Exemplar-based Face Recognition from Video *

Volker Krueger and Shaohua Zhou

University of Maryland, Center for Automation Research
A.V. Williams Building
College Park, MD 20742
USA
vok@cfar.umd.edu

Abstract. A new exemplar-based probabilistic approach for face recognition in video sequences is presented. The approach has two stages: First, *Exemplars*, which are selected representatives from the raw video, are automatically extracted from gallery videos. The exemplars are used to summarize the gallery video information. In the second part, exemplars are then used as centers for probabilistic mixture distributions for the tracking and recognition process. Probabilistic methods are attractive in this context as they allow a systematic handling of uncertainty and an elegant way for fusing temporal information.
Contrary to some previous video-based approaches, our approach is not limited to a certain image representation. It rather enhances known ones, such as the PCA, with temporal fusion and uncertainty handling. Experiments demonstrate the effectiveness of each of the two stages. We tested this approach on more than 100 training and testing sequences, with 25 different individuals.

Keywords: Surveillance, Video-based Face Recognition, Exemplar-based Learning

1 Introduction

Face recognition has been a major research topic in recent years. Among the most successful approaches are [19; 14; 20]. 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 or variations in facial expressions, recognition rates drop drastically. The major reason is, that all the 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. An alternative to the *still-to-still* approach is to select appropriate images from a video, i.e. track the person under consideration and wait for a good shot. This approach

* This work was supported by the DARPA Human ID Grant program under the ONR Grant N00014-00-1-0908.

is essentially equivalent to the *still-to-still* techniques and success is therefore limited (see Sec. 2 for related literature).

Clearly, an alternative to representing each individual by a single mug-shot would be to use a set of mug-shots per individual that cover the possible variatinos between frames: *multiple-stills-to-stills*. This approach has, however, two major problems:

1. How should one select the right mug-shots?
2. What happens if the conditions in the probe images still happen to be different from those in the gallery mug-shots?

To solve these problems we propose the *video-to-video* technique. In this setting, gallery and probe sets consist of videos, instead of single mug-shots, i.e., each individual is represented by a video, ideally showing a variety of views of that person, and the individual is to be recognized in a video where he/she of course also shows a wide variety of views.

Our aim, therefore, is to develop a paradigm that allows to

1. learn the probabilistic settings of each individual from a gallery video and
2. test a multiple number of those settings as hypotheses' in a probe video.

The use of *exemplars* offers a good way to tackle the first problem [17; 3]. Exemplar-based models are constructed directly from the gallery videos, thus preventing the need to set up complex intermediate 2D, 3D or feature-based representations. We solve the second problem with an enhancement of the well-known Condensation method [7; 21]. Here, the identity is treated as a state that is to be estimated and that remains constant over time.

The paper is organized as follows: We will give an overview of the related literature in the next section. Sec. 3 introduces some preliminaries. In Sec. 4 we will introduce our method for exemplar learning. The recognition method is introduced in Sec. 5. We present experimental results in Sec. 6 and final remarks are in Sec. 7.

2 Related Literature

Nearly all video-based recognition systems apply still-image-based recognition to selected good frames. The face images are warped into frontal views whenever pose and depth information about the faces is available [1].

In [6; 13; 18] RBF (Radial Basis Function) networks are used for tracking and recognition purposes. In [6], the system uses an RBF (Radial Basis Function) network for recognition. Since no warping is done, the RBF network has to learn the individual variations as well as possible transformations. The performance appeares to varies widely, depending on the size of the training data but has not been thoroughly evaluated. In [13] face tracking is based on a 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. [18] present a fully automatic person authentication system. The

sustem uses video break, face detection, and authentication modules and cycles over successive video images until a high recognition confidence is reached. During operation, the face is tracked, face images are normalized and then used for authentication with an RBF network. This system was tested on three image sequences; the first was taken indoors with one subject present, the second was taken outdoors with two subjects, and the third was taken outdoors with one subject in stormy conditions. Perfect results were reported on all three sequences, as verified against a database of 20 still face images.

In [9], a generic approach to simultaneous object tracking and verification is proposed. The approach is based on posterior probability density estimation using sequential Monte Carlo methods [2; 7; 8; 10]. Tracking is formulated as a probability density propagation problem and the algorithm also provides verification results. However, no systematic recognition evaluation was done.

