Apperance-based tracking and face identification in video sequences

José Miguel Buenaposada[1], Juan Bekios[2], and Luis Baumela[3]

[1] Dept. de Ciencias de la Computación, Universidad Rey Juan Carlos
Calle Tulipán s/n, 28933, Móstoles, Spain
josemiguel.buenaposada@urjc.es
[2] Dept. de Ingeniería de Sistemas y Computación, Universidad Católica del Norte
Av. Angamos 0610, Antofagasta, Chile
juan.bekios@ucn.cl
[3] Dept. de Inteligencia Artificial, Universidad Politécnica de Madrid
Campus Montegancedo s/n, 28660 Boadilla del Monte, Spain
lbaumela@fi.upm.es

Abstract. We present a technique for face recognition in videos. We are able to recognise a face in a video sequence, given a single gallery image. By assuming that the face is in an approximately frontal position, we jointly model changes in facial appearance caused by identity and illumination. The identity of a face is described by a vector of appearance parameters. We use an angular distance to measure the similarity of faces and a probabilistic procedure to accumulate evidence for recognition along the sequence. We achieve 93.8% recognition success in a set of 65 sequences of 6 subjects from the LaCascia and Sclaroff database.

1 Introduction

Face recognition (FR) is perhaps one of the oldest challenges of computer vision. Although the first results date back to the early 70s, in the last 15 years research in this field has grown significantly caused by the commercial importance of its applications.

Although it can be solved by humans in an effortless way, FR is a daunting task for a computer since the appearance of a face may change dramatically depending on face orientation, illumination, facial expression, occlusions, etc. Traditionally FR has been solved in a static way, *still-to-still*, e.g. [1–4], and only more recently the problem of FR in video sequences has attracted attention [5–8]. This interest may be attributed to the increasing importance of FR in surveillance. The poor quality of video in these type of applications make the problem of recognising faces in video an even more challenging task.

Two main approaches to FR have been introduced in the literature, holistic and feature-based. Holistic approaches use a whole face region as input for recognition [6, 7, 4]. Feature-based approaches use the location of a certain set of points in the image and local statistics as input data [1]. Some authors report that feature-based techniques are less stable an accurate than holistic approaches

[9]. In this paper we will introduce an appearance-based holistic approach for FR in videos.

Two problems must be considered to recognise faces in videos. First, locating the face image in each frame of the sequence. Appearance-based approaches are usually very sensitive to geometric or photometric face misalignment. So, a fundamental prerequisite for recognition is accurately registering the face in each frame of the sequence. A key element here is the existence of a good model to represent all possible sources of appearance variation. In second place, the face must be recognised by accumulating evidence for recognition over the whole sequence.

In this paper we propose a technique for face recognition in videos. We use an image-based approach to represent the appearance of a frontal face and to model the changes caused by illumination and identity. By jointly modelling the appearance changes of illumination and identity we achieve a tightly integrated face tracking system in which the identity information contributes to tracking and vice-versa. We use a recently introduced subspace-based minimisation procedure [10] to efficiently fit the face model to each frame in the image sequence. Finally, we use an angular distance to measure the similarity of faces and a probabilistic procedure to accumulate evidence for recognition along the sequence.

Our recognition technique is most closely related to two previous results [6, 7]. Like in [6] we consider the problem of *still to video* face recognition (the gallery of images is composed of a frontal picture of each individual). They use a particle filter to track the face and recognise identity, whereas we use an efficient Gauss-Newton minimisation procedure. On the other hand, in [7], the problem of *video-to-video* face recognition is considered (both the gallery and the input data are videos). They build a non-linear manifold for each subject in the gallery and compute a probabilistic distance of the input set of images to each stored gallery sequence. Although they consider changes in face orientation and illumination, both sources of appearance change are not separated in the manifold. So the recogniser will probably fail if the combination of orientation and illumination in the input sequence is different from that in the gallery sequence.

