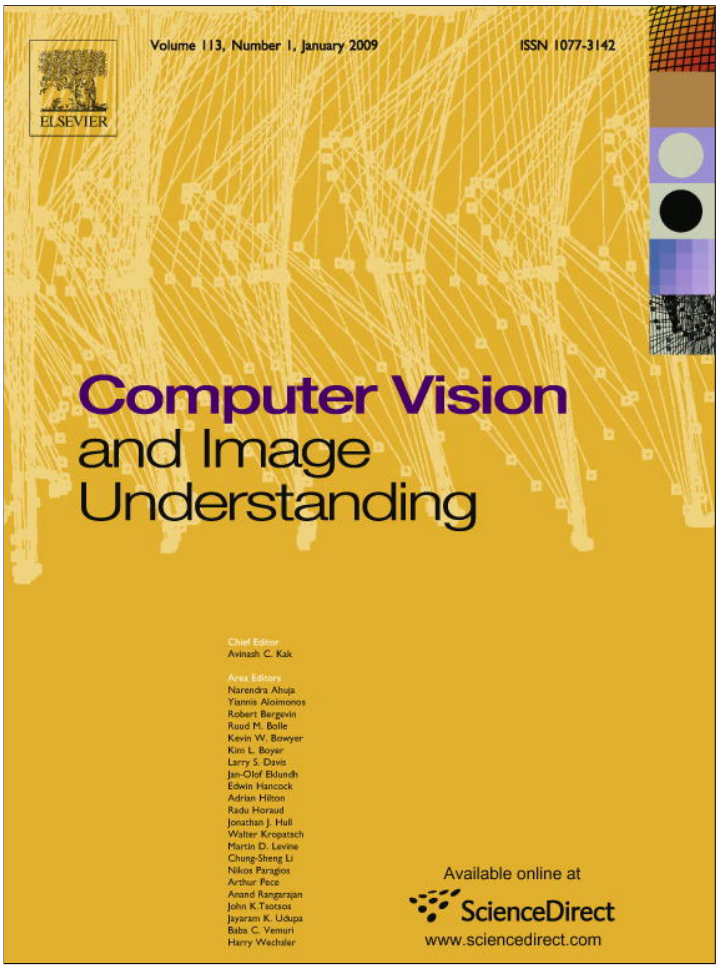


Computer Vision and Image Understanding 113 (2009) 113–125

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A pose-wise linear illumination manifold model for face recognition using video

Ognjen Arandjelović [a,*], Roberto Cipolla [b]

[a] Trinity College, University of Cambridge, Cambridge, CB2 1TQ, United Kingdom
[b] Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, United Kingdom

ARTICLE INFO

Article history:
Received 7 February 2007
Accepted 23 July 2008
Available online 8 August 2008

Keywords:
Face recognition
Manifolds
Illumination
Pose
Robustness
Invariance
Video

ABSTRACT

The objective of this work is to recognize faces using video sequences both for training and novel 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. There are three major areas of novelty: (i) illumination generalization is achieved by combining coarse histogram correction with fine illumination manifold-based normalization; (ii) pose robustness is achieved by decomposing each appearance manifold into semantic Gaussian pose clusters, comparing the corresponding clusters and fusing the results using an RBF network; (iii) a fully automatic recognition system based on the proposed method is described and extensively evaluated on 600 head motion video sequences with extreme illumination, pose and motion pattern variation. On this challenging data set our system consistently demonstrated a very high recognition rate (95% on average), significantly outperforming state-of-the-art methods from the literature.

1. Introduction

For decades, the personal identification task had shown progress by employing technological means like *secret knowledge*, such as Personal Identification Numbers, and by using *personal possessions*, such as Identity Cards and Radio Frequency Identification chips. As opposed to these means which are generally easy targets for fraud, biometric modalities like facial geometry, ear form and iris are universal and consistent over time.

Automatic face recognition (AFR) has long been established as one of the most active research areas in computer vision [1]. In spite of a 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% [2], while in less controlled environments the performance degrades even further [3]. We believe that the main reason for the apparent discrepancy between results reported in the literature and those observed in the real world is that the assumptions that most AFR methods rest upon are hard to satisfy in practice (see Section 2).

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 user to move the head in front of a mounted camera.

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.

2. Related previous work

Good general reviews of recent AFR literature can be found in [1,4,5]. In this section, we focus on AFR literature that deals specifically with recognition from image sequences, and with invariance to pose and illumination.

2.1. Recognition from multiple-image input

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 [6–8]. In contrast to these, a number of methods that do not use temporal information have been proposed. Recent ones include statistical [9,10] and principal angle-based methods with underlying simple linear [11], kernel-based [12] or Gaussian mixture-based [13] models. By their very nature, these are inherently invariant to changes in head motion pattern.

* Corresponding author.
E-mail address: oa214@eng.cam.ac.uk (O. Arandjelović).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.07.010

Other algorithms implement the "still-to-video" scenario [14,15], not taking full advantage of sequences available for training.

2.2. Recognition under varying illumination

Illumination invariance for AFR, while perhaps the most significant challenge for AFR [16] 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 [17,18]. Others rely on availability of ample training data but achieve limited generalization [9,19].

Two influential generative model-based approaches for illumination-invariant single-shot recognition are the illumination cones [20,21] and the 3D morphable model [22,23]. 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 Section 4 for data quality). Similar limitations apply to the related method of Riklin-Raviv and Shashua [24]. On the other hand, the 3D morphable model is easily extended to video-based recognition, but it requires (in our case prohibitively) high resolution [18], 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 (e.g. glasses, facial hair).

2.3. Recognition across pose

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 [22,25]. This is a view-independent representation. A second class of algorithms consists of global, parametric models, such as the eigenspace method [26] that estimates a single parametric (typically linear) subspace from all the views for all the objects (also see [27]). In AFR tests, such methods are usually outperformed by methods from the third class: view-based techniques e.g. the view-based eigenspaces [28] (also [6,7]), 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 method proposed in this paper.

3. Recognition from face motion manifolds

A video sequence of a moving face carries information about its 3D shape and texture. In terms of recognition, this information can be used either explicitly, by recovering parameters of a generative model (e.g. as in [22]), or implicitly by modelling face appearance and trying to achieve invariance to extrinsic causes of its variation (e.g. as in [17]). In this paper we employ the latter approach, as more suited for low-resolution input data [18] (see Section 4 for typical data quality).

Concepts in this paper heavily rely on the notion of *face manifolds*. Under the standard rasterized representation of an image, images of a given size can be viewed as points in a Euclidean *image space* $\mathbb{R}^D$, its dimensionality D being equal to the number of pixels. However, the surface and texture of a face is mostly smooth making its appearance constrained and confining it to an embedded face manifold of dimension $d \ll D$ [9,29], as conceptually illustrated in Fig. 1.

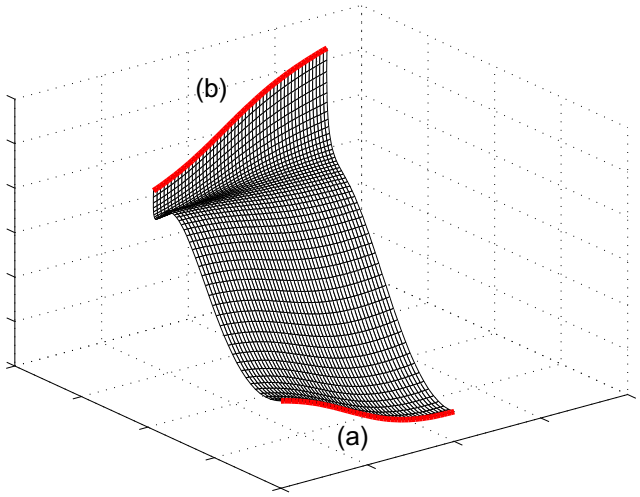


Fig. 1. Shown is a face appearance manifold, conceptually depicted as 2-dimensional, embedded in a 3-dimensional principal component space. In this paper we explicitly separate motion-affected appearance changes which give rise to (a,b) *face motion manifolds (FMMs)*, shown as 1-dimensional in red, and illumination-affected appearance changes which in turn define *face illumination manifolds*, shown in grey as connecting two FMMs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the proposed method, face manifold [9,29] are modelled using at most three Gaussian pose clusters describing small face motion around different head poses. Given two such manifolds, first (i) the pose clusters are determined, then (ii) those corresponding in pose are compared and finally, (iii) the results of pairwise cluster comparisons are combined to give a unified measure of similarity of the manifolds themselves. Each of the steps, aimed at achieving robustness to a specific set of nuisance parameters, is described in detail next.

3.1. Face registration

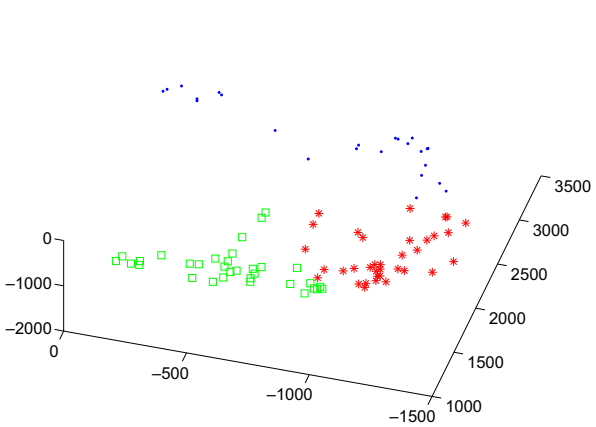
It can be observed that the corresponding variations due to head motion, i.e. pose changes, are highly nonlinear, see Fig. 2a and b. A part of the difficulty of recognition from appearance manifolds is then contained in the problem of what is an appropriate way of representing them, in a way suitable for the analysis of the effects of varying illumination or pose.