In [16], a system called *PersonSpotter* is described. This system is able to capture, track and recognize a person walking toward or passing a stereo CCD camera. It has several modules, including a head tracker, and a landmark finder. The landmark finder uses a dense graph consisting of 48 nodes learned from 25 example images to find landmarks such as eyes and nose tip. An elastic graph matching scheme is employed to identify the face.

A multimodal based person recognition system is described in [1]. This system consists of a face recognition module, a speaker identification module, and a classifier fusion module. The most reliable video frames and audio clips are selected for recognition. 3D information about the head is used to detect the presence of an actual person as opposed to an image of that person. Recognition and verification rates of 100% were achieved for 26 registered clients.

3 Preliminaries

Before delving into details about exemplar learning and recognition, we will introduce some terminology borrowed from the FERET evaluation protocol [15]. A *Gallery* $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$ is here a set of videos. Each V_i is associated with a single individual, i.e., N individuals $\mathcal{N} = \{1, 2, \ldots, N\}$, are represented in the Gallery $\mathcal{V}$.

A *Probe set* $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$ is a set of M probe videos which are used for testing.

3.1 Objects, Faces and Exemplars

In our framework a *face* is defined to be a gray value image that has been suitably processed. We therefore treat faces in an appearance-based 2D manner. An *exemplar* is a selected "representative", extracted directly from raw video.

3.2 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.

For example, $\mathcal{T}_\alpha$ represents the similarity transforms, if with $\alpha = (\mathbf{c}, \theta, s)$

$$\mathcal{T}_\alpha\{Z(\mathbf{x})\} = Z(T_\alpha(\mathbf{x})) = Z(s\mathbf{R}(\theta)\mathbf{x} + \mathbf{c}) \ . \tag{2}$$

The set of possible transformations $\mathcal{A}$ has to be pre-defined in our framework.

3.3 Likelihood Measure

Let $F = \{f_1, f_2 \ldots, f_N\}$ be a set of faces, 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

$$\begin{aligned} p(Z|X) &\equiv p(Z|\alpha, i) \\ &\propto z \exp -\frac{1}{2\sigma^2} d(Z, \mathcal{T}_\alpha\{f_i\}) \ , \end{aligned} \tag{3}$$

Eq. (3) 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. (3) 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.

Clearly, the computation of this posterior joint probability does not depend on the specific choice of certain distance function d. The choice of a suitable d depends rather on the choice of image representation which may be chosen from a large variety (PCA, LDA, bunchgraph) that have proven to be useful for face recognition. σ and the normalizing constant z have do be chosen with respect to the chosen distance measure d.

4 Learning Exemplars from Video

In order to realize *video-to-video* recognition, a probabilistic model needs to be learned from each gallery video V. For this we take an approach which is similar to the ones proposed in [17; 3]. These two approaches have in common that they try to find a set of exemplars that describe the set of training images best, i.e., that minimize the expected distance between the given set of images $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ and a set of exemplars (cluster centers) $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$:

$$E\left\{d(\mathcal{Z}, \mathcal{C})\right\} \ . \tag{4}$$

In other words let $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ be the sequence of video images. It is being searched for a set of exemplars $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ for that video such that

$$p(Z_t) = \sum_{c \in \mathcal{C}} \int_{\mathcal{A}} p(Z_t|\alpha, c) p(\alpha|c) p(c) d\alpha \tag{5}$$

is maximal for all t.

In [17], a k-means clustering is applied to minimize Eq. (5), in [3] an EM approach is used. In spite of the essential differences between these two approaches they have for our purposes the common draw-back that

- they find k exemplars, where k has to be given in advance. For face recogntition this draw-back is essential: Clearly, in Eq. (4) the distance measure d may be chosen arbitrarily and for face recognition it is wise to choose one of the well evaluated ones (PCA, LDA, ...) [15]. Thresholds and variances for each of these measures that minimize mis-classification are known and considering them asks for a dynamic choice of the number of clusters rather than a static one.
- they have to store the training data in order to compute the clusters which becomes difficult for long video streams.