2 Face alignment

In this section we describe the face alignment used in our FR algorithm. The face is initially located using a face detection algorithm, in our case we use the well known Viola-Jones procedure [11]. Face detection algorithms provide a rough estimate of the location and scale of the face (geometrical alignment) which does not suffice for face recognition. We use the a model-based face alignment procedure to accurately locate the face and compensate illumination effects.

2.1 The face model

We will assume that faces are in frontal view. In this case the major changes in the appearance of a face are caused by identity and illumination variations. Our model is based on a first order approximation to the gray value of face pixels.

Let $I(\mathbf{x}, t)$ be the image acquired at time t, where $\mathbf{x}$ is a vector representing the co-ordinates of a point in the image, and let $\mathbf{I}(\mathbf{x}, t)$ be a vector storing the brightness values of $I(\mathbf{x}, t)$. The first order approximation to the grey value of pixel $\mathbf{x}$ can be expressed as $\bar{\mathbf{I}}_d(\mathbf{x}) + [\mathtt{B}_i \mathbf{c}_{i,t}](\mathbf{x}) + [\mathtt{B}_d \mathbf{c}_{d,t}](\mathbf{x})$, where $\mathtt{B}_i$ and $\mathtt{B}_d$ are the Jacobian matrices representing the derivative of the grey value of pixel $\mathbf{x}$ w.r.t. changes in illumination and identity respectively. These matrices may also be interpreted as the basis of the illumination and identity subspaces. Vectors $\mathbf{c}_i$ and $\mathbf{c}_d$ are respectively the illumination and identity appearance parameters. $\bar{\mathbf{I}}_d(\mathbf{x})$ is the average image representing the point in illumination and identity space in which the first order expansion is made.

The rigid motion of the face is modelled by $f(\mathbf{x}, \boldsymbol{\mu})$, being $\boldsymbol{\mu}$ the vector of rigid motion parameters. So, the brightness constancy equation is

$$\mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t) = \bar{\mathbf{I}}_d(\mathbf{x}) + [\mathtt{B}\mathbf{c}_t](\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{F}, \tag{1}$$

where $\mathtt{B} = [\mathtt{B}_i | \mathtt{B}_d]$, $\mathbf{c}_t^\top = (\mathbf{c}_{i,t}^\top, \mathbf{c}_{d,t}^\top)^\top$, $k = \dim(\mathbf{c}_t)$, and $\mathcal{F}$ represents the set of pixels of the face used for alignment. This first order approximation implies that illumination and identity subspaces are independent. This assumption will simplify the training of the model. Instead of having to use image sequences in which all combinations of illuminations and identities appear, both models will be trained independently.

We train this generic appearance model with the PIE and FERET databases. Matrix $\mathtt{B}_i$ is estimated by selecting the nine[4] directions with highest variance of the eigenspace spanning the set of frontal images of the PIE database. Here, each illumination is averaged across all identities. The result is an average image for each illumination direction (see Fig. 1).

Fig. 1. Aligned images used to build the illumination subspace.

We build the basis of the identity subspace, $\mathtt{B}_d$, using frontal images from the FERET database [13]. We have chosen 781 images from the *fa* FERET gallery (see Fig. 2) displaying a neutral facial expression. Again we choose the 66 directions with highest variance as the components of the basis of the identity subspace (see Fig. 3), once removed the illumination component.

[4] Nine components suffice to represent 97% of the energy in the image[12].

Fig. 2. Some sample aligned images used to build the identity subspace.

Fig. 3. Mean of illumination and identity training images (first image in first row). Illumination subspace basis vectors (remaining images in first row). Most significative components of the identity subspace basis (second row).

