Pattern Recognition ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant

Ognjen Arandjelović*, Roberto Cipolla

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

ARTICLE INFO

Article history:
Received 29 January 2011
Received in revised form
8 June 2012
Accepted 24 June 2012

Keywords:
Manifold
Illumination
Pose
Motion
Invariance
Generic

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. The central contribution is an illumination invariant, which we show to be suitable for recognition from video of loosely constrained head motion. In particular there are three contributions: (i) we show how a photometric model of image formation can be combined with a statistical model of generic face appearance variation to exploit the proposed invariant and generalize in the presence of extreme illumination changes; (ii) we introduce a video sequence "re-illumination" algorithm to achieve fine alignment of two video sequences; and (iii) we use the smoothness of geodesically local appearance manifold structure and a robust same-identity likelihood to achieve robustness to unseen head poses. We describe a fully automatic recognition system based on the proposed method and an extensive evaluation on 323 individuals and 1474 video sequences with extreme illumination, pose and head motion variation. Our system consistently achieved a nearly perfect recognition rate (over 99.7% on all four databases).

1. Introduction

Automatic face recognition 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 state-of-the-art methods has been 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%, while in less controlled environments the performance degrades even further [1,2]. 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 face matching algorithms rest upon are difficult to satisfy in practice.

1.1. Recognition from video sequences and image sets

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 [3–5]. In contrast to these, a number of methods that do not use temporal information have been proposed. Recent ones include statistical [6,7] and principal angle-based methods with underlying simple linear [8,9], kernel-based [10] or Gaussian mixture-based [11] models. By their very nature, these are inherently invariant to changes in head motion pattern. Other algorithms implement the "still-to-video" scenario [12,13] and do not take full advantage of sequences available for training.

1.2. Recognition across illumination

Illumination invariance, perhaps the most significant challenge for automatic face recognition [14], 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 [15,16]. Others rely on the availability of ample training data but achieve limited generalization [6,17].

Two influential generative model-based approaches for illumination-invariant single-shot recognition are the illumination cones [18,19] and the 3D morphable model [20–22]. Both 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

* Corresponding author. Tel.: +44 7598 999612
E-mail addresses: oa214@eng.cam.ac.uk, ognjen.arandjelovic@gmail.com (O. Arandjelović).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.06.024

Table 1
A qualitative comparison of advantages and disadvantages of the two main groups of face recognition methods in the literature.

	Appearance-based [29,3,30,31]	**Model-based** [32,22,20,16]
Advantages	Well-understood off-the-shelf statistical methods readily applied Can be used for poor quality and low resolution input	Explicit modelling and recovery of personal and extrinsic variables Prior, domain-specific knowledge is used
Disadvantages	Poor generalization to unseen pose, illumination etc. No (or little) use of domain-specific knowledge	High quality data required Time-consuming model parameter estimation User intervention is often required for initialization Difficult to model complex illumination effects – fitting becomes increasingly ill-conditioned

pose which is difficult to achieve in practice (see Section 4 for typical data quality). Similar limitations apply to the related method of Riklin–Raviv and Shashua [23]. 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 [16], 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).

1.3. 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 [20,24]. This is a view-independent representation. A second class of algorithms consists of global, parametric models, such as the eigenspace method [25] that estimates a single parametric (typically linear) subspace from all the views for all the objects (also see [26]). In comparative face recognition evaluation trials, such methods are usually outperformed by methods from the third class: view-based techniques e.g. the view-based eigenspaces [27] (also [3,4,28]), 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.

1.4. Problem statement

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 a face recognition 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 and a generic illumination invariant, learnt off-line. Specifically, we show that the combined effects of face shape and illumination can be effectively learnt using a mixture of Probabilistic Principal Component Analyzers (PPCA) [33] from a small, unlabelled set of video sequences of faces in randomly varying lighting conditions, while a novel manifold-based "re-illumination" 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. We demonstrate that the manner in which generative and discriminative elements are interlaced in the proposed method succeeds in inheriting the strengths of both groups of approaches, which are summarized in Table 1.

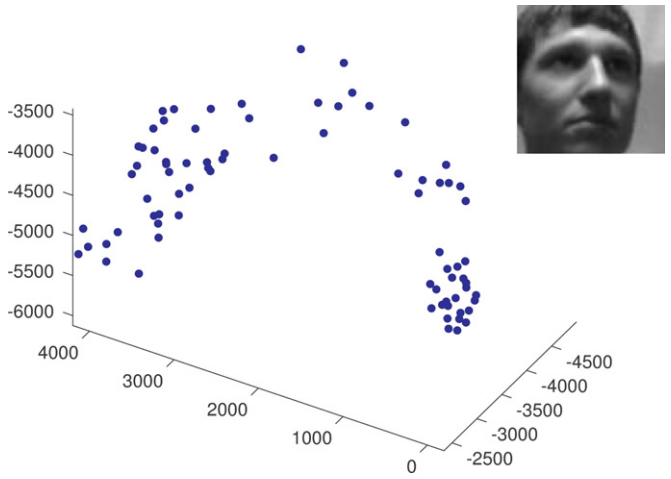


Fig. 1. Samples from a manifold of face appearance corresponding to a single head motion sequence. Images in the set were first converted into vectors by column-wise rasterization (as in [35], for example) of their greyscale pixel values and then displayed projected to the first three linear principal components computed using the entire image set. A typical manifold sample displayed as an image is shown in the top-right corner.

2. Face motion and other manifolds

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*, its dimensionality being equal to the number of pixels D. However, the surface and texture of a face is mostly smooth making its appearance constrained and approximately confining it to an embedded *face manifold* of dimension d, where usually $d \ll D$ [6,34]. Formally, the distribution of observed face images $\mathbf{x}$ of the subject i can be written as the integral

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

where p_n is the noise distribution, $f^{(i)} : \mathbb{R}^d \to \mathbb{R}^D$ the embedding function, $\tilde{\mathbf{x}}$ an intrinsic face descriptor and $p_F^{(i)}(\tilde{\mathbf{x}})$ the corresponding probability density function over $\mathbb{R}^d$. Fig. 1 illustrates the validity of the notion on an example of a set of face images extracted from a

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

head motion video sequence. For the proposed method, the crucial properties of face appearance manifolds are their (i) $\mathcal{C}^0$ continuity and (ii) approximate smoothness ($\mathcal{C}^1$ continuity).

2.1. Synthetic re-illumination of face motion manifolds

One of the key ideas of this paper is the *re-illumination* of video sequences. Our goal is to take two input sequences of faces and produces 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 re-illuminated version of the original face. Our algorithm is therefore global, unlike most of the previous methods which use a sparse set of detected salient points for registration, e.g. [36,37,8]. We found that these fail on our data set due to the severity of illumination conditions (see Section 4). The two stages of the proposed algorithm are described in detail next.

2.1.1. Stage 1: pose matching

Consider two motion sequences of a person's face in different illuminations[1]:

$$\{\mathbf{x}_i\}^{(1)} = \{\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{N_1}^{(1)}\} \quad \text{and} \quad \{\mathbf{x}_i\}^{(2)} = \{\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{N_2}^{(2)}\}. \tag{2}$$

Then, for each $\mathbf{x}_i^{(1)}$ we are interested in finding $\mathbf{x}_{c(i)}^{(2)}$ that corresponds to it best in terms of head pose. Finding the unknown mapping $c : \{1, \ldots, N_1\} \to \{1, \ldots, N_2\}$ on a frame-by-frame basis is difficult in the presence of extreme illumination changes and when face images are of low resolution. Instead, we exploit the smoothness of face appearance manifolds by formulating the problem as a minimization task with the fitness function $f(c)$ taking on the following form:

$$f(c) = f_{match}(c) + \omega \cdot f_{reg}(c), \tag{3}$$

$$f(c) = \underbrace{\sum_j d_E(\mathbf{x}_j^{(1)}, \mathbf{x}_{c(j)}^{(2)})^2}_{\text{Matching term}} + \omega \underbrace{\sum_j \sum_k \frac{d_G^{(2)}\left(\mathbf{x}_{c(j)}^{(2)}, \mathbf{x}_{c(n(j,k))}^{(2)}; \{\mathbf{x}_j\}^{(2)}\right)}{d_G^{(1)}\left(\mathbf{x}_j^{(1)}, \mathbf{x}_{n(j,k)}^{(1)}; \{\mathbf{x}_j\}^{(1)}\right)}}_{\text{Regularization term}}, \tag{4}$$

where $n(i,j)$ is the j-th of K nearest neighbours of face i, d_E a pose dissimilarity function, $d_G^{(k)}$ a geodesic distance estimate along the appearance manifold corresponding to the sequence k, and ω a relative weighting constant. The first term is easily understood as a penalty for the dissimilarity of matched poses. The latter is a regularizing term that 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. This is illustrated conceptually in Fig. 2.

2.1.2. Regularization

The manifold-oriented nature of the regularizing function $f_{reg}(c)$ in Equation (4) has significant advantages over alternatives that use some form of temporal smoothing. Firstly, it is unaffected by changes on the motion pattern of the user (i.e. sequential ordering of $\{\mathbf{x}_i\}^{(j)}$). On top of the inherent benefit (a person's motion should not affect recognition), this is important for several practical reasons, the most important of which are:

- face images need not originate from a single sequence—multiple sequences are easily combined together by computing the union of their frame sets, and

[1] Note that illumination may arbitrarily vary *within* each sequence as well. No aspect of our method requires unchanging illumination within a sequence.