Being inspired by the probabilistically interpreted RBF neural network approach [11], we propose an online technique to learn the exemplars: At each time step t, $p(Z_t|\alpha, c)$ of Eq. (5) is maximized. If $p(Z_t|\alpha, c) < \rho$ for some ρ (which depends on the choice of d) then Z_t is added to the set of exemplars.

4.1 Learning Algorithm

In this section we will discuss the details about our online-learning approach. For this let $\mathcal{Z} = \{\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_N\}$ be the training images, and $\mathcal{C} = \{c_1, \ldots, c_K\}$ be a set of already located exemplars.

1. The first step is the alignment or tracking step: a cluster i and a deformation $\alpha \in \mathcal{A}$ is found such that $d(\mathcal{T}_\alpha\{c_i\}, Z_t)$ is minimized:

$$\begin{aligned} \alpha_t \longleftarrow \arg\min_\alpha \min_i d(\mathcal{T}_\alpha^{-1}\{Z_t\}, c_i) \text{ and} \\ i_t \longleftarrow \arg\min_i d(\mathcal{T}_{\alpha_t}^{-1}\{Z_t\}, c_i) \end{aligned} \tag{6}$$

2. The second step generates a new cluster center, if necessary: if

$$p(Z_t|\alpha_t, c_{i_t}) < \rho$$

then

$$\mathcal{C} \longleftarrow \mathcal{C} \cup \{\widehat{\mathcal{T}_\alpha^{-1}\{Z_t\}}\} ,$$

where $\widehat{\mathcal{T}_\alpha^{-1}\{Z_t\}}$ is the subimage of $\mathcal{T}_\alpha^{-1} Z_t$ on which the computation of the distance d the first step (6) was based.
Count the number of times, $count(i_t) = count(i_t) + 1$, that cluster c_{i_t} approximaged image Z_t best.
3. Repeat steps one and two until all video frames are processed.
4. Compute the mixture weights $\pi_i \propto count(i)$.

The result of this learning procedure is

1. a set $\mathcal{C} = \{c_1 \ldots, c_K\}$ of aligned exemplars c_i
2. a prior π_i for each of the exemplars c_i.

Clearly, the more careful the set $\mathcal{A}$ is chosen, fewer exemplars are generated. Allowing $\mathcal{A}$, e.g., to compensate only for translations, exemplars are generated to compensate scale changes and rotation.

Given a gallery $\mathcal{V}$ of videos, the above has to be carried out for each video.

During recognition, as will be explained in the next section, the exemplars are used as centers of mixture models.

The above learning algorithm is motivated by the online learning approaches for artificial neural networks (ANNs) [4; 12] and clearly, all kinds of enhancements can be imagined (topology preserving maps, neighborhood relations, etc.). An online learning algorithm for exemplars used during testing could allow, in a bootstrapping manner, to learn new exemplars from probe videos.

In [18] a similar learning approach was presented. In contrary to our work, face images are not normalized with respect to $\mathcal{A}$ which results in a far larger number of clusters. In [6] a 'Unit Face' RBF model is porposed where for each individual, a single RGF network is trained. The authors have also investigated different geometrical normalizations and have tested preprocessing such as the application of a 'difference of Gaussians' or Gabor wavelets.

The goal of both above works was to build a representation of a face intensity by using an RBF network. We want to make clear once more, that this is exectly what we do not want! Our intention is, to chose a well-known face representation in *advance* (such as, e.g., PCA). *Then*, we learn the different exemplars of a single face. The advantage is that this way we inherit the "face recognition capabilities" of the PCA, LDA, etc. techniques and recognition rates can thus be predicted. This representation can be viewed as an "appearance-based 3D model", where affine tracking is used to compensate for the missing calibration information.

5 Tracking and Recognizing in Video

In this section we discuss the recognition of individuals in videos. After the application of the learning algorithm in the previous section, we have a set of exemplars $\mathcal{C}^i$ for each individual $i \in \mathcal{N}$ in the Gallery $\mathcal{V}$.