2.2 Model fitting

We fit the previous model to a target image by estimating the motion, $\boldsymbol{\mu}$, and appearance, $\mathbf{c}$, parameters which minimise $E(\boldsymbol{\mu}, \mathbf{c}) = ||\mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t) - \bar{\mathbf{I}}_d - [\mathtt{B}\mathbf{c}_t](\mathbf{x})||^2$. This is efficiently achieved by making a Taylor series expansion of $\mathbf{I}$ at $(\boldsymbol{\mu}_t, \mathbf{c}_t, t)$, producing a new error function

$$E(\delta\boldsymbol{\mu}, \delta\mathbf{c}) = ||\mathtt{M}\delta\boldsymbol{\mu} + \mathbf{I}(f(\mathbf{x}, \boldsymbol{\mu}_t), t + \delta t) - \bar{\mathbf{I}}_d - \mathtt{B}(\mathbf{c}_t + \delta\mathbf{c})||^2, \tag{2}$$

where $\mathtt{M} = \left[\frac{\partial \mathbf{I}(f(\mathbf{x},\boldsymbol{\mu}),t)}{\partial \boldsymbol{\mu}} \Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t} \right]$ is the $N \times n$ ($n = \dim(\boldsymbol{\mu})$) Jacobian matrix of $\mathbf{I}$.

An efficient solution for estimating the minimum of (2) is given by

$$\delta\boldsymbol{\mu} = -(\Sigma^\top \Lambda_{M1} \Sigma)^{-1} \Sigma^\top \Lambda_{M2} \mathcal{E}; \qquad \delta\mathbf{c} = \Lambda_B[\mathtt{M}\delta\boldsymbol{\mu} + \mathcal{E}], \tag{3}$$

where Λ_B, Λ_{M1} and Λ_{M2} are constant matrices that can be precomputed off-line and Σ is a matrix that depends on $\boldsymbol{\mu}_t$ and $\mathbf{c}_t$ [10].

3 Face identification

In this section we introduce a probabilistic procedure for FR by accumulating evidence along the image sequence. The identity of one subject is represented by the vector $\mathbf{c}_d$ of appearance parameters in the identity subspace. We use a cosine distance to measure the similarity between the identity parameters of the gallery image and those of input image sequence provided by the alignment algorithm.

Finally, we use a probabilistic procedure to accumulate evidence for recognition along the image sequence.

Let $\mathtt{I}_1, \ldots, \mathtt{I}_t$ be a temporally ordered image sequence of a face and $\mathbf{x}_1, \ldots, \mathbf{x}_t$ be the be the temporally ordered set of co-ordinates of the face in the identity subspace, which we will denote $\mathcal{X}_{1:t}$. Let $G_t = \{g_1, g_2, \ldots, g_c\}$ be a discrete random variable representing the probability of the c gallery images at time t and X_t be a continuous random associated to the co-ordinates in the identity subspace of the image acquired at time t. We will denote by $P(g_i) \equiv P(G_t = g_i)$ the probability that the discrete random variable G_t takes value g_i and by $p(\mathbf{x}) \equiv p(X_t = \mathbf{x})$ the probability density function (p.d.f.) of the continuous variable $\mathbf{x}$ at time t.

The facial identity $g(t)$ at time instant t is obtained as the maximum of the posterior distribution of G_t given the sequence of images up to time t

$$g(t) = \arg\max_i \{P(G_t = g_i|\mathcal{X}_{1:t})\}. \tag{4}$$

We will estimate the posterior distribution using a recursive Bayesian filter. For the first image in the sequence the problem can be immediately solved by

$$P(G_1|\mathbf{x}_1) = \frac{p(\mathbf{x}_1|G_1)P(G_1)}{p(\mathbf{x}_1)} \propto p(\mathbf{x}_1|G_1)P(G_1),$$

where $P(G_1)$ represents our prior knowledge of the probabilities of facial expressions.