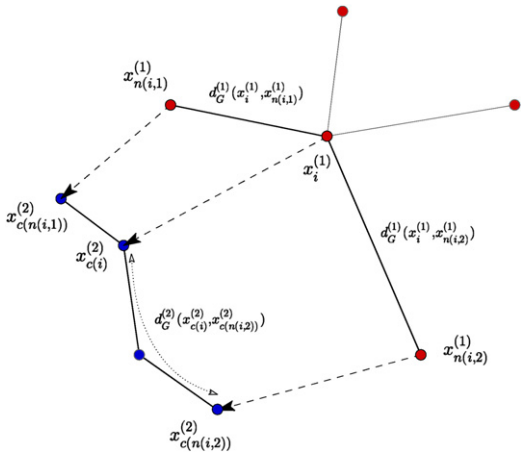


Fig. 2. Manifold-to-manifold pose matching: geodesic distances between neighbouring faces on the domain manifold and the corresponding faces on the codomain manifold are used to regularize the solution.

- regularization works even if there are bursts of missed or incorrect face detections (see Section 4).

To understand the form of the regularizing function note that the mapping function c only affects the numerator of each summation term in $f_{reg}(c)$. Its effect is then to penalize cases in which neighbouring faces of the domain manifold map to geodesically distant faces on the codomain manifold. The penalty is further weighted by the inverse of the original geodesic distance $d_G^{(1)}(\mathbf{x}_j^{(1)}, \mathbf{x}_{n(j,k)}^{(1)}; \{\mathbf{x}_j\}^{(1)})$ to place more emphasis on local pose agreement.

2.1.2.1. Pose-Matching Function: The performance of function d_E in Equation (4 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:

$$d_E(\mathbf{x}_j^{(1)}, \mathbf{x}_{c(j)}^{(2)}) = \|\mathbf{X}_j^{(1)} - \mathbf{X}_{c(j)}^{(2)}\|_2. \tag{5}$$

Note that the 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, shown in Fig. 3, motivated by the success of this representation in object-configuration matching across other computer vision applications, e.g. [38,39].

2.1.2.2. Optimizing the Frame Correspondence Function: Exact minimization of the fitness function in Eq. (4) over all functions c is an NP-complete problem. However, since the final synthesis of novel faces (Stage 2) involves an entire geodesic neighbourhood 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) [40] 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. 4(b), with the i-th gene in a chromosome representing the value of $c(i)$. GA parameters were determined

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

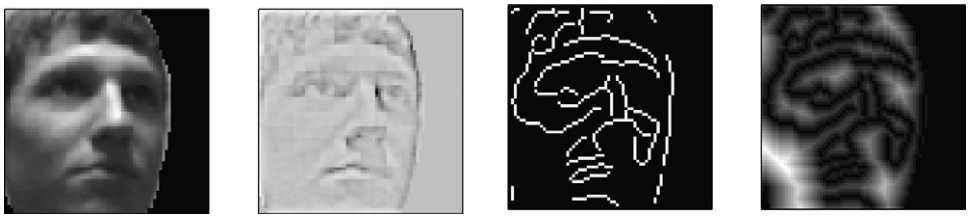
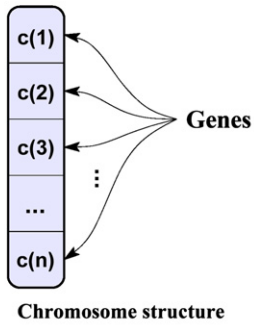


Fig. 3. Left-to-right: original image, the image after high-pass filtering, Canny-detected edges and the final pose-signature as a distance-transformed edge map.

a

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

b



c

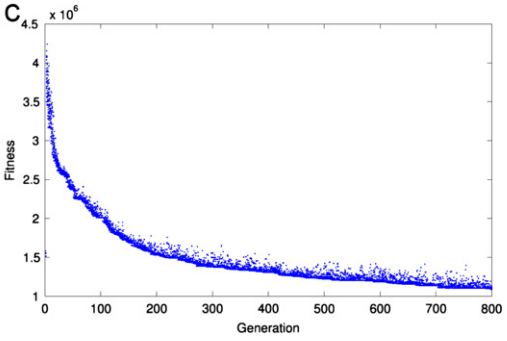


Fig. 4. (a) Parameters of the proposed GA optimization, (b) the corresponding chromosome structure, and (c) population fitness of Eq. (4) in a typical evolution. Each data point (dot in the plot) represents the fitness of a single individual in the population of the corresponding generation (thus the number of data points for each abscissa value is 20 which is the population size). Maximal generation count of 200 (maximal abscissa value) was chosen empirically as a trade-off between accuracy and matching speed.

experimentally by optimizing the algorithm's performance on a small training data set. A summary is given in Fig. 4(a) and (c).

2.1.2.3. Estimating Geodesic Distances: The expression for the fitness function in Equation (4 involves geodesic distances along manifolds. Due to the nonlinearity of face appearance manifolds [6,4], these are not well approximated by the corresponding Euclidean distances in the image space. Thus we estimate the geodesic distance between every two faces lying on an appearance manifold using Floyd's algorithm [41] on a constructed undirected graph whose nodes correspond to face images (also see [42]). Then, if $\mathbf{x}_i$ is one of the K nearest neighbours of $\mathbf{x}_j$.[2]

$$d_G(\mathbf{x}_i,\mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2. \tag{6}$$

Otherwise

$$d_G(\mathbf{x}_i,\mathbf{x}_j) = \min_k [d_G(\mathbf{x}_i,\mathbf{x}_k) + d_G(\mathbf{x}_k,\mathbf{x}_j)]. \tag{7}$$

2.1.3. Stage 2: fine re-illumination

Having computed a pose-matching function c^*, we turn to the problem of re-illuminating face images $\mathbf{x}_i^{(1)}$. We exploit the smoothness of pose-signature manifolds (which was ensured by distance-transforming face edge maps), illustrated in Fig. 5, by computing $\mathbf{y}_i^{(1)}$, the re-illuminated face $\mathbf{x}_i^{(1)}$, as a linear combination of K nearest-neighbour frames of $\mathbf{x}_{c^*(i)}^{(2)}$:

$$\mathbf{y}_i^{(1)} = \sum_{k=1}^{K} \alpha_k \mathbf{x}_{n(c^*(i),k)}^{(2)}. \tag{8}$$

Linear combining coefficients $\alpha_1, \ldots \alpha_K$ are found from the corresponding pose-signatures by solving the following constrained minimization problem:

$$\{\alpha_j\} = \arg\min_{\{\alpha_j\}} \left\| \mathbf{X}_i^{(1)} - \sum_{k=1}^{K} \alpha_k \mathbf{X}_{n(c^*(i),k)}^{(2)} \right\|_2 \tag{9}$$

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 the target illumination), which are then combined in the same fashion to synthesize a re-illuminated face, as shown in Figs. 6 and 7. We restrict the set of frames used for re-illumination to the K-nearest neighbours for two reasons. Firstly, the computational time of using all faces would make this highly unpractical. Secondly, the nonlinearity of both face appearance manifolds and pose-signature manifolds, demands that only the faces in the local, Euclidean-like neighbourhood are used.

Optimization of the expression in Equation (9 is readily performed by differentiating the quadratic term corresponding

[2] Note that the converse does not hold as $\mathbf{x}_i$ being one of the K nearest neighbours of $\mathbf{x}_j$ does not imply that $\mathbf{X}_j$ is one of the K nearest neighbours of $\mathbf{x}_i$. Therefore the edge relation of this graph is a superset of the "in K-nearest neighbours" relation on $\mathbf{x}$s.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

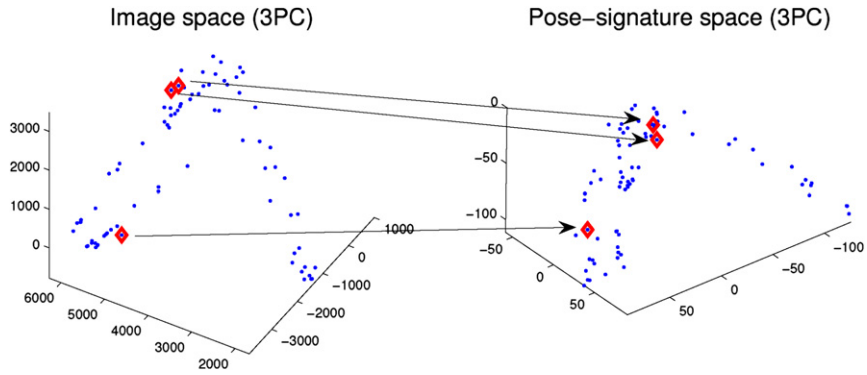


Fig. 5. A face motion manifold in the input image space and the corresponding pose-signature manifold (both shown in their respective 3D principal subspaces). Much like the original appearance manifold, the latter is continuous and smooth, as ensured by distance transforming the face edge maps. While not necessarily similar globally, the two manifolds retain the same *local* structure, which is crucial for the proposed fine re-illumination algorithm.

