



A methodology for rapid illumination-invariant face recognition using image processing filters

Ognjen Arandjelović^{a,*}, Roberto Cipolla^b

^aTrinity College, University of Cambridge, Trinity Street, Cambridge CB2 1TQ, UK

^bDepartment of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK

ARTICLE INFO

Article history:

Received 20 February 2007

Accepted 11 June 2008

Available online 5 November 2008

Keywords:

Face recognition
Illumination
Invariance
Filters
Video
Image processing

ABSTRACT

Achieving illumination invariance in the presence of large pose changes remains one of the most challenging aspects of automatic face recognition from low resolution imagery. In this paper, we propose a novel recognition methodology for their robust and efficient matching. The framework is based on outputs of simple image processing filters that compete with unprocessed greyscale input to yield a single matching score between two individuals. Specifically, we show how the discrepancy of the illumination conditions between query input and training (gallery) data set can be estimated implicitly and used to weight the contributions of the two competing representations. The weighting parameters are representation-specific (i.e. filter-specific), but not gallery-specific. Thus, the computationally demanding, learning stage of our algorithm is offline-based and needs to be performed only once, making the added online overhead minimal. Finally, we describe an extensive empirical evaluation of the proposed method in both a video and still image-based setup performed on five databases, totalling 333 individuals, over 1660 video sequences and 650 still images, containing extreme variation in illumination, pose and head motion. On this challenging data set our algorithm consistently demonstrated a dramatic performance improvement over traditional filtering approaches. We demonstrate a reduction of 50–75% in recognition error rates, the best performing method-filter combination correctly recognizing 97% of the individuals.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

In this paper, we are interested in the problem of accurately recognizing human faces in the presence of large and unpredictable illumination changes. Our aim is to do this in a setup realistic for most practical applications, that is, without overly constraining the conditions in which data is acquired. Most often, this means that the amount of available training data is limited and the image quality (spatial resolution) low.

In conditions such as these, invariance to changing lighting is perhaps the most significant practical challenge for face recognition algorithms. The illumination setup in which recognition is performed is in most cases impractical to control, its physics difficult to accurately model and recover, with face appearance differences due to varying illumination often larger in magnitude than those differences between individuals [1]. Additionally, the nature of most real-world applications is such that prompt, often real-time system response is needed, demanding appropriately efficient as well as robust matching algorithms.

In this paper, we describe a novel framework for rapid recognition under varying illumination, based on simple image-filtering

techniques. The proposed methodology is very general and we demonstrate that it offers a dramatic performance improvement when used with a wide range of filters and different baseline matching algorithms, without sacrificing their computational efficiency. The framework is based around two parallel pipelines: one matching unprocessed input imagery, the other matching filtered data. These are fused *at the decision level*, relative contributions of the two representations being conditioned on the similarity of illumination conditions between the compared data sets. By formulating the problem of estimating this similarity in a discriminative manner, the entirety of online computation is performed in the closed-form, making the proposed sequence matching extremely efficient.

1.1. Paper organization

The remainder of this paper is organized as follows: in the next section, we review relevant previous work and emphasize its key limitations that are addressed by our work; an overview of the use of image processing filters in face recognition is given in Section 2.1. Section 3 describes each of the main components of the proposed system in detail: the main premise of the paper is introduced in Section 3 and the proposed solution in Sections 3.1 and 3.1.1. Our empirical evaluation methodology, data sets used and the performance of the proposed algorithm are reported and

* Corresponding author.

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

discussed in Section 4. The paper is concluded with a summary and an outline of promising directions for future research in Section 5.

2. Relevant previous work

Illumination-invariant face recognition is a very active research area, as witnessed by the diversity in the approaches proposed in the literature. It is out of scope of this paper to present a detailed review of the whole body of work on this topic. Instead, we focus on a number of methods that typify the most influential research directions and, indeed, their main limitations that we seek to improve on. For recent surveys of the face recognition field, one may start from [2,3].

Representations in face recognition: The choice of representation, that is, the model used to describe a person's face is central to the problem of automatic face recognition. Consider the components of a generic face recognition system schematically shown in Fig. 1.

A number of influential approaches in the literature employ suitably complex, *generative* facial and scene models that allow for explicit separation of extrinsic and intrinsic variables which affect the observed appearance. The appeal of these methods is clear: after intrinsic face parameters are estimated, their classification to one of the known classes is typically straightforward. On the other hand, although a rather diverse variety of models has been proposed, in most cases their inherent complexity makes parameter recovery impossible using a closed-form expression (“*Model parameter recovery*” in Fig. 1). Rather, fitting is performed through an iterative optimization scheme. A large subclass of this group are 3D model-based approaches, epitomized by the *3D Morphable Model* of Blanz and Vetter [4–6]. The shape and texture of a query face are recovered through gradient descent by minimizing the discrepancy between the observed and predicted appearance. Gradient descent is also employed in a rather different, sparse approach of *Elastic Bunch Graph Matching* [7–9], in which it is the placements of fiducial features, corresponding to bunch graph nodes, and the locations of local texture descriptors that are estimated. In contrast, the *Generic Shape-Illumination Manifold* method operates on the appearance manifold level, using a genetic algorithm to estimate a manifold-to-manifold mapping that preserves pose [10]. Another popular group are photometric stereo approaches [11–16]. While the early applications of the “illumination cone constraint” [11] on the appearance of a single Lambertian surface were limited by training data requirements, promising results are reported by the more recent, generalized photometric stereo methods [13,14,16] which in addition exploit class-specific constraints of face shape and albedo.

One of the main limitations of most generative-model group of methods arises due to the existence of local minima in the model

fitting stage, of which there are usually many [17]. The key problem is that if the estimated model parameters correspond to a local minimum, classification is performed not merely on noise-contaminated but rather entirely *incorrect* data. An additional unappealing feature of these methods is that it is also not possible to determine if model fitting had failed in such a manner.

The alternative approach is to employ a simple face appearance model and put greater emphasis on the classification stage whereby illumination is effectively learnt through a discriminative rather than generative framework. This general direction has several advantages which make it attractive from a practical standpoint. First, model parameter estimation can now be performed as a closed-form computation, which is not only more efficient, but also void of the issue of fitting failure such that can happen in an iterative optimization scheme. This allows for more powerful statistical classification, thus clearly separating well understood and explicitly modelled stages in the image formation process, from those that are more easily learnt implicitly from training exemplars. This is the methodology followed in this paper.

2.1. Image processing filters

Most relevant to the material presented in this paper are illumination-normalization methods that can be broadly described as quasi illumination-invariant *image filters*. Amongst the most used ones in the face recognition literature are high-pass [18] and locally-scaled high-pass filters [19], directional derivatives [1] [20–22], Laplacian-of-Gaussian filters [1], region-based gamma intensity correction filters [23] [24], edge-maps [1] and various wavelet-based filters [25–27]. These are most commonly based on simple image formation models, for example modelling illumination as a spatially low-frequency band of the Fourier spectrum and identity-based information as high-frequency [18,28], see Fig. 2. Methods of this group can be applied in a straightforward manner to either single or multiple-image face recognition and are often extremely efficient. However, due to the simplistic nature of the underlying models, in general they do not perform well in the presence of extreme illumination changes [1].

While developing faster, more robust and discriminative filters is still an area of active ongoing research [29], the primary focus of this paper is different and can be seen as complementary. In essence, the question we are asking is if one can do better with *existing* filters by a more careful use of all the available data.

3. Proposed method details

The framework proposed in this paper is motivated by our previous research and the findings first published in [10]. Four face recog-

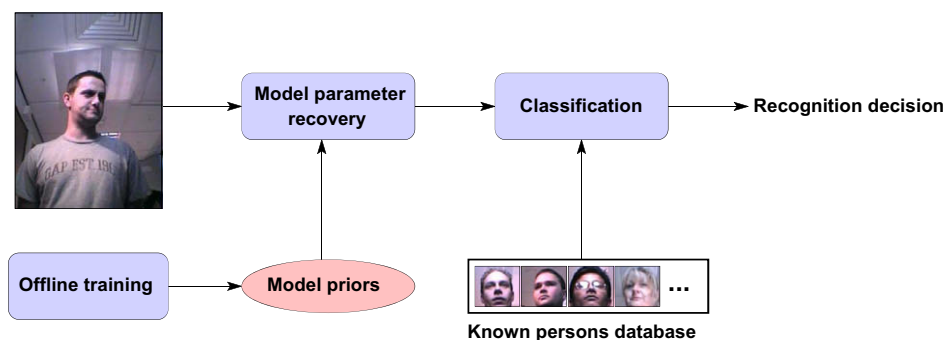


Fig. 1. A diagram of the main components of a generic face recognition system. The “Model parameter recovery” and “Classification” stages can be seen as mutually complementary: (i) a complex model that explicitly separates extrinsic (pose, illumination,...) and intrinsic (shape and albedo) appearance variables places most of the workload on the former stage, while the classification of the representation becomes straightforward; in contrast, (ii) simplistic models have to resort to more statistically sophisticated approaches to matching.

