Face Recognition from Video Using the Generic Shape-Illumination Manifold

Ognjen Arandjelović and Roberto Cipolla

Department of Engineering, University of Cambridge, CB2 1PZ, UK

Abstract. In spite of over two decades of intense research, illumination and pose invariance remain prohibitively challenging aspects of face recognition for most practical applications. The objective of this work is to recognize faces using video sequences both for training and recognition input, in a realistic, unconstrained setup in which lighting, pose and user motion pattern have a wide variability and face images are of low resolution. In particular there are three areas of novelty: (i) we show how a photometric model of image formation can be combined with a statistical model of generic face appearance variation, learnt offline, to generalize in the presence of extreme illumination changes; (ii) we use the smoothness of geodesically local appearance manifold structure and a robust same-identity likelihood to achieve invariance to unseen head poses; and (iii) we introduce an accurate video sequence "reillumination" algorithm to achieve robustness to face motion patterns in video. We describe a fully automatic recognition system based on the proposed method and an extensive evaluation on 171 individuals and over 1300 video sequences with extreme illumination, pose and head motion variation. On this challenging data set our system consistently demonstrated a nearly perfect recognition rate (over 99.7%), significantly outperforming state-of-the-art commercial software and methods from the literature.

1 Introduction

Automatic face recognition (AFR) has long been established as one of the most active research areas in computer vision. In spite of the large number of developed algorithms, real-world performance of AFR has been, to say the least, disappointing. Even in very controlled imaging conditions, such as those used for passport photographs, the error rate has been reported to be as high as 10% [6], while in less controlled environments the performance degrades even further. We believe that the main reason for the apparent discrepancy between results reported in the literature and observed in the real world is that the assumptions that most AFR methods rest upon are hard to satisfy in practice.

In this paper, we are interested in recognition using *video sequences*. This problem is of enormous interest as video is readily available in many applications, while the abundance of information contained within it can help resolve some of the inherent ambiguities of single-shot based recognition. In practice, video data can be extracted from surveillance videos by tracking a face or by instructing a cooperative to move the head in front of a mounted camera.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 27–40, 2006.

We assume that both the training and novel data available to an AFR system is organized in a database where a sequence of images for each individual contains some variability in pose, but is not obtained in scripted conditions or in controlled illumination. The recognition problem can then be formulated as taking a sequence of face images from an unknown individual and finding the best matching sequence in the database of sequences labelled by the identity.

Our approach consists of using a weak photometric model of image formation with offline machine learning for modelling *manifolds* of faces. Specifically, we show that the combined effects of face shape and illumination can be effectively learnt using Probabilistic PCA (PPCA) [40] from a small, unlabelled set of video sequences of faces in randomly varying lighting conditions, while a novel manifold-based "reillumination" algorithm is used to provide robustness to pose and motion pattern. Given a novel sequence, the learnt model is used to decompose the face appearance manifold into albedo and shape-illumination manifolds, producing the classification decision by robust likelihood estimation.

2 Previous Work

Good general reviews of recent AFR literature can be found in [5, 46]. In this section, we focus on AFR literature that deals specifically with recognition from image sequences, and with invariance to pose and illumination.

Compared to single-shot recognition, face recognition from image sequences is a relatively new area of research. Some of the existing algorithms that deal with multi-image input use temporal coherence within the sequence to enforce prior knowledge on likely head movements [26, 27, 47]. In contrast to these, a number of methods that do not use temporal information have been proposed. Recent ones include statistical [3, 35] and principal angle-based methods with underlying simple linear [16], kernel-based [45] or Gaussian mixture-based [24] models. By their very nature, these are inherently invariant to changes in head motion pattern. Other algorithms implement the "still-to-video" scenario [28, 31], not taking full advantage of sequences available for training.

Illumination invariance, while perhaps the most significant challenge for AFR [1] remains a virtually unexplored problem for recognition using video. Most methods focus on other difficulties of video-based recognition, employing simple preprocessing techniques to deal with changing lighting [4, 13]. Others rely on availability of ample training data but achieve limited generalization [3, 37].

Two influential generative model-based approaches for illumination-invariant single-shot recognition are the illumination cones [7, 18] and the 3D morphable model [10]. Both of these have significant shortcomings in practice. The former is not readily extended to deal with video, assuming accurately registered face images, illuminated from several well-posed directions for each pose which is difficult to achieve in practice (see §5 for data quality). Similar limitations apply to the related method of Riklin-Raviv and Shashua [34]. On the other hand, the 3D morphable model is easily extended to video-based recognition, but it requires a (in our case prohibitively) high resolution [13], struggles with non-Lambertian effects (such as specularities) and multiple light sources, and has convergence

problems in the presence of background clutter and partial occlusion (glasses, facial hair).