In the proposed method, face motion manifolds are represented in piece-wise linear manner by a set of semantic Gaussian *pose clusters*, see Fig. 2b and c. Seeing that each cluster describes a locally linear mode of variation, this approach to modelling manifolds becomes increasingly difficult as their intrinsic dimensionality is increased. Therefore, it is advantageous to normalize the raw, input frames as much as possible so as to minimize this dimensionality. In this first step of our method, this is done by *registering* faces i.e. by warping them to have a set of salient facial features aligned (for related approaches see [17,30]).

We compute warps that align each face with a canonical frame using four point correspondences: the locations of pupils (2) and nostrils (2). These are detected using a two-stage feature detector of Fukui and Yamaguchi [31][1]. Briefly, in the first stage, shape matching is used to rapidly remove a large number of locations in

[1] We thank the authors for kindly providing us with the original code of their algorithm.

(a) Input video sequence



(b) Face Motion Manifold

(c) Clusters

Fig. 2. A typical input video sequence of random head motion performed by the user (a) and the corresponding face motion manifold (b). Shown is the projection of affine-registered data (see Section 3.1) to the first three linear principal components. Note that while highly nonlinear, the manifold is continuous and smooth. Different poses are marked in different styles (red stars, blue dots and green squares). Examples of faces from the three clusters can be seen in (b) (also affine-registered and cropped). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

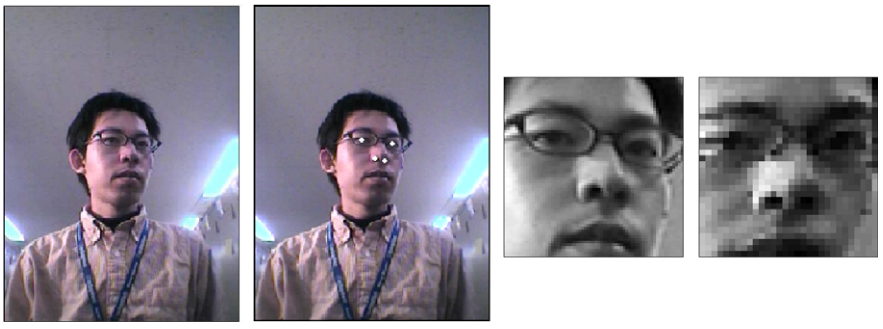
the input image that do not contain features of interest. Out of the remaining, 'promising' features, true locations are chosen using the appearance-based, distance from feature space criterion. We found that the described method reliably detected pupils and nostrils across a wide variation in illumination conditions and pose.

From the four point correspondences between the locations of the facial features and their canonical locations (we chose canonical locations to be the mean values of true feature locations) we compute optimal affine warps on a per-frame basis. Since four correspondences over-determine the affine transformation parameters (eight equations with six unknown parameters), we estimate them in the minimum L_2 error sense. Finally, the resulting images are cropped, so as to remove background clutter, and resized to the uniform scale of 30×30 pixels. An example of a face registered and cropped in the described manner is shown in Fig. 3 (also see Fig. 2c).

3.2. Pose-invariant recognition

Achieving invariance to varying pose is one of the most challenging aspects of face recognition and yet a prerequisite condition for most practical applications. This problem is complicated further by variations in illumination conditions, which inevitably occur due to movement of the user relative to the light sources.

We propose to handle changing pose in two, complementary stages: (i) in the first stage an appearance manifold is *decomposed* to Gaussian pose clusters, effectively reducing the problem to recognition under a small variation in pose parameters; (ii) in the second stage, fixed-pose recognition results are *fused* using a neural network, trained offline. The former stage is addressed next, while the latter is the topic of Section 3.4.1.



(a) Original (b) Detections (c) Cropped (d) Registered

Fig. 3. Feature-based face localization and registration: (a) original input frame (resolution 320×240 pixels), (b) superimposed detections of the two pupils and nostrils (as white circles), (c) cropped face regions with background clutter removed, and (d) the final affine registered and cropped image of the face (30×30 pixels).

3.2.1. Defining pose clusters

Inspection of manifolds of registered faces in random motion around the fronto-parallel face shows that they are dominated by the first nonlinear principal component. This principal component corresponds to lateral head rotation, i.e. changes in the face yaw, see Fig. 2a and b. The reason for this lies in the greater smoothness of the the face surface in the vertical than in the horizontal direction—pitch changes ("nodding") are largely compensated for by using the affine registration described in Section 3.1. This is not the case with significant yaw changes, when self-occlusion occurs.

Therefore, the centres of Gaussian clusters used to linearize an appearance manifold correspond to different yaw angle values. In this work we describe manifolds using three Gaussian clusters, corresponding to the frontal face orientation, face left and face right, see Fig. 2a–c. The choice of the number of clusters was determined by fitting a Gaussian Mixture Model to a small number of training sequences, manually selected to ensure that each contains the full range of head motion modelled, and examining the optimal number of components as determined using the Minimum Description Length criterion. This procedure is very much like in [9].

3.2.2. Finding pose clusters

As the extent of lateral rotation, as well as the number of frames corresponding to each cluster, can vary between video sequences, a generic clustering algorithm, such as the k-means algorithm, is unsuitable for finding the three Gaussians.

With prior knowledge of the semantics of clusters, we decide on a single face image membership on a frame-by-frame basis. We show that this can be done in a very simple and rapid manner from already detected locations of the four characteristic facial features: the pupils and nostrils, see Section 3.1.

The proposed method relies on motion parallax based on inherent properties of the shape of faces. Consider the anatomy of a human head shown in profile view in Fig. 4a. It can be seen that the nostrils are further away than the pupils from the vertical axis defined by the neck. Hence, assuming no head roll takes place, as the head rotates laterally, nostrils travel a longer projected path in the image. In other words, the midpoint between the nostrils in the image drifts relative to the midpoint of the pupils in the direction of head rotation. Using this observation, we define the quantity η as follows:

$$\eta = x_e^c - x_n^c \tag{1}$$

where x_e^c and x_n^c are the mid-points between, respectively, the eyes and the nostrils:

$$x_e^c = \frac{x_{e1} + x_{e2}}{2} \quad x_n^c = \frac{x_{n1} + x_{n2}}{2}. \tag{2}$$

It can now be understood that η approximates the discrepancy between distances travelled by the mid-points between the eyes and nostrils, measured from the frontal face orientation. Finally, we normalize η by dividing it by the distance between the eyes, to obtain $\hat{\eta}$, a scale-invariant parallax measure:

$$\hat{\eta} = \frac{\eta}{\|x_{e1} - x_{e2}\|} = \frac{x_e^c - x_n^c}{\|x_{e1} - x_{e2}\|} \tag{3}$$

3.2.3. Learning the parallax model

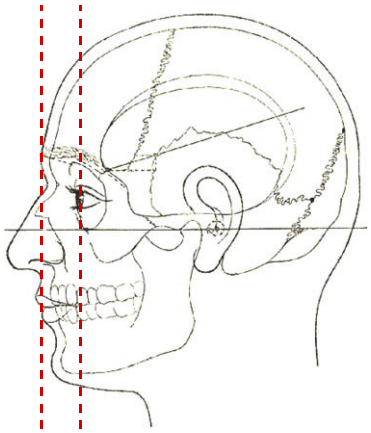
In our method, discrete poses used for linearizing appearance manifolds are automatically learnt from a small training corpus of video sequences of faces in random motion. To learn the model, we took 20 sequences of 100 frames each, acquired at 10 fps, and computed the value of $\hat{\eta}$ for each registered face. We then applied the k-means clustering algorithm [32] on the obtained set of parallax measure values and fitted a 1D Gaussian to each, see Fig. 4b.

To apply the learnt model, a frame in our method is classified to the maximal likelihood pose. In other words, when a novel face is to be classified to one of the three pose clusters (i.e. head poses), we evaluate pose likelihood given each of the learnt distributions and classify it to the one giving the highest probability of the observation. Fig. 5 shows the proportions of faces belonging to each pose cluster.

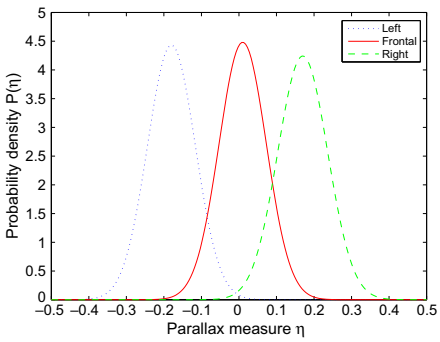
3.3. Illumination-invariant recognition

Illumination variation of face patterns is extremely complex due to varying surface reflectance properties, face shape, and type and distance of lighting sources. Hence, in such a general setup, this is a difficult problem to approach in a purely discriminative fashion.