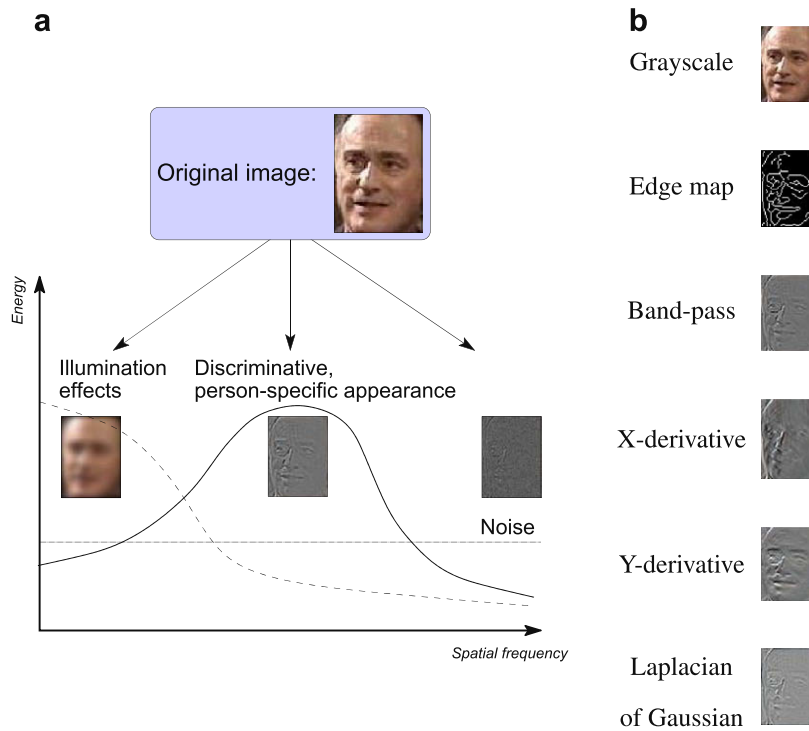


Fig. 2. (a) One of the simplest generative model used for face recognition: images are assumed to consist of the low-frequency band that mainly corresponds to illumination changes, mid-frequency band which contains most of the discriminative, personal information and white noise. (b) The results of several most popular image filters operating under the assumption of the frequency model.

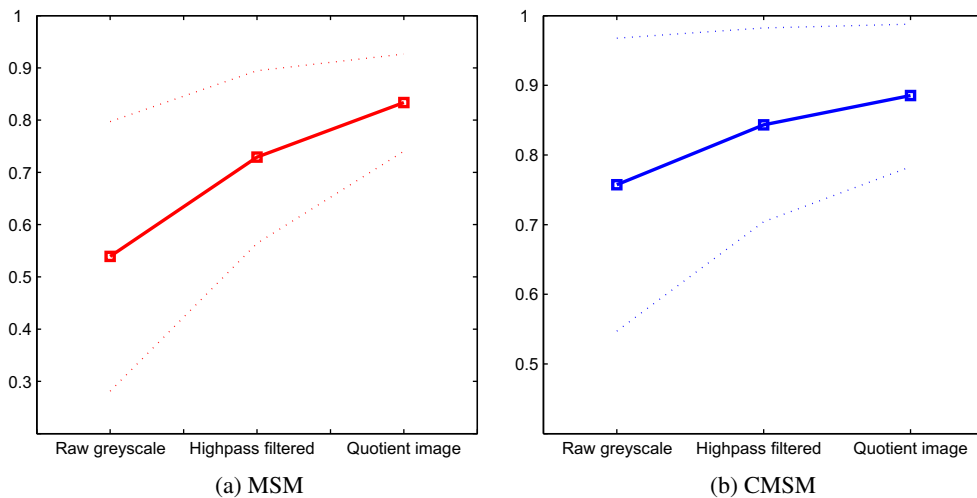


Fig. 3. Performance of the (a) Mutual Subspace Method and the (b) Constrained Mutual Subspace Method using raw greyscale imagery, high-pass (HP)-filtered imagery and the Self-Quotient Image (QI), evaluated on over 1300 video sequences with extreme illumination, pose and head motion variation (as reported in [10]). Shown are the average performance and \pm one standard deviation intervals.

nition algorithms, the *Generic Shape-Illumination* method [10], the *Constrained Mutual Subspace Method* [30], the commercial system *Facelt* and a *Kullback–Leibler Divergence*-based matching method, were evaluated on a large database using (i) raw greyscale imagery, (ii) high-pass (HP) filtered imagery and (iii) the Self-Quotient Image (QI) representation [19]. Both the high-pass and even further Self-Quotient Image representations produced an improvement in recognition for all methods over raw grayscale, as shown in Fig. 3, which is consistent with previous findings in the literature [1,18,28,19].

Of importance to this work is that it was also examined in which cases these filters help and how much depending on the

data acquisition conditions. It was found that recognition rates using greyscale and either the HP or the QI filter negatively correlated (with $\rho \approx -0.7$), as illustrated in Fig. 4. This finding was observed consistently across the result of the four algorithms, all of which employ mutually drastically different underlying models.

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 when no normalization is actually needed. The observed phenomenon is well understood in the context of energy of intrinsic and extrinsic image differences and noise (see [31] for a thorough discussion). Higher than average

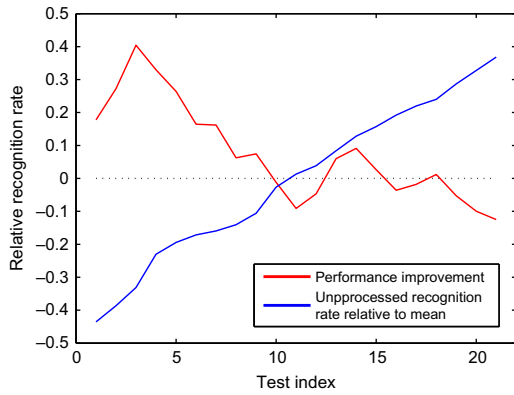


Fig. 4. A plot of the performance improvement with HP and QI filters against the performance of unprocessed, raw imagery across different illumination combinations used in training and test. The tests are shown in the order of increasing raw data performance for easier visualization.

recognition rates for raw input correspond to small changes in imaging conditions between training and test, and hence lower energy of extrinsic variation. In this case, filters cannot increase signal-to-noise ratio, and generally decrease it, worsening the algorithm performance, see Fig. 5(a). On the other hand, when the imaging conditions between training and test data are very different, normalization of extrinsic variation is the dominant filtering effect and performance is improved, see Fig. 5(b).

This is an important observation: it suggests that the performance of a method that uses either of the representations can be increased if the extent of change of illumination conditions

between new data and training data is known. In this paper, we propose a novel, learning-based framework to do this.

3.1. Adaptive framework

In this section, our goal is to first extract information on the change in illumination conditions in which data was acquired, and then use it to optimally exploit raw and filtered imagery in casting the recognition decision. Explicit recovery of the lighting setup description is difficult: the space of parameters is generally very large (illumination sources can vary in number, position, direction and type) and their interaction with potentially complex scene surface properties make the problem ill-posed. Instead, we propose to infer the change, or rather the magnitude of its effects on the observed appearance, directly from image-based comparisons.

Let $\{\mathcal{X}_1, \dots, \mathcal{X}_N\}$ be a database of known individuals, \mathcal{X}_0 the query input corresponding to one of the gallery classes, and $\rho(\mathcal{X}_0, \mathcal{X}_i) \in [0, 1]$ and $F(\mathcal{X}_i)$, respectively, a given similarity function and a quasi illumination-invariant filter. Note that we make no assumptions on the nature of training or test data. Each \mathcal{X}_i can either be a single image, an image set or a sequence. Equally, we do not concern ourselves with the form of $\rho(\mathcal{X}_0, \mathcal{X}_i)$ – the aim of this paper is not to devise a better way of comparing two face images (say). Rather, given a way to perform such a comparison, we are interested in optimally combining appearance and filter-based similarities. However, we do restrict ourselves to considering only the problem of illumination invariance. As argued previously (see Section 1) this is a problem causing greater practical difficulties than varying pose. Thus, we assume that the robustness to pose is accomplished either by virtue of variability in data itself (e.g. as in [32]), or through a pose-invariant $\rho(\mathcal{X}_0, \mathcal{X}_i)$.