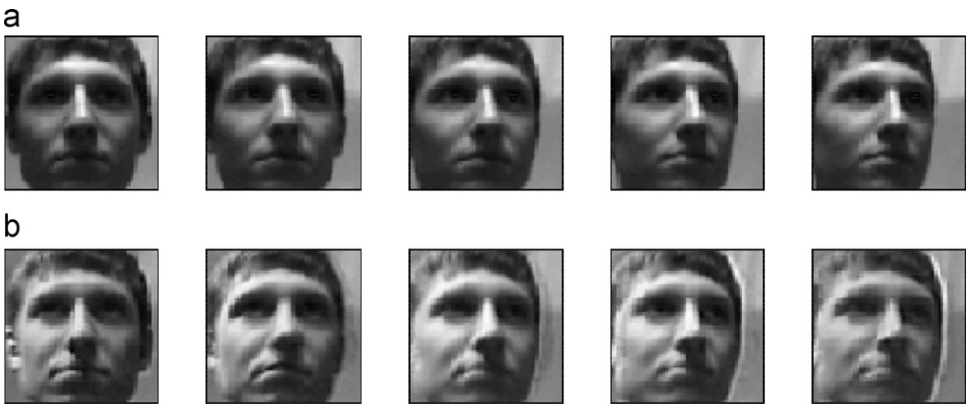


Fig. 6. (a) Original images from a novel video sequence and (b) the result of re-illumination using the proposed genetic algorithm with nearest neighbour-based reconstruction.

to the square of the objective function, giving:

$$[\alpha_2\ \alpha_3\ \ldots\ \alpha_K]^T = \mathbf{R}^{-1}\mathbf{t}, \tag{10}$$

where:

$$\mathbf{R}(j,k) = (\mathbf{X}^{(2)}_{n(c^*(i),1)} - \mathbf{X}^{(2)}_{n(c^*(i),j)})^T (\mathbf{X}^{(2)}_{n(c^*(i),1)} - \mathbf{X}^{(2)}_{n(c^*(i),k)})$$

and

$$\mathbf{t}(j) = (\mathbf{X}^{(2)}_{n(c^*(i),1)} - \mathbf{X}^{(2)}_{n(c^*(i),j)})^T (\mathbf{X}^{(2)}_{n(c^*(i),1)} - \mathbf{X}^{(1)}_{i}). \tag{11}$$

3. 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 face recognition 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 $\mathbf{x}$(j) is a linear function of the albedo $\mathbf{a}(j)$ of the corresponding 3D point:

$$\mathbf{x}(j) = \mathbf{a}(j) \cdot \mathbf{s}(j), \tag{12}$$

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 [43], the main difference being that we make no further assumptions on the functional form of $\mathbf{s}(.)$. Note that the commonly used Lambertian reflectance model (e.g. see [20,19,23]) is a special case of Eq. (12) [18]:

$$\mathbf{s}(j) = \sum_i \max(\mathbf{n}_j \cdot \mathbf{L}_i, 0), \tag{13}$$

where $\mathbf{n}_i$ is the corresponding surface normal and $\{\mathbf{L}_i\}$ the intensity-scaled illumination directions at the point.

The image formation model introduced in Eq. (12) 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 Eq. (12):

$$\Delta \log \mathbf{x}(j) = \log \mathbf{s}_2(j) - \log \mathbf{s}_1(j) \equiv \mathbf{d}_s(j). \tag{14}$$

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. This is the shape-illumination manifold (SIM) corresponding to a particular pair of video sequences.

3.1. The generic shape-illumination manifold

A crucial assumption of our work is that the shape-illumination manifold of all possible illuminations and head poses is *generic for human faces* (generic SIM, or G-SIM). This is motivated by a number of independent results reported in the literature that

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

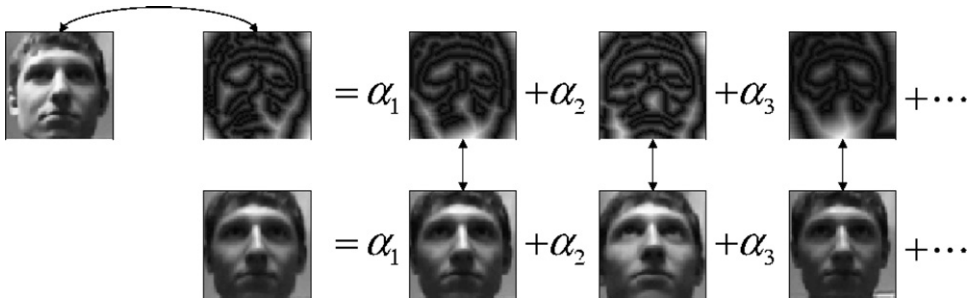


Fig. 7. Face re-illumination: the coefficients for linearly combining face appearance images (bottom row) are computed using the corresponding pose-signatures (top row). Also see Fig. 5.

have shown face shape to be less discriminating than albedo across different models [44,45] or have reported good results in synthetic re-illumination of faces using the constant-shape assumption [23]. In the context of face manifolds this means that the effects of *illumination and shape* can be learnt off-line from a training corpus containing typical modes of pose and illumination variation.

It is worth emphasizing the key difference in the proposed off-line learning from previous approaches in the literature which try to learn the *albedo* of human faces. Since off-line 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 [23], for example, the authors demonstrate generalization to albedos in the rational span of those in the off-line 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 significant generalization [6].

3.1.1. Training data organization

The proposed face recognition method consists of two training stages—a one-time off-line learning performed using *off-line 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.

3.2. Off-line stage: learning generic shape-illumination effects

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 re-illumination algorithm in Section 2.1. Then from Eq. (14), samples from the generic shape-illumination manifold can be computed by logarithm-transforming all images and subtracting those corresponding in identity and pose:

$$\mathbf{d} = \log \mathbf{x}_i^{(j,p)} - \log \mathbf{x}_i^{(j,q)}. \tag{15}$$

Provided that training data contains typical variations in pose and illumination (i.e. that the probability density function 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), already proven successful in a variety of face recognition algorithms [46,47,6,48]. In the proposed framework, this representation is motivated by: (i) the assumed low-dimensional manifold model in Eq. (1), (ii) its compactness and (iii) the existence of incremental model parameter estimation algorithms (e.g. [49,50]). Thus:

$$\mathcal{G}(\mathbf{d};\boldsymbol{\Theta}) = \sum_{q=1}^{Q} [\alpha_q \cdot G(\mathbf{d};\mu_q,\Sigma_q)], \tag{16}$$

where $G(\mathbf{d};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q)$ is a multivariate Gaussian function in $\mathbb{R}^D$, with the mean $\boldsymbol{\mu}_q$ and the covariance matrix $\boldsymbol{\Sigma}_q$:

$$G(\mathbf{d};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q) = \exp\{-\tfrac{1}{2}(\mathbf{d}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q^{-1}(\mathbf{d}-\boldsymbol{\mu}_q)\}, \tag{17}$$

and $\boldsymbol{\Theta}$ the set of all mixture parameters. By construction, the covariance matrix can be written as a sum of a full covariance in at most d_{GSIM} directions and an isotropic complementary covariance, uniform across different components in the mixture:

$$\boldsymbol{\Sigma}_q = \overbrace{\hat{\mathbf{P}}_q \hat{\Lambda}_q \hat{\mathbf{P}}_q^T}^{\text{principal subspace}} + \overbrace{\hat{\rho}\hat{\mathbf{C}}_q\mathbf{C}_q^T}^{\text{complementary subspace}} \tag{18}$$

where

$$\mathbf{P}_q^T\mathbf{C}_q = \mathbf{0} \quad \mathbf{P}_q^T\mathbf{P}_q = \mathbf{1}^{(d_{\mathrm{GSIM}})} \quad \mathbf{C}_q^T\mathbf{C}_q = \mathbf{1}^{(D-d_{\mathrm{GSIM}})}, \tag{19}$$

and $\mathbf{1}^{(d_{\mathrm{GSIM}})}$ and $\mathbf{1}^{(D-d_{\mathrm{GSIM}})}$ are identity matrices of dimensions, respectively, $d_{\mathrm{GSIM}} \times d_{\mathrm{GSIM}}$ and $(D-d_{\mathrm{GSIM}}) \times (D-d_{\mathrm{GSIM}})$. We estimate the multivariate Gaussian components using the Expectation Maximization (EM) algorithm [40], initialized by K-means clustering. Automatic model order selection is performed using the well-known Minimum Description Length criterion [40,51] 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. From 6123 G-SIM samples computed from 100 video sequences, we obtained $Q=12$ mixture components, each with a $d_{\mathrm{GSIM}} = 6$-dimensional principal subspace.

3.3. Model application: matching a novel query sequence

The discussion so far has concentrated on off-line training and building an illumination model for faces – the generic shape-illumination manifold. Central to the proposed algorithm was a method for re-illuminating a face motion sequence of a person with another sequence of the *same* person—see Section 2.1. We now show how the same method can be used to compute a similarity between two unknown individuals, given a single training sequence for each and the probability density capturing the structure of the generic shape-illumination manifold.

Let gallery data consist of face 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 $\mathcal{G}(\mathbf{d};\boldsymbol{\Theta})$ a mixture of Probabilistic PCA corresponding to the generic SIM. Using the re-illumination algorithm of Section 2.1, the novel sequence can be re-illuminated with each $\{\mathbf{x}_i\}^{(j)}$ from the gallery, producing samples $\{\mathbf{d}_i\}^{(j)}$. We assume these to be identically and independently distributed according to a density corresponding to a *postulated* subject-specific SIM. We then compute the probability of these under

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

Fig. 8. An example of "re-illumination" results when the two compared sequences do not correspond to the same individual: the target sequence is shown on the left, the output of our algorithm on the right. Most of the frames do not contain face which correspond in pose.