Broadly speaking, there are three classes of algorithms aimed at achieving pose invariance. The first, a model-based approach, uses an explicit 2D or 3D model of the face, and attempts to estimate the parameters of the model from the input [10, 23]. This is a view-independent representation. A second class of algorithms consists of global, parametric models, such as the eigenspace method [30] that estimates a single parametric (typically linear) subspace from all the views for all the objects (also see [29]). In AFR tests, such methods are usually outperformed by methods from the third class: view-based techniques e.g. the view-based eigenspaces [32] (also [26, 27]), in which a separate subspace is constructed for each pose. These algorithms usually require an intermediate step in which the pose of the face is determined, and then recognition is carried out using the estimated view-dependent model. A common limitation of these methods is that they require a fairly restrictive and labour-intensive training data acquisition protocol, in which a number of fixed views are collected for each subject and appropriately labelled. This is not the case with the proposed method.

3 Face Motion (and Other) Manifolds

Concepts in this paper heavily rely on the notion of face *manifolds*. Briefly, under the standard rasterized representation of an image, images of a given size can be viewed as points in a Euclidean *image space*, its dimensionality being equal to the number of pixels D. However, the surface and texture of a face is mostly smooth making its appearance quite constrained and confining it to an embedded *face manifold* of dimension $d \ll D$ [3, 9]. Formally, the distribution of observed face images of the subject i can be written as the integral:

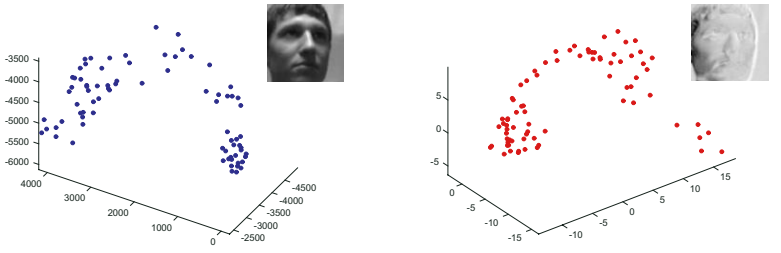
$$p^{(i)}(\mathbf{X}) = \int p_F^{(i)}(\mathbf{x}) p_n(f_i(\mathbf{x}) - \mathbf{X}) d\mathbf{x}. \tag{1}$$

where p_n is the noise distribution, $f^{(i)} : \mathbb{R}^d \to \mathbb{R}^D$ the embedding function and $\mathbf{x}$ an intrinsic face descriptor. Fig. 1 (a) illustrates the validity of the notion on an example of a face motion image sequence. For the proposed method, the crucial properties are their (i) continuity and (ii) smoothness.

3.1 Synthetic Reillumination of Face Motion Manifolds

One of the key ideas of this paper is the *reillumination* of video sequences. Our goal is to take two input sequences of faces and produce a third, synthetic one, that contains the same poses as the first in the illumination of the second.

The proposed method consists of two stages. First, each face from the first sequence is matched with the face from the second that corresponds to it best in terms of pose. Then, a number of faces close to the matched one are used to finely reconstruct the reilluminated version of the original face. Our algorithm is therefore global, unlike most of the previous methods which use a sparse set



(a) Face Motion Manifold (FMM) (b) Shape-Illumination Manifold

Fig. 1. Manifolds of (a) face appearance and (b) albedo-free appearance i.e. the effects of illumination and pose changes, in a single motion sequence. Shown are projections to the first 3 linear principal components, with a typical manifold sample on the top-right.

of detected salient points for registration, e.g. [4, 8, 16]. We found that facial feature localization using trained Support Vector Machines (similar to [4, 8]), as well as algorithms employed in commercial systems FacePass® [15, 41] and FaceIt® [22] failed on data sets used for evaluation in this paper (see §5) due to the severity of illumination conditions. We next describe the two stages of the proposed algorithm in detail.

Stage 1: Pose Matching. Let $\{\mathbf{X}_i\}^{(1)}$ and $\{\mathbf{X}_i\}^{(2)}$ be two motion sequences of a person's face in two different illuminations. Then, for each $\mathbf{X}_i^{(1)}$ we are interested in finding $\mathbf{X}_{c(i)}^{(2)}$ that corresponds to it best in pose. Finding the unknown mapping c on a frame-by-frame basis is difficult. Instead, we formulate the problem as a minimization task with the fitness function taking the form:

$$f(c) = \sum_j d_E\left(\mathbf{X}_j^{(1)}, \mathbf{X}_{c(j)}^{(2)}\right)^2 + \omega \sum_j \sum_k \frac{d_G^{(2)}\left(\mathbf{X}_{c(j)}^{(2)}, \mathbf{X}_{c(n(j,k))}^{(2)}; \{\mathbf{X}_j\}^{(2)}\right)}{d_G^{(1)}\left(\mathbf{X}_j^{(1)}, \mathbf{X}_{n(j,k)}^{(1)}; \{\mathbf{X}_j\}^{(1)}\right)} \qquad (2)$$

where $n(i, j)$ is the j-th of K nearest neighbours of face i, d_E a pose dissimilarity function and $d_G^{(k)}$ a geodesic distance estimate along the FMM of sequence k. The first term is easily understood as a penalty for dissimilarity of matched pose-signatures. The latter enforces a globally good matching by favouring mappings that map geodesically close points from the domain manifold to geodesically close points on the codomain manifold.

Pose-matching function: The performance of function d_E in (2) at estimating the goodness of a frame match is crucial for making the overall optimization scheme work well. Our approach consists of filtering the original face image to produce a quasi illumination-invariant *pose-signature*, which is then compared with other pose-signatures using the Euclidean distance. Note that these signatures are *only* used for frame matching and thus need not retain any power of discrimination between individuals – all that is needed is sufficient pose information. We use a distance-transformed edge map of the face image as a pose-signature,

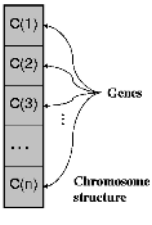
(a) Original (b) Reilluminated

Fig. 2. (a) Original images from a novel video sequence and (b) the result of reillumination using the proposed genetic algorithm with nearest neighbour-based reconstruction

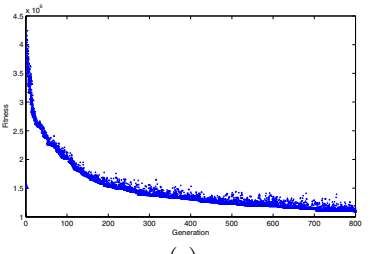
motivated by the success of this representation in object-configuration matching across other computer vision applications, e.g. [17, 38].

Minimizing the fitness function: Exact minimization of the fitness function (2) over all functions c is an NP-complete problem. However, since the final synthesis of novel faces (Stage 2) involves an entire geodesic neighbouring of the paired faces, it is inherently robust to some non-optimality of this matching. Therefore, in practice, it is sufficient to find a good match, not necessarily the optimal one.

We propose to use a genetic algorithm (GA) [12] as a particularly suitable approach to minimization for our problem. GAs rely on the property of many optimization problems that sub-solutions of good solutions are good themselves. Specifically, this means that if we have a globally good manifold match, then local matching can be expected to be good too. Hence, combining two good matches is a reasonable attempt at improving the solution. This motivates the chromosome structure we use, depicted in Fig. 3 (a), with the i-th gene in a chromosome being the value of $c(i)$. GA parameters were determined experimentally from a small training set and are summarized in Fig. 3 (b,c).



Property	Value
Population size	20
Elite survival no.	2
Mutation (%)	5
Migration (%)	20
Crossover (%)	80
Max. generations	200



(a) (b) (c)

Fig. 3. (a) The chromosome structure used in the proposed GA optimization, (b) its parameters and (c) population fitness (see (2)) in a typical evolution. Maximal generation count of 200 was chosen as a trade-off between accuracy and matching speed.

Estimating geodesic distances: The definition of the fitness function in (2) involves estimates of geodesic distances along manifolds. Due to the nonlinearity of FMMs [3, 27] it is not well approximated by the Euclidean distance. We estimate the geodesic distance between every two faces from a manifold using the Floyd's algorithm on a constructed undirected graph whose nodes correspond to face images (also see [39]). Then, if $\mathbf{X}_i$ is one of the K nearest neighbours of $\mathbf{X}_j$:

$$d_G(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|_2 \,. \tag{3}$$

Otherwise:

$$d_G(\mathbf{X}_i, \mathbf{X}_j) = \min_k \left[d_G(\mathbf{X}_i, \mathbf{X}_k) + d_G(\mathbf{X}_k, \mathbf{X}_j) \right]. \tag{4}$$

Stage 2: Fine Reillumination. Having computed a pose-matching function c^*, we exploit the smoothness of FMMs by computing $\mathbf{Y}_i^{(1)}$, the reilluminated frame $\mathbf{X}_i^{(1)}$, as a linear combination of K nearest-neighbour frames of $\mathbf{X}_{c^*(i)}^{(2)}$. Linear combining coefficients $\alpha_1, \ldots \alpha_K$ are found from the corresponding pose-signatures by solving the following constrained minimization problem:

$$\{\alpha_j\} = \arg\min_{\{\alpha_j\}} \left\| \mathbf{x}_i^{(1)} - \sum_{k=1}^{K} \alpha_k \mathbf{x}_{n(c^*(i),k)}^{(2)} \right\|_2 \tag{5}$$

subject to $\sum_{k=1}^{K} \alpha_k = 1.0$, where $\mathbf{x}_i^{(j)}$ is the pose-signature corresponding to $\mathbf{X}_i^{(j)}$. In other words, the pose-signature of a novel face is first reconstructed using the pose-signatures of K training faces (in target illumination), which are then combined in the same fashion to synthesize a reilluminated face, see Fig. 2 and 4. Optimization of (5) is readily performed by differentiation.