5.1 Tracking and Recognition in the Bayesian Framework

We can now compute the observation likelihoods as in Eq. 3 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

$$p(X_t|Z_1, \ldots, Z_t) \tag{7}$$

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

$$\begin{aligned} &p(X_t|Z_1, Z_2, \ldots, Z_t) \equiv p_t(\alpha_t, i_t) \\ &= \sum_{i_{t-1}} \int_{\alpha_{t-1}} p(Z_t|\alpha_t, i_t) p(\alpha_t, i_t|\alpha_{t-1}, i_{t-1}) p_{t-1}(\alpha_{t-1}, i_{t-1}) \; . \end{aligned} \tag{8}$$

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

$$p(i_t|Z_1,\ldots,Z_t) = \int_{\alpha_t} p(\alpha_t, i_t|Z_1,\ldots,Z_t) \ . \tag{9}$$

Maximizing (9) leads to the desired identity.

5.2 Exemplars as Mixture Centers

To take into account a set of exemplars $\mathcal{C}^i = \{c^i_1,\ldots,c^i_{K_i}\}$ for each individual i, we refine Eq. (3):

$$\begin{aligned} p(Z|X) &\equiv p(Z|\alpha, i) \\ &\propto \sum_{c\in\mathcal{C}^i} p(Z|\alpha, i, c)p^i(c) \qquad (10) \\ &\propto \sum_{c\in\mathcal{C}^i} z \exp\left[-\frac{1}{2\sigma^2} d(Z, T_\alpha\{c\})\right] \pi^i_c \ . \qquad (11) \end{aligned}$$

Here, the *exemplars* in $\mathcal{C}^i$ are used as the mixture center of a joint distribution and $p^i(c) = \pi^i_c$ is the prior for mixture center c of individual i.

5.3 Dynamic Model

In Eq. (8)

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

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 [17].

5.4 Computation of Posterior Distribution

We have used a particle method to efficiently compute $p_t(i_t, \alpha_t|Z_t)$ [21; 2; 7; 8; 10], where i_t, α_t depicts the hypothesised identity and transformation of the individual in the video. In [7] only the transformation α_t was estimated, in [21] the special case was discussed where each individual is presented by only a single exemplar. In our case, however, we have a Gallery of $N = |\mathcal{N}|$ persons and each person i is represented by a set of exemplars $\mathcal{C}_i = \{c^i_1,\ldots,c^i_{K_i}\}$.

In order to efficiently use the Condensation method, it has to be adapted for our needs. Using Condensation the posterior probability distribution $p_t(i_t, k_t, \alpha_t|Z_t)$

(where i refers to the individual and k to the exemplar number) is represented by a set of M indexed and weighted particles

$$\left\{\left(i^{(m)}, j^{(m)}, \alpha^{(m)}, w^{(m)}\right)\right\}_{m=1\ldots M}^{t} \quad . \tag{12}$$

Here, $i^{(m)}$ refers to the identity, $j^{(m)}$ to the exemplar, $\alpha^{(m)}$ to the deformation and $w^{(m)}$ to the weight. Note that we have, for better readability, indexed the entire set with t, instead of each component. Since all exemplars per person are aligned, we do not have to treat the different exemplars for a single person seperately. We can therefore increase efficiency if we rewrite set (12):

$$\left\{\begin{bmatrix} i^{(m)}, 1, \alpha^{(m)}, w_1^{(m)} \\ \vdots \\ i^{(m)}, K_{i^{(m)}}, \alpha^{(m)}, w_{K_i}^{(m)} \end{bmatrix}\right\}_{m=1..M'}^{t} \quad . \tag{13}$$

Set (13) is a set of $K_{i^{(m)}} \times 4$ dimensional matrices, and each matrix represents one particle, where $K_{i^{(m)}} = \left|\mathcal{C}^{i^{(m)}}\right|$. We can now easily marginalizing over $\mathcal{C}^{i^{(m)}}$ to compute the posteriori probability $p_t(i_t, \alpha_t | Z_t)$: We get with

$$\hat{w}^{(m)} = \sum_{k=1}^{K_{i^{(m)}}} \pi_k^{i^{(m)}} {w_k}^{(m)} \tag{14}$$

a new set of weighted sample vectors:

$$\left\{\left(i^{(m)}, \alpha^{(m)}, \hat{w}^{(m)}\right)\right\}_{m=1\ldots M'}^{t} \quad . \tag{15}$$

In Eq. (14), $\pi_k^{i^{(m)}}$ is the prior of exemplar k of person $i^{(m)}$.

To compute the identity from the particle set (15) we marginalize over α in the same manner. See [21] for a detailed discussion of convergance speed and convergence properties of this particle method.