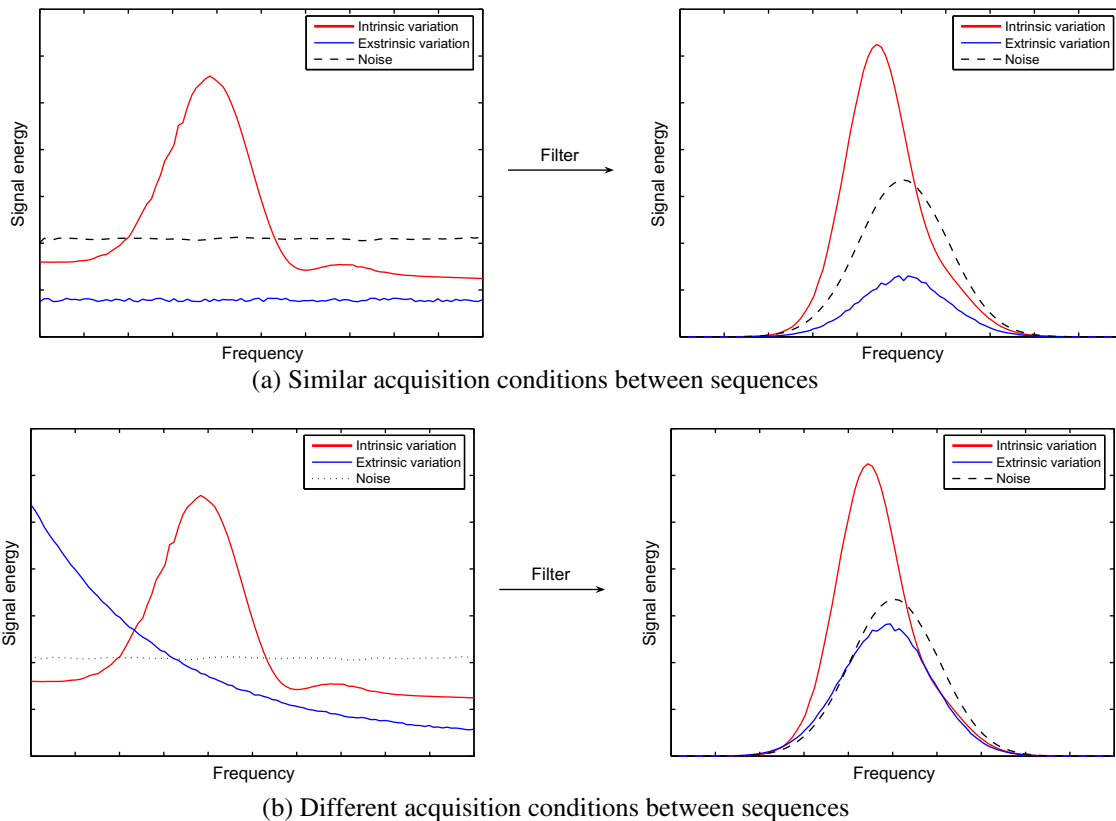


Fig. 5. A conceptual illustration of the distributions of intrinsic, extrinsic and noise signal energies across frequencies in the cases when training and test data acquisition conditions are (a) similar and (b) different, before (left) and after (right) band-pass filtering.

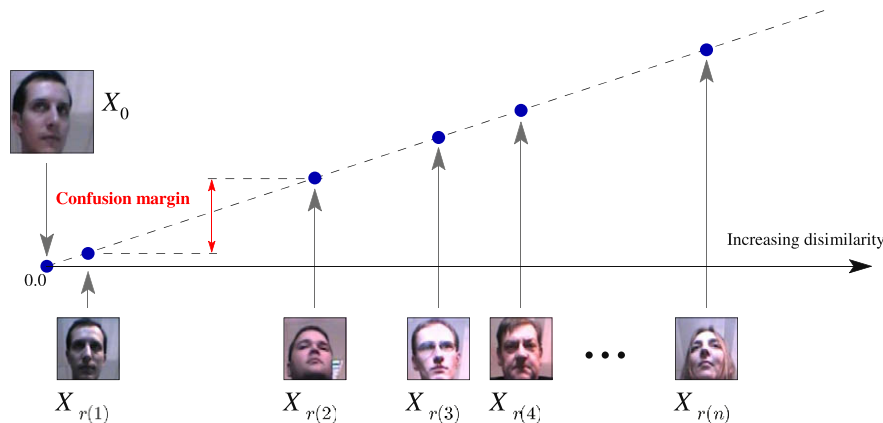


Fig. 6. Our offline algorithm implicitly accounts for the difference in illuminations conditions between the query and a gallery entry. This is done by learning the dependency of the optimal weighting of decisions based on unfiltered and filtered data, as a function of the *confusion margin*. Confusion margin, as illustrated, is defined as the matching confidence difference between the top two gallery matches given a query.

We propose to express the degree of belief $\eta(\mathcal{X}_0, \mathcal{X}_i) \in [0, 1]$ that two face sets \mathcal{X}_0 and \mathcal{X}_i belong to the same person as a weighted combination of similarities between the corresponding unprocessed and filtered data:

$$\eta(\mathcal{X}_0, \mathcal{X}_i) = \alpha^* \rho(\mathcal{X}_0, \mathcal{X}_i) + (1 - \alpha^*) \rho(F(\mathcal{X}_0), F(\mathcal{X}_i)) \quad (1)$$

In the light of the previous results and discussion, α^* ought to be large (closer to 1.0) when the query and the corresponding gallery data are acquired in similar illuminations, and small (closer to 0.0) when in very different ones. Thus, α^* is not a constant but rather a function of the similarity function ρ , filter F , as well as data. We show that α^* can be effectively learnt as

$$\alpha^* \equiv \alpha^*(\mu), \quad (2)$$

where μ is the *confusion margin*. We define the confusion margin as the difference between the similarities of the query data \mathcal{X}_0 and the two gallery individuals \mathcal{X}_i most similar to it. Formally, if $r(i)$ is the index of the i th best match for \mathcal{X}_0 using the similarity function ρ :

$$\mu = \rho(\mathcal{X}_0, \mathcal{X}_{r(1)}) - \rho(\mathcal{X}_0, \mathcal{X}_{r(2)}), \quad (3)$$

as illustrated in Fig. 6. Intuitively, given a sensible similarity function ρ , the confusion margin quantifies the confidence that the top ranking match using unprocessed imagery is indeed the correct one. To see why this is the case, let us consider the two extreme cases, when query data and the corresponding gallery data are acquired in (i) the same illumination conditions and (ii) extremely different illumination conditions.

(i) Same illumination conditions

In this case, the similarity of the query data and the corresponding gallery data is nearly perfect (close to 1.0) and the correct match by identity is retrieved first. We can then write $|1 - \rho(\mathcal{X}_0, \mathcal{X}_{r(1)})| \ll |1 - \rho(\mathcal{X}_0, \mathcal{X}_{r(2)})|$ and from the definition (3) it is clear that the confusion margin is large.

(ii) Extremely different illumination conditions

In this case, intrapersonal appearance changes due to illumination will be greater than interpersonal variations [1], and consequently neither of the top matches will correspond in identity to the query data. Thus, both $\rho(\mathcal{X}_0, \mathcal{X}_{r(1)})$ and $\rho(\mathcal{X}_0, \mathcal{X}_{r(2)})$ will be relatively small. In addition, as they are by definition constrained to lie in the domain $[0, 1]$, their difference must be small too.

While the values that weighting function α^* assumes at the extrema of the region on which it is defined are now clear, its behaviour between them is not as obvious. We address this issue next.

3.1.1. Learning the α -function

The value of $\alpha^*(\mu)$ can then be interpreted as statistically the optimal choice of the mixing coefficient α for the confusion margin μ . Formalizing this we can write

$$\alpha^*(\mu) = \arg \max_{\alpha} p(\alpha | \mu), \quad (4)$$

or, equivalently

$$\alpha^*(\mu) = \arg \max_{\alpha} \frac{p(\alpha, \mu)}{p(\mu)}. \quad (5)$$

Under the assumption of a uniform prior on the confusion margin, $p(\mu)$

$$p(\alpha | \mu) \propto p(\alpha, \mu), \quad (6)$$

and

$$\alpha^*(\mu) = \arg \max_{\alpha} p(\alpha, \mu). \quad (7)$$

Thus, the problem of inferring the optimal weighting function α^* is in fact reduced to the estimation of a 2D probability density function.

Proposed methodology: To learn the α -function $\alpha^*(\mu)$ as defined in (4), we first need an estimate $\hat{p}(\alpha, \mu)$ of the joint probability density $p(\alpha, \mu)$ as per (7). The main difficulty of this problem is of practical nature: in order to obtain an accurate estimate using one of the many off-the-shelf density estimation techniques, a prohibitively large training database would be needed to ensure a well sampled distribution. Instead, we propose a heuristic alternative which, we will show, allows us to do this from a small offline training corpus which comprises individuals imaged in various illumination conditions. It is important to emphasize that this data is used only for the purpose of estimating the α -function and has no relation (or rather, need not have any) to the set of gallery individuals. This logical separation of training data and the fact that offline training needs to be performed only once, means on the one hand (i) that we can collect an offline corpus containing representative lighting variation, while on the other (ii) allowing us to perform recognition with both gallery and query data acquired in a single arbitrary illumination condition each.

The key idea that makes it possible for us to accurately estimate $p(\alpha, \mu)$ while drastically reducing the amount of offline training

data, lies in the use of domain specific knowledge of the properties of $p(\alpha, \mu)$, employed to constrain the estimation process. Our algorithm is based on an iterative incremental update of the density, initialized as a uniform density over a discrete grid on the domain $\alpha, \mu \in [0, 1]$, as in Fig. 8. Then, using the offline data corpus, we iteratively simulate matching of an “unknown”, query person against a set of provisional gallery individuals which are randomly drawn from the offline training database.