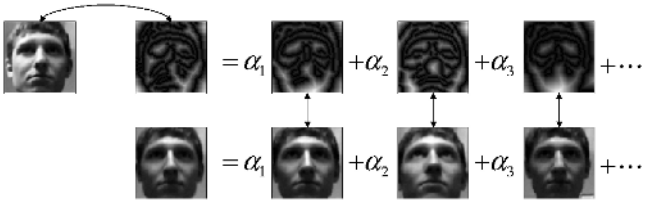


Fig. 4. Face reillumination: the coefficients for linearly combining face appearance images (bottom row) are computed using the corresponding pose-signatures (top row)

4 The Shape-Illumination Manifold

In most practical applications, specularities, multiple or non-point light sources significantly affect the appearance of faces. We believe that the difficulty of dealing with these effects is one of the main reasons for poor performance of most AFR systems when put to use in a realistic environment. In this work we make a very weak assumption on the process of image formation: the only assumption made is that the intensity of each pixel is a linear function of the albedo $a(j)$ of the corresponding 3D point:

$$X(j) = a(j) \cdot s(j) \tag{6}$$

where $\mathbf{s}$ is a function of illumination, shape and other parameters not modelled explicitly. This is similar to the reflectance-lighting model used in Retinex-based algorithms [25], the main difference being that we make no further assumptions on the functional form of $\mathbf{s}$. Note that the commonly-used (e.g. see [10, 18, 34]) Lambertian reflectance model is a special case of (6) [7]:

$$s(j) = \max(\mathbf{n}_j \cdot \mathbf{L}, 0) \tag{7}$$

where $\mathbf{n}_i$ is the surface normal and $\mathbf{L}$ the intensity-scaled illumination direction.

The image formation model introduced in (6) leaves the image pixel intensity as an unspecified function of face shape or illumination parameters. Instead of formulating a complex model of the geometry and photometry behind this function (and then needing to recover a large number of model parameters), we propose to learn it implicitly. Consider two images, $\mathbf{X}_1$ and $\mathbf{X}_2$ of the same person, in the same pose, but different illuminations. Then from (6):

$$\Delta \log X(j) = \log s_2(j) - \log s_1(j) \equiv d_s(j) \tag{8}$$

In other words, the difference between these logarithm-transformed images is not a function of face albedo. As before, due to the smoothness of faces, as the pose of the subject varies the difference-of-logs vector $\mathbf{d}_s$ describes a manifold in the corresponding embedding vector space. These is the Shape-Illumination manifold (SIM) corresponding to a particular pair of video sequences, see Fig. 1 (b).

The Generic SIM: A crucial assumption of our work is that the Shape-Illumination Manifold of all possible illuminations and head poses is *generic for human faces* (gSIM). This is motivated by a number of independent results reported in the literature that have shown face shape to be less discriminating than albedo across different models [11, 20] or have reported good results in synthetic reillumination of faces using the constant-shape assumption [34]. In the context of face manifolds this means that the effects of *illumination and shape* can be learnt offline from a training corpus containing typical modes of pose and illumination variation.

It is worth emphasizing the key difference in the proposed offline learning from previous approaches in the literature which try to learn the *albedo* of human faces. Since offline training is performed on persons not in the online gallery, in the case when albedo is learnt it is necessary to have means of generalization i.e. learning what *possible* albedos human faces can have from a small subset. In [34], for example, the authors demonstrate generalization to albedos in the rational span of those in the offline training set. This approach is not only unintuitive, but also without a meaningful theoretical justification. On the other hand, previous research indicates that illumination effects can be learnt *directly* without the need for generalization [3].

Training data organization: The proposed AFR method consists of two training stages – a one-time offline learning performed using *offline training data* and a stage when *gallery data* of known individuals with associated identities is collected. The former (explained next) is used for learning the generic face shape contribution to face appearance under varying illumination, while the latter is used for subject-specific learning.