Now, if we have a temporal sequence $\mathcal{X}_{1:t}$, we can then update G_t as

$$P(G_t|\mathcal{X}_{1:t}) = \frac{p(\mathbf{x}_t|G_t, \mathcal{X}_{1:t-1})p(G_t, \mathcal{X}_{1:t-1})}{p(\mathcal{X}_{1:t})}.$$

If we assume that measurements depend only on the current identity, then $p(X_t|G_t, \mathcal{X}_{1:t-1}) = p(X_t|G_t)$ and, hence,

$$P(G_t|\mathcal{X}_{1:t}) \propto p(X_t|G_t)P(G_t|\mathcal{X}_{1:t-1}),$$

where $P(G_t|\mathcal{X}_{1:t-1})$ is the prediction of G_t, given the data up to time instant $t-1$. This probability can be estimated as

$$P(G_t|\mathcal{X}_{1:t-1}) = \sum_{i=1}^{c} P(G_t, G_{t-1} = g_i|\mathcal{X}_{1:t-1}) = \sum_{i=1}^{c} P(G_t|g_i, \mathcal{X}_{1:t-1})P(g_i|\mathcal{X}_{1:t-1}).$$

If we assume that our system is Markovian (G_t depends only on G_{t-1}), then

$$P(G_t|\mathcal{X}_{1:t-1}) = \sum_{i=1}^{c} P(G_t|G_{t-1} = g_i)P(G_{t-1} = g_i|\mathcal{X}_{1:t-1}),$$

where $P(G_t|G_{t-1})$ would represent the identity transition probability. Of course, the identity of a subject does not change in a sequence, so $P(G_t|G_{t-1}) = \delta(G_t, G_{t-1})$, where $\delta(a, b)$ is the Kronecker delta function.

$p(\mathbf{x}|g_i)$ represents the p.d.f. of identity parameters for an image of gallery subject g_i. Given that $\cos(\mathbf{x}, \mathbf{y})$ gives a measure of similarity between vectors $\mathbf{x}$ and $\mathbf{y}$, we make the following approximation

$$p(\mathbf{x}|g_i) \approx |\cos(\mathbf{x}, \mathbf{x}_{g_i})|,$$

where $\mathbf{x}_{g_i}$ are the identity parameters associated to gallery subject g_i.

4 Experiments

In this section we describe some results of the test that we have performed with the algorithm described above. We have processed 71 sequences of 6 subjects from the LaCascia and Sclaroff database[5] [14]. The database consists of mug-shots videos of 6 subjects labelled as: *jam*, *jim*, *llm*, *ssm*, *vam* and *mll*. There are sequences with uniform illumination (8 for *jam*, 9 for *jim*, 9 for *llm*, 9 for *ssm*, 9 for *vam*) and with varying illumination (9 for *jam*, 9 for *ssm* and 9 for *mll*). See Fig. 4 for sample images of the 6 subjects in the database.

Fig. 4. The 6 subjects in the Sclaroff and LaCascia image sequence database, from left to right: *jam*, *jim*, *llm*, *ssm*, *vam* and *mll*

.

4.1 Tracking results

We have used the alignment algorithm described in section 2 to process the 71 sequences in the LaCascia and Sclaroff database. The rigid face motion model is a rotation, translation and scale of the face texture, $f(\mathbf{x}, \boldsymbol{\mu}) = s\mathtt{R}(\theta)\mathbf{x} + \mathbf{t}$, being $\boldsymbol{\mu} = (s, \theta, t_x, t_y)$. With Viola and Jones' face detector [11] we locate the face in the first image of each sequence and track the face motion with the image alignment procedure. In 65 of the 71 sequences the tracking was perfect from the first to the last frame (200 frames per sequence). Only in 6 sequences the face was lost at the end of the sequence. All tracking failures are caused by strong rotations of the face out of the camera plane.

