Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 2475–2484
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Boosted manifold principal angles for image set-based recognition

Tae-Kyun Kim*, Ognjen Arandjelović, Roberto Cipolla

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

Received 29 June 2006; received in revised form 31 October 2006; accepted 29 December 2006

Abstract

In this paper we address the problem of classifying vector *sets*. We motivate and introduce a novel method based on comparisons between corresponding vector subspaces. In particular, there are two main areas of novelty: (i) we extend the concept of principal angles between linear subspaces to manifolds with arbitrary nonlinearities; (ii) it is demonstrated how boosting can be used for application-optimal principal angle fusion. The strengths of the proposed method are empirically demonstrated on the task of automatic face recognition (AFR), in which it is shown to outperform state-of-the-art methods in the literature.

Keywords: Face recognition; Manifolds; Image set; Principal angle; Canonical correlation analysis; Boosting; Nonlinear subspace; Illumination; Pose; Robustness; Invariance

1. Introduction

Many computer vision tasks can be cast as learning problems over vector *sets*. In object recognition, for example, a set of vectors may represent a variation in an object's appearance—be it due to camera pose changes, non-rigid deformations or variation in illumination conditions. The objective of this work is to classify a novel set of vectors to one of the training classes, each also represented by a vector set. In this paper, learning concepts will be illustrated on sets of face appearance images using the AFR paradigm, although the reader should note that no domain-specific information is actually used.

1.1. Previous work

Most of the previous work on matching vector or image sets exploits their semantics to a certain degree, for example by modelling temporal coherence between consecutive vectors i.e. by matching sequences. By their nature, these methods are of little relevance to the work presented in this paper, so we do not address them here. Broadly speaking, in the recent literature we recognize two groups of approaches to learning over sets of vectors: statistical and principal-angle based.

1.1.1. Statistical methods

Statistical learning approaches rely on the assumption that vectors $\mathbf{x}$ of the ith class are independently and identically (i.i.d.) drawn samples from $p^{(i)}(\mathbf{x})$. The problem of set matching then becomes that of estimating each underlying probability density and comparing two such estimates. In the work of Shakhnarovich et al. [1], densities $p^{(i)}(\mathbf{x})$ are modelled as multivariate Gaussians, estimated with probabilistic principal component analysis (PCA) [2] and compared using the Kullback–Leibler (KL) divergence [3]. Arandjelović et al. criticized this approach for its insufficiently expressive modelling and proposed a kernel-based method to implicitly model nonlinear, but intrinsically low-dimensional manifolds of faces [4]. In this work, the authors also argue against the use of KL divergence due to its asymmetry and demonstrate a superior performance of the resistor–average distance [5] on the task of AFR under mildly varying imaging conditions. In Ref. [6], a Gaussian mixture model (GMM) is proposed for high-dimensional density estimation. The advantage of this approach over the previously mentioned kernel method lies in its more principled modelling of densities confined to nonlinear manifolds; however this benefit comes at the cost

* Corresponding author. Tel.: +44 791 3925831.
E-mail address: tkk22@eng.cam.ac.uk (T.-K. Kim).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.12.030

of increased difficulty of divergence computation, performed using a Monte–Carlo algorithm.

1.1.2. Principal angle-based methods

Principal angles are minimal angles between vectors of two subspaces (see Section 2). Since the concept of principal angles was first introduced by Hotelling in Ref. [7], it has been applied in various fields [8–10]. Of most relevance to the work addressed in this paper is the mutual subspace method (MSM) of Yamaguchi et al. [11]. In MSM the sum of cosines of the first (i.e. smallest) few principal angles[1] is used as a similarity measure between linear subspaces used to compactly characterize vector sets. MSM has been successfully used for face recognition [11] and ship identification [12] (for evaluation results also see Refs. [4,6]). In the related works [32,13], vector sets are projected to the linear subspace that attempts to maximize the separation (in terms of principal angles) between vector spaces corresponding to different classes, under the assumption of their linearity.