Our method for compensating for illumination changes is based on the observation that on a coarse level most of the variation can be described by the *dominant light* direction e.g. 'strong light from the left'. Such variations are addressed much more easily. We will also demonstrate that it is the case that *once normalized* at this, coarse level, the learning of residual illumination changes is signif-



(a)Parallax



(b) Parallax measure distributions

Fig. 4. (a) A schematic illustration of the motion parallax used for coarse pose clustering of input faces (the diagram is based on a figure taken from [33]). (b) The distributions of the scale-normalized parallax measure $\hat{\eta}$ defined in (3) for the three pose clusters on the offline training data set. Good separation is demonstrated.

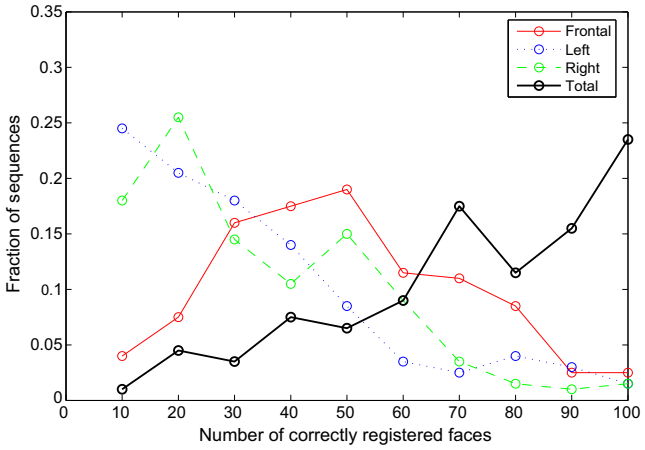


Fig. 5. Histograms of the number of correctly registered faces using four point correspondences between detected facial features (pupils and nostrils) for each of the three discrete poses and in total for each sequence.

icantly simplified as well. This motivates the two-stage, per-pose illumination normalization employed in the proposed method:

(1) *Coarse level:* Region-based gamma intensity correction (GIC), followed by
(2) *Fine level:* Illumination subspace normalization.

The algorithm is summarized in Fig. 6 while its details are explained in the sections that follow.

3.3.1. Gamma intensity correction

Gamma Intensity Correction (GIC) is a well-known image intensity histogram transformation technique that is used to compensate for global brightness changes [34]. It transforms pixel values (normalized to lie in the range [0.0, 1.0]) by exponentiation so as to best match a *canonically illuminated image*. This form of the operator is motivated by non-linear exposure-image intensity response of the photographic film that it approximates well over a wide range of exposure. Formally, given an image I and a canonically illuminated image I_C, the gamma intensity corrected image I^* is defined as follows:

$$I^*(x,y) = I(x,y)^{\gamma^*}, \tag{4}$$

where γ^* is the optimal gamma value and is computed using

$$\gamma^* = \arg\min_{\gamma} \|\mathbf{I}^{\gamma} - \mathbf{I}_C\| \tag{5}$$

$$= \arg\min_{\gamma} \sum_{x,y} [I(x,y)^{\gamma} - I_C(x,y)]^2. \tag{6}$$

This is a nonlinear optimization problem in 1D. In our implementation of the proposed method it is solved using the Golden Section search with parabolic interpolation, see [35] for details.

3.3.2. Region-based gamma intensity correction

Gamma intensity correction can be used across a wide range of types of input to correct for *global* brightness changes. However, in the case of objects with a highly variable surface normal, such as faces, it is unable to correct for the effects of side lighting. This is recognized as one of the most difficult problems in AFR [16].

Input: pose clusters $\mathcal{C}_1 = \{\mathbf{x}_i^{(1)}\}, \mathcal{C}_2 = \{\mathbf{x}_i^{(2)}\}$,
face regions mask $\mathbf{r}$,
mean face (for pose) $\mathbf{m}$,
pose illumination subspace basis matrix $\mathbf{B}_I$.

Output: pose cluster $\hat{\mathcal{C}}_1$ normalized to $\mathcal{C}_2$.

---

(1) **Coarse normalization:** *Per-frame region-based GIC*

$$\forall i.\ \mathbf{x}_i^{(1)} = \text{region\_GIC}(\mathbf{r}, \mathbf{m}, \mathbf{x}_i^{(1)}),$$

(2) **Coarse normalization:** *Per-frame region-based GIC*

$$\forall i.\ \mathbf{x}_i^{(2)} = \text{region\_GIC}(\mathbf{r}, \mathbf{m}, \mathbf{x}_i^{(2)})$$

(3) **Fine normalization:** *Per-frame illumination subspace compensation*

$$\forall i.\ \hat{\mathbf{x}}_i^{(1)} = \mathbf{B}_I \mathbf{a}_i^* + \mathbf{x}_i^{(1)}$$
$$\text{where } \mathbf{a}_i^* = \arg\min_{\mathbf{a}_i} D_{MAH}\left[\mathbf{B}_I \mathbf{a}_i + \mathbf{x}_i^{(1)} - \langle \mathcal{C}_2 \rangle;\ \mathcal{C}_2\right]$$

(4) **Normalized cluster:** *The result is cluster $\mathcal{C}_1$ normalized to $\mathcal{C}_2$*

$$\hat{\mathcal{C}}_1 = \{\hat{\mathbf{x}}_i^{(1)}\}$$

Fig. 6. Illumination compensation overview: coarse appearance changes due to illumination variation are normalized using region-based gamma intensity correction, while the residual variation is modelled using a linear, pose-specific illumination subspace, learnt offline. Local manifold shape is employed as a constraint in the second, 'fine' stage of normalization in the form of the Mahalanobis distance for the computation of the optimal additive illumination subspace component.

Region-based GIC proposes to overcome this problem by dividing the image (and hence implicitly the imaged object/face as well) into regions corresponding to surfaces with a near-constant surface normal. Regular gamma intensity correction is then applied to each region separately, see Fig. 7.

An undesirable result of this method is that it tends to produce artificial intensity discontinuities at region boundaries [36]. This occurs due to discontinuities in the computed gamma values between neighbouring regions. We propose to first Gaussian-blur the obtained gamma value map image $\mathbf{\Gamma}^*$:

$$\mathbf{\Gamma}^*_S = \mathbf{\Gamma}^* * \mathbf{G}_{\sigma=2}, \tag{7}$$

before applying it to an input image to give the final, region-based gamma corrected output $\mathbf{I}^*_S$:

$$\forall x, y . I^*_S(x,y) = I(x,y)^{\Gamma^*_S(x,y)} \tag{8}$$

This method almost entirely remedies the problem with boundary artefacts, as illustrated in Fig. 7. Note that because smoothing is performed on the gamma map, not the processed image, the artefacts are removed without any loss of discriminative, high frequency detail, see Fig. 8.

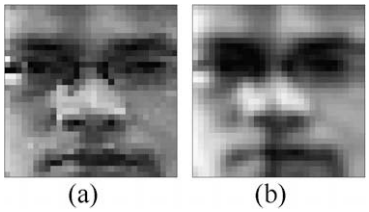


Fig. 8. (a) Seamless output of the proposed smooth region-based GIC. Boundary artefacts are removed without blurring of the image. Contrast this with the output of the original region-based GIC, after Gaussian smoothing of the output (b). Image quality is significantly reduced, with boundary edges still clearly visible.

3.3.3. Pose-specific illumination subspace normalization

After region-based GIC is applied to all images, it is assumed that the lighting variation for each of the pose clusters can be modelled using a linear, *pose-specific illumination subspace*. Given a reference and a novel cluster corresponding to the same pose, each frame of the novel cluster is normalized for the illumination change. This is done by adding a vector from the pose illumination subspace to the frame so that its distance from the reference cluster's centre is minimized.

3.3.4. Learning the model

We define a pose-specific illumination subspace to be a linear manifold that explains *intra-personal* appearance variations due to illumination changes across a narrow range of poses. In other words, this is the principal subspace of the within-class scatter.

Formalizing the definition above, given that $\mathbf{x}^k_{i,j}$ is the k-th of $N_f(i,j)$ frames of person i under the illumination j (out of $N_l(i)$), the within-class scatter matrix is:

$$\mathbf{S}_B = \sum_{i=1}^{N_p} \sum_{j=1}^{N_l(i)} \sum_{k=1}^{N_f(i,j)} (\mathbf{x}^k_{i,j} - \bar{\mathbf{x}}_i)(\mathbf{x}^k_{i,j} - \bar{\mathbf{x}}_i)^{\mathrm{T}}, \tag{9}$$

where N_p is the total number of training individuals and $\bar{\mathbf{x}}_i$ is the mean face of the person in the range of considered poses:

$$\bar{\mathbf{x}}_i = \frac{\sum_{j=1}^{N_l(i)} \sum_{k=1}^{N_f(i,j)} \mathbf{x}^k_{i,j}}{\sum_j N_f(i,j)}. \tag{10}$$

The pose-specific illumination subspace basis $\mathbf{B}_I$ is then computed by eigen decomposition of $\mathbf{S}_B$ as the principal subspace explaining 90% of data energy variation.