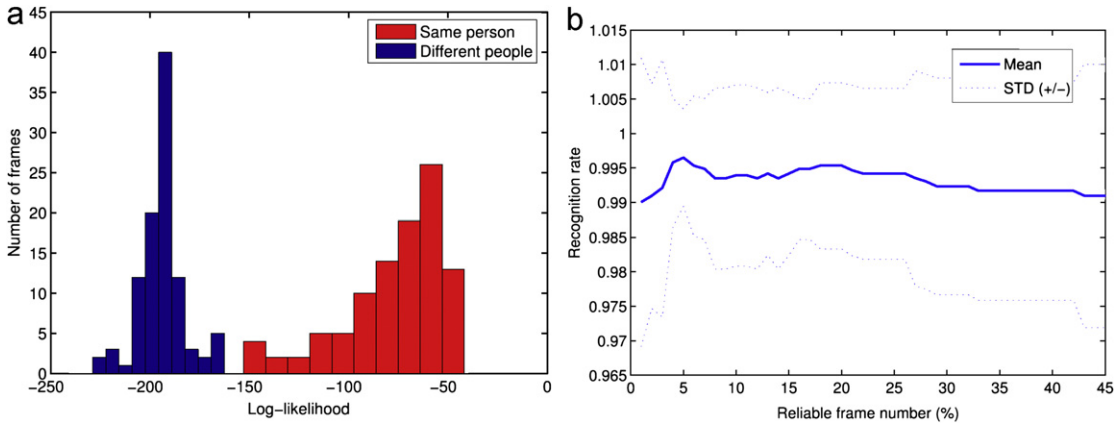


Fig. 9. (a) Histograms of intra-personal likelihoods across frames of a sequence when two sequences compared correspond to the same (red) and different (blue) people. (b) Recognition rate as a function of the number of frames deemed 'reliable'. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

$\mathcal{G}(\mathbf{d};\boldsymbol{\Theta})$

$$p_i^{(j)} = \mathcal{G}(\mathbf{d}_i^{(j)};\boldsymbol{\Theta}). \tag{20}$$

When $\{\mathbf{x}_i\}^{(0)}$ and $\{\mathbf{x}_i\}^{(j)}$ correspond in identity, from the way the generic SIM is learnt, it can be seen that the probabilities $p_i^{(j)}$ will be large. The more interesting question arises when the two compared sequences do not correspond to the same person. In this case, the re-illumination algorithm will typically fail to produce a meaningful result—the output will not correspond in pose to the target sequence, as illustrated on an example in Fig. 8. Consequently, the observed appearance difference will have a low probability under the hypothesis that it is caused purely by an illumination change. A similar result is obtained if the two individuals share sufficiently similar facial lines and poses are correctly matched. In this case it is the differences in face surface albedo that are not explained well by the generic SIM, producing low $p_i^{(j)}$ in Eq. (20).

3.3.1. *Varying pose and robust likelihood*

Instead of basing the classification of $\{\mathbf{x}_i\}^{(0)}$ on the likelihood corresponding to observing the *entire* set $\{\mathbf{d}_i\}^{(j)}$ in Eq. (20), we propose a more robust measure. To appreciate the need for additional robustness, consider the histograms in Fig. 9(a). It can be observed that the likelihood of the most similar faces in an inter-personal comparison, in terms of the expression in Eq. (20), approaches that of the most *dissimilar* faces in an *intra-personal* comparison (sometimes even exceeding it). This occurs when the correct gallery 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 using Eq. (9). In our method, the robustness to these, unseen modes of pose variation is achieved by considering the mean log-likelihood of only the most likely faces. In our experiments we used the top 15% of the faces, but we found the algorithm to exhibit little sensitivity to the exact choice of this number, as Fig. 9(b) shows. A summary of the proposed algorithms is shown in Fig. 10.

4. Empirical evaluation

Methods in this paper were evaluated on four databases, containing in total 323 people and 1474 sequences, resulting in 117,271 automatically detected faces. We summarize the data sets:

CamFace: The University of Cambridge face motion database contains 100 individuals of varying age and ethnicity. For each person in the database we collected seven video sequences of the person in arbitrary motion (significant translation, yaw and pitch, negligible roll), each in a different illumination setting, at 10 fps and in 320×240 pixel resolution (face size ≈ 60 pixels)[3]; see

[3] A thorough description of the University of Cambridge face database with examples of video sequences is available at http://mi.eng.cam.ac.uk/~oa214/.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

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

Fig. 10. Summary of the proposed learning (off-line) and recognition algorithms.

Figs. 11(a), 12(a) and 13, as well as [52] for a detailed description of the database.

ToshFace: This database was kindly provided to us by Toshiba Corporation. It contains 60 individuals of varying age, mostly male Japanese, and 10 sequences per person. Each sequence was acquired at 10 fps and in 320×240 pixel resolution, and corresponds to a different illumination setting in which the subject performed uncontrolled body and head pose changes, (face size ≈ 60 pixels), as illustrated in Figs. 11(b) and 12(b).

FaceVideo: This database is freely available[4] and described in [53]. It contains 11 individuals with 2 sequences per person, little variation in illumination, but extreme and uncontrolled variations in pose and motion, acquired at 25 fps and 160×120 pixel resolution (face size ≈ 45 pixels), see Fig. 11(c).

Faces96: This is the most challenging subset of the University of Essex face database, also freely available.[5] It contains 152 individuals, most of whom are 18–20 years old, and a single 20-frame sequence per person in 196×196 pixel resolution (face size ≈ 80 pixels). The users were asked to approach the camera while performing arbitrary head motion. Although the illumination was kept constant throughout each sequence, there is some variation in the manner in which faces were lit due to the change in the relative position of the user with respect to the lighting sources, as shown in Fig. 11(d).

4.1. Automatic data extraction

The discussion so far focused on recognition using fixed-scale face images. Our system uses the cascaded detector of Viola and Jones [54] for localization of faces in cluttered images, which are then rescaled to the uniform resolution of 50×50 pixels (approximately the average size of detected faces). Depending on the severity of the illumination in which data was acquired as well as the poses assumed by the user, a varying number of faces is extracted from a single sequence, see Fig. 14.

4.2. Methods and representations

We compared the performance of our recognition algorithm with and without the robust likelihood of Section 3.3 (i.e. using only the most reliable vs. all detected and re-illuminated faces) to that of:

- Commercial system FaceIt® by Identix [55] (the best performing software in the 2003 Face Recognition Vendor Test [56]; also see [57]).
- Mutual Subspace Method, based on canonical correlation analysis (CCA) constrained to low-dimensional linear subspaces, [8,58].[6]
- Constrained Mutual Subspace Method [8], a discriminative form of Canonical Correlation Analysis (C-CCA), used in Toshiba's commercial system FacePass® [59].[7]
- The probability density-based algorithm of Shakhnarovich et al. (G-KLD) which models personal appearance variations by multivariate normal distributions and uses the Kullback-Leibler divergence to measure their similarity [7]. (See Footnote 7.)

In all tests, both training data for each person in the gallery, as well as query data, consisted of only a single sequence. Off-line training of the proposed algorithm was performed using 20 individuals in five illuminations from the CamFace data set—we emphasize that these were not used as query input for the evaluations reported in this section.

For *CamFace*, *ToshFace* and *FaceVideo* databases, we trained our algorithm using a single sequence per person and tested against a single other query sequence per person, acquired in a different session (for *CamFace* and *ToshFace* different sessions correspond to different illumination conditions). Since *Faces96* database contains only a single sequence per person, we used the frames 1–10 of each for training and frames 11–20 for test. Seeing that each video sequence in this database shows a person walking to

[4] See http://synapse.vit.iit.nrc.ca/db/video/faces/cvglab.
[5] See http://cswww.essex.ac.uk/mv/allfaces/faces96.html.
[6] http://cswww.essex.ac.uk/mv/allfaces/faces96.html.
[7] The algorithm was re-implemented in consultation with the authors.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

