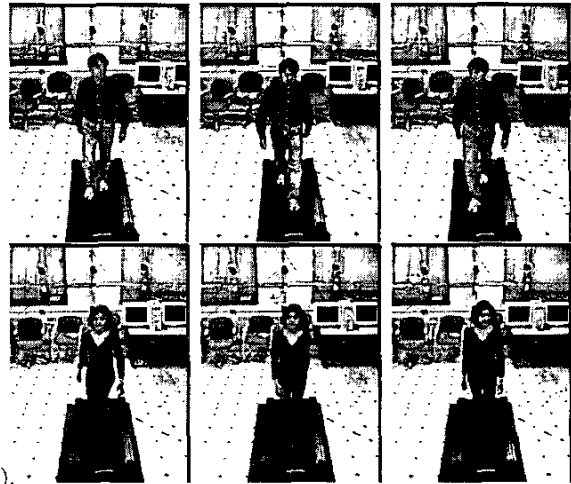


Fig. 3 Sample images from the videos (*slowWalk*) in Database-2.

face sample was cropped by hand. Based on this sample, the training process was initiated. An examples of an automatically extracted exemplar set is shown in Fig. 4 (extracted from a *slowWalk* video of subject 04079). The leftmost exemplar is the hand-extracted one.

During testing, these exemplar galleries were used to compute, over time, the posterior probabilities $p_t(n_t|Z_{0:t})$. Fig. 2 shows, how this probability develops over time. The dashed line refers to the correct hypothesized identity, and the other five curves refer to the probabilities of the top matching identities other than the true one. One can see that the dashed line (true hypothesis) increases quickly to 1. In order to consider *all* the frames of the video, we restart the

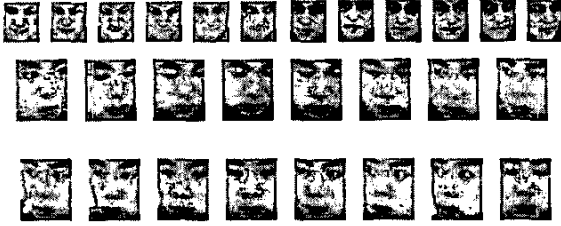


Fig. 4 Exemplars of different persons in the gallery videos (*slowWalk*).

algorithm after convergence. Recognition is established by that identity to which the SIS converged most often.

The major problems that we encountered during our experiments were that in many cases the subjects looked down in one of the videos which caused $\approx 50\%$ of the encountered mis-classifications.

The overall recognition results for one and two gallery videos are summarized in Table 2. The 'g' indicates which videos were used in the gallery. The gallery contained 25 different individuals; however, for the "carrying" video set, only 24 different individuals were available.

slow	fast	incline	carrying
g	100%	96%	92%
92%	g	100%	96%
100%	96%	g	96%
88%	96%	92%	g
g	g	100%	96%
g	100%	g	100%
g	100%	96%	g
100%	g	g	96%
100%	g	100%	g
100%	100%	g	g

Table 2

Overall recognition rates in percent for $\sigma = 0.4$. The 'g' indicates the video used as gallery.

5. CONCLUSIONS

In this paper we have presented a systematic method for face recognition in video. We have used this method for galleries of still as well as for videos. A particle method provides a numerical solution. We have tested our proposed method on a large dataset.

6. REFERENCES

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear pro-

jection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[2] D. Blackburn, M. Bone, and P. Phillips. Facial recognition vendor test 2000: Evaluation report, 2000.

[3] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83:704–740, 1995.

[4] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *Proc. Int. Conf. on Audio- and Video-based Biometric Person Authentication*, pages 176–181, Washington, D.C., March 22–23, 1999.

[5] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10:197–209, 2000.

[7] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *J. Opt. Soc. Am.*, pages 1742–1733, 1997.

[8] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.

[9] A. Howell and H. Buxton. Towards unconstrained face recognition from image sequences. In *Proc. British Machine Vision Conference*, pages 455–464, Edinburgh, 1996.

[10] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. J. of Computer Vision*, 1998.

[11] Y. Li, S. Gong, and H. Liddell. Constructing structures of facial identities on the view sphere using kernel discriminant analysis. In *Proc of the 2nd Intl. Workshop on SCTV*, 2001.

[12] J. Liu and R. Chen. Sequential monte carlo for dynamic systems. *Journal of the American Statistical Association*, 93:1031–1041, 1998.

[13] D. Lowe. Radial basis function networks. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 779–782. MIT-Press, 1995.

[14] S. McKenna and S. Gong. Non-intrusive person authentication for access control by visual tracking and face recognition. In *Proc. Int. Conf. on Audio- and Video-based Biometric Person Authentication*, pages 177–183, Crans-Montana, Switzerland, 1997.

[15] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1103, 2000.

[16] M. Turk and A. Pentland. Eigenfaces for recognition. *Int. Journal of Cognitive Neuroscience*, 3:71–89, 1991.

[17] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen. Automatic video-based person authentication using th rbf network. In *Proc. Int. Conf. on Audio- and Video-based Biometric Person Authentication*, pages 85–92, Crans-Montana, Switzerland, 1997.

[18] L. Wiskott, J. M. Fellous, N. Krüger, and C. v. d. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19:775–779, 1997.

[19] W. Zhao, R. Chellappa, R. Rosenfeld, and P. Phillips. Face recognition: A literature review. Technical Report CAR-TR948, University of Maryland, Center for Automation Research (CFAR), 2000.