For offline learning of illumination subspaces we used 10 s video sequences of 20 individuals, each in five illumination conditions, acquired at 10 fps. The first few basis vectors learnt in the described manner are shown as images in Fig. 9.

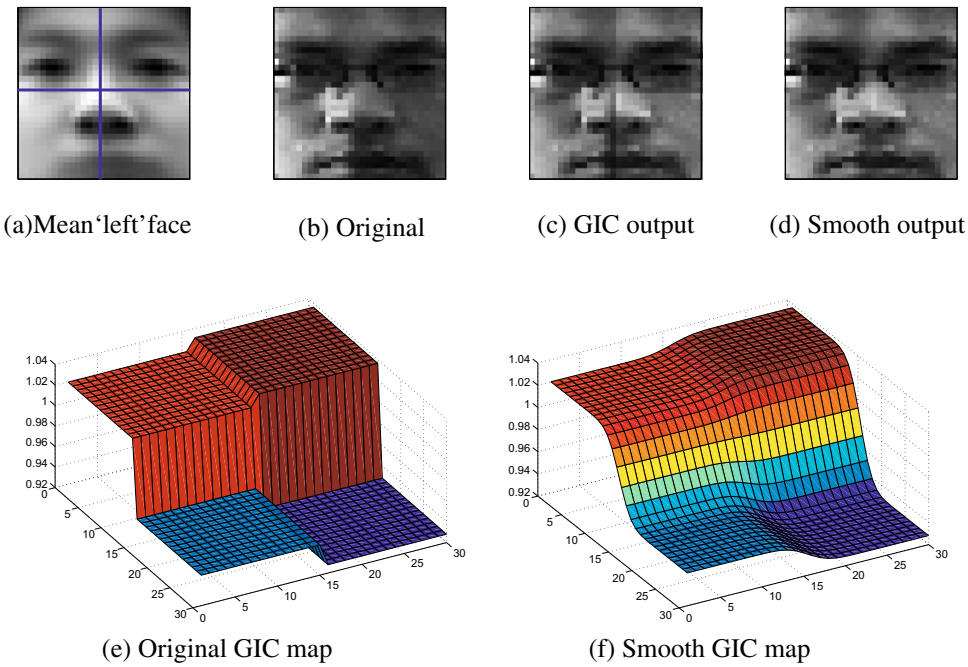


Fig. 7. Coarse illumination normalization: canonical illumination image and the regions used in region-based GIC (a), original unprocessed face image (b), region-based GIC corrected image without smoothing (c), region-based GIC corrected image with smoothing (d), original gamma value map (e) and smoothed gamma value map (f). Notice artefacts at region boundaries in the gamma corrected image (c). The output of the proposed smooth region-based GIC in (d) does not have the same problem. Finally, note that the coarse effects of the strong side lighting in (b) have been greatly removed.

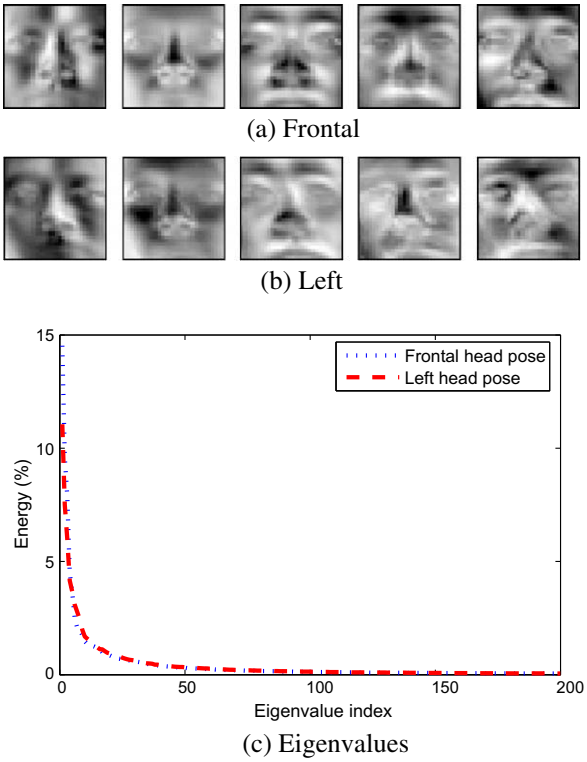


Fig. 9. Shown as images are the first 5 bases of pose-specific illumination subspaces for the (a) frontal and (b) left head orientations. The distribution of energy for pose-specific illumination variation across principal directions is shown in (c).

3.3.5. Employing the model

Let $\mathscr{C}_1 = \{\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{N_1}^{(1)}\}$ and $\mathscr{C}_2 = \{\mathbf{x}_1^{(2)} \ldots, \mathbf{x}_{N_2}^{(2)}\}$ be two corresponding pose clusters of different appearance manifolds, previously preprocessed using the region-based gamma correction algorithm described in Section 3.3.1. Cluster $\mathscr{C}_1$ is then illumination-normalized with respect to $\mathscr{C}_2$ (we will therefore refer to $\mathscr{C}_2$ as the *reference cluster*), under the null assumption that the identities of the two people they represent are the same. The normalization is performed on a frame-by-frame basis, by adding a vector $\mathbf{B}_I\mathbf{a}_i^*$ from the estimated pose-specific illumination subspace:

$$\forall i. \hat{\mathbf{x}}_i^{(1)} = \mathbf{B}_I\mathbf{a}_i^* + \mathbf{x}_i^{(1)} \tag{11}$$

where we define $\mathbf{a}_i^*$ as:

$$\mathbf{a}_i^* = \arg\min_{\mathbf{a}_i} \|\mathbf{B}_I\mathbf{a}_i + \mathbf{x}_i^{(1)} - \langle\mathscr{C}_2\rangle\|, \tag{12}$$

and $|\cdots|$ is a vector norm and $\langle\mathscr{C}_2\rangle$ the mean face of cluster $\mathscr{C}_2$. We then define cluster $\mathscr{C}_1$ normalized to $\mathscr{C}_2$ to be $\hat{\mathscr{C}}_1 = \{\hat{\mathbf{x}}_1^{(1)} \ldots \hat{\mathbf{x}}_{N_1}^{(1)}\}$. This form is directly motivated by the definition of a pose-specific subspace.

To understand the next step, which is the choice of the vector norm in (12), it is important to notice in the definition of the pose-specific illumination subspace, that the basis $\mathbf{B}_I$ explains not only appearance variations caused by illumination: reflectance properties of faces used in training (e.g. their albedos), as well as subjects' pose changes also affect it. This is especially the case as we do not make the common assumption that surfaces of faces are Lambertian, or that light sources are point lights at infinity.

The significance of this observation is that the subspace of a dimensionality sufficiently high to explain the modelled phenomenon (illumination changes) will, undesirably, also be able to explain 'distracting' phenomena, such as differing identity. The problem is therefore that of *constraining* the region of interest in the subspace to that which is most likely to be due to illumination changes for a particular individual. For this purpose we propose to exploit the local structure of appearance manifolds, which are smooth. We do this by employing the Mahalanobis distance (using the probability density corresponding to the reference cluster) when computing the illumination subspace correction for each novel frame using (12). Formally:

$$\mathbf{a}_i^* = \arg\min_{\mathbf{a}_i}(\mathbf{B}_I\mathbf{a}_i + \mathbf{x}_i^{(1)} - \langle\mathscr{C}_2\rangle)^{\mathrm{T}}\mathbf{B}_2\mathbf{\Lambda}_2^{-1}\mathbf{B}_2^{\mathrm{T}}(\mathbf{B}_I\mathbf{a}_i + \mathbf{x}_i^{(1)} - \langle\mathscr{C}_2\rangle), \tag{13}$$

where $\mathbf{B}_2$ and $\mathbf{\Lambda}_2$ are, respectively, reference cluster's orthonormal basis and the diagonalized covariance matrix. We found that the use of the Mahalanobis distance, as opposed to the usual Euclidean distance, achieved better explanation of novel images when the person's identity was the same, and worse when it was different, achieving better inter-to-intra class separation.

This quadratic minimization problem is solved by differentiation and the minimum is achieved for:

$$\mathbf{a}_i^* = \left(\mathbf{B}_I^{\mathrm{T}}\mathbf{B}_2\mathbf{\Lambda}_2^{-1}\mathbf{B}_2^{\mathrm{T}}\mathbf{B}_I\right)^{-1}\mathbf{B}_I^{\mathrm{T}}\mathbf{B}_2\mathbf{\Lambda}_2^{-1}\mathbf{B}_2^{\mathrm{T}}(\langle\mathscr{C}_2\rangle - \mathbf{x}) \tag{14}$$

Examples of registered and cropped face images before and after illumination normalization can be seen in Fig. 10a.

3.3.6. Practical considerations

The computation of the optimal value $\mathbf{a}^*$ using (14) involves inversion and Principal Component Analysis (PCA) on matrices of size $D \times D$, where D is the number of pixels in a face image (in our case equal to 900, see Section 3.1). Both of these operations put high demands on computer resources. To reduce the computational overhead, we exploit the assumption that the data modelled is of much lower dimensionality than D.