6 Experiments

We have tested our *video-to-video* paradim on 100 video sequences of 25 different individuals.

The video sequences show the individuals walking on a tread-mill. We simulated different walking styles to assure a variety of conditions that are likely to appear in real life: *Walking slowly*, *walking fast*, *inclining* and *carrying an object*. Therefore, four videos per person are available. The subjects were not aware that their images were being acquired for face recognition studies and so did not move their heads unnaturally. During recording of the videos, illumination conditions were not altered. Each video constists of 300 frames (480×640 pixels per frame) captured at 30 Hz.

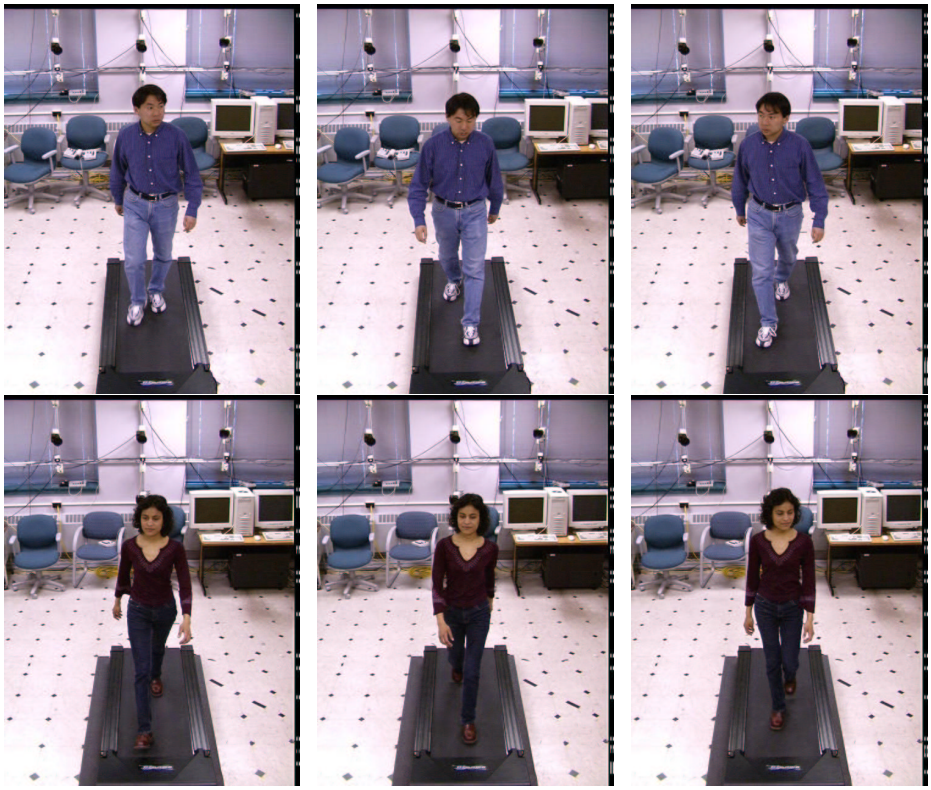

Fig. 1. The figure shows example images of the videos (*slowWalk*).

Some example images of the videos (*slowWalk*) are shown in Fig. 1.

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

In the experiments we used one of the video types as gallery videos for training while the remaining ones were used as probes for testing.

For each gallery video, a first face sample was cropped by hand. Based on this sample, the training process was started. Four examples of automatically extracted exemplar sets are shown in Fig. 2 (extracted from the videos *slowWalk*). The top row shows the exemplars of subjects 04006 and 04079 (six exemplars each). The leftmost exemplars of each of the two sets are the handextracted ones. Rows three and four of Fig. 2 shows the exemplars of subject 04015, rows five and six the exemplars of subject 04022. The top left exemplars of each of the two sets are again the handextracted ones. Clearly, the number of generated exemplars depends on the variety of different views that are apparent in the gallery video. To generate these exemplars, we set $\rho = 0.65$ and standard deviation per pixel to $\sigma = 0.4$. Increase of σ to $\sigma = 0.5$ roughtly decreased the number of exemplars by a factor of two.