An iteration: Specifically, in each iteration we randomly choose \mathcal{Y}_0 , i.e. a particular individual in a specific illumination which serves as a query, and similarly, a set of individuals $\{\mathcal{Y}_1, \dots, \mathcal{Y}_M\}$ which in the simulation have the role of gallery data. Using the similarity metric $\rho(\cdot)$ and the filter $F(\cdot)$, we can compute the similarity between \mathcal{Y}_0 and each \mathcal{Y}_i , which also gives us the value of the confusion margin μ , as well as the similarity between $F(\mathcal{Y}_0)$ and each $F(\mathcal{Y}_i)$. The combined similarity score $\eta(\mathcal{Y}_0, \mathcal{Y}_i)$ is then computed for each possible value of α or, remembering that alpha is restricted to a discrete grid $\alpha = k\Delta\alpha$, for each value of $k \in [0 \dots 1/\Delta\alpha]$:

$$\eta_i \equiv \eta(\mathcal{Y}_0, \mathcal{Y}_i) = k\Delta\alpha\rho(\mathcal{Y}_0, \mathcal{Y}_i) + (1 - k\Delta\alpha)\rho(F(\mathcal{Y}_0), F(\mathcal{Y}_i)) \quad (8)$$

Since the ground truth identities of all persons in the offline database are known, we can quantify how well a particular value of α performed by considering the resulting similarity η_c of \mathcal{Y}_0 and the correct match amongst the $\{\mathcal{Y}_i\}$, and the similarity $\eta_{r(1)}$ of \mathcal{Y}_0 and the individual actually deemed most similar in (8), $\mathcal{Y}_{r(1)}$. The greater the ratio $\eta_c/\eta_{r(1)}$, the better the choice of α is. Density $\hat{p}(\alpha, \mu)$ is then incremented proportionally to $\eta_c/\eta_{r(1)}$ to reflect higher confidence in the particular value of α for the observed μ .

The proposed offline learning algorithm is summarized using pseudo-code in Fig. 7 with a typical evolution of $\hat{p}(\alpha, \mu)$ shown in Fig. 8. The final stage of the offline learning in our method involves imposing the monotonicity constraint on $\alpha^*(\mu)$ and smoothing of the result, see Fig. 9.

On the fusion constraints: As a final theoretical point, we wish to discuss the space of functions that our fusion is constrained to by the Eq. (1) and the proposed algorithm for learning the mixing function α^* .

On the surface, (1) appears to be a linear combination of similarities of unprocessed and filtered data, $\rho(\mathcal{X}_0, \mathcal{X}_i)$ and $\rho(F(\mathcal{X}_0), F(\mathcal{X}_i))$. However, it is important to recognize that α^* is not constant but rather a function of the confusion margin μ , which itself is by definition dependent on $\rho(\mathcal{X}_0, \mathcal{X}_1), \dots, \rho(\mathcal{X}_0, \mathcal{X}_N)$. On the one hand, this observation makes rigorous theoretical analysis very difficult but, on the other, provides the functional flexibility that is necessary given that no assumptions are made on the form of ρ . Put differently, any monotonically non-decreasing function of ρ could be substituted for ρ , since this transformation would only affect the *the magnitude* of the confusion margin, but not the gallery ordering $r(i)$ in (3). As a consequence, assuming adequate training data and noting the non-parametric nature of the proposed algorithm for the estimation of α^* , our fusion algorithm would still converge to the effectively same solution for the fusion η .

4. Empirical evaluation

In the preceding sections we dealt with theoretical aspects of the proposed framework. To empirically test the main premises of our work and the effectiveness of the described methodology, we evaluated its performance on 1662 video sequences of head motion, as well as 650 still images, collected from five databases totalling 333 individuals:

Cambridge Face Database (CamFace), with 100 individuals of varying age and ethnicity, and equally represented genders. For each person in the database we collected 7 video sequences of the person in arbitrary motion (significant translation, yaw and

Input: training data $D(\text{person}, \text{illumination})$,
 filtered data $F(\text{person}, \text{illumination})$,
 similarity function ρ .

Output: estimate $\hat{p}(\alpha, \mu)$.

1: Initialization

$$\hat{p}(\alpha, \mu) = 0$$

2: Simulated matching iteration

for all illuminations i, j and persons p

3: Confusion margin

$$\mu = \rho(D(p, i), D(r_1, j)) - \rho(D(p, i), D(r_2, j))$$

4: Iteration

for all $k = 0, \dots, 1/\Delta\alpha$, $\alpha = k\Delta\alpha$

5: Quantify performance of α

$$\delta(k\Delta\alpha) = \frac{\alpha\rho(D(p,i), D(p,j)) + (1-\alpha)\rho(F(p,i), F(p,j))}{\max_{p \neq q} [\alpha\rho(D(p,i), D(q,j)) + (1-\alpha)\rho(F(p,i), F(q,j))]}$$

6: Update density estimate

$$\hat{p}(k\Delta\alpha, \mu) = \hat{p}(k\Delta\alpha, \mu) + \delta(k\Delta\alpha)$$

7: Smooth the output

$$\hat{p}(\alpha, \mu) = \hat{p}(\alpha, \mu) * \mathbf{G}_{\sigma=0.05}$$

8: Normalize to unit integral

$$\hat{p}(\alpha, \mu) = \hat{p}(\alpha, \mu) / \int_{\alpha} \int_{\mu} \hat{p}(\alpha, \mu) d\mu d\alpha$$

Fig. 7. A summary of the main part of the proposed offline training algorithm used to estimate the joint probability density function $p(\alpha, \mu)$. Note that for the reasons of clarity and brevity, at places a somewhat different notation is used in the pseudo-code above from that in the main text; please refer to the algorithm header.

pitch, negligible roll), each in a different multiple light source illumination setting, see Fig. 10 (a) and 11, at 10 fps and 320×240 -pixel resolution (face size ≈ 60 pixels).¹

Toshiba Face Database (ToshFace), kindly provided to us by Toshiba Corp. This database contains 60 individuals of varying age, mostly male Japanese, and 10 sequences per person. Each sequence corresponds to a different multiple light source illumination setting, at 10 fps and 320×240 pixel resolution (face size ≈ 60 pixels), see Fig. 10(b).

Face Video Database, freely available from <http://synapse.vit.iit.nrc.ca/db/video/faces/cvglab> and described in [33]. Briefly, it contains 11 individuals and 2 sequences per person, little variation in illumination, but extreme and uncontrolled variations in pose and motion, acquired at 25 fps and 160×120 -pixel resolution (face size ≈ 45 pixels), see Fig. 10(c).

Faces96, the most challenging subset of the University of Essex face database, freely available from <http://cswww.essex.ac.uk/mv/allfaces/faces96.html>. It contains 152 individuals, most 18–20 years old and a single 20-frame sequence per person in

¹ 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/>.