Formalizing the model of low-dimensional face manifolds, we assume that an image $\mathbf{y}$ of subject i's face is drawn from the probability density $p_F^{(i)}(\mathbf{y})$ within the face space, and embedded in the image space by means of a mapping function $f^{(i)} : \mathbb{R}^d \to \mathbb{R}^D$. The resulting point in the D-dimensional space is further perturbed by noise drawn from a noise distribution p_n (note that the noise operates in the image space) to form the observed image $\mathbf{x}$. Therefore the distribution of the observed face images of the subject i is given by the integral:

$$p^{(i)}(\mathbf{x}) = \int p_F^{(i)}(\mathbf{y})p_n(f_i(\mathbf{y}) - \mathbf{x})\mathrm{d}\mathbf{y} \tag{15}$$

This model is then used in two stages:

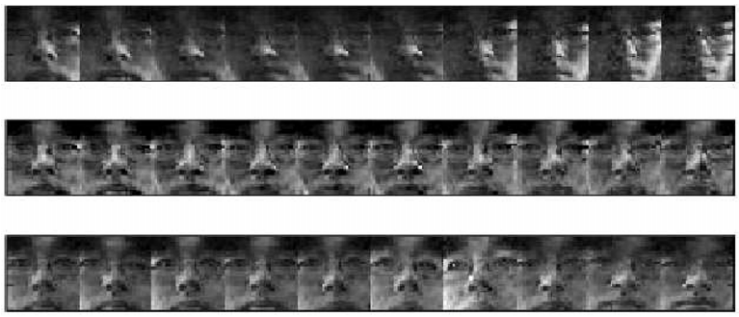
(1) Pose-specific PCA dimensionality reduction.
(2) Exact computation of the linear principal subspace and rapid estimation of the complementary subspace of a pose cluster.

Specifically, we first perform a linear projection of all images in a specific pose cluster to a *pose-specific face subspace* that explains 95% of data variation in a specific pose. This achieves data dimensionality reduction from 900 to 250.

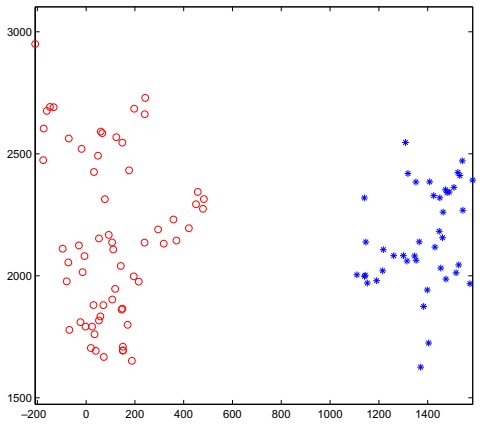
Referring back to (15), to additionally speed up the process, we estimate the intrinsic dimensionality of face manifolds (defined as explaining 95% of within-cluster data variability) and assume that all other variation is due to isotropic Gaussian noise p_n. Hence, we can write the basis of the PCA subspace corresponding to the reference cluster as consisting of principal and complementary subspaces [37] represented by orthonormal basis matrices, respectively $\mathbf{V}_P$ and $\mathbf{V}_C$:

$$\mathbf{B}_2 = [\mathbf{V}_P\mathbf{V}_C] \tag{16}$$

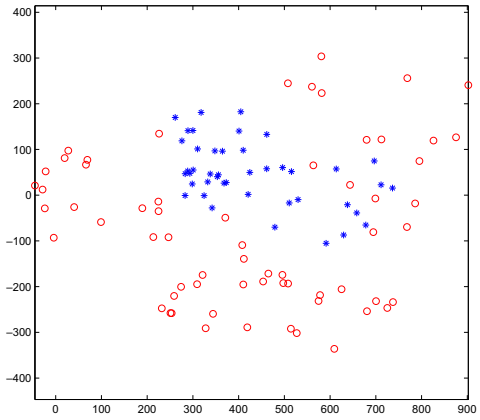
where $\mathbf{V}_P \in \mathbb{R}^{250\times6}$ and $\mathbf{V}_C \in \mathbb{R}^{250\times244}$. The principal subspace and the associated eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_6$ are rapidly computed, e.g. using [38]. The isotropic noise covariance and the complementary subspace basis are then estimated in the following manner:



(a) Illumination normalization



(b) Original clusters



(c) Normalized clusters

Fig. 10. In (a) are respectively, top to bottom, shown the original registered and cropped face images from an input video sequence, the same faces after the proposed illumination normalization and a sample from the *reference* video sequence. The effects of strong side lighting have been greatly removed, while at the same time a high level of detail is retained. The corresponding data from the two sequences, before and after illumination compensation are shown under (b) and (c). Shown are their projections to the first two principal components. Notice that initially the clusters were completely non-overlapping. Illumination normalization has adjusted the location of the centre of the blue cluster, but has also affected its spread. After normalization, while overlapping, the two sets of patterns are still distributed quite differently. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

$$\lambda_n = \omega \sum_{i=1}^{6} \lambda_i \quad \mathbf{V}_C = \text{null}(\mathbf{V}_P) \tag{17}$$

where the nullspace of the principal subspace is computed using QR-decomposition [35], while the value of ω is estimated from a small training corpus; we obtained $\omega \approx 2.2\text{e}{-4}$. The diagonalized covariance matrix is then simply:

$$\mathbf{\Lambda}_2 = \text{diag}(\lambda_1, \ldots, \lambda_6, \overbrace{\lambda_n, \ldots, \lambda_n}^{244}) \tag{18}$$

3.4. Comparing normalized pose clusters

Having illumination normalized one face cluster to match another, we want to compute a similarity measure between them, a *distance*, expressing our degree of belief that they belong to the same person.

At this point it is instructive to examine the effects of the described method for illumination normalization on the face patterns. Two clusters before and after one has been normalized, are shown in Fig. 10b and c. An interesting artefact can be observed: the spread of the normalized cluster is significantly reduced. This is easily understood by referring back to (11) and (12) and noticing that the normalization is performed frame-by-frame, trying to make each normalized face as close as possible to the reference cluster's mean, i.e. a *single point*. For this reason, dissimilarity measures between probability densities common in the literature, such as the Bhattacharyya distance, the Kullback-Leibler divergence [9,10] or the Resistor-Average distance [39], are not suitable choices. Instead, we propose to use the simple Euclidean distance between normalized cluster centres:

$$D(\mathscr{C}_1, \mathscr{C}_2) = \frac{\sum_{i=1}^{N_1} \hat{\mathbf{x}}_i^{(1)}}{N_1} - \frac{\sum_{j=1}^{N_2} \mathbf{x}_j^{(2)}}{N_2}. \tag{19}$$

3.4.1. Inter-manifold distance

The last stage in the proposed method is the computation of an inter-manifold distance, or an inter-manifold dissimilarity measure, based on the distances between corresponding pose clusters. There are two main challenges in this problem: (i) depending on the poses assumed by the subjects, one or more clusters, and hence the corresponding distances, may be void; (ii) different poses are not equally important, or discriminative, in terms of face recognition [40].

Writing $\mathbf{d}$ for the vector containing the three pose cluster distances, we want to classify a novel appearance manifold to the gallery class giving the highest probability of corresponding to it in identity, $P(s|\mathbf{d})$. Then, using Bayes' theorem:

$$P(s|\mathbf{d}) = \frac{p(\mathbf{d}|s)P(s)}{p(\mathbf{d})} = \frac{p(\mathbf{d}|s)P(s)}{p(\mathbf{d}|s)P(s) + p(\mathbf{d}|\neg s)P(\neg s)} \tag{20}$$

$$= \frac{1}{1 + p(\mathbf{d}|\neg s)P(\neg s)/p(\mathbf{d}|s)P(s)} \tag{21}$$

Assuming that the ratio of same-identity to differing-identities priors $P(\neg s)/P(s)$ is constant across individuals, it is clear that classifying to the class with the highest $P(s|\mathbf{d})$ is equivalent to classifying to the class with the highest *likelihood ratio*:

$$\mu(\mathbf{d}) = \frac{p(\mathbf{d}|s)}{p(\mathbf{d}|\neg s)} \tag{22}$$

3.4.2. Learning pose likelihood ratios

Writing $\mathbf{d} = [D_1, D_2, D_3]^T$, we assume statistical independence between pose cluster distances:

$$p(\mathbf{d}|s) = \prod_{i=1}^{3} p(D_i|s) \quad p(\mathbf{d}|\neg s) = \prod_{i=1}^{3} p(D_i|\neg s) \tag{23}$$

We propose to learn likelihood ratios $\mu(D_i) = p(\mathbf{d}|s)/p(\mathbf{d}|\neg s)$ offline, from a small data corpus labelled by the identity, in two stages. First, (i) we obtain a Parzen window estimate of intra- and inter-personal pose distances by comparing all pairs of training appearance manifolds; then (ii) we refine the estimates using a Radial Basis Functions (RBF) artificial neural network trained for each pose.

A Parzen window-based [32] estimate of $\mu(D)$ for the frontal head orientation, obtained by directly comparing appearance manifolds as described in Sections 3.1–3.4 is shown in Fig. 11a. In the proposed method, this, and the similar likelihood ratio estimates for the other two head poses are not used directly for recognition as they suffer from an important limitation: the estimates are ill-defined in domain regions sparsely populated with training data. Specifically, an artefact caused by this problem can be observed by noting that the likelihood ratios are not monotonically decreasing. What this means is that *more distant* pose clusters can result in a higher chance of classifying two sequences as originating from the *same individual*.