MSM-based methods have two major shortcomings: the limited capability of modelling nonlinear pattern variations and the *ad hoc* fusion of information contained in different principal angles. The assumption of linearity of modelled vector subspaces is important, both because it means that MSM is incapable of differentiating between two nonlinear manifolds embedded in the same linear space and because of the sensitivity of such estimate to particular data variation [4]. In Ref. [14] Wolf and Shashua show how principal angles between nonlinear subspaces can be computed using the "kernel trick" [15]. However, the reported evaluation was performed on a database of a rather small size, making it difficult to judge the performance of their method. Additionally, as in all kernel approaches, finding the optimal kernel function is a difficult problem.

An attractive feature of MSM-based methods is their computational efficiency: principal angles between linear subspaces can be computed rapidly [16], while the estimation of linear subspaces can be performed in an incremental manner [17–20].

1.1.3. Densities vs. subspaces

As a conclusion to this section, we would like to briefly discuss the advantages and disadvantages of the two learning approaches: one which learns densities confined to low-dimensional subspaces and the other which learns the subspaces themselves. In many computer vision applications, due to different data acquisition conditions, the frequency of occurrence of a particular pattern can vary arbitrarily between the training stage and a novel input to the system.[2] In this case, subspace learning techniques are more applicable as they effectively place a uniform prior over a space of possible pattern variation. On the other hand, if there is a reason to believe that training and novel data share some statistical properties, density-based methods may produce better results. In AFR work of Arandjelović et al. [6], for example, the authors note that anatomical constraints and the constraints of the imaging setup make certain head poses more likely than others, therefore opting for a statistical approach to recognition. The point to take is that neither of the two approaches is inherently the right one, but that the choice between the two is dictated by a particular problem.

[1] In statistics, the cosines of canonical angles are termed canonical correlations.

[2] The term "arbitrarily" should be taken in practical terms i.e. given the parameters which one can realistically expect to model, control or affect.

2. Boosted manifold principal angles (BoMPA)

In this work (the earlier conference version appeared in Ref. [33]), we are interested in discriminating between abstract classes represented as vector sets without any knowledge of what the data represents. Before tackling this problem, it is important to recognize the difficulties of comparing vector sets common to its different semantic instances:

- *Expressiveness*: Pattern changes across and within modelled vector sets often exhibit significant nonlinearities. Seeing that differences within a class can oftentimes be greater than between classes (in Euclidean distance sense), it is important to use a model flexible enough to capture this complex variation, see Fig. 1 for an example. In Section 2.3 we achieve this by moving away from the typically used parametric models and formulate a method that uses canonical correlations and Gaussian mixtures matching.
- *Graceful degradation*: The exact vectors used as an input (either as training or test) to a practical system can be expected to vary from time to time, depending on the exact data acquisition protocol employed. In particular, sometimes more and sometimes less data is available. In the context of face recognition, for example, this may be because the user has not assumed certain poses or because face detection has failed. Graceful degradation refers to slow decay in performance of a learning algorithm as less and less data is available. Our canonical angles-based framework is already exhibiting this property in that only the most similar and discriminating regions of two subspaces are actually compared (see Sections 2.1 and 2.2). Further robustness is achieved by our extension of the similarity function to nonlinear manifolds in Section 2.3 by discarding all but the most reliable matching linear patches.
- *Robustness to noise*: Noise is very much an inherent problem in any practical application. In computer vision, for example, vector patterns considered may represent appearance images—these are affected by noise sources such as quantum, quantization or due to spatial discretization. Our assumption of intrinsically low-dimensional pattern variations within a set, corrupted by isotropic Gaussian noise, are captured well using probabilistic PCA in Section 2.3.
- *Numerical stability and efficiency*: Closely related to the previously mentioned issue of noise in data are numerical issues pertaining to the implementation of a particular algorithm. It is an imperative for a practical algorithm to be numerically stable and, often, be time efficient. These issues are discussed in Sections 2.3 and 3.