4.1 Offline Stage: Learning the Generic SIM (gSIM)

Let $\mathbf{X}_i^{(j,k)}$ be the i-th face of the j-th person in the k-th illumination, same indexes corresponding in pose, as ensured by the proposed reillumination algorithm

Fig. 5. Learning complex illumination effects: Shown is the variation along the 1st mode of a single PPCA space in our SIM mixture model. Cast shadows (e.g. from the nose) and the locations of specularities (on the nose and above the eyes) are learnt as the illumination source moves from directly overhead to side-overhead.

in §3.1. Then from (8), samples from the generic Shape-Illumination manifold can be computed by logarithm-transforming all images and subtracting those corresponding in identity and pose:

$$\mathbf{d} = \log \mathbf{X}_i^{(j,p)} - \log \mathbf{X}_i^{(j,q)} \tag{9}$$

Provided that training data contains typical variations in pose and illumination (i.e. that the p.d.f. confined to the generic SIM is well sampled), this becomes a standard statistical problem of high-dimensional density estimation. We employ the Gaussian Mixture Model (GMM). In the proposed framework, this representation is motivated by: (i) the assumed low-dimensional manifold model (1), (ii) its compactness and (iii) the existence of incremental model parameter estimation algorithms (e.g. [21]).

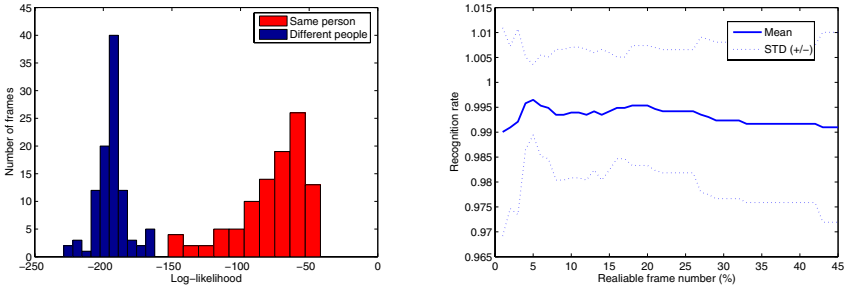
Briefly, we estimate multivariate Gaussian components using the Expectation Maximization (EM) algorithm [12], initialized by k-means clustering. Automatic model order selection is performed using the well-known Minimum Description Length criterion [12] while the principal subspace dimensionality of PPCA components was estimated from eigenspectra of covariance matrices of a diagonal GMM fit, performed first. Fitting was then repeated using a PPCA mixture. We obtained 12 components, each with a 6D principal subspace. Fig. 5 shows an example of subtle illumination effects learnt with this model.

4.2 Robust Likelihood for Novel Sequence Classification

Let gallery data consist of sequences $\{\mathbf{X}_i\}^1, \ldots, \{\mathbf{X}_i\}^N$, corresponding to N individuals, $\{\mathbf{X}_i\}^0$ be a novel sequence of one of these individuals and $\mathcal{G}(\mathbf{x}; \boldsymbol{\Theta})$ a Mixture of Probabilistic PCA corresponding to the generic SIM. Using the reillumination algorithm of §3.1, the novel sequence can be reilluminated with each from the gallery, producing samples $\{\mathbf{d}_i\}$, assumed identically and independently distributed, from a *postulated* subject-specific SIM. We compute the probability of these observations under $\mathcal{G}(\mathbf{x}; \boldsymbol{\Theta})$:

$$p_i = \mathcal{G}(\mathbf{d}_i; \boldsymbol{\Theta}) \tag{10}$$

Instead of classifying $\{\mathbf{X}_i\}^0$ using the likelihood given the *entire* set of observations $\{\mathbf{d}_i\}$, we propose a more robust measure. To appreciate the need for robustness, consider the histograms in Fig. 6 (a). It can be observed that the probability of the most similar faces in an inter-personal comparison, in terms of (10), approaches that of the most *dissimilar* faces in an *intra-personal* comparison (sometimes even exceeding it). This occurs when the correct gallery



(a) Histograms (b) Recognition

Fig. 6. (a) Histograms of intra-personal likelihoods across frames of a sequence when two sequences compared correspond to the same (red) and different (blue) people. (b) Recognition rate as a function of the number of frames deemed 'reliable'.

