Appearance-based 3-D Face Recognition from Video

Volker Krüger[1], Ralph Gross[2] and Simon Baker[2]

[1]University of Maryland, Center for Automation Research
A.V. Williams Building
College Park, MD 20742
[2] The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract. In this work we present an appearance-based 3-D Face Recognition approach that is able to recognize faces in video sequences, independent from face pose. For this we combine *eigen light-fields* with probabilistic propagation over time for evidence integration. Eigen light-fields allow us to build an appearance based 3-D model of an object; probabilistic methods for evidence integration are attractive in this context as they allow a systematic handling of uncertainty and an elegant way for fusing temporal information. Experiments demonstrate the effectiveness of our approach. We tested this approach successfully on more than 20 testing sequences, with 74 different individuals.

1 Introduction

Face recognition has been a major research topic in recent years. Among the most successful approaches are [21; 12; 22]. The techniques have been thoroughly evaluated in the FERET-Protocol [15] and produce acceptable recognition rates in ideal conditions. However, if ideal conditions are not met, e.g., in case of out-of-plane rotation, recognition rates drop drastically. The major reason is, that the above recognition approaches use the *still-to-still* technique: gallery and probe sets contain still face images (mug-shots), and recognition rates are high only if geometrical and photometrical conditions of the test images in the probe set match those in the gallery set. To solve these problems a *video-to-video* technique has been proposed [8]. In this setting, gallery and probe sets consist of videos, instead of mug-shots, i.e., each individual is represented by a video ideally showing a variety of views, and the individual is to be recognized from a video where he/she also shows a wide variety of views. In this approach, exemplars are learned that summarize the visible 3-D variations of the face in the video, their priors as well as their dynamics. Matching is done by evidence integration over time; a particle method is used to analytically estimate the probability density function over the set of known individuals.

The set of exemplars that are learned from the training videos represent an appearance-based 3-D representation of the face. This representation is built incrementally and depends heavily on the training video: slight variations in the video lead to completely different representations. This hinders a common representation of the face space; the consequence is that one needs to test each single face as a hypothesis. A more systematic way of building an appearance-based 3-D model is therefore important. In this paper we propose to use *eigen*

light-fields (ELFs) [4], which were previously used to build a view-independent *still-to-still* face representation. In this paper we combine the advantages of the ELFs with the probabilistic evidence integration over time of [8]. The challenge is to use ELFs for low resolution video data instead of high resolution still images.

The remainder of this paper is organized as follows: Sec. 2 introduces some preliminaries. In Sec. 3 we introduce eigen light-fields. The recognition method is discussed in Sec. 4. We conclude with experimental results in Sec. 5 and final remarks are in Sec. 6.

2 Preliminaries

Before delving into details about ELFs and evidence integration, we will introduce some terminology borrowed from the FERET evaluation protocol [15]. A *Gallery* $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$ is a set of image sets. Each V_i is associated with a single individual, i.e., N individuals $\mathcal{N} = \{1, 2, \ldots, N\}$, are represented in the Gallery $\mathcal{V}$. The gallery contains the exemplars against which the probe set is matched. A *Probe set* $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$ is a set of M probe videos which are used for testing.

2.1 Geometric and Photometric Transformations

An image Z may undergo a geometrical or photometrical transformation

$$\tilde{Z} = \mathcal{T}_\alpha\{Z\} \tag{1}$$

for $\alpha \in \mathcal{A}$, where $\mathcal{A}$ is the set of possible transformations. The set of possible transformations $\mathcal{A}$ has to be pre-defined in our framework.

2.2 Likelihood Measure

Let $F = \{f_1, f_2 \ldots, f_N\}$ be a set of face images, with $\mathcal{N} = \{1, 2, \ldots, N\}$. Let further $X \in \mathcal{A} \times \mathcal{N}$ be a random variable. This random variable defines the transformation $\mathcal{T}_\alpha$ and the number i of a face $f_i \in F$. Thus, having observed a video image Z, the observation likelihood for a hypothesis $X = (\alpha, i)$, is given by

$$\begin{aligned} p(Z|X) &\equiv p(Z|\alpha, i) \\ &\propto z \exp -\frac{1}{2\sigma^2} d(Z, \mathcal{T}_\alpha\{f_i\}) \ , \end{aligned} \tag{2}$$

Eq. (2) computes the probability that the observation Z shows the face of an individual i, while the face f_i undergoes the transformation α. Here, $d(\cdot, \cdot)$ is a suitable distance function. In face recognition, one usually deals with the inner face region of the subject, rather than the entire image. We therefore interpret Eq. (2) such that $\mathcal{T}_\alpha\{f_i\}$ is compared to a subimage of Z where the position and scale of the subimage is specified by α. If $\mathcal{A}$ is the set of affine deformation, our-of-plane rotation cannot be modeled adequately. Such transformations have to be coped with in a different manner. To do so, we use as the distance function d the eigen light-fields, that will be introduced in the next section.