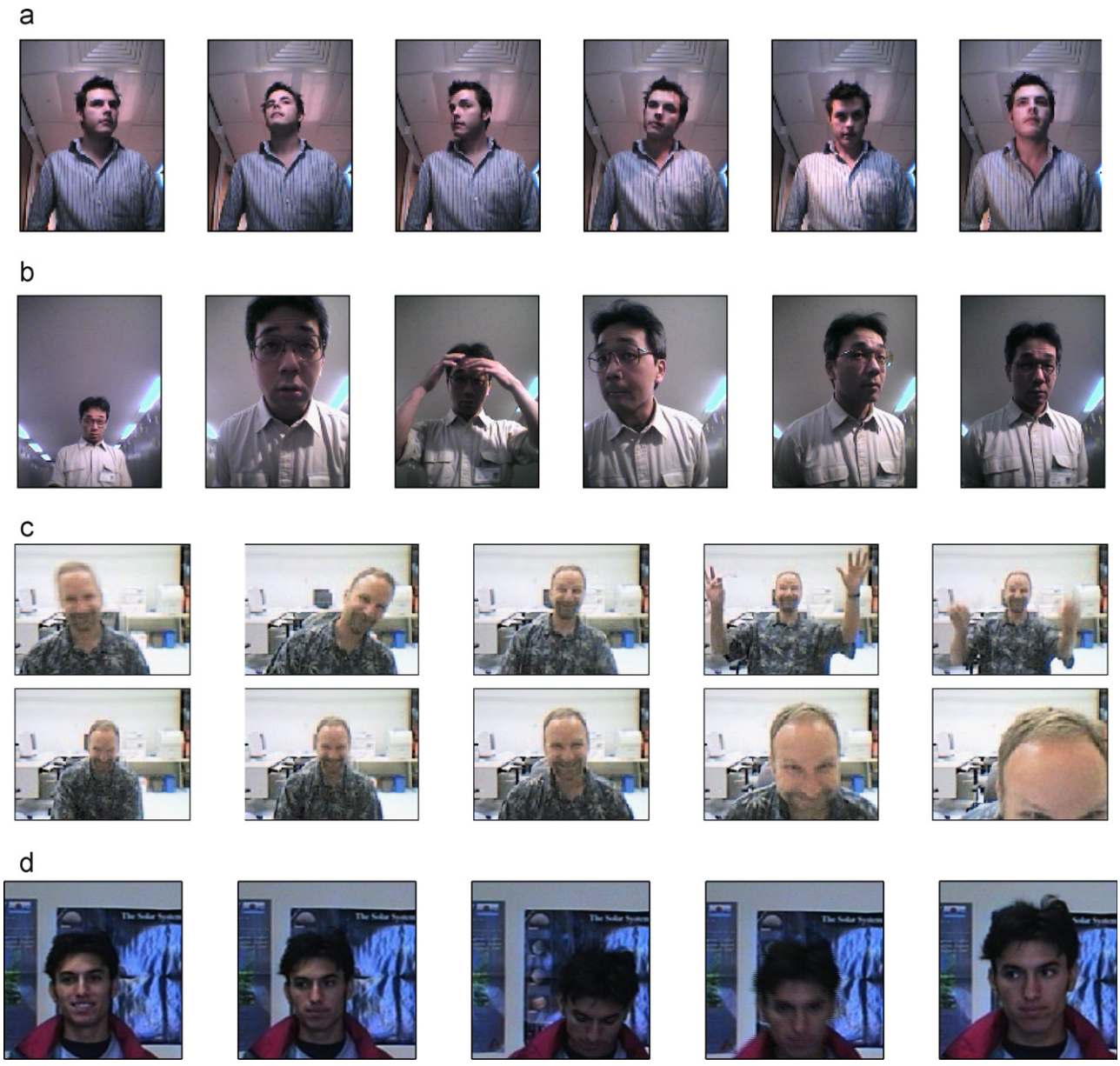


Fig. 11. Frames from typical video sequences in the databases used for evaluation. (a) CamFace. (b) ToshFace. (c) FaceVideo. (d) Faces96.

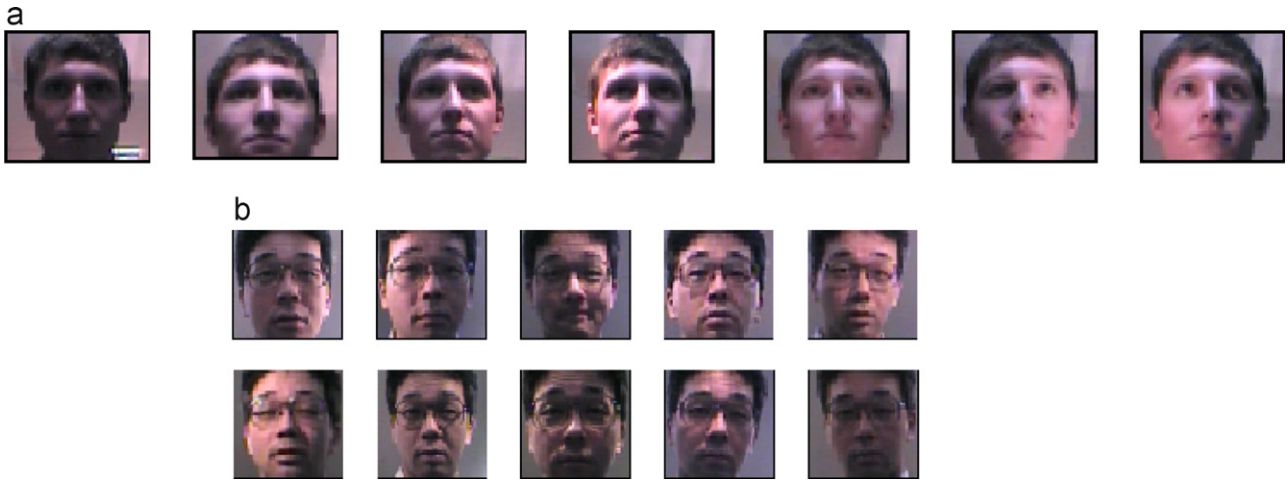


Fig. 12. (a) CamFace and (b) ToshFace illuminations. Also see Fig. 13.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

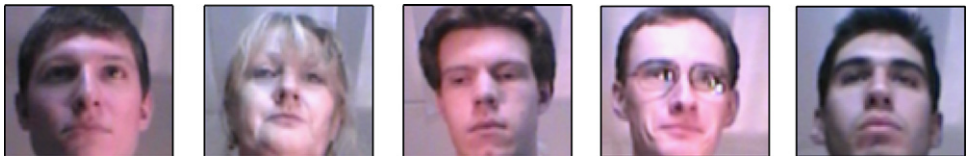


Fig. 13. Five different individuals in the illumination setting number 6. In spite of the same spatial arrangement of light sources, their effect on the appearance of faces changes significantly due to variations in people's heights, the *ad lib* chosen position relative to the camera etc.

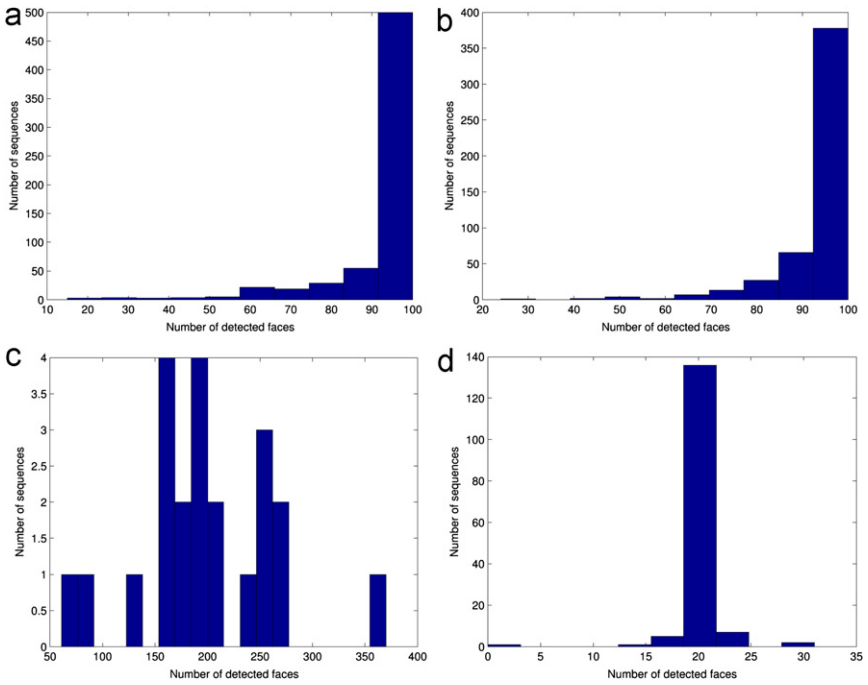


Fig. 14. Histograms of the number of detected faces per video sequence for the four databases used in evaluation. All sequences in CamFace, ToshFace and Faces96 data sets are of equal duration, resulting in roughly unimodal histograms, while the duration of sequences in FaceVideo varies from 9 s to 21 s, producing a more varied number of detections.

the camera, this division maximizes the variation in illumination, scale and pose between training and test, thus maximizing the recognition challenge. The methods were evaluated using three face representations:

- raw appearance images $\mathbf{x}$,
- Gaussian high-pass filtered images, successfully used for face matching in [60,36,61] amongst others:

$$\mathbf{x}_H = \mathbf{x} - (\mathbf{x} * \mathbf{G}_{\sigma=1.5}), \tag{21}$$

  where $*$ denotes convolution, and
- local intensity-normalized high-pass filtered images – similar to the Self Quotient Image [62] (also see [63,23,60]):

$$\mathbf{x}_Q = \mathbf{x}_H \oslash (\mathbf{x} - \mathbf{x}_H), \tag{22}$$

  where $\oslash$ denotes element-wise matrix division.

Background clutter was suppressed using a weighting mask $\mathbf{m}_F$, produced by feathering the mean face outline $\mathbf{m}$:

$$\mathbf{m}_F = \mathbf{m} * \exp\left\{-\frac{r^2(x,y)}{8}\right\}. \tag{23}$$

This simple form of background suppression was adopted from [15] where it was successfully applied in the context of clustering of face appearance sets. A typical result of applying the mask is shown in Fig. 15.

4.3. Results

A summary of experimental results is shown in Table 2. The proposed algorithm greatly outperformed other methods, achieving nearly perfect recognition (99.7+%) on all four databases. This is an extremely high recognition rate for such unconstrained conditions (see Fig. 11), 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 Kullback–Leibler divergence-based 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 query [7]. Additionally, it is important to note the excellent performance of our algorithm on the Japanese database, even though off-line training was performed using Caucasian individuals only.