Algorithm 1: Offline training	**Algorithm 2:** Recognition (online)
Input: database of sequences $\{\mathbf{X}_i\}^j$ **Output:** model of gSIM $\mathcal{G}(\mathbf{d};\boldsymbol{\Theta})$	**Input:** sequences $\{\mathbf{X}_i\}^G$, $\{\mathbf{X}_i\}^N$ **Output:** same-identity likelihood ρ
1: gSIM iteration **for** all j, k	**1: Reilluminate using** $\{\mathbf{X}_i\}^G$ $\{\mathbf{Y}_i\}^N = \text{reilluminate}\big(\{\mathbf{X}_i\}^N\big)$
2: **Reilluminate using** $\{\mathbf{X}_i\}^k$ $\{\mathbf{Y}_i\}^j = \text{reilluminate}\big(\{\mathbf{X}_i\}^j\big)$	**2: Postulated SIM samples** $\mathbf{d}_i = \log \mathbf{X}_i^N - \log \mathbf{Y}_i^N$
3: **Add gSIM samples** $\mathbb{D} = \mathbb{D} \bigcup (\{\mathbf{Y}_i\}^j - \{\mathbf{X}_i\}^j)$	**3: Compute likelihoods of** $\{\mathbf{d}_i\}$ $p_i = \mathcal{G}(\mathbf{d}_i;\boldsymbol{\Theta})$
4: Computed gSIM samples **end for**	**4: Order** $\{\mathbf{d}_i\}$ **by likelihood** $p_{s(1)} \geq \cdots \geq p_{s(N)} \geq \ldots$
5: GMM $\mathcal{G}$ **from gSIM samples** $\mathcal{G}(\mathbf{d};\boldsymbol{\Theta})$ =EM_GMM($\mathbb{D}$)	**5: Inter-manifold similarity** ρ $\rho = \sum_{i=1}^{N} \log p_{s(i)}/N$

Fig. 7. A summary of the proposed offline learning and recognition algorithms

sequence contains poses that are very dissimilar to even the most similar ones in the novel sequence, or vice versa (note that small dissimilarities are extrapolated well from local manifold structure in (5)). In our method, the robustness to these, unseen modes of pose variation is achieved by considering the mean log-likelihood given only the most probable faces. In our experiments we used the top 15% of faces, but we found the algorithm to exhibit little sensitivity to the exact choice of this number, see Fig. 6 (b). A summary of proposed algorithms is shown in Fig. 7.

5 Empirical Evaluation

Methods in this paper were evaluated on three databases:

- **FaceDB100**, with 100 individuals of varying age and ethnicity, and equally represented genders. For each person in the database we collected 7 video sequences of the person in arbitrary motion (significant translation, yaw and pitch, negligible roll), each in a different illumination setting, see [2, 3] and Fig. 8, at 10fps and 320×240 pixel resolution (face size ≈ 60 pixels).
- **FaceDB60**, kindly provided to us by Toshiba Corp. This database contains 60 individuals of varying age, mostly male Japanese, and 10 sequences per person. Each sequence corresponds to a different illumination setting, at 10fps and 320×240 pixel resolution (face size ≈ 60 pixels), see [2].
- **FaceVideoDB**, freely available and described in [19]. Briefly, it contains 11 individuals and 2 sequences per person, little variation in illumination, but extreme and uncontrolled variations in pose and motion, acquired at 25fps and 160×120 pixel resolution (face size ≈ 45 pixels).

Data acquisition: The discussion so far focused on recognition using fixed-scale face images. Our system uses a cascaded detector [42] for localization of faces in cluttered images, which are then rescaled to the unform resolution of 50×50 pixels (approximately the average size of detected faces).

Methods and representations: We compared the performance of our recognition algorithm with and without the robust likelihood of §4.2 (i.e. using only the most reliable vs. all detected faces) to that of:

- State-of-the-art commercial system FaceIt® by Identix [22] (the best performing software in the most recent Face Recognition Vendor Test [33]),
- Constrained MSM (CMSM) [16] used in a state-of-the-art commercial system FacePass® [41],
- Mutual Subspace Method (MSM) [16], and
- KL divergence-based algorithm of Shakhnarovich *et al.* (KLD) [35].

In all tests, both training data for each person in the gallery, as well as test data, consisted of only a single sequence. Offline training of the proposed algorithm was performed using 20 individuals in 5 illuminations from the FaceDB100 – we

(a) FaceDB100

(b) FaceDB60

Fig. 8. Different illumination conditions in databases FaceDB100 and FaceDB60

emphasize that these were not used as test input for the evaluations reported in this section. The methods were evaluated using 3 face representations:

- raw appearance images $\mathbf{X}$,
- Gaussian high-pass filtered images – already used for AFR in [4, 14]:

$$\mathbf{X}_H = \mathbf{X} - (\mathbf{X} * \mathbf{G}_{\sigma=1.5}), \tag{11}$$

- local intensity-normalized high-pass filtered images – similar to the Self Quotient Image [43]:

$$\mathbf{X}_Q = \mathbf{X}_H / (\mathbf{X} - \mathbf{X}_H), \tag{12}$$

  the division being element-wise.