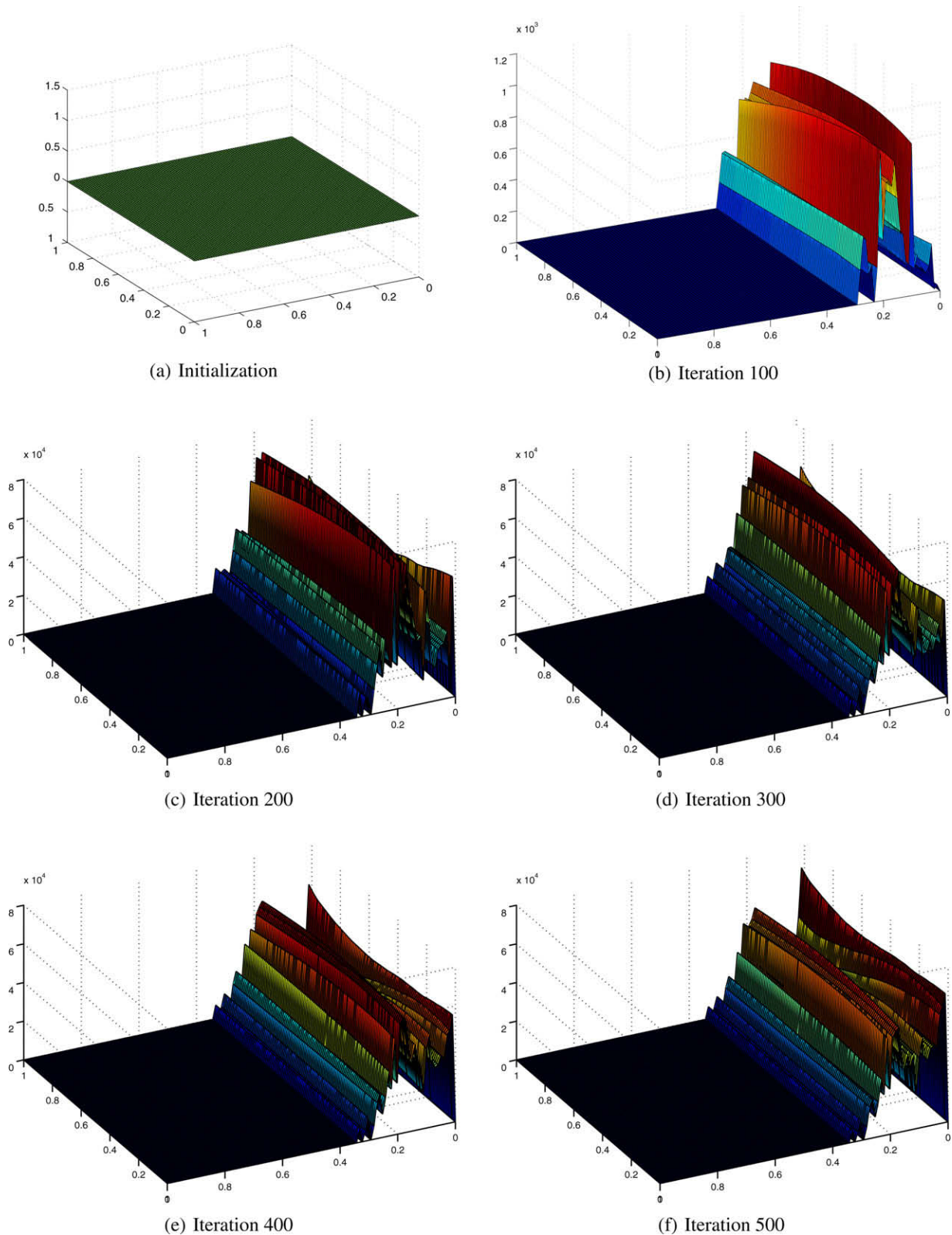


Fig. 8. The estimate of the joint density $p(\alpha, \mu)$ through 500 iterations for a band-pass filter used for the evaluation of the proposed framework in Section 4.1.

196 × 196-pixel resolution (face size ≈ 80 pixels). The users were asked to approach the camera while performing arbitrary head motion. Although the illumination 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, see Fig. 10(d).

Yale Face Database B (YaleDB), freely available from <http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html> and described in [12]. It contains 5850 single light source images of 10 subjects, each under 576 viewing conditions (9 poses and 64 illumination conditions) and an additional ambient light only image per pose, see Fig. 10(d). Since the focus of this paper is specifically on

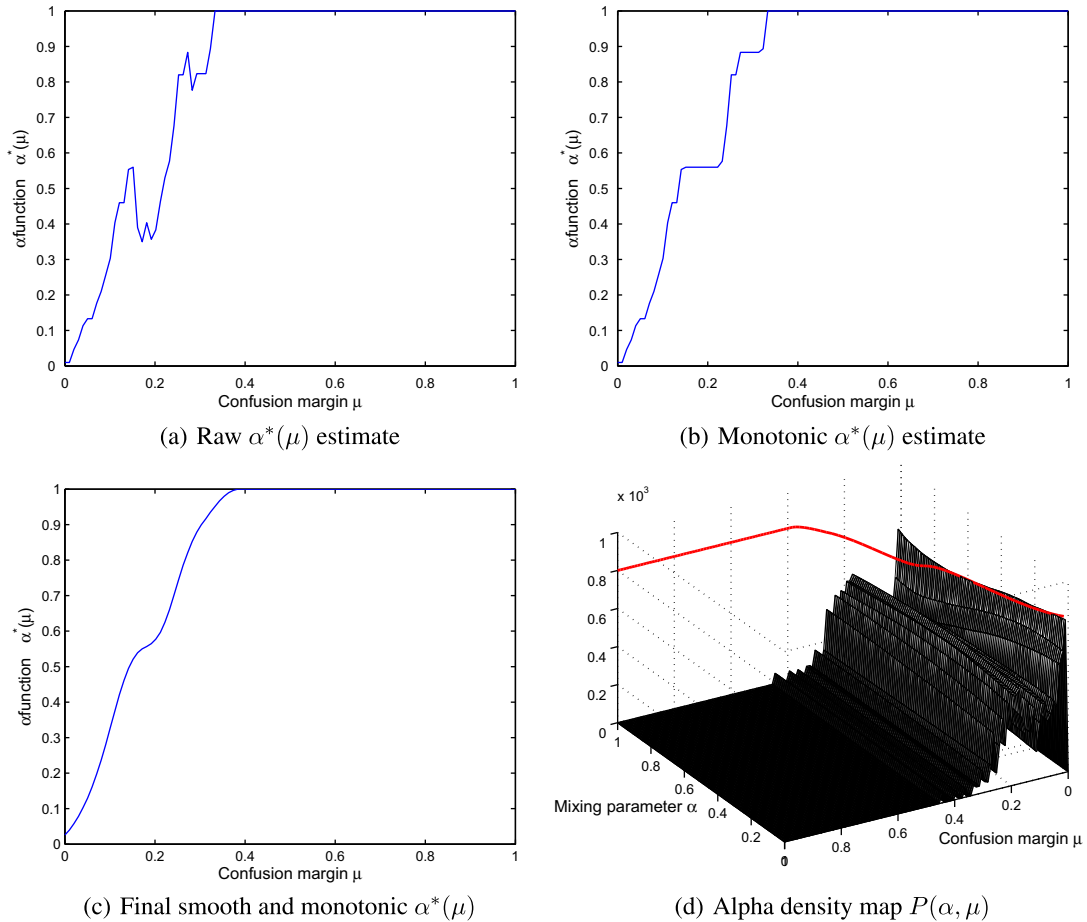


Fig. 9. Typical estimates of the α -function plotted against confusion margin μ . The estimate shown was computed using 40 individuals in five illumination conditions for a Gaussian high-pass filter. As expected, α^* assumes low values for small confusion margins and high values for large confusion margins (see (1)).

illumination invariance in recognition, we used only the 650 image subset of frontal views only.

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 the camera, this division maximizes the variation in illumination, scale and pose between training and test, thus maximizing the recognition challenge. For tests performed on *YaleDB* only a single image per subject was used both for training and querying.

Offline training, that is, the estimation of the α -function (see Section 3.1.1) was performed using 40 individuals and five illuminations from the *CamFace* database. We emphasize that these were not used as test input for the evaluations reported in the following section.

Data acquisition: The discussion so far focused on recognition using fixed-scale face images. For all video-based databases we used a cascaded detector [34] for localization of faces across scale in cluttered images. The detector was trained using roughly roughly fronto-parallel views. Detected faces were rescaled to the uniform resolution of 50×50 pixels.

Although the cascaded face detector was successful on *YaleDB* as well, we decided on a different approach to data extrac-

tion. Specifically, we manually localized the centres of eyes and the mouth, and then affine registered all faces to the same geometric frame (as in e.g. [18]). This was done because unlike in the other four data sets, in *YaleDB* pose was strictly controlled for each subject but not between subjects. Since our face detector is not rotation invariant, there would have been a possibility of learning to discriminate between poses, rather than individuals themselves.

Methods and representations. The proposed framework was evaluated using the following filters (illustrated in Fig. 12):

- Gaussian high-pass-filtered images [18,28] (HP):

$$F_1(\mathbf{X}) \equiv \mathbf{X}_H = \mathbf{X} - (\mathbf{X} * \mathbf{G}_{\sigma=1.5}), \quad (9)$$

- Local intensity-normalized high-pass-filtered images—similar to the Self-Quotient Image [19] (QI):

$$F_2(\mathbf{X}) \equiv \mathbf{X}_Q = \mathbf{X}_H / \mathbf{X}_L \equiv \mathbf{X}_H / (\mathbf{X} - \mathbf{X}_H), \quad (10)$$

the division being elementwise,

- Distance-transformed edge map [10,35] (ED):

$$F_3(\mathbf{X}) \equiv \mathbf{X}_{ED} = \text{DistanceTransform}[\mathbf{X}_E] \quad (11)$$

$$\equiv \text{DistanceTransform}[\text{Canny}(\mathbf{X})], \quad (12)$$

- Laplacian-of-Gaussian [1] (LG):

$$F_4(\mathbf{X}) \equiv \mathbf{X}_L = \mathbf{X} * \nabla \mathbf{G}_{\sigma=3}, \quad (13)$$