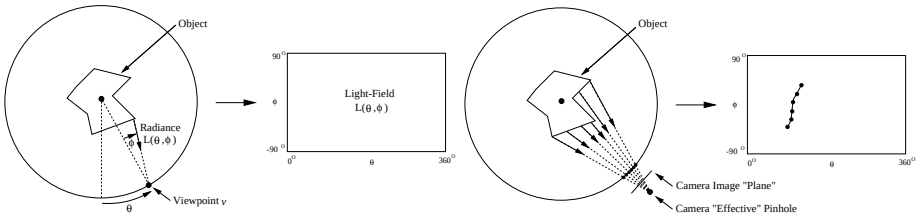

Fig. 1. Left image: An illustration of the 2D light-field of a 2D object [10]. The object is conceptually placed within a circle. The angle to the viewpoint v around the circle is measured by the angle θ, and the direction that the viewing ray makes with the radius of the circle is denoted ϕ. For each pair of angles θ and ϕ, the radiance of light reaching the viewpoint from the object is then denoted by $L(\theta, \phi)$, the *light-field*. Although the light-field of a 3D object is actually 4D, we will continue to use the 2D notation of this figure in this paper for ease of explanation. **Right image:** The 1D image of a 2D object corresponds to a curve (surface for a 2D image of a 3D object) in the light-field. Each pixel in the image corresponds to a ray in space through the camera pinhole and the location of the pixel on the image plane. In general this ray intersects the light-field circle at a different point for each pixel. As the pixel considered "moves" in the image plane, the point on the light-field circle therefore traces out a curve in θ-ϕ space. This curve is a straight vertical line iff the "effective pinhole" of the camera lies on the circle used to define the light-field.

3 Appearance-based 3-D Representation with Eigen Light-Fields

3.1 Object Light-Fields

The *plenoptic function* [1] or *light-field* [10] is a function which specifies the radiance of light in free space. It is a 5D function of position (3D) and orientation (2D). In addition, it is also sometimes modeled as a function of time, wavelength, and polarization, depending on the application in mind. Assuming that there is no absorption or scattering of light through the air [14], the light-field is actually only a 4D function, a 2D function of position defined over a 2D surface, and a 2D function of direction [3; 10]. In 2D, the light-field of a 2D object is actually 2D rather, than the 3D that might be expected. See Figure 1,left, for an illustration.

3.2 Eigen Light-Fields

Suppose we are given a collection of light-fields $L_i(\theta, \phi)$ where $i = 1, \ldots, N$. See Figure 1,left, for the definition of this notation. If we perform an eigen-decomposition of these vectors using Principal Components Analysis (PCA), we obtain $d \leq N$ eigen light-fields $E_i(\theta, \phi)$ where $i = 1, \ldots, d$. Then, assuming that the eigen-space of light-fields is a good representation of the set of light-fields under consideration, we can approximate any light-field $L(\theta, \phi)$ as:

$$L(\theta, \phi) \approx \sum_{i=1}^{d} \lambda_i E_i(\theta, \phi) \tag{3}$$

where $\lambda_i = \langle L(\theta, \phi), E_i(\theta, \phi) \rangle$ is the inner (or dot) product between $L(\theta, \phi)$ and $E_i(\theta, \phi)$. This decomposition is analogous to that used in face and object recognition [19; 13]; it is just performed on the entire light-field rather than on images.

3.3 Estimating Light-Fields from Images

Capturing the complete light-field of an object is a difficult task, primarily because it requires a huge number of images [3; 10]. In most object recognition scenarios it is unreasonable to expect more than a few images of the object; often just one. As shown in Figure 1, right, however, any image of the object corresponds to a curve (for 3D objects, a surface) in the light-field. One way to look at this curve is as a highly occluded light-field; only a very small part of the light-field is visible.