In Fig. 5 we display some results from the *llm5* sequence, in which the *llm* subject moves in front of the camera under uniform illumination. The estimated position of the face is overlayed in red. To the right side of each result image we show four smaller images: the rectified image estimated from the motion

[5] http://www.cs.bu.edu/groups/ivc/data.php

parameters ($I(f(\mathbf{x}, \boldsymbol{\mu}_t), t + \delta t)$) on the left-upper side, the image reconstructed with the face model ($I_d(\mathbf{x}) + [\mathtt{B}_i\mathbf{c}_{i,t}](\mathbf{x}) + [\mathtt{B}_d\mathbf{c}_{d,t}](\mathbf{x})$) on the right-upper side, the illumination reconstructed image ($I_d(\mathbf{x}) + [\mathtt{B}_i\mathbf{c}_{i,t}](\mathbf{x})$) on the left-lower side and the identity subspace reconstructed image ($I_d(\mathbf{x}) + [\mathtt{B}_d\mathbf{c}_{d,t}](\mathbf{x})$) on the right-lower side. Note that the reconstruction of the face with the subspace identity parameters is near perfect although the subject was not in the identity subspace training images.

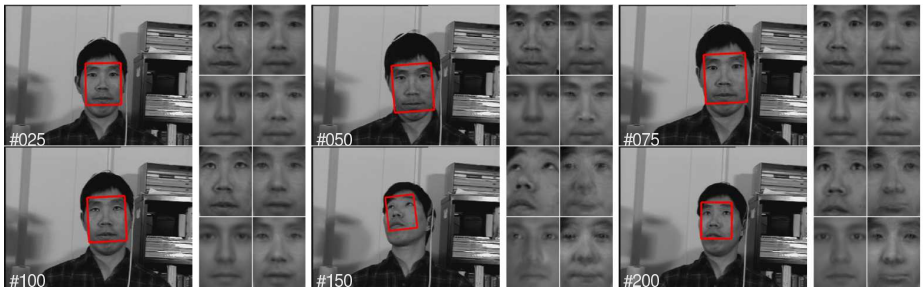

Fig. 5. Alignment results for *llm5* sequence.

In Fig. 6 we show some results from the *jal4* sequence corresponding to subject *jam* moving in front of the camera under varying illumination. The tracker can cope with illumination and head motion while estimating the identity parameters with no problem. Here the more out of plane the rotation of the head is, the worst is the identity reconstruction.

Fig. 6. Alignment result for *jal4* sequence.

In Fig. 7 we show the results of the system in a similar situation. In sequence *ssl6* the subject *ssm* also moves with head rotations out of camera plane under varying illumination. In this case, the identity reconstruction is still qualitatively quite good.

Fig. 7. Alignment result for *ssl6* sequence.

4.2 Identity recognition

Once we have shown the performance of the image alignment algorithm, we test the quality of FR using the identity subspace parameters. First of all, we have chosen one image of each subject as gallery for face identification (See Fig. 8). Gallery images are the result of the image alignment process performed in one image of the sequences *jam1*, *jim2*, *llm1*, *ssm1*, *vam1* and *mll1*. These sequences images are not used for evaluation. The result of this process is that we have a training video (although we are only using one image from it) and 65 test sequences.

Fig. 8. Probe gallery images. From left to right subjects *jam*, *jim*, *llm*, *ssm*, *vam* and *mll*

.

After processing all 65 sequences we get the confusion matrix in table 1. The overall recognition rate is 93.85% quite remarkable given that we are using a single image as gallery and there are pose and illumination changes in the test sequences.

In Fig. 9 is displayed a successful identification, in spite of important illumination changes in the sequence. The identification is correct from the beginning because the head is not rotating out of plane and therefore the illumination appearance model can cope with the changing conditions. In Fig. 10 we show a partially failed identification. The performance of the FR algorithm is correct until the rotation of the head clearly violates the frontal image assumption.

	jam	*jim*	*llm*	*ssm*	*vam*	*mll*	total
jam	100.0	12.5	0	0	0	0	
jim	0	75.0	0	0	0	12.5	
llm	0	0	100.0	0	0	0	
ssm	0	0	0	100.0	12.5	0	
vam	0	0	0	0	87.5	0	
mll	0	12.5	0	0	0	87.5	
total							93.85

Table 1. Confusion matrix (expressed in percentage) of the FR experiment.