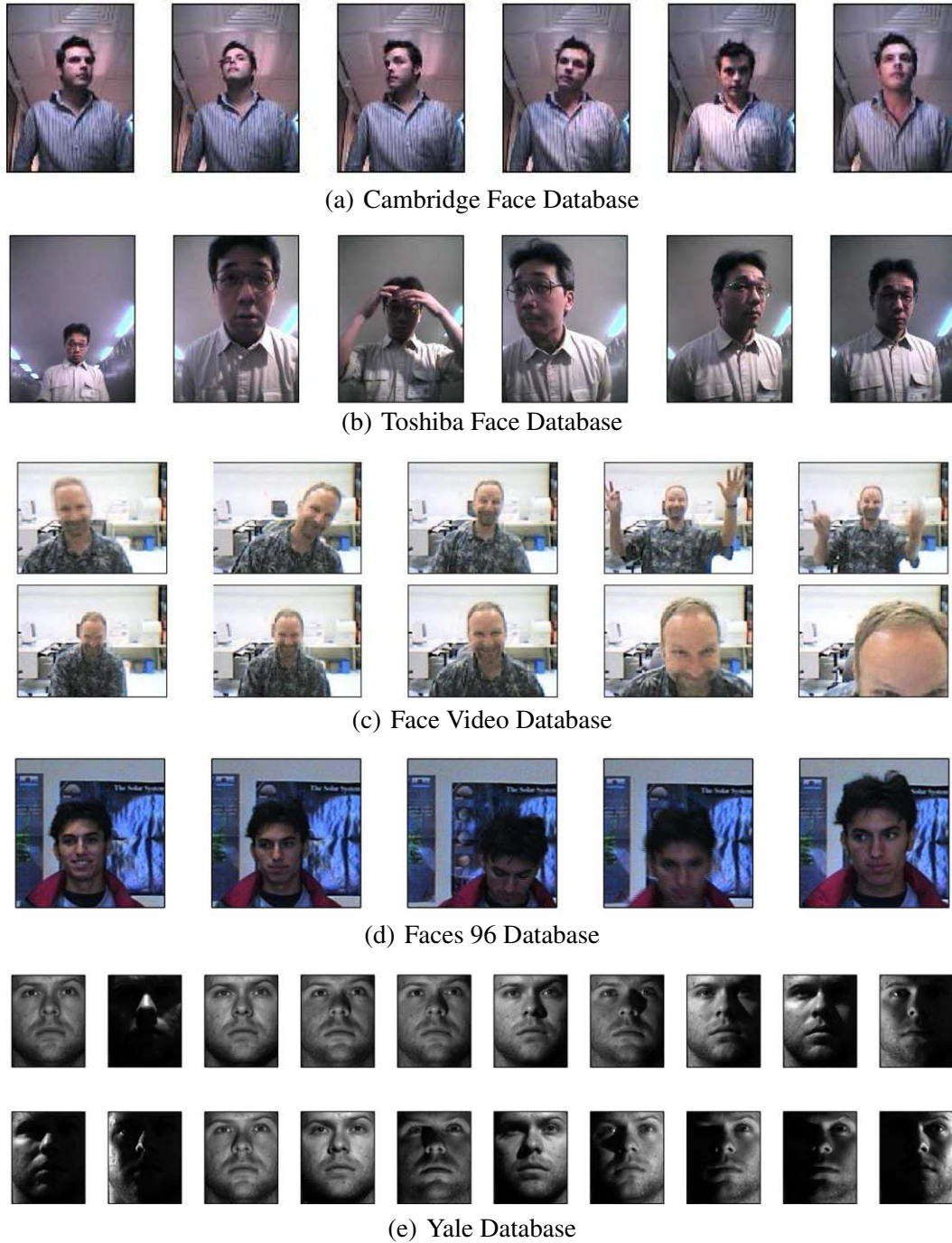


Fig. 10. Representative samples from the five data sets used for evaluation in this paper; (a)–(d) are collections of video sequences and the frames shown for each are from a single sequence, while for (e) we show a cross section through the illumination variation present in the data set.

where $*$ denotes convolution, and

- directional grey-scale derivatives [1,20] (DX, DY):

$$F_5(\mathbf{X}) \equiv \mathbf{X}_x = \mathbf{X} * \frac{\partial}{\partial x} \mathbf{G}_{\sigma_x} = 3 \quad (14)$$

$$F_6(\mathbf{X}) \equiv \mathbf{X}_y = \mathbf{X} * \frac{\partial}{\partial y} \mathbf{G}_{\sigma_y} = 3. \quad (15)$$

Video sequence matching For baseline classification on the video-based databases, we used two *canonical correlations*-based [36,37] methods which have gained considerable attention in recent literature on face recognition from video. These are:

- *Mutual Subspace Method (MSM)* of Yamaguchi and Fukui [30], and
- *Constrained MSM (CMSM)* [30] used in a state-of-the-art commercial system FacePass[®] [38].

These were chosen as fitting the main premise of the paper, due to their efficiency, numerical stability and generalization robustness [39]. We now briefly summarize the procedure in which the two methods were used for matching.

We represent each face image as a raster-ordered pixel array, and each sequence of detected faces as a data matrix $\mathbf{d} \in \mathbb{R}^{D \times N}$, each

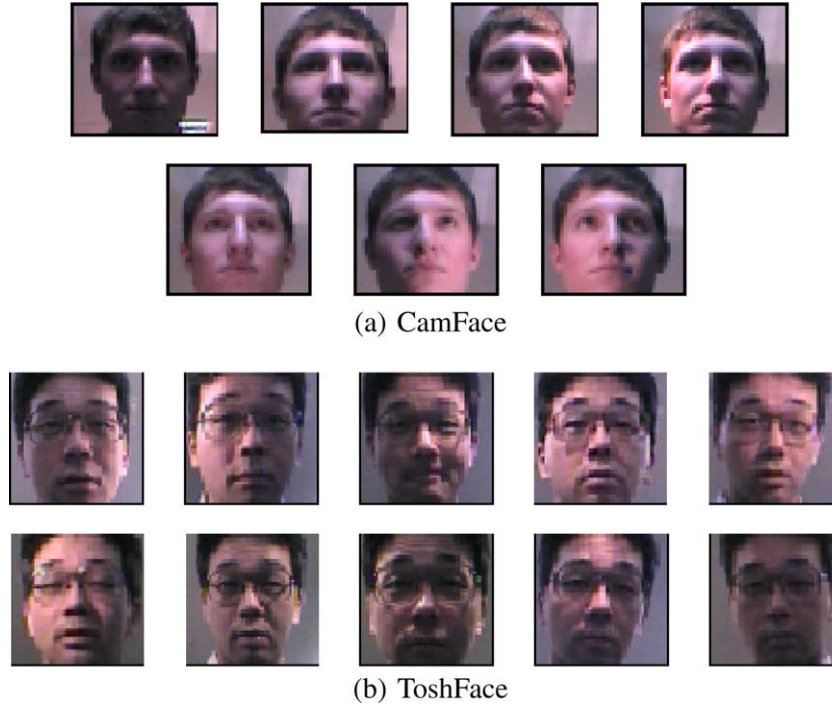


Fig. 11. An illustration of variation of lighting conditions in our databases. Shown are two faces lit with (a) illuminations 1–7 from database FaceDB100 and (b) illuminations 1–10 from database FaceDB60. It is important to emphasize that the actual appearance effects of the same illumination setup were different between different individuals (or even between different session of the same individual) due to *ad lib* chosen body position with respect to the mounted camera.

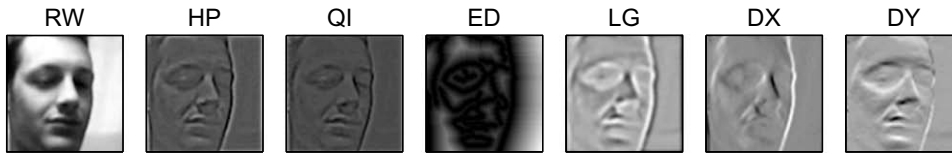


Fig. 12. Examples of the evaluated face representations: raw greyscale input (RW), high-pass-filtered data (HP), the Quotient Image (QI), distance-transformed edge map (ED), Laplacian-of-Gaussian-filtered data (LG) and the two principal axis derivatives (DX and DY).

column corresponding to a single image. Principal Component Analysis (PCA) of the cross-correlation matrix $\mathbf{C} = \mathbf{d}\mathbf{d}^T$ is used to extract the main modes of appearance variation, as the eigenvectors corresponding to the largest eigenvalues. Note that PCA was applied *without* mean subtraction. We used 6D subspaces, as sufficiently expressive to on average explain over 90% of data variation within intrinsically low-dimensional face appearance changes within a set.

In MSM, the similarity between two subspaces (each corresponding to a face motion video sequence \mathbf{d}_i) is computed as the mean of the first three principal correlations $\omega_{1..3}$ between them:

$$\rho(\mathbf{d}_i, \mathbf{d}_j) = \frac{1}{3} \sum_{k=1..3} \omega_k \quad (16)$$

If \mathbf{B}_i and \mathbf{B}_j are orthonormal basis matrices corresponding to the subspaces, then writing the Singular Value Decomposition (SVD) of the matrix $\mathbf{B}_i^T \mathbf{B}_j$:

$$\mathbf{M} = \mathbf{B}_i^T \mathbf{B}_j = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \quad (17)$$

The k th canonical correlation ω_k is then given by the k th singular value of \mathbf{M} i.e. $\Sigma_{k,k}$, and the k th pair of principal vectors \mathbf{u}_k and \mathbf{v}_k by, respectively, $\mathbf{B}_i \mathbf{u}_k$ and $\mathbf{B}_j \mathbf{v}_k$ [40], see Fig. 13. In CMSM, the computation of canonical correlations is preceded with linear projection of the two subspaces to the *Constraint Subspace*, see the original publication for more detail [30]. In our empirical evaluation, we estimate the Constraint Subspace using gallery training data.