It was argued in [4] that the eigen coefficients λ_i can be estimated from such an occluded view. An algorithm used in [4] solves for λ_i as the least squares solution of:

$$L(\theta, \phi) - \sum_{i=1}^{d} \lambda_i E_i(\theta, \phi) \;=\; 0 \tag{4}$$

where there is one such equation for each pair of θ and ϕ that are un-occluded in $L(\theta, \phi)$. Assuming that $L(\theta, \phi)$ lies *completely within the eigen-space* and that enough pixels are un-occluded, then the solution of Equation (4) will be exactly the same as that obtained using the inner product [4].

Since there are d unknowns $(\lambda_1 \ldots \lambda_d)$ in Equation (4), at least d un-occluded light-field pixels are needed to over-constrain the problem, but more may be required due to linear dependencies between the equations. In practice, $2 - 3$ times as many equations as unknowns are typically required to get a reasonable solution [9]. Given an image $I(m, n)$, the following is then an algorithm for estimating the eigen light-field coefficients λ_i:

Algorithm 1: Eigen Light-Field Estimation

1. For each pixel (m, n) in $I(m, n)$ compute the corresponding light-field angles $\theta_{m,n}$ and $\phi_{m,n}$.
2. Find the least-squares solution (for $\lambda_1 \ldots \lambda_d$) to the set of equations:

$$I(m, n) - \sum_{i=1}^{d} \lambda_i E_i(\theta_{m,n}, \phi_{m,n}) \;=\; 0 \tag{5}$$

   where m and n range over their allowed values.

Although we have described this algorithm for a single image $I(m, n)$, any number of images can obviously be used. The extra pixels from the other images are simply added in as additional constraints on the unknown coefficients λ_i in Equation (5). Algorithm 1 can be used to estimate a light-field from a collection of images. Once the light-field has been estimated, it can then be used to render new images of the same object under different poses (See also [20]). It was shown in [4] that the algorithm correctly re-renders a given object assuming a Lambertian reflectance model.

4 Tracking and Recognizing in Video

In this section we discuss the recognition of individuals in videos. After the generation of ELFs from the image sets $\mathcal{V}_i$ in the previous section, we have a vector of eigen values for each individual $i \in \mathcal{N}$ in the Gallery $\mathcal{V}$.

4.1 Tracking and Recognition in the Bayesian Framework

We can now compute the observation likelihoods as in Eq. 2 and we can track and identify individuals in the video: Let $X_t = (\alpha_t, i_t) \in \mathcal{A} \times \mathcal{N}$ be a random variable. We want to find X_t such that the joint distribution

$$p(X_t|Z_1, \ldots, Z_t) \tag{6}$$

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

$$\begin{aligned} &p(X_t|Z_1, Z_2, \ldots, Z_t) \equiv p_t(\alpha_t, i_t) \\ &= \sum_{i_{t-1}} \int_{\alpha_{t-1}} p(Z_t|\alpha_t, i_t) p(\alpha_t, i_t|\alpha_{t-1}, i_{t-1}) p_{t-1}(\alpha_{t-1}, i_{t-1}) \ . \end{aligned} \tag{7}$$

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

$$p(i_t|Z_1, \ldots, Z_t) = \int_{\alpha_t} p(\alpha_t, i_t|Z_1, \ldots, Z_t) \ . \tag{8}$$

Maximizing (8) leads to the desired identity.

In Eq. (7)

$$p(X_t|X_{t-1}) \equiv p(\alpha_t, i_t|\alpha_{t-1}, i_{t-1})$$

defines the probability of the state variable to change from X_{t-1} to X_t. The transformation α_t may change according to a dynamic model. The identity i, however, is assumed to be constant over time, i.e., it is assumed that the identity of the tracked person does not change over time. Learning of a dynamic model has been discussed in [18].

We have used a particle method to efficiently compute $p_t(i_t, \alpha_t|Z_t)$ [23; 2; 6; 7; 11], where i_t, α_t depicts the hypothesised identity and transformation of the individual in the video. In [6] only the transformation α_t was estimated, in [23] the special case was discussed where each individual is presented by only a single exemplar. In [8] this was generalized to the case of several exemplars for each individual. Since the ELFs offer a common 3-D representation for each face, we use the more efficient particle method of [23].

5 Experiments

We used the CMU PIE database [16] as the training set to build eigen light-fields for our experiments and as part of the gallery. The database consists of 68 subjects. The images were preprocessed as explained in [4] and were then downsampled to a height of 38 pixels. In Fig. 2 the set of available views in the training set is shown. For testing we have used CMU Mobo Database [5]. We needed to select a subset of 6 individuals and 20 videos as the facial views in the remaining videos were not consistent with the 3-D model as defined by the eigen light-fields: in those videos, the individuals looked either up or down, a pose which was not modeled by our ELFs (see Fig. 2 for the possible views). We therefore extracted manually the inner face regions from the selected 20 videos of the individuals for additional training. The complete Gallery therefore consisted of 74 individuals (6 from the MoBo database and 68 from the PIE database).