As expected, when plain likelihood was used instead of the robust version proposed in Section 3.3, the recognition rate was lower, but still significantly higher than that of other methods. The high performance of non-robust G-SIM is important as an estimate of the expected recognition rate in the "still-to-video" scenario of the proposed method. We conclude that our algorithm's performance seems very promising in this setup as well. An inspection of the Receiver–Operator Characteristic curves of the two methods in Fig. 16(a) shows an even more drastic improvement.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

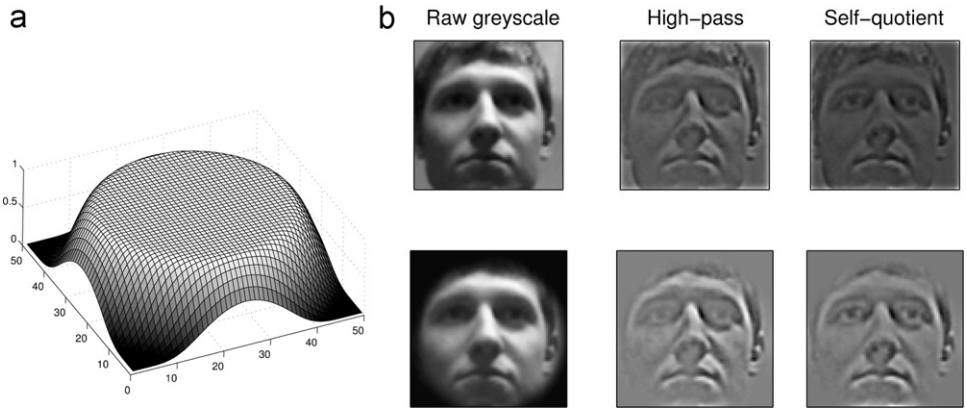


Fig. 15. (a) The weighting mask used to suppress background clutter. (b) The three face representations used in evaluation, shown as images, before (top row) and after (bottom row) the weighting mask was applied.

Table 2
Average recognition rates (%) and their standard deviations (where applicable).

	G-SIM, rob.	G-SIM	FaceIt	C-CCA	CCA	G-KLD
CamFace						
$\mathbf{x}$	**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
ToshFace						
$\mathbf{x}$	**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
FaceVideo						
$\mathbf{x}$	**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
Faces96						
$\mathbf{x}$	**100.0**	100.0	94.1	100.0	90.9	51.0
$\mathbf{x}_H$	–	–	–	100.0	94.0	27.8
$\mathbf{x}_Q$	–	–	–	100.0	99.3	28.5

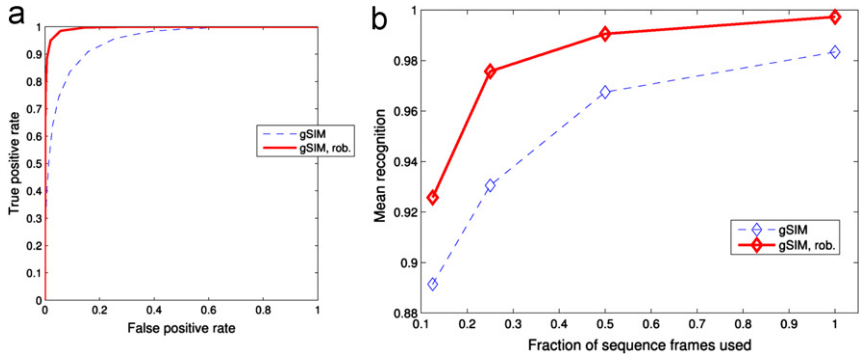


Fig. 16. (a) The Receiver-Operator Characteristic (ROC) curves of the G-SIM method, with and without the robust likelihood proposed in Section 3.3 estimated from CamFace and ToshFace, and (b) the variation in the mean recognition rate across the two data sets as a function of the amount of available data used for training and querying the algorithm.

This is an insightful observation: it shows that the use of the proposed robust likelihood yields less variation in the estimated similarity between individuals across different sequences.

To see how our algorithm copes with a progressively decreasing number of poses as well as a reduced density of face appearance manifold samples, we repeated the experiments, this time using only the first 50%, 25% and 12.5% of frames in each video sequence. The result, summarized in Fig. 16(b), shows a very gradual degradation of recognition performance. Robust G-SIM consistently achieved a higher recognition rate than non-robust G-SIM, its performance also decaying more slowly as the amount of training data is reduced. Even when only the first 12.5% of sequence frames are used (i.e. on average about 11 frames per sequence), it correctly recognized in 92.5% of the cases.

Finally, note that the standard deviation of our algorithm's performance across different training and query illuminations is much lower than that of other methods, showing less dependency on the exact imaging conditions used for data acquisition.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

4.3.1. Representations

Both the high-pass and even further Self Quotient Image representations generally produced an improvement in recognition accuracy over raw grayscale. This is consistent with previous findings in the published literature [14,36,61,62]. Performance degraded only in the case of *Faces96* data set and Gaussian appearance models matched using Kullback–Leibler divergence. The likely reason for this stems from a low number of faces per sequence (usually 10) available both for training and as a query in this data set, limiting the robustness of the corresponding probability density function estimates in the presence of noise amplified by filtering, as explained in further detail next.

In contrast to previous empirical studies of image filters in face recognition, in this paper we also sought to investigate the source of variation in the discriminative gain achieved with their use. 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/query data combination. Then the vector $\Delta\mathbf{s}_R \equiv \mathbf{s}_R - \bar{\mathbf{s}}_R$ contains relative recognition rates to its average on raw input, and $\Delta\mathbf{s} \equiv \mathbf{s}_F - \mathbf{s}_R$ the improvement with a particular 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 $\phi \approx 136^\circ$.

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 [65] for a thorough discussion). Higher than average recognition rates for raw input correspond to small changes in imaging conditions between training and query, and hence lower energy of extrinsic variation. In this case the training and query data sets are already normalized to have the same illumination and the two filters can only decrease the signal-to-noise ratio, thereby worsening the recognition performance. On the other hand, when the imaging conditions between training and query are very different, normalization of extrinsic variation is the dominant factor and the performance is improved, as illustrated in Fig. 17.

This is an important observation, as it suggests that the performance of a method that uses either of the representations can be increased further in a very straightforward manner by detecting the difficulty of recognition conditions. This is exploited in [60].

4.3.2. Imaging conditions

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 [66]. 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 $\rho = 0.96$.

4.3.3. Re-illumination and topological similarity of manifolds

The presented empirical analysis indirectly but strongly supports the validity of principles underlying the proposed recognition method. As a particularly interesting novelty in our approach, we pursued a more detailed and direct examination of the implicit assumption of topological similarity of face appearance and the corresponding pose-signature manifolds, used for sequence re-illumination in Section 2.1 and, specifically, Eq. (9).

For each face $\mathbf{x}_i^{(1)}$ in a sequence, we first computed its optimal reconstruction as a linear combination of its K-nearest neighbours $\sum_{k=1}^{K} \alpha_k \mathbf{x}_{n(i,k)}^{(1)}$, in a manner similar to that in Eq. (9), and measured the relative reconstruction error e_{IM} in the image space

$$e_{\mathrm{IM}} = \frac{\left\| \mathbf{x}_i^{(1)} - \sum_{k=1}^{K} \alpha_k \mathbf{x}_{n(i,k)}^{(1)} \right\|_2}{\|\mathbf{x}_i^{(1)}\|_2}. \tag{24}$$

The same coefficients $\alpha_1, \ldots, \alpha_K$ were then used to hypothesize a reconstruction of the corresponding pose-signature and thus the reconstruction error e_{SG} in the pose-signature space:

$$e_{\mathrm{SG}} = \frac{\left\| \mathbf{X}_i^{(1)} - \sum_{k=1}^{K} \alpha_k \mathbf{X}_{n(i,k)}^{(1)} \right\|_2}{\|\mathbf{X}_i^{(1)}\|_2}. \tag{25}$$

The pose-signature reconstruction error was found to be, quite expectedly, somewhat higher than that of the corresponding appearance, but consistently of the same order of magnitude. This is illustrated in Fig. 18.

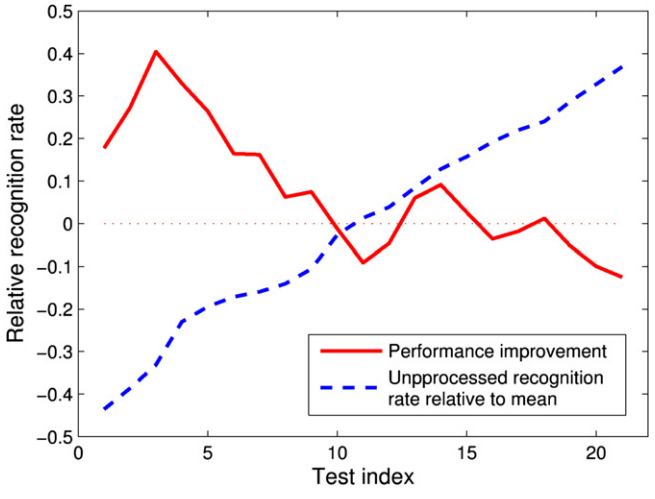


Fig. 17. Shown is the measured recognition performance improvement with high-pass and quotient image filters against the performance of unprocessed, raw imagery across different illumination combinations used in training and query. The two strongly negatively correlate. Queries are shown in the order of increasing raw data performance for easier visualization.