Still image matching: Since pose was strictly controlled in this data set, we compared two still images of faces extracted from *YaleDB* using a simple matching procedure, analogous to that previously employed on video. If \mathbf{x}_i and \mathbf{x}_j are two images as raster-ordered pixel arrays, their similarity is computed as the cosine of the angle between them:

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} \quad (18)$$

For brevity, we shall refer to this as the COS distance.

4.1. Results

To establish baseline performance, we performed recognition on all data sets using raw greyscale data first. A summary is shown in Table 1. As these results illustrate, *CamFace*, *ToshFace* and *YaleDB* were found to be very challenging, primarily due to extreme variations in illumination between training and query data, as well *within* sequences in the case of the former two databases. The performance on *Face Video* and *Faces96* databases was significantly better. This can be explained by noting that the first major source of appearance variation present in these sets, the scale, is normalized for in the data extraction stage; the remainder of the appearance variation is dominated by pose changes, to which MSM and CMSM are particularly robust to [32,39]. This confirms the premise that varying illumination indeed does represent the main difficulty

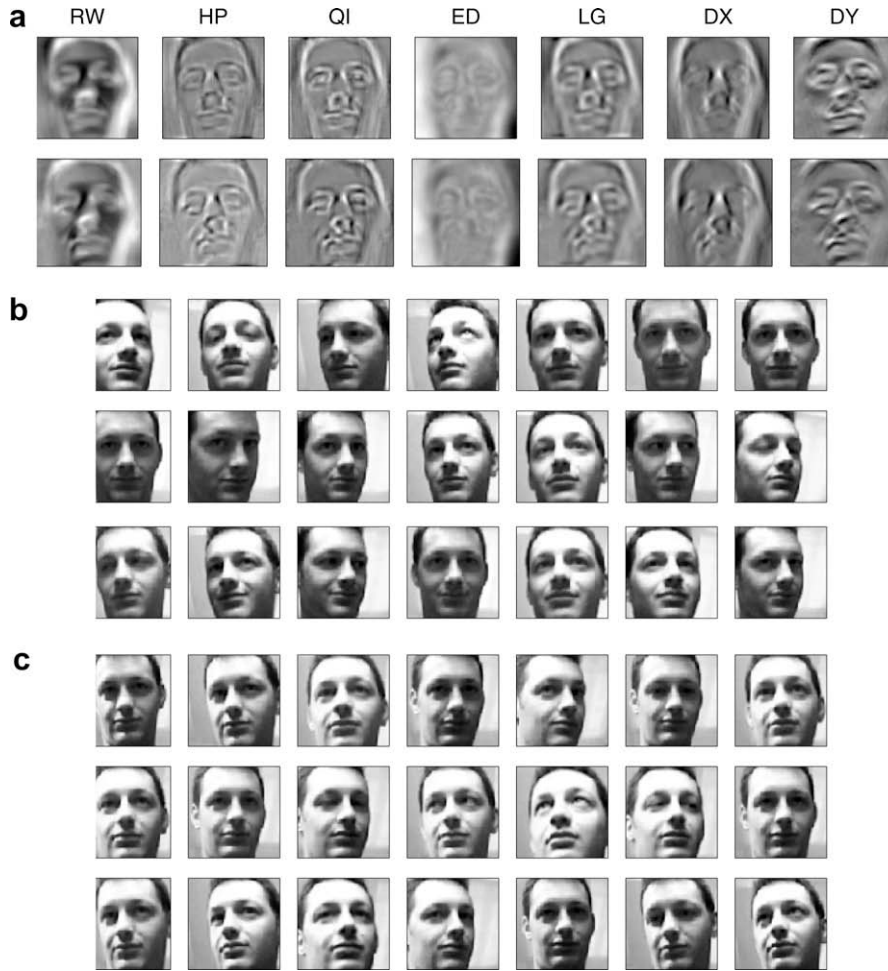


Fig. 13. (a) The first pair of principal vectors (top and bottom) corresponding to the sequences (b) and (c) (every 4th detection is shown for compactness), for each of the seven representations used in the empirical evaluation described in this paper. A higher degree of similarity between the two vectors indicates a greater degree of illumination invariance of the corresponding filter.

Table 1
Recognition rates (mean/STD, %).

| | CamFace | ToshFace | FaceVideoDB | Faces96 | YaleDB | mean |
|------|-----------|-----------|-------------|---------|-----------|------|
| CMSM | 73.6/22.5 | 79.3/18.6 | 91.9 | 100.0 | — | 87.8 |
| MSM | 58.3/24.3 | 46.6/28.3 | 81.8 | 90.1 | — | 72.7 |
| COS | — | — | — | — | 66.9/35.4 | — |

in achieving robust face recognition when the data acquisition set-up is unconstrained.

Next we evaluated the two methods with each of the six filter-based face representations. The recognition results for the *CamFace*, *ToshFace*, *Faces96* and *YaleDB* databases are shown in blue in Fig. 14, while the results on the *Face Video* data set are separately shown in Table 2 for the ease of visualization. Confirming the first premise of this work as well as corroborating previous research findings, all of the filters, except for the distance transformed edge map, produced an improvement in average recognition rates when used with methods which provide little additional illumination invariance themselves (MSM and the COS distance). The failure of the distance transformed edge map to provide a consistent improvement can be attributed to its high sensitivity to cast shadows and effectively a complete loss of any albedo information. The usefulness of this representation was already argued in the literature as being better suited to pose, rather than identity discrimination [10,41,42].

Little interaction between method/filter combinations was found, the Quotient Image performing the best, with the horizontal intensity derivative and the Laplacian-of-Gaussian producing comparable results and bringing the average recognition errors down to the region of about 10%.

Finally, in the last set of experiments, we employed each of the 6 filters in the proposed data-adaptive framework. The recognition results are shown in red in Fig. 14 and in Table 2 for the *Face Video* database. The proposed method produced a dramatic performance improvement in the case of *all filters*, reducing the average recognition error rate to only 3% in the case of CMSM/Quotient Image combination. This is a very high recognition rate for such unconstrained conditions (see Fig. 10), small amount of training data per gallery individual and the degree of illumination variation between training and query data. An improvement in the robustness to illumination changes can also be seen in the significantly reduced standard deviation of the recognition, as shown in Fig. 14. Finally, it should be emphasized that the demonstrated improvement is obtained with a negligible increase in the computational cost as all time-demanding learning is performed offline.

4.2. Failure modes

In the discussion of failure modes of the described framework, it is necessary to distinguish between errors introduced by a *particular* image processing filter used, and the fusion algorithm itself. As