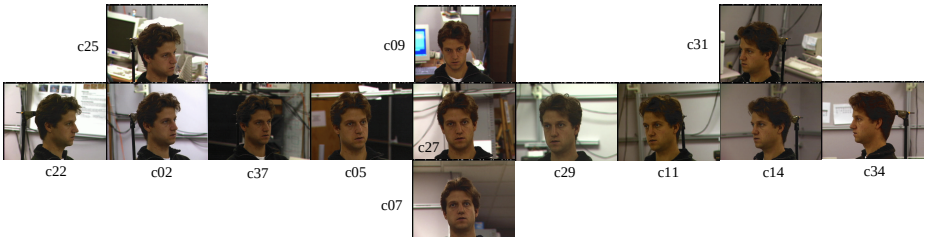

Fig. 2. The pose variation in the PIE database [17]. The pose varies from full left profile (c34) to full frontal (c27) and on to full right profile (c22). The 9 cameras in the horizontal sweep are each separated by about 22.5°. The 4 other cameras include 1 above (c09) and 1 below (c07) the central camera, and 2 in the corners of the room (c25 and c31), typical locations for surveillance cameras.

Between two and four face images were extracted from the videos of each of the 6 individuals. The face images had a height of between 30 and 38 pixels. Smaller images were scaled to a consistent hight of 38 pixels. Using a small number of low-resolution video images results in quite noisy eigenvectors that can hardly be used for recognition based on still images (see below).

The video sequences in the MoBo database show the individuals walking on a tread-mill. Different walking styles were used to assure a variety of conditions that are likely to appear in real life: *slow walk*, *fast walk*, *incline walk* and *walking while carrying a ball*. Therefore, four videos per person are available. During the recording of the videos the illumination conditions were not altered. Each video consists of 300 frames (480×640 pixels per frame) captured at 30 Hz.

Some example images of the videos (*slowWalk*) are shown in Fig. 3.

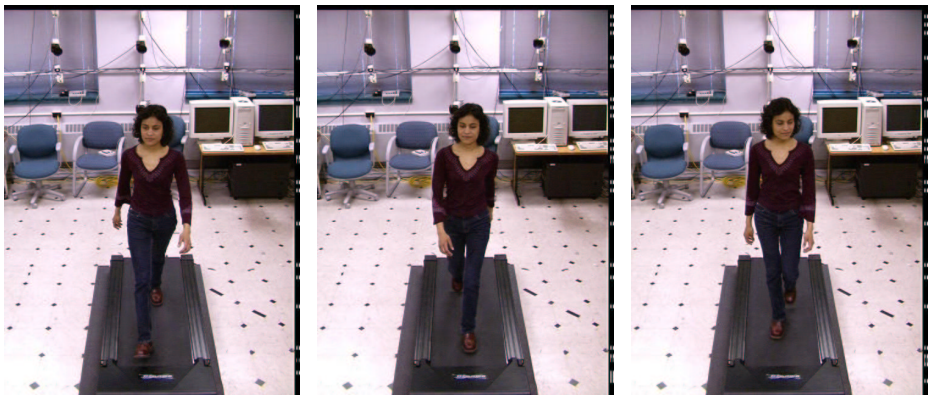

Fig. 3. The figure shows example images of one of the videos (*slowWalk*).

The inner face regions in these videos are between 30×30 and 40×40 pixels.

During testing, the ELFs were used to compute, over time, the posteriori probabilities $p_t(i_t|Z_t)$. It is interesting to see, how the posteriori probabilities develop over time. Examples for this can be seen in Fig. 5. The dashed line refers to the correct hypothesized identity, 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 one.

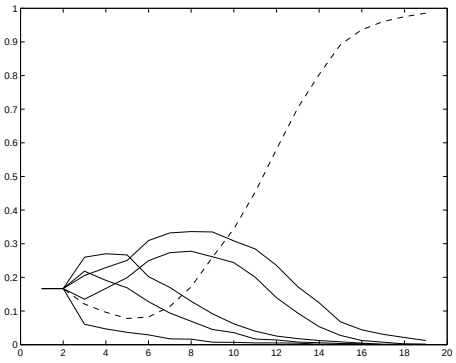

Fig. 4. The figure shows how the posterior probabilities $p_t(i_t|Z_t)$ develop over time. The dashed curve refers to the correct hypothesis. The y-axis refers to the believe that at a given time t a given hypothesis is the correct one.