5.1 Results

A summary of experimental results is shown in Table 1. The proposed algorithm greatly outperformed other methods, achieving a nearly perfect recognition (99.7+%) on all 3 databases. This is an extremely high recognition rate for such unconstrained conditions, small amount of training data per gallery individual and the degree of illumination, pose and motion pattern variation between different sequences. This is witnessed by the performance of Simple KLD method which can be considered a proxy for gauging the difficulty of the task, seeing that it is expected to perform well if imaging conditions are not greatly different between training and test [35]. Additionally, it is important to note the excellent performance of our algorithm on the Japanese database, even though offline training was performed using Caucasian individuals only.

As expected, when plain likelihood was used instead of the robust version proposed in §4.2, the recognition rate was lower, but still significantly higher than that of other methods. The high performance of non-robust gSIM is important as an estimate of the expected recognition rate in the "still-to-video" scenario of the proposed method. We conclude that the proposed algorithm's performance seems very promising in this setup as well. Finally, note that the standard deviation of our algorithm's performance across different training and test illuminations is much lower than that of other methods, showing less dependency on the exact imaging conditions used for data acquisition.

Table 1. Average recognition rates (%) and their standard deviations (if applicable)

		gSIM, rob.	gSIM	FaceIt	CMSM	MSM	KLD
Face DB100	$\mathbf{X}$	**99.7/0.8**	97.7/2.3	64.1/9.2	73.6/22.5	58.3/24.3	17.0/8.8
	$\mathbf{X}_H$	–	–	–	85.0/12.0	82.8/14.3	35.4/14.2
	$\mathbf{X}_Q$	–	–	–	87.0/11.4	83.4/8.4	42.8/16.8
Face DB60	$\mathbf{X}$	**99.9/0.5**	96.7/5.5	81.8/9.6	79.3/18.6	46.6/28.3	23.0/15.7
	$\mathbf{X}_H$	–	–	–	83.2/17.1	56.5/20.2	30.5/13.3
	$\mathbf{X}_Q$	–	–	–	91.1/8.3	83.3/10.8	39.7/15.7
Face VideoDB	$\mathbf{X}$	**100.0**	91.9	91.9	91.9	81.8	59.1
	$\mathbf{X}_H$	–	–	–	100.0	81.8	63.6
	$\mathbf{X}_Q$	–	–	–	91.9	81.8	63.6

Representations: Both the high-pass and even further Self Quotient Image representations produced an improvement for all methods over raw grayscale. This is consistent with previous findings in the literature [1, 4, 14, 43].

Unlike in previous reports of performance evaluation of these filters, we also ask the question of *when* they help and how much in *each case*. To quantify this, consider "performance vectors" $\mathbf{s}_R$ and $\mathbf{s}_F$, corresponding to respectively raw and filtered input, whose each component is equal to the recognition rate of a method on a particular training/test data combination. Then the vector $\Delta\mathbf{s}_R \equiv \mathbf{s}_R - \overline{\mathbf{s}}_R$ contains relative recognition rates to its average on raw input, and $\Delta\mathbf{s} \equiv \mathbf{s}_F - \mathbf{s}_R$ the improvement with the filtered representation. We then considered the *angle* ϕ between vectors $\Delta\mathbf{s}_R$ and $\Delta\mathbf{s}$, using both the high-pass and Self Quotient Image representations. In both cases, we found the angle to be $\phi \approx 136°$. This is an interesting result: it means that while on average both representations increase the recognition rate, they actually *worsen* it in "easy" recognition conditions. The observed phenomenon is well understood in the context of energy of intrinsic and extrinsic image differences and noise (see [44] for a thorough discussion). Higher than average recognition rates for raw input correspond to small changes in imaging conditions between training and test, and hence lower energy of extrinsic variation. In this case, the two filters decrease the SNR, worsening the performance. On the other hand, when the imaging conditions between training and test are very different, normalization of extrinsic variation is the dominant factor and performance is improved.

This is an important observation: it suggests that the performance of a method that uses either of the representations can be increased further in a straightforward manner by detecting the difficulty of recognition conditions, see [2].

Imaging conditions: Finally, we were interested if the evaluation results on our database support the observation in the literature that some illumination conditions are intrinsically more difficult for recognition than others [36]. An inspection of the performance of the evaluated methods has shown a remarkable correlation in relative performance across illuminations, despite the very different models used for recognition. We found that relative recognition rates across illuminations correlate on average with $\rho = 0.96$.

6 Summary and Conclusions

The proposed method for AFR from video has been demonstrated to achieve a nearly perfect recognition on 3 databases containing extreme illumination, pose and motion pattern variation, significantly outperforming state-of-the-art commercial software and methods in the literature.

The main direction for future work is to make a further use of offline training data, by taking into account probabilities of both intra- and inter-personal differences confined to the gSIM. This is the focus of our current work. Additionally, we would like to improve the computational efficiency of the method by representing each FMM by a strategically chosen set of sparse samples.