To overcome the problem of insufficient training data, we train a two-layer RBF-based neural network for each of the discrete poses used in approximating face motion manifolds, see Fig. 11c. In its basic form, this means that the estimate $\hat{\mu}(D_i)$ is given by the following expression:

$$\hat{\mu}(D_i) = \sum_j \alpha_j \mathcal{G}(D_i; \boldsymbol{\mu}_j, \sigma_j) \tag{24}$$

where

$$\mathcal{G}(D_i; \boldsymbol{\mu}_j, \sigma_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(D_i - \boldsymbol{\mu}_j)^2}{2\sigma^2} \tag{25}$$

In the proposed method, this is modified so as to enforce prior knowledge on the functional form of $\mu(D_i)$ in the form of its monotonicity:

$$\hat{\mu}^*(D_i) = \max_{\delta > D_i} \left\{ \sum_j \alpha_j \mathcal{G}(D_i; \boldsymbol{\mu}_j, \sigma_j), \hat{\mu}(\delta) \right\} \tag{26}$$

Finally, to ensure that the networks are trained using reliable data (in the context of training sample density in the training domain), we use only local peaks of Parzen window-based estimates. Results using a network with six second-layer neurons, each with the spread of $\sigma_j = 60$, see (26), are summarized in Figs. 11 and 12.

4. Experimental evaluation

Methods in this paper were evaluated on a database of video sequences kindly provided to us by Toshiba Corporation (from here on referred to as *FaceDB60*). This database contains 60 individuals of varying age, mostly male and Japanese, and 10 sequences per person. Each sequence corresponds to a different illumination setting, acquired for 10 s at 10 fps and 320×240 pixel resolution (face size ≈60–120 pixels), see Fig. 13. Typical variations in pose and expression within a single sequence are illustrated in Fig. 14, while Fig. 15 shows different illumination conditions both within and across different sequences.[2]

To establish baseline performance, we compared our recognition algorithm to:

- Kernel Principal Angles (KPA) of Wolf and Shashua [12].[3]
- *Mutual Subspace Method* (MSM) of Fukui and Yamaguchi [11], used in a state-of-the-art commercial system FacePass® [41].
- *KL divergence-based algorithm* of Shakhnarovich et al. (KLD) [10].
- *Majority vote* across all pairs of frames using *Eigenfaces* (MVE) of Turk and Pentland [42].

In the KL divergence-based method we used principal subspaces that explain 85% of data variation energy. In MSM we set the dimensionality of linear subspaces to nine and used the first three principal angles for recognition, as suggested by the authors in [11]. For the Eigenfaces method, the 22-dimensional eigenspace used explained 90% of total training data energy. The methods were evaluated using three face representations:

- raw appearance images $\mathbf{X}$,
- Gaussian high-pass filtered images—used for face recognition in [17,43]:

$$\mathbf{X}_H = \mathbf{X} - (\mathbf{X} * \mathbf{G}_{\sigma=1.5}), \tag{27}$$

- local intensity-normalized high-pass filtered images—similar to the Self Quotient Image [44] (also see [45]):

$$\mathbf{X}_Q = \mathbf{X}_H ./ (\mathbf{X} - \mathbf{X}_H) \tag{28}$$

the division being element-wise.

Offline training, i.e. learning of the pose-specific illumination subspaces and likelihood ratios, was performed using 20 randomly chosen individuals in five illumination settings, for a total of 100 sequences. These were used for neither gallery data nor test input for the evaluation reported in this section.

Recognition performance of the proposed system was assessed by training it with the remaining 40 individuals in a single illumination setting, and using the rest of the data as test input. In all tests, both training data for each person in the gallery, as well as test data, consisted of only a single sequence.

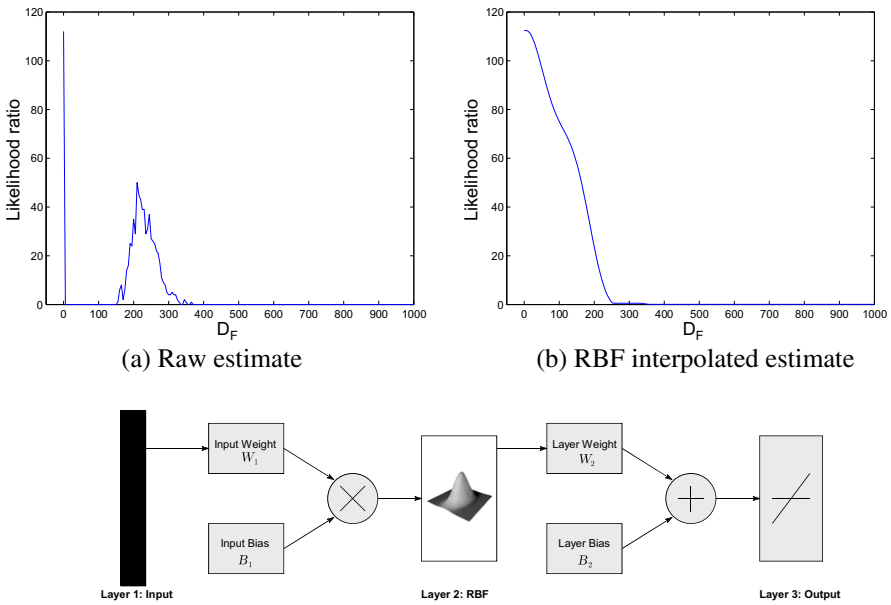
4.1. Results

The performance of the proposed method is summarized in Table 1. We tabulated the recognition rates achieved across different combinations of illuminations used for training and test input, so as to illustrate its degree of sensitivity to the particular choice of data acquisition conditions. An average rate of 95.2% was achieved, with a mean standard deviation of only 4.7%. Therefore, we conclude that the proposed method is successful in recognition across illumination, pose and motion pattern variation, with high robustness to the exact imaging setup used to provide a set of gallery videos.

This conclusion is further corroborated by Fig. 16a, which shows cumulative distributions of inter- and intra-personal manifold distances (see Section 3.4.1) and Fig. 16b which plots the Receiver-Operator Characteristic of the proposed algorithm. Good class separation can be seen in both, illustrating the suitability of our method for verification (one-against-one matching) applications: less than 0.5% false positive rate is attained for 91.2% true positive rate. Additionally, it is important to note that good separation is maintained across a wide range of distances, as can be seen in Fig. 16a from low gradients of inter- and intra-class distributions (e.g. on the interval between 1.0 and 15.0). This is significant as it implies that the interclass threshold choice is not very numerically sensitive: by choosing a threshold in the middle of this range, we can

[2] Also see urlhttp://mi.eng.cam.ac.uk/oa214/ for more information on this database and examples of video sequences.

[3] We used the original authors' implementation.



(a) Raw estimate (b) RBF interpolated estimate

(c) RBF network architecture

Fig. 11. Likelihood ratio corresponding to the frontal head pose obtained from the training corpus using Parzen windows (a) and the RBF network-based likelihood ratio (b). The corresponding RBF network architecture is shown in (c). Note that the initial estimate (a) is not monotonically decreasing, while (b) is.

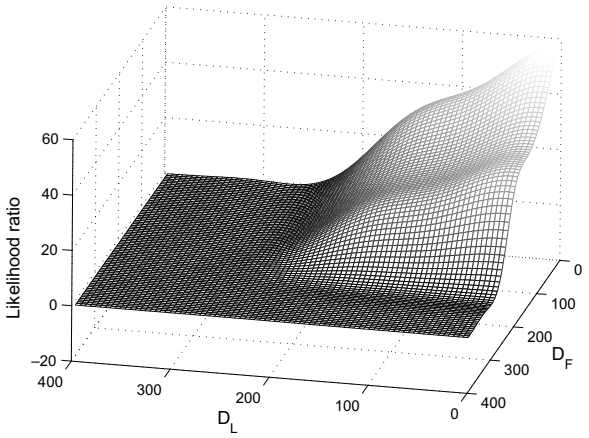
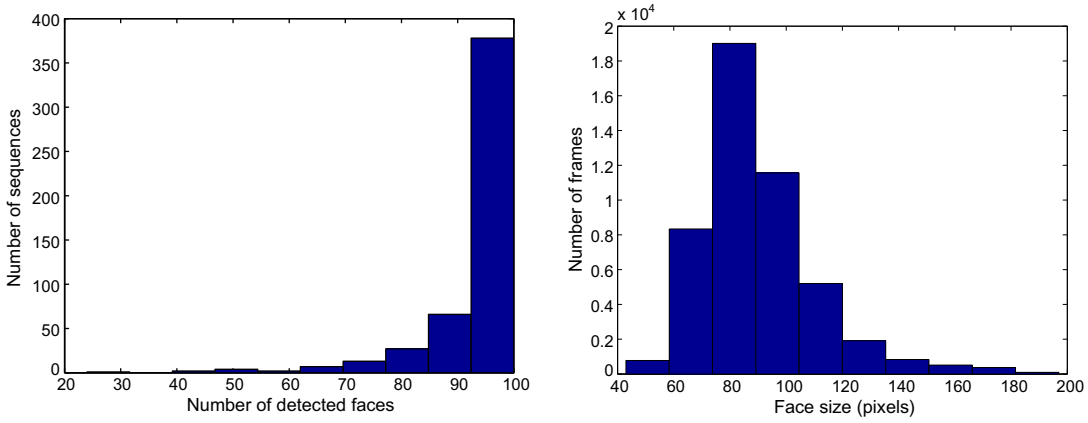


Fig. 12. Joint RBF network-based likelihood ratio for the frontal and left head orientations.

expect the recognition performance to generalize well to different data sets.