Of the 20 videos tested, recognition was successful in 13 cases, i.e. the true hypothesis had maximal probability after convergence. In 4 cases, the true hypothesis was the second highest probability during the evidence integration process, i.e. in 17 cases the true hypothesis was among the top two matches. In the remaining three videos, recognition failed. After an average time of 15 frames the particle method had converged.

We also tested the recognition without evidence integration: Testing on all single images of the videos, recognition only succeeded in two cases. This shows the importance of evidence integration for recognition when dealing with noisy observations.

Video images from our test data were converted from color to gray value images, but no further processing was done. The set of deformations $\mathcal{A}$ included scale and translation. Shear and rotation were not considered.

Top matches	13 (out of 20)
Second best matches	4 (out of 20)
still-to-still	2 (out of 1500)

Table 1. The Table summarizes the recognition results: In 13 out of twenty videos the individuals were correctly identified. In four out of twenty, the correct match was only the second best match. Applying the ELF approach without evidence integration lead to 2 correct matches on all images of the videos.

6 Conclusion

In this paper we presented a novel approach for appearance based face recognition across pose. We used eigen light-fields to build a 3-D model of faces. The advantage of ELFs is that once a 3-D model is built from a generic training set, one only needs two to four views of a new and before unseen face to be able to recognize this face from a new and previously unseen view.

This property has been shown in [4] with a large number of experiments. In [4], however, the face sets consisted of high resolution images and the faces did not show any facial expressions. In this paper we examined how this method scales to face images as small as 30×38 pixels with strong variations in appearance due to facial expressions. The resulting noisy feature vectors could not have been used for *still-to-still* recognition. We solved this problem by integrating the evidence of identity over time by applying Bayesian propagation [8]. Using this approach, experiments showed more stable recognition results.

As it is difficult to draw general conclusions from a database of only 20 videos, we currently evaluate our approach on 40 newly recorded sequences.

References

1. E.H. Adelson and J. Bergen. The plenoptic function and elements of early vision. In Landy and Movshon, editors, *Computational Models of Visual Processing*. MIT Press, 1991.
2. A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10:197–209, 2000.
3. S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. In *SIGGRAPH*, 1996.
4. R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, May 21-22, 2002.
5. Ralph Gross and Jianbo Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.
6. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. J. of Computer Vision*, 1998.
7. G. Kitagawa. Monta carlo filter and smoother for non-gaussian nonlinear state space models. *J. Computational and Graphical Statistics*, 5:1–25, 1996.
8. V. Krueger and S. Zhou. Exemplar-based face recognition from video. In *Proc. European Conf. on Computer Vision*, Copenhagen, Denmark, June 27-31, 2002.
9. A. Leonardis and H. Bischof. Dealing with occlusions in the eigenspace approach. In *Proceedings of CVPR*, 1996.
10. M. Levoy and M. Hanrahan. Light field rendering. In *Proc. of SIGGRAPH*, 1996.
11. J.S. Liu and R. Chen. Sequential monte carlo for dynamic systems. *Journal of the American Statistical Association*, 93:1031–1041, 1998.
12. B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:696–710, 1997.
13. H. Murase and S.K. Nayar. Visual learning and recognition of 3-D objects from appearance. *Int. J. of Computer Vision*, 14:5–24, 1995.
14. S.K. Nayar and S. Narasimhan. Vision in bad weather. In *Korfu, Greece*, 1999.
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. T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (pie) database. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, May 21-22, 2002.
17. T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *Proc. of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
18. K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. Int. Conf. on Computer Vision*, volume 2, pages 50–59, Vancouver, Canada, 9-12 July, 2001.
19. M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. of CVPR*, 1991.
20. T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. on PAMI*, 19(7):733–741, 1997.
21. L. Wiskott, J. M. Fellous, N. Krüger, and C. v. d. Malsburg. Face recognition and gender determination. In *Proc. Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 26-28, 1995.
22. W. Zhao, R. Chellappa, and N. Nandhakumar. Discriminant analysis fo principal components for face recognition. In *Nara, Japan, April 14-16*, pages 336–341, 1998.
23. S. Zhou, V. Krüger, and R. Chellappa. Face recognition from video: A CONDENSATION approach. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, May 21-22, 2002.