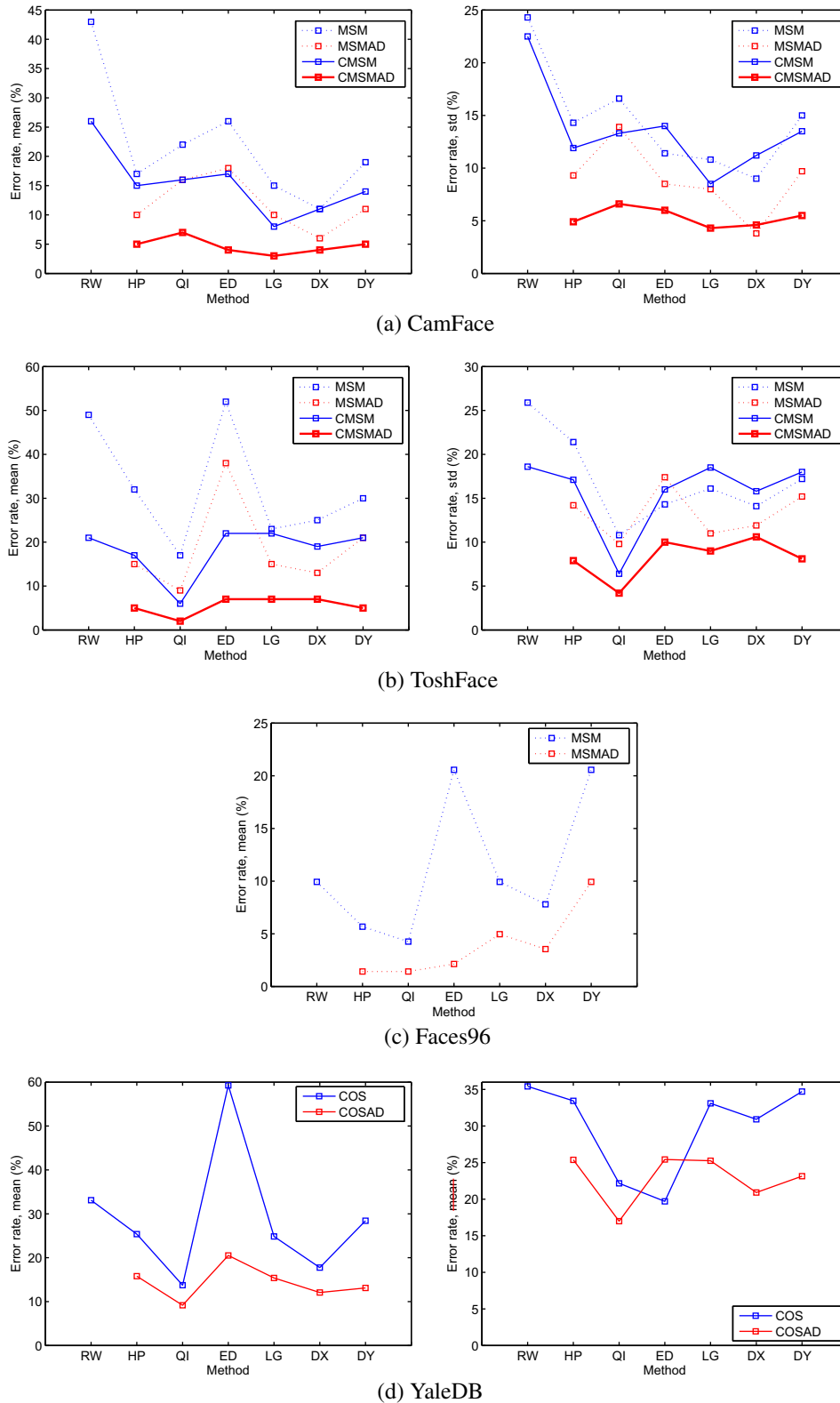


Fig. 14. Error rate statistics. The proposed framework (-AD suffix) dramatically improved recognition performance on all method/filter combinations, as witnessed by the reduction in both error rate averages and their standard deviations. The results of CSM on *Faces96* are not shown as it performed perfectly on this data set.

generally recognized across literature (e.g. see [1]), qualitative inspection of incorrect recognitions using filtered representations indicates that the main difficulties are posed by those illumination effects which most significantly deviate from the underlying frequency model (see Section 2.1) such as: cast shadows, specularities

(especially commonly observed for users with glasses) and photo-sensor saturation.

On the other hand, any failure modes of our fusion framework were difficult to clearly identify, due to such a low frequency of erroneous recognition decisions. Even these were in virtually all

Table 2
FaceVideoDB, mean error (%).

| | RW | HP | QI | ED | LG | DX | DY |
|---------|------|------|------|------|------|------|------|
| MSM | 0.00 | 0.00 | 0.00 | 0.00 | 9.09 | 0.00 | 0.00 |
| MSM-AD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CMSM | 0.00 | 9.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CMSM-AD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

of the cases due to overly confident decisions in the filtered pipeline. Overall, this makes the methodology proposed in this paper extremely promising as a robust and efficient way of matching face appearance image sets, and suggests that future work should concentrate on developing appropriately robust image filters that can deal with more complex illumination effects.

5. Conclusions

In this paper we described a novel framework for automatic face recognition in the presence of varying illumination, primarily applicable to matching face sets or sequences. The framework is based on simple image processing filters that compete with unprocessed greyscale input to yield a single matching score between individuals. By performing all numerically consuming computation offline, our method both (i) retains the matching efficiency of simple image filters, but (ii) with a greatly increased robustness, as all online processing is performed in closed-form. Evaluated on a large, real-world data corpus, the proposed framework was shown to be successful in both video and still image-based recognition across a wide range of variation in illumination.

As suggested by our experimental results, the main direction for future work is to develop more robust image processing filters without a great loss of computational efficiency. Specifically, we are investigating the use of colour invariants and photometric camera models with the aim of achieving unified detection and elimination of specularities, cast shadows and points of photo-sensor saturation.

Acknowledgments

We thank Trinity College Cambridge and the Toshiba Corporation for their kind support for our research, the volunteers whose face videos were entered in our face database and Cambridge Commonwealth Trust.

References

- [1] 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* 19 (7) (1997) 721–732.
- [2] S. Kong, J. Heo, B. Abidi, J. Paik, M. Abidi, Recent advances in visual and infrared face recognition—a review, *Computer Vision and Image Understanding* 97 (1) (2005) 103–135.
- [3] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, *ACM Computing Surveys* 35 (4) (2004) 399–458.
- [4] V. Blanz, T. Vetter. A morphable model for the synthesis of 3D faces, in: *Proc. Conference on Computer Graphics (SIGGRAPH)*, 1999, pp. 187–194.
- [5] S. Romdhani, V. Blanz, T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions, in: *Proc. European Conference on Computer Vision (ECCV)*, 2002, pp. 3–19.
- [6] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (9) (2003) 1063–1074.
- [7] D.S. Bolme. Elastic Bunch Graph Matching. Master's Thesis, Colorado State University, 2003.
- [8] C. Kotropoulos, A. Tefas, I. Pitas, Frontal face authentication using morphological elastic graph matching, *IEEE Transactions on Image Processing* 9 (4) (2000) 555–560.
- [9] L. Wiskott, J.-M. Fellous, N. Krüger, C. von der Malsburg, Face recognition by elastic bunch graph matching, *Intelligent Biometric Techniques in Fingerprint and Face Recognition* (1999) 355–396.
- [10] O. Arandjelović, R. Cipolla, Face recognition from video using the generic shape-illumination manifold, in: *Proc. European Conference on Computer Vision (ECCV)*, vol. 4, 2006, pp. 27–40.
- [11] 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* 28 (3) (1998) 245–260.
- [12] A.S. Georghades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (6) (2001) 643–660.
- [13] S. Zhou, G. Aggarwal, R. Chellappa, D. Jacobs, Appearance characterization of linear Lambertian objects, generalized photometric stereo, and illumination-invariant face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2) (2007) 230–245.
- [14] M. Nishiyama, O. Yamaguchi. Face recognition using the classified appearance-based quotient image, in: *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 49–54.
- [15] 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* 23 (2) (2001) 139–219.
- [16] R. Basri, D. Jacobs, I. Kemelmacher, Photometric stereo with general, unknown lighting, *International Journal of Computer Vision* 72 (3) (2007) 239–257.
- [17] S. Romdhani, T. Vetter. Efficient, robust and accurate fitting of a 3D morphable model, in: *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2003, pp. 59–66.
- [18] 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, 2005, pp. 860–867.
- [19] 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.
- [20] M. Everingham, and A. Zisserman, Automated person identification in video, in: *Proc. IEEE International Conference on Image and Video Retrieval (CIVR)*, 2004, pp. 289–298.
- [21] Y. Gao, M.K.H. Leung, Face recognition using line edge map, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (6) (2002) 764–779.
- [22] B. Takács, Comparing face images using the modified Hausdorff distance, *Pattern Recognition* 31 (12) (1998) 1873–1881.
- [23] O. Arandjelović, R. Cipolla. An illumination invariant face recognition system for access control using video, in: *Proc. IAPR British Machine Vision Conference (BMVC)*, 2004, pp. 537–546.
- [24] 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.
- [25] C. Garcia, G. Zikos, G. Tziritas. A wavelet-based framework for face recognition, in: *Proc. European Conference on Computer Vision (ECCV)*, 1998, pp. 1–7.
- [26] B. Kepenekci, Face Recognition Using Gabor Wavelet Transform. Ph.D. Thesis, The Middle East Technical University, 2001.
- [27] P. Jonathon Phillips, Matching pursuit filters applied to face identification, *IEEE Transactions on Image Processing* 7 (8) (1998) 1150–1164.
- [28] 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.
- [29] B.V.K. Vijayakumar, X. Chunyan, S. Marios, Correlation filters for large population face recognition, in: *Proc. SPIE Conference on Biometric Technology for Human Identification*, 2007, p. 6539.
- [30] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, *International Symposium of Robotics Research* (2003).
- [31] X. Wang, X. Tang. Unified subspace analysis for face recognition, in: *Proc. IEEE International Conference on Computer Vision (ICCV)*, vol. 1, 2003, pp. 679–686.
- [32] 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, pp. 581–588.
- [33] D.O. Gorodnichy. Associative neural networks as means for low-resolution video-based recognition, in: *Proc. International Joint Conference on Neural Networks*, 2005.
- [34] P. Viola, M. Jones, Robust real-time face detection, *International Journal of Computer Vision* 57 (2) (2004) 137–154.
- [35] J. Canny. A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8 (6) (1986) 679–698.
- [36] R. Gittins, *Canonical Analysis: A Review with Applications in Ecology*, Springer-Verlag, Berlin, 1985.
- [37] H. Hotelling, Relations between two sets of variates, *Biometrika* 28 (1936) 321–372.
- [38] Toshiba, Facepass. Available at: <www.toshiba.co.jp/mmlab/tech/w31e.htm>.
- [39] T.-K. Kim, O. Arandjelović, R. Cipolla, Boosted manifold principal angles for image set-based recognition, *Pattern Recognition* 40 (9) (2007) 2475–2484.
- [40] Å. Björck, G.H. Golub, Numerical methods for computing angles between linear subspaces, *Mathematics of Computation* 27 (123) (1973) 579–594.
- [41] D.M. Gavrila. Pedestrian detection from a moving vehicle, in: *Proc. European Conference on Computer Vision (ECCV)*, vol. 2, 2000, pp. 37–49.
- [42] B. Stenger, A. Thayananthan, P.H.S. Torr, R. Cipolla. Filtering using a tree-based estimator, in: *Proc. IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2003, pp. 1063–1070.