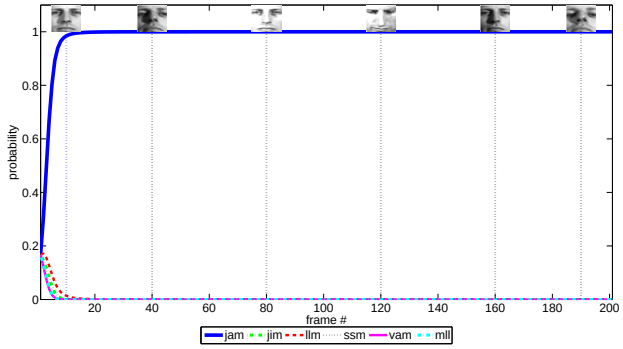

Fig. 9. A successful recognition using *jal4* sequence

5 Conclusions

We have introduced a fully automatic algorithm for face detection, tracking and recognition. It uses a single frontal image as probe gallery. It performs remarkably well in presence of illumination changes and translational and/or rotational motion in camera plane. Performance degrades when the frontal image assumption is violated. We are working towards improving the performance of the system for out of camera plane face rotations and to make *video-to-video* FR.

Acknowledgements

The authors gratefully acknowledge funding from the Spanish *Ministerio de Educación y Ciencia* under contract `TRA2005-08529-C02-02`. They also thank the anonymous reviewers for their comments and Lacascia and Sclaroff for providing the image sequences data base.

References

1. Wiskott, L., Fellous, J.M., von der Malsburg, C.: Face recognition by elastic bunch graph matching. Trans. on PAMI **19**(7) (July 1997) 775–779

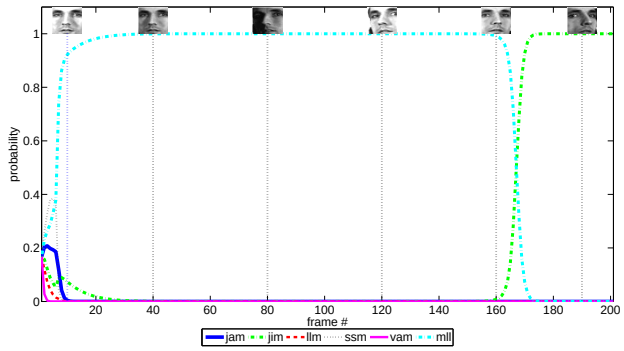

Fig. 10. A partial recognition failure with *mll4*

2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Trans. on PAMI **19**(7) (July 1997) 711–720
3. Martinez, A.: Recognizing imprecisely located, partially ocluded and expression variant faces from a single sample per class. Trans. on PAMI **24**(6) (June 2002) 748–763
4. Zhang, L., Samaras, D.: Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. Trans. on PAMI **28**(3) (March 2006) 351–363
5. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding **91**(1-2) (July 2003) 214–245
6. Zhou, S., Chellappa, R., Moghaddam, B.: Visual tracking and recognition using appearance-adaptive models in particle filters. Trans. on IP **13**(11) (November 2004) 1491–1506
7. Lee, K.C., Kriegman, D.: Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In: Proc. of CVPR. Volume I. (2005) 852 – 859
8. Xu, Y., Roy Chowdhury, A., Patel, K.: Pose and illumination invariant face recognition in video. In: Proc. of CVPR. (2007) 1–7
9. Brunelli, T., Poggio, T.: Face recognition: features versus templaes. Trans. on PAMI **15**(10) (October 1993) 1042–1052
10. Buenaposada, J.M., Muñoz, E., Baumela, L.: Efficiently estimating facial expression and illumination in appearance-based tracking. In: Proc. BMVC. Volume I. (2006) 57–66
11. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (May 2004) 137–154
12. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. Trans. on PAMI **25**(2) (February 2003) 218–233
13. Phillips, P., Moon, H., Rauss, P., Rizvi, S.: The feret evaluation methodology for face recognition algorithms. Trans. on PAMI **22**(10) (October 2000) 1090–1104
14. La Cascia, M., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3d models. Trans. on PAMI **22**(4) (April 2000) 322–336