During testing, these exemplar galleries were used to compute, over time, the posteriori probabilities $p_t(i_t|Z_t)$. It is interesting to see, how the posteriori

Fig. 2. The figure shows the exemplars of a person in a gallery video. In this example, *slowWalk*-videos were used as gallery videos.

probabilities develope over time. Examples for this can be seen in Fig. 3. 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 in the left and the middle plot, that the dashed line (true hypothesis) increases quickly to one. The left plot shows an example of p_t within the first 20 frames. Here, at frame $t = 18$ the particle method had converged. In order to consider *all* the frames of the video, we restart the algorithm after convergence. Recognition is established by that identity, to which the SIS converged most often.

Examples illustrating the robustness as well as of the limits of our approach are shown in Figs. 2, 3 and 5: Due to the severe differences between the gallery exemplars (derived from "slowWalk") in Fig. 2 (5th and 6th row) and the sample images from the probe video in Fig. 4, the recognition of subject 04022 was not successful. On the other hand, in spite of the differences between the gallery exemplars and the probe video, subject 04079 was always recognized successfully (see Fig. 3, right). The major problems that we encountered during our experiments were:

1. Subjects appear severely different in the gallery video and in the probe videos: This was the case for about 50% of the failed experiments.

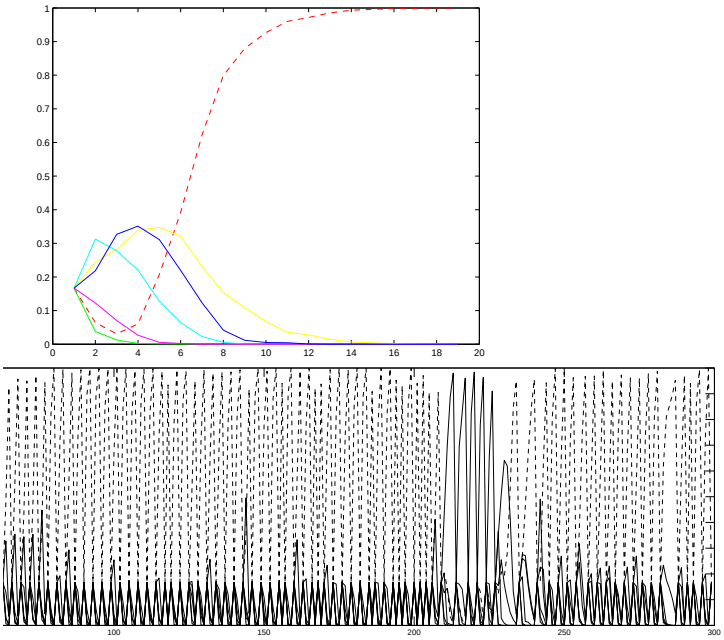

Fig. 3. The figure shows two typical probability evolutions of two successful recognitions. The graphs plot the top 5 matches, the dashed line refers to the true hypothesis. The x-axis refers to the time t. The top graph shows the curve (subject 04006) for the first 20 frames, the bottom graph (subject 04079) shows the curve for the entire video (gallery: slowWalk; probe: fastWalk).

2. Subjects looked down while walking: This was the case for roughly 10 subjects (Fig. 6). In some cases, where the subject looked down in the Gallery as well as in the Probe, this wasn't secessarily a problem. However, in cases, where this happened in either the probe or the gallery (see Fig. 6, left), this led to mis-classification.

Clearly, both problems can be solved by using more gallery videos. We have therefore done a second set of experiments, where two videos were used in the gallery, while the testing was carried out on the remaining two videos. The overall recognition results for one and two gallery videos are summarized in Table 1. 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.

In [17] it is demonstrated that the dynamic information can also be learned. We have done extensive experiments to incorporate facial and dynamic information. However, we have observed, that the dynamic information of persons can change severely with walking speed. Therefore, we have not used that information.

Fig. 4. The figure shows sample frames 1, 35, 81, 100 of a probe video. One observes large differences to the gallery. In this case recognition was *not* successful.

Fig. 5. The figure shows sample frames 1, 9, 40, 72 of a probe video. One observes large differences to the gallery. In this case, however, recognition *was* successful.