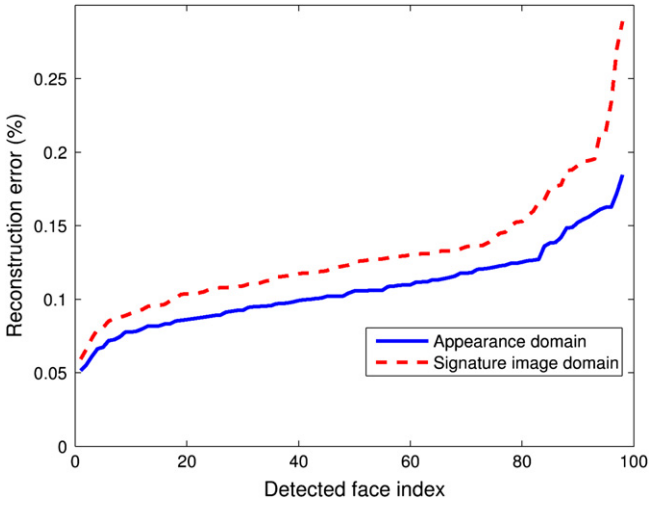


Fig. 18. Topological similarity of face motion and pose-signature manifolds was evaluated by comparing the reconstruction error of a face as a linear combination of its K-nearest neighbours in the appearance space and the reconstruction error of the reconstruction of the corresponding pose-signatures, using the same linear combination of its neighbourhood.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

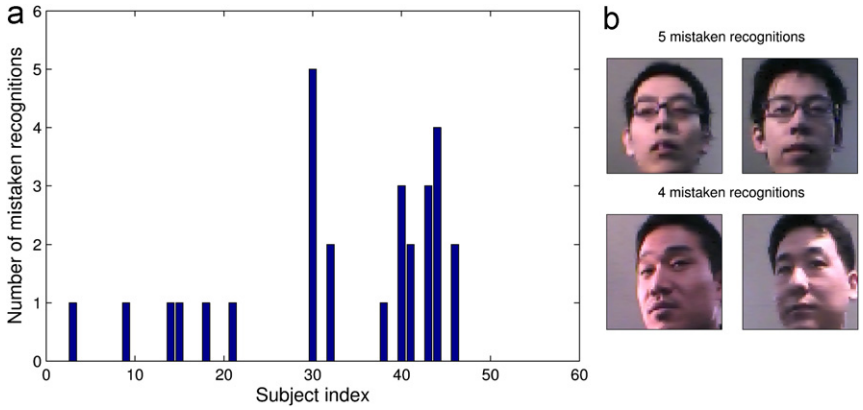


Fig. 19. (a) A histogram of non-robust G-SIM recognition failures across individuals in the ToshFace data set. The majority of errors are repeated, of which two of the most common ones are shown in (b). Visual inspection of these suggests that these individuals are indeed inherently similar in appearance.

4.3.4. Faces and individuals

Finally, in a similar manner as previously for different illumination conditions, we were interested to see if certain individuals were more difficult for recognition than others. In other words, are incorrect recognitions roughly equally distributed across the database, or does a relatively small number of people account for most? Our robust algorithm failed in too few cases to make a statistically significant observation, so we instead looked at the performance of the non-robust G-SIM which failed at about an order of magnitude greater frequency.

A histogram of recognition errors across individuals in ToshFace is shown Fig. 19(a), showing that most errors were indeed repeated. It is difficult to ascertain if this is a consequence of an inherent similarity between these individuals or a modelling limitation of our algorithm. A subjective qualitative inspection of the individuals most commonly confused, shown in Fig. 19(b), tends to suggest that the former is the dominant cause.

4.4. Computational complexity

The proposed method consists of two complementary algorithms. The first considers the problem of one-time off-line learning of the shape-illumination effects for human faces, while the second one concerns the application of the learnt model in assigning the identity to a novel face image set. It is the performance of the latter which is critical in practice so in this section we focus solely on its analysis.

The application of the shape-illumination invariant comprises the following:

1. K-nearest neighbour computation for each face image,
2. geodesic distance estimation between all pairs of images,
3. pose correspondence optimization using a GA,
4. re-illumination based on pose correspondence, and
5. robust computation of likelihood of the same identity.

We use the following notation: N is the number of face images in a set, K the number of neighbourhood faces used for re-illumination, N_{gen} the maximal number of generations in the genetic algorithm iteration, N_{chr} the number of chromosomes in each generation and N_{comp} the number of Gaussian components in the mixture capturing the distribution of shape-illumination effects $\mathcal{G}(\mathbf{d};\boldsymbol{\Theta})$.

4.4.1. Asymptotic complexity

In *step*1, to determine the exact K-nearest neighbourhood of each face in a set, the distance to all other faces must be computed – which requires exactly N comparisons – and the result ordered to find the smallest K, which has the computational load proportional to NK. Thus, the entire process is $\Theta(N^2K)$.[8]

In *step*2, the estimation of geodesic distances involves the initialization of elementary distances within all K-neighbourhoods, which is $\Theta(NK)$, and an application of Floyd's algorithm [41], which is $\Theta(N^3)$. As $K \ll N$, the complexity of *step*2 is the same as that of Floyd's algorithm.

In a generation of the genetic algorithm applied in *step*3, the computation of the similarity between pairs of matching pose-signatures given for every chromosome is $\Theta(N)$. The look-up of geodesic distances in all K-neighbourhoods (required for imposing the smoothness constraint) is $\Theta(NK)$. The total complexity of *step*3 is thus $\Theta(N_{gen}N_{chr}NK)$.

*Step*4 refines re-illumination results using the pose matching pairs estimated by the genetic algorithm and their K-neighbourhoods. It consists of a single $K \times K$ matrix inversion for each of K images in a set, giving the total complexity of $O(NK^3)$.

Finally, in *step*5, the likelihoods corresponding to all face images are computed in $\Theta(N_{comp}N)$, which are then ordered in further average $\Theta(N \log N)$ time. In principle, N_{comp} is a constant determined by the nature of illumination-shape effects (or, in practice, by fitting an optimal mixture to the sample from the corresponding distribution), so the complexity of *step*5 simplifies to $O(N \log N)$.

Treating everything but N as a constant, the overall asymptotic complexity of the algorithm is $\Theta(N^3)$. A summary is presented in Fig. 20(a).

4.4.2. Empirical performance

We next profiled an implementation of the algorithm written in Matlab on an Intel Pentium 4 PC, with a 3.2 GHz CPU and 2GB RAM. In all experiments only N, the number of faces per set, was varied. Sets of sizes 25, 50, 100, 200, 400 and 800 were used. Mean computation times for different stages of the matching algorithm (estimated from 100 executions of independently drawn identity-illumination combinations for the sets matched) are shown in Fig. 20(b). In this range of N, the measured asymptote slopes were typically lower than predicted, which was especially noticeable for the most demanding computation of geodesic distances. The most likely reason for this phenomenon is the presence of large proportionality constants, associated with Matlab's *for*-loops and data allocation routines.

[8] We use the standard notation whereby a function $f(N)$ is said to be $O(g(N))$ if and only if $\exists r > 0, N_0 \forall N > N_0 |f(N)| \leq |r \cdot g(N)|$. Similarly, $f(N)$ is said to be $\Theta(g(N))$ if and only if $\exists r_1, r_1 > 0, N_0 \forall N > N_0 |r_1 \cdot g(N)| < |f(N)| < |r_2 \cdot g(N)|$.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

a

Algorithm stage	Complexity
K-nearest neighbours	$\Theta(N^2 K)$
geodesic distances	$\Theta(N^3)$
genetic algorithm	$\Theta(N_{gen} N_{chr} N K)$
fine re-illumination	$O(N K^3)$
robust likelihood	$O(N \log N)$

b

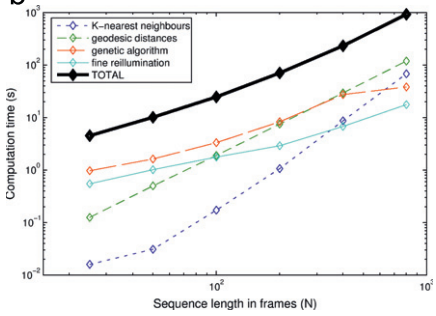


Fig. 20. (a) Asymptotic complexity of different stages of the proposed online, novel sequence recognition and (b) the measured times of our Matlab implementation.

5. Summary and conclusions

In this paper we described a novel algorithm for face recognition that uses video to achieve invariance to illumination, pose and user motion pattern variation. We introduced the concept of the generic shape-illumination manifold as a model of illumination effects on faces and showed how it can be learnt off-line from a small training corpus. This was made possible by the proposed "re-illumination" algorithm which is used extensively both in the off-line and online stages of the method.

Our method was demonstrated to achieve a nearly perfect recognition on four databases containing extreme variation in acquisition conditions. It was compared to and has significantly outperformed state-of-the-art commercial software and methods in the literature. Furthermore, an analysis of a large-scale performance evaluation (i) showed that the method is promising for image-to-sequence matching, (ii) suggested a direction of research to improve image filtering for illumination invariance, and (iii) confirmed that certain illuminations and individuals are inherently particularly challenging for recognition.