4.1.1. Pose clusters

One of the main premises that this work rests on is the idea that illumination and pose robustness in recognition can be achieved by decomposing an appearance manifold into a set of pose ranges (see Section 3.2.1) which are, after being processed independently, probabilistically combined (see Section 3.4.1). We investigated the discriminating power of each of the three pose clusters used in the proposed context by performing recognition using the inter-cluster distance defined in Section 3.4. Table 2 show a summary of the results. High recognition rates were achieved even using only a single pose cluster. Furthermore, the proposed method for integrating cluster distance into a single inter-manifold distance can be seen to improve the average performance of the most discriminative pose. In the described recognition framework, side poses contributed more discriminative information to the distance



(a) (b)

Fig. 13. *Detected faces:* Histograms of (a) the number of detected faces across sequences in the entire database FaceDB60 (10 sequences for each of the 60 individuals in the database) and (b) the detected face sizes (assumed square).

Fig. 14. *Input data:* Frames from a typical input video sequence used for evaluation of methods in this paper. Notice the presence of cast shadows and overall extreme imaging conditions: pose, illumination and even occlusion, in the form of facial wear (glasses) and hands. The size of the face area is also greatly variable.

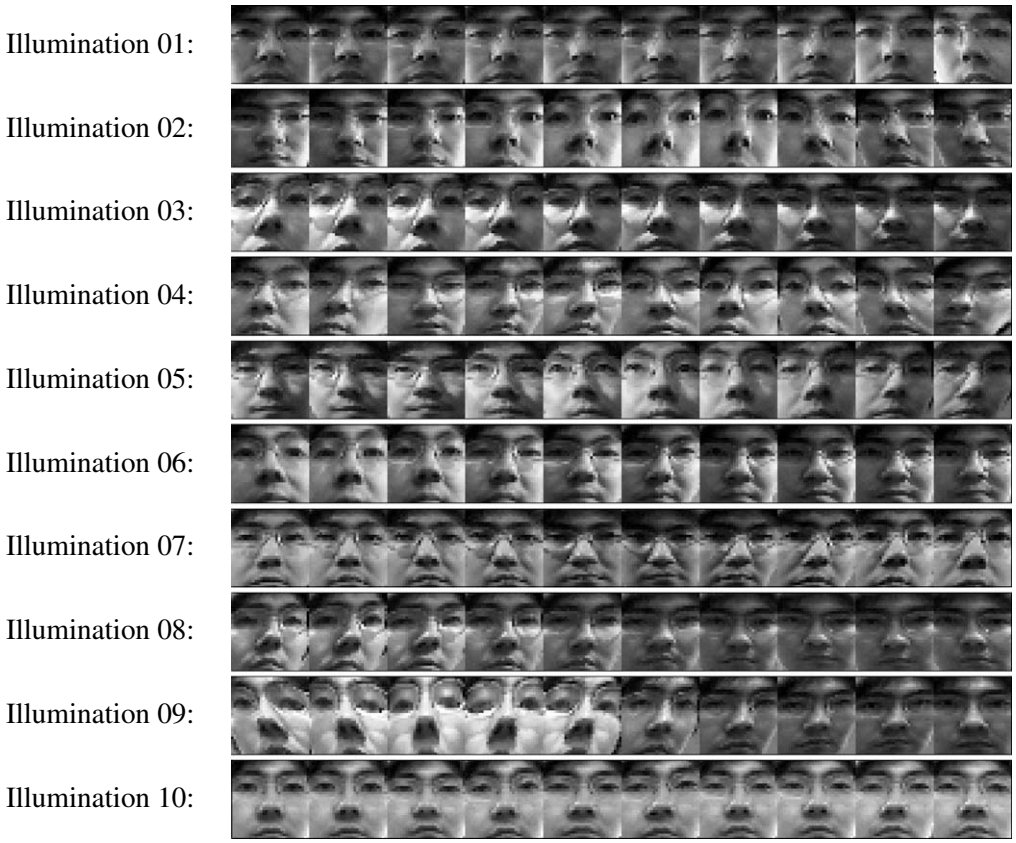


Fig. 15. Registered and automatically cropped faces (30×30 pixels) from typical sequences used for the comparison of recognition methods in this paper. All frames are of the same person, in frontal pose, each row corresponding to one of 10 different illumination conditions used for the evaluation. Cast shadows and specularities are common. Notice extreme illumination changes both between and *within* sequences.

Table 1
Recognition performance (%) of the proposed method using different illuminations for training and test input

	IL. 1	IL. 2	IL. 3	IL. 4	IL. 5	Mean	std
IL. 1	100	90	95	95	90	94	4.2
IL. 2	95	95	95	95	90	94	2.2
IL. 3	95	95	100	95	100	97	2.7
IL. 4	95	90	100	100	95	96	4.2
IL. 5	100	80	100	95	100	95	8.7
Mean	97	90	98	96	95	95.2	4.5

Excellent results are demonstrated with little dependence of the recognition rate on the data acquisition conditions.

than the frontal pose (in spite of a lower average number of side faces per sequence, see Fig. 5 in Section 3.1), as witnessed by both a higher average recognition accuracy and lower standard deviation of recognition. It is interesting to observe that this is in agreement with the finding that appearance in a roughly semi-profile head pose is inherently most discriminative for AFR [40].

4.1.2. Other algorithms

The result of the comparison with the other evaluated methods is shown in Table 3. The proposed algorithm outperformed others by a significant margin when raw data was used. Kernel Principal Angles, majority vote using Eigenfaces and the KL divergence algorithm performed with statistically insignificant difference, while MSM showed least robustness to the extreme changes in illumination conditions. It is interesting to note that all three algorithms achieved perfect recognition when training and test sequences were acquired in same illumination conditions. Thus, the observed performance improvement with both the high-pass and even further Self Quotient Image representations is unsurprising. This also highlights the need for an explicit illumination model, even if ample motion data is available. Furthermore, the consistently superior performance of the KPA in comparison to the MVE, KLD and MSM, as well as the overall best performance of the proposed algorithm, supports another aspect of our work which concerns the amodelling of pose-affected nonlinearities in face manifolds. Overall, our combination of illumination and pose models outperformed the best competing method/

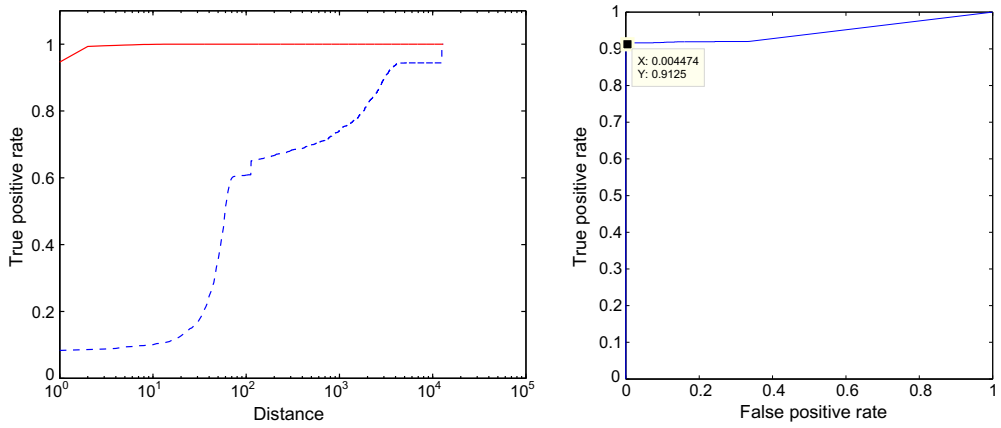


Fig. 16. Cumulative distributions of intra-personal (dashed line) and inter-personal (solid line) distances (a). Good separability is demonstrated. The corresponding ROC curve can be seen in (b)—less than 0.5% of false positive rate is attained for 91% true positive rate. The corresponding distance threshold choice is numerically well-conditioned, as witnessed by close-to-zero derivatives of the plots in (a) at the corresponding point.

Table 2
A comparison of identification statistics for recognition using each of the pose-specific cluster distances separately and the proposed method for combining them using an RBF-based neural network

Measure	Manifold distance	Front clusters distance	Side clusters distance
Mean	95	90	93
std	4.7	5.7	3.6

In addition to the expected performance improvement when using all over only some poses, it is interesting to note different contributions of side and frontal pose clusters, the former being more discriminative in the context of the proposed method.