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

In a further experiment, we have used our method for training and testing on video data from a surveillance camera. The used camera was a *Philips EnviroDome* camera, mounted on the cieling of the lobby of our building. The camera has a resolution of 640×480 interlaced pixels. Images of an example view of that camera are shown in Fig. 7. We have acquired a gallery video for training and a probe video for testing. Training results from the gallery video are shown in Fig. 8. One can see, that not only facial information is learned but also the image noise that is due to interlacing and scaling[1]. Recognition in this test was successful on the part of the probe video, where the contrast was high (i.e. when the subject in the video passed unter a cieling light). For low contrast the recognition failed, however. A thorough evaluation of our method on such camera data is being undertaken.

[1] During training, to assure a lossless representation, we enlarged the face images to 80×80 pixels.

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**

slow	fast	incline	carrying
g	96%	92%	88%
92%	**g**	92%	92%
96%	96%	**g**	96%
88%	88%	83%	**g**
g	**g**	96%	92%
g	100%	**g**	100%
g	96%	96%	**g**
100%	**g**	**g**	96%
92%	**g**	96%	**g**
100%	96%	**g**	**g**

Table 1. Overall recognition rates in percent for $\sigma = 0.4$ (left), and $\sigma = 0.5$ (right). The 'g' indicates the video used as gallery.

Fig. 6. Images show failure examples, where the galleries were not sufficient to recognize the subjects.

7 Conclusion

The method presented takes advantage of the probabilistic framework for the automatic generation of exemplar-based models from video sequences and for the tracking and recognition in video sequences. One major power of this technique is its independence from the choice of the distance function d:

- This allows to add temporal fusion to the different well known face representations (see [15]).
- Also, as the distance function d measures the uncertainty in a recognition process, it assures at the same time that enough exemplars for a successful recognition under a variety of conditions are generated.

In order to show that our paradigm is able to recognize faces, one needs to work with much larger face databases; What we *can* show, however, is that the system is able to generate automatically an appropriate number of good exemplars by taking into account the distance function d. Once the exemplars are generated, they could theoretically be used as gallery images in a *multiple-*

Fig. 7. The Figure shows sample views from our surveillance camera.

Fig. 8. The Figure shows the exemplars learned from the video shown in Fig. 7.

still-to-multiple-still (or *still-to-multiple-still*) face recognition approach. Here, recognition is based on the pairwise computation of the distance measure $d(\cdot, \cdot)$.

If we use one of the image representations that are already well known for face recognition (such as, e.g., the PCA) and an appropriate distance measure, we can consider the techniques that were tested in the FERET test [15] as a "baseline" for our paradigm. That means, we can predict the recognition results for our paradigm and the FERET-performances are a lower bound. We have presented in this work a face recognition application. However, it appeares that the presented method should be applicable also to other recognition-from-video problems such as street-sign recognition from a driving car[5]

References

1. 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.

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. B. Frey and N. Jojic. Learning graphical models if images, videos and their spatial transformations. In *Proc. Conf. Uncertainty in Artificial Intelligence*, 2000.
4. B. Fritzke. Growing cell structures – a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7:1441–1460, 1995.
5. D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Proc. Int. Conf. on Computer Vision*, pages 87–93, Korfu, Greece, 1999.
6. A. Howell and H. Buxton. Towards unconstrained face recognition from image sequences. In *Proc. British Machine Vision Conference*, pages 455–464, Edinburgh, 1996.
7. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. J. of Computer Vision*, 1998.
8. G. Kitagawa. Monta carlo filter and smoother for non-gaussian nonlinear state space models. *J. Computational and Graphical Statistics*, 5:1–25, 1996.
9. B. Li and R. Chellappa. Simultanious tracking and verification via sequential posterior estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, June 13-15, 2000.
10. J. Liu and R. Chen. Sequential monte carlo for dynamic systems. *Journal of the American Statistical Association*, 93:1031–1041, 1998.
11. D. Lowe. Radial basis function networks. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 779–782. MIT-Press, 1995.
12. T. Martinez and K. Schulten. Topology representing networks. *Neural Networks*, 7:505–522, 1994.
13. 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.
14. B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:696–710, 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. J. Steffens, E. Elagin, and H. Neven. Personspotter – fast and robust system for human detection, tracking and recognition. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, pages 516–521, Nara, Japan, April 14-16, 1998.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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. Accepted for publication.