There are several avenues for future work that we would like to explore. Firstly, we would like to make further use of off-line training data, by constructing the G-SIM while taking into account probabilities of both intra- and inter-personal differences. Additionally, we would like to improve the computational efficiency of the method, e.g. by representing each FMM by a strategically chosen set of sparse samples. Finally, we are evaluating the performance of image-to-sequence matching and looking into increasing its robustness, in particular to pose.

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] P.J. Phillips, K.W. Bowyer, P.J. Flynn, A.J. O'Toole, W.T. Scruggs, C.L. Schott, M. Sharpe, Face Recognition Vendor Test 2006 and Iris Challenge Evaluation 2006 Large-Scale Results, DIANE Publishing Company, March 2007.
[3] K. Lee, D. Kriegman. Online learning of probabilistic appearance manifolds for videobased recognition and tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, pp. 852–859.
[4] K. Lee, M. Ho, J. Yang, D. Kriegman. Video-based face recognition using probabilistic appearance manifolds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, June 2003, pp. 313–320.
[5] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (1) (2003) 214–245.
[6] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell. Face recognition with image sets using manifold density divergence, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, June 2005, pp. 581–588.
[7] G. Shakhnarovich, J.W. Fisher, and T. Darrel. Face recognition from long-term observations, in: Proceedings of the European Conference on Computer Vision (ECCV), vol. 3, May–June 2002, pp. 851–868.
[8] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, International Symposium of Robotics Research, 2003.
[9] O. Arandjelović, Recognition from appearance subspaces across image sets of variable scale, in: Proceedings of the IAPR British Machine Vision Conference (BMVC), September (2010).
[10] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, Journal of Machine Learning Research (JMLR) 4 (10) (2003) 913–931.
[11] T-K. Kim, O. Arandjelović, R. Cipolla, Boosted manifold principal angles for image set-based recognition, Pattern Recognition 40 (September(9)) (2007) 2475–2484.
[12] Y. Li, S. Gong, H. Liddell, Modelling faces dynamically across views and over time, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 1, 2001, pp. 554–559.
[13] 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.
[14] 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 (TPAMI) 19 (7) (1997) 721–732.
[15] O. Arandjelović and R. Cipolla. Automatic cast listing in feature-length films with anisotropic manifold space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, June 2006, pp. 1513–1520.
[16] M. Everingham, A. Zisserman, Automated person identification in video, in: Proceedings of the IEEE International Conference on Image and Video Retrieval CIVR, 2004, pp. 289–298.
[17] J. Sivic, M. Everingham, A. Zisserman, Person spotting: video shot retrieval for face sets, in: Proceedings of the IEEE International Conference on Image and Video Retrieval (CIVR), 2005, pp. 226–236.
[18] 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 (July/August(3)) (1998) 245–260.
[19] A.S. Georghiades, D.J. Kriegman, P.N Belhumeur, Illumination cones for recognition under variable lighting: Faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, 1998, pp. 52–58.
[20] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Transactions on Pattern Analysis and Machine Intelligence TPAMI 25 (September(9)) (2003) 1063–1074.
[21] 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2005, pp. 206–216.
[22] L. Zhang, D. Samaras, Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 28 (March(3)) (2006) 351–363.
[23] 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 (TPAMI) 23 (2) (2001) 219–239.
[24] B. Kepenekci, Face Recognition using Gabor Wavelet Transform, Ph.D. Thesis, The Middle East Technical University, 2001.
[25] 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.
[26] X. Liu, T. Chen. Video-based face recognition using adaptive hidden Markov models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2003, pp. 340–345.
[27] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, 1994, pp. 84–91.
[28] O. Arandjelović, R. Cipolla, An illumination invariant face recognition system for access control using video, in: Proceedings of the IAPR British Machine Vision Conference (BMVC), September 2004, pp. 537–546.
[29] L. Zhang, S. Shan, X. Chen, W. Gao, Histogram of Gabor phase patterns (HGPP): a novel object representation approach for face recognition, IEEE Transactions on Image Processing 16 (1) (2007) 57–68.
[30] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, H.-J. Zhang, Multilinear discriminant analysis for face recognition, IEEE Transactions on Image Processing 16 (1) (2007) 212–220.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024

[31] X. Jiang, B. Mandal, A. Kot, Extending the feature vector for automatic face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 30 (3) (2008) 383–394.
[32] M.D. Levine, Y. Yu, State-of-the-art of 3D facial reconstruction methods for face recognition based on a single 2D training image per person, Journal of the Korea Information Science Society 30 (10) (2009) 908–913.
[33] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Computation 11 (2) (1999) 443–482.
[34] 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.
[35] O. Arandjelović, R. Cipolla, An information-theoretic approach to face recognition from face motion manifolds, Image and Vision Computing (special issue on Face Processing in Video) 24 (June(6)) (2006) 639–647.
[36] O. Arandjelović, A. Zisserman. Automatic face recognition for film character retrieval in feature-length films, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, June 2005, pp. 860–867.
[37] 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004, pp. 848–854.
[38] D.M. Gavrila. Pedestrian detection from a moving vehicle, in: Proceedings of the European Conference on Computer Vision (ECCV), vol. 2, 2000, pp. 37–49.
[39] B. Stenger, A. Thayananthan, P.H.S. Torr, R. Cipolla. Filtering using a tree-based estimator, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, 2003, pp. 1063–1070.
[40] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wily & Sons Inc., New York, 2nd edition, 2000.
[41] T.H. Cormen, C.E. Leiserson, R.L. Rivest, Introduction to Algorithms, MIT Press, June 1990.
[42] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[43] R. Kimmel, M. Elad, D. Shaked, R. Keshet, I. Sobel, A variational framework for retinex, International Journal of Computer Vision (IJCV) 52 (1) (2003) 7–23.
[44] I. Craw, N.P. Costen, T. Kato, S. Akamatsu, How should we represent faces for automatic recognition? IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 21 (1999) 725–736.
[45] R. Gross, I. Matthews, S.Baker, Generic vs. person specific active appearance models, in: Proceedings of the IAPR British Machine Vision Conference (BMVC), 2004, pp. 457–466.
[46] R. Gross, J. Yang, A. Waibel. Growing Gaussian mixture models for pose invariant face recognition, in: Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), vol. 1, September 2000, pp. 1088–1091.
[47] X. Wang, X. Tang. Bayesian face recognition based on Gaussian mixture models, in: Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), vol. 4, 2004, pp. 142–145.
[48] O. Arandjelović, R. Cipolla, A pose-wise linear illumination manifold model for face recognition using video, Computer Vision and Image Understanding (CVIU) 113 (January(1)) (2009) 113–125.
[49] O. Arandjelović, R. Cipolla. Incremental learning of temporally-coherent Gaussian mixture models, in: Proceedings of the IAPR British Machine Vision Conference (BMVC), vol. 2, September 2005, pp. 759–768.
[50] P. Hall, D. Marshall, R. Martin, Merging and splitting eigenspace models, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22 (September(9)) (2000) 1042–1048.
[51] P. Grünwald,The Minimum Description Length Principle, MIT Press, June 2007.
[52] O. Arandjelović, Computationally efficient application of the generic shape-illumination invariant to face recognition from video, Pattern Recognition 45 (January(1)) (2012) 92–103.
[53] D.O. Gorodnichy, Associative neural networks as means for low-resolution video-based recognition, in: Proceedings of the International Joint Conference on Neural Networks, 2005.
[54] P. Viola, M. Jones, Robust real-time face detection, International Journal of Computer Vision (IJCV) 57 (May(2)) (2004) 137–154.
[55] Identix Ltd. Faceit, 〈http://www.FaceIt.com/〉.
[56] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, J.M. Bone, FRVT 2002: Overview and summary, Technical Report, National Institute of Justice, March (2003).
[57] J. Heo, B. Abidi, J. Paik, M. A. Abidi. Face recognition: evaluation report for FaceItr, in: Proceedings of the International Conference on Quality Control by Artificial Vision, 5132:551–558, 2003.
[58] K. Maeda, O. Yamaguchi, K. Fukui, Towards 3-dimensional pattern recognition, Statistical Pattern Recognition August 3138 (2004) 1061–1068.
[59] Toshiba, Facepass, 〈www.toshiba.co.jp/mmlab/tech/w31e.htm〉.
[60] O. Arandjelović, R. Cipolla, A methodology for rapid illumination-invariant face recognition using image processing filters, Computer Vision and Image Understanding (CVIU) 113 (February(2)) (2009) 159–171.
[61] A. Fitzgibbon and A. Zisserman, On affine invariant clustering and automatic cast listing in movies, in: Proceedings of the European Conference on Computer Vision ECCV, 2002, pp. 304–320.
[62] H. Wang, S.Z. Li, Y. Wang, Face recognition under varying lighting conditions using self quotient image, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR), May 2004, pp. 819–824.
[63] M. Nishiyama, O. Yamaguchi, Face recognition using the classified appearance-based quotient image, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR), 2006, pp. 49–54.
[65] X. Wang, X. Tang. Unified subspace analysis for face recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 1, 2003, pp. 679–686.
[66] T. Sim, S. Zhang, Exploring face space, in: Proceedings of the IEEE Workshop on Face Processing in Video, 2004, p. 84.

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 science. He is a Fellow of the Cambridge Overseas Trust and a winner of multiple best research paper awards.

Please cite this article as: O. Arandjelović, R. Cipolla, Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant, Pattern Recognition (2012), http://dx.doi.org/10.1016/j.patcog.2012.06.024