References

1. Y Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *PAMI*, 19(7):721–732, 1997.
2. O. Arandjelović and R. Cipolla. A new look at filtering techniques for illumination invariance in automatic face recogition. *FG*, 2006.
3. O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *CVPR*, 2005.
4. O. Arandjelović and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. *CVPR*, 1:860–867, 2005.
5. W. A. Barrett. A survey of face recognition algorithms and testing results. *Systems and Computers*, 1:301–305, 1998.
6. BBC. Doubts over passport face scans. *BBC Online, UK Edition*, October 2004.
7. P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *IJCV*, 28(3):245–260, 1998.
8. T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D.A Forsyth. Names and faces in the news. *CVPR*, 2:848–854, 2004.
9. M. Bichsel and A. P. Pentland. Human face recognition and the face image set's topology. *Computer Vision, Graphics and Image Processing*, 59(2):254–261, 1994.
10. V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *PAMI*, 25(9):1063–1074, 2003.
11. I. Craw, N. P. Costen, T. Kato, and S. Akamatsu. How should we represent faces for automatic recognition? *PAMI*, 21:725–736, 1999.
12. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wily & Sons, Inc., New York, 2nd edition, 2000.
13. M. Everingham and A. Zisserman. Automated person identification in video. *CIVR*, pages 289–298, 2004.
14. A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. *ECCV*, pages 304–320, 2002.
15. K. Fukui and O. Yamaguchi. Facial feature point extraction method based on combination of shape extraction and pattern matching. *Systems and Computers in Japan*, 29(6):2170–2177, 1998.
16. K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *Int'l Symp. of Robotics Research*, 2003.
17. D. M. Gavrila. Pedestrian detection from a moving vehicle. *ECCV*, 2:37–49, 2000.
18. A. S. Georghiades, D. J. Kriegman, and P. N Belhumeur. Illumination cones for recognition under variable lighting: Faces. *CVPR*, pages 52–59, 1998.
19. D. O. Gorodnichy. Associative neural networks as means for low-resolution video-based recognition. *International Joint Conference on Neural Networks*, 2005.
20. R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *BMVC*, 2004.
21. P. Hall, D. Marshall, and R. Martin. Merging and splitting eigenspace models. *PAMI*, 22(9):1042–1049, 2000.
22. Identix. Faceit. `http://www.FaceIt.com/`.
23. B. Kepenekci. *Face Recognition Using Gabor Wavelet Transform.* PhD thesis, The Middle East Technical University, 2001.
24. T. Kim, O. Arandjelović, and R. Cipolla. Learning over sets using boosted manifold principal angles (BoMPA). *BMVC*, 2005. (to appear).
25. R. Kimmel, M. Elad, D. Shaked, R. Keshet, and I. Sobel. A variational framework for retinex. *IJCV*, 52(1):7–23, 2003.

26. K. Lee and D. Kriegman. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. *CVPR*, 1:852–859, 2005.
27. K. Lee, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *CVPR*, 1:313–320, 2003.
28. Y. Li, S. Gong, and H. Liddell. Modelling faces dynamically across views and over time. *ICCV*, 1:554–559, 2001.
29. X. Liu and T. Chen. Video-based face recognition using adaptive hidden Markov models. *CVPR*, 1:340–345, 2003.
30. H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *IJCV*, 14:5–24, 1995.
31. S. Palanivel, B. S. Venkatesh, and B Yegnanarayana. Real time face recognition system using autoassociative neural network models. *ASSP*, 2:833–836, 2003.
32. A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. *CVPR*, 84–91, 1994.
33. P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. FRVT 2002: Overview and summary. *Technical report, National Institute of Justice*, March 2003.
34. T. Riklin-Raviv and A. Shashua. The quotient image: Class based re-rendering and recognition with varying illuminations. *PAMI*, 23(2):219–139, 2001.
35. G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *ECCV*, 3:851–868, 2002.
36. T. Sim and S. Zhang. Exploring face space. *Face Processing in Video*, 2004.
37. J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. *CIVR*, 2005.
38. B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Filtering using a tree-based estimator. *ICCV*, 2:1063–1070, 2003.
39. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
40. M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.
41. Toshiba. Facepass. *http://www.toshiba.co.jp/mmlab/tech/w31e.htm*.
42. P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2), 2004.
43. H. Wang, S. Z. Li, and Y. Wang. Face recognition under varying lighting conditions using self quotient image. *FG*, pages 819–824, 2004.
44. X. Wang and X. Tang. Unified subspace analysis for face recognition. *ICCV*, 2003.
45. L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *JMLR*, 4(10):913–931, 2003.
46. W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2004.
47. S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(1):214–245, 2003.