Table 3
Average recognition rates (%) of the compared methods across different illumination conditions used for training and test

Representation	Measure	Proposed	KPA	MVE	KLD	MSM
$\mathbf{X}$	Mean	**95**	49	43	39	24
	std	**4.7**	25.0	31.9	32.5	38.9
$\mathbf{X}_H$	mean	–	61	43	48	53
	std	–	18.9	31.9	21.1	19.1
$\mathbf{X}_Q$	mean	–	**88**	43	57	79
	std	–	**11.2**	31.9	19.5	12.7

The performance of the proposed method is by far the best, both in terms of the average recognition rate and its variance.

representation combination, with the corresponding average recognition errors of 5% and 12%.

4.1.3. Failure modes

Finally, we investigated the main failure modes of our algorithm. An inspection of failed recognitions suggests that the largest difficulty was caused by significant user motion to and from the camera. During the data acquisition, for some of the illumination conditions the dominant light sources were relatively close to the user (from ≈0.5 m). This invalidated the implicit assumption that illumination conditions were unchanging within a single video sequence i.e. that the main cause of appearance changes in images was head rotation. Some examples of very differently illuminated faces within a single sequence can be seen in Fig. 15.

Another limitation of the method was observed in cases when only few faces were clustered to a particular pose, either because of facial feature detection failure or because the user did not spend enough time in a certain range of head poses. The noisy estimate of the corresponding cluster density in (16) propagated the estimation error to illumination normalized images and finally to the overall manifold distance, reducing separation between classes.

5. Summary and conclusions

In this paper we introduced a novel algorithm for face recognition from video, robust to changes in illumination, pose and the motion pattern of the user. This was achieved by combining person-specific face motion appearance manifolds with generic pose-specific illumination manifolds, which were assumed to be linear. Integrated into a fully automatic practical system, the method has demonstrated a high degree of robustness in realistic and uncontrolled data acquisition conditions—specifically to changes in illumination, pose and the motion pattern of the user. We described an extensive empirical evaluation and a comparison with state-of-the-art algorithms in the literature. On average the system correctly recognized in 95% of the cases, exhibiting little sensitivity to the imagining conditions used for data acquisition and consistently outperforming other methods.

We intend to investigate several improvements to the method. Firstly, by employing a more sophisticated reflectance model, we hope to implicitly model nonlinearities in the pose-specific illumination subspaces. Another possible improvement we are considering is the use of quasi illumination-invariant image filters for precise pose matching between faces from two manifolds.

References

[1] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys 35 (4) (2004) 399–458.
[2] British Broadcasting Corporation, Doubts over passport face scans, BBC News Online, UK Edition, October 21, 2004. Available from: <http://news.bbc.co.uk/1/hi/uk/3762398.stm>.
[3] Boston Globe, Face recognition fails in Boston airport, July 2002.
[4] W.A. Barrett, A survey of face recognition algorithms and testing results, Systems and Computers 1 (1998) 301–305.
[5] T. Fromherz, P. Stucki, M. Bichsel, A survey of face recognition, MML Technical Report, (97.01), 1997.
[6] K. Lee, D. Kriegman, Online learning of probabilistic appearance manifolds for video-based recognition and tracking, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, pp. 852–859.
[7] K. Lee, M. Ho, J. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2003, pp. 313–320.
[8] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (1) (2003) 214–245.
[9] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, 581–588.

[10] G. Shakhnarovich, J.W. Fisher, T. Darrel, Face recognition from long-term observations, in: Proc. European Conference on Computer Vision (ECCV), vol. 3, 2002, pp. 851–868.
[11] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, International Symposium of Robotics Research, 2003.
[12] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, Journal of Machine Learning Research (JMLR) 4 (10) (2003) 913–931.
[13] T. Kim, O. Arandjelović, R. Cipolla, Learning over sets using boosted manifold principal angles (BoMPA), in: Proc. IAPR British Machine Vision Conference (BMVC), vol. 2, September 2005, pp. 779–788.
[14] Y. Li, S. Gong, H. Liddell, Modelling faces dynamically across views and over time, in: Proc. IEEE International Conference on Computer Vision (ICCV), vol. 1, 2001, pp. 554–559.
[15] S. Palanivel, B.S. Venkatesh, B. Yegnanarayana, Real time face recognition system using autoassociative neural network models, Acoustics, Speech and Signal Processing 2 (2003) 833–836.
[16] Y. Adini, Y. Moses, S. Ullman, Face recognition: the problem of compensating for changes in illumination direction, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 19 (7) (1997) 721–732.
[17] O. Arandjelović, A. Zisserman, Automatic face recognition for film character retrieval in feature-length films, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, June 2005, pp. 860–867.
[18] M. Everingham, A. Zisserman, Automated person identification in video, in: Proc. IEEE International Conference on Image and Video Retrieval (CIVR), 2004, pp. 289–298.
[19] J. Sivic, M. Everingham, A. Zisserman, Person spotting: video shot retrieval for face sets, in: Proc. IEEE International Conference on Image and Video Retrieval (CIVR), 2005, pp. 226–236.
[20] P.N. Belhumeur, D.J. Kriegman, What is the set of images of an object under all possible illumination conditions?, International Journal of Computer Vision (IJCV) 28 (3) (1998) 245–260
[21] A.S. Georghiades, D.J. Kriegman, P.N. Belhumeur, Illumination cones for recognition under variable lighting: faces, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1998, pp. 52–58.
[22] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 25 (9) (2003) 1063–1074.
[23] L. Zhang, S. Wang, D. Samaras, Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2005, pp. 206–216.
[24] T. Riklin-Raviv, A. Shashua, The quotient image: class based re-rendering and recognition with varying illuminations, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23 (2) (2001) 139–219.
[25] B. Kepenekci, Face Recognition Using Gabor Wavelet Transform, PhD Thesis, The Middle East Technical University, 2001.
[26] H. Murase, S. Nayar, Visual learning and recognition of 3-D objects from appearance, International Journal of Computer Vision (IJCV) 14 (1) (1995) 5–24.
[27] X. Liu, T. Chen, Video-based face recognition using adaptive hidden Markov models, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2003, pp. 340–345.
[28] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994, pp. 84–91.
[29] M. Bichsel, A.P. Pentland, Human face recognition and the face image set's topology, Computer Vision, Graphics and Image Processing: Image Understanding 59 (2) (1994) 254–261.
[30] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E. Learned-Miller, D.A. Forsyth, Names and faces in the news, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004, pp. 848–854.
[31] K. Fukui, O. Yamaguchi, Facial feature point extraction method based on combination of shape extraction and pattern matching, Systems and Computers in Japan 29 (6) (1998) 2170–2177.
[32] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., John Wily & Sons, Inc., New York, 2000.
[33] H. Gray, Anatomy of the Human Body, 20th ed., Lea & Febiger, Philadelphia, 1918.
[34] R. Gonzalez, R. Woods, Digital Image Processing, Addison-Wesley Publishing Company, 1992.
[35] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, 2nd ed., Cambridge University Press, 1992.
[36] S. Shan, W. Gao, B. Cao, D. Zhao, Illumination normalization for robust face recognition against varying lighting conditions, in: Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 157–164.
[37] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society 3 (61) (1999) 611–622.
[38] J. Baglama, D. Calvetti, L. Reichel, Iterative methods for the computation of a few eigenvalues of a large symmetric matrix, BIT 36 (3) (1996) 400–440.
[39] D.H. Johnson, S. Sinanović, Symmetrizing the Kullback-Leibler distance, Technical Report, Rice University, 2001.
[40] T. Sim, S. Zhang, Exploring face space, in: Proc. IEEE Workshop on Face Processing in Video, 2004, p. 84.
[41] Toshiba, Facepass. Available from: <www.toshiba.co.jp/mmlab/tech/w31e.htm>.
[42] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[43] A. Fitzgibbon, A. Zisserman, On affine invariant clustering and automatic cast listing in movies, in: Proc. European Conference on Computer Vision (ECCV), 2002, pp. 304–320.
[44] H. Wang, S.Z. Li, Y. Wang, Face recognition under varying lighting conditions using self quotient image, in: Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FGR), 2004, pp. 819–824.
[45] O. Arandjelović, R. Cipolla, A new look at filtering techniques for illumination invariance in automatic face recognition, in: Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FGR), April 2006, pp. 449–454.

Ognjen Arandjelović: Ognjen Arandjelović is a Research Fellow at Trinity College, Cambridge. He graduated top of his class from the Department of Engineering Science at the University of Oxford (M.Eng.). In 2007 he was awarded the Ph.D. degree from the University of Cambridge. His main research interests are computer vision and machine learning, and their application in other scientific disciplines. He is a Fellow of the Cambridge Overseas Trust.

Roberto Cipolla: Roberto Cipolla received the BA degree (engineering) from the University of Cambridge in 1984 and the MSE degree (electrical engineering) from the University of Pennsylvania in 1985. In 1991, he was awarded the D.Phil. degree (computer vision) from the University of Oxford. His research interests are in computer vision and robotics and include the recovery of motion and 3D shape of visible surfaces from image sequences, visual tracking and navigation, robot hand-eye coordination, algebraic and geometric invariants for object recognition and perceptual grouping, novel man-machine interfaces using visual gestures, and visual inspection. He has authored three books, edited six volumes, and coauthored more than 200 papers.