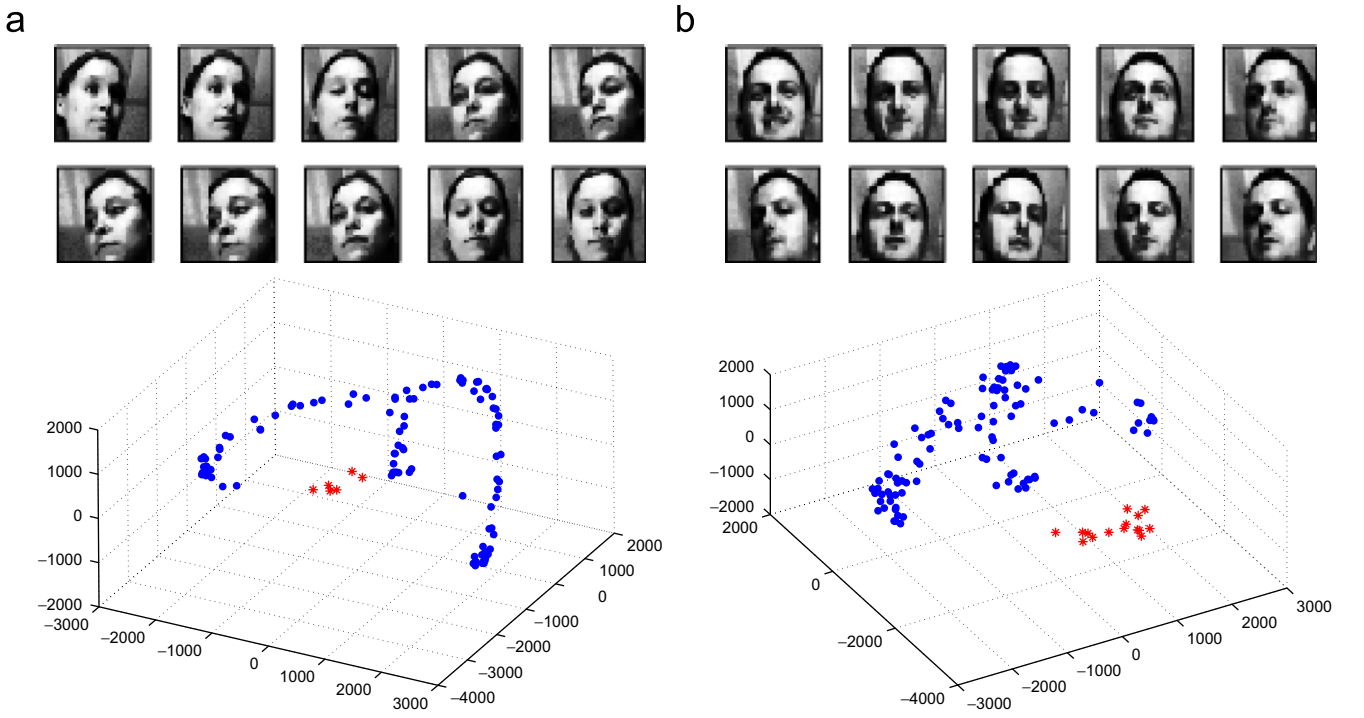

Fig. 1. *Face vector sets:* 10 samples of two typical face sets used to illustrate concepts proposed in this paper (top) and the corresponding patterns in the 3D principal component subspaces (bottom), estimated from data. The sets capture appearance changes of faces of two different individuals as they performed unconstrained head motion in front of a fixed camera. The corresponding pattern variations (blue circles) are highly nonlinear, with a number of outliers present (red stars).

We will often refer back to these four requirements throughout the paper, using them to motivate different features of the proposed method.

2.1. Principal angles

Principal, or canonical, angles $0 \leqslant \theta_1 \leqslant \cdots \leqslant \theta_D \leqslant (\pi/2)$ between two D-dimensional linear subspaces U_1 and U_2 are uniquely defined as the minimal angles between any two vectors of the subspaces:

$$\cos\theta_i = \max_{\mathbf{u}_i \in U_1} \max_{\mathbf{v}_i \in U_2} \mathbf{u}_i^{\mathrm{T}} \mathbf{v}_i \tag{1}$$

subject to:

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = 1, \quad \mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = 0, \quad j = 1, \ldots, i-1. \tag{2}$$

We will refer to $\mathbf{u}_i$ and $\mathbf{v}_i$ as the ith pair of *principal vectors*. Intuitively, the first pair of principal vectors corresponds to the most similar modes of variation of two linear subspaces; every next pair to the most similar modes orthogonal to all previous ones. This concept is illustrated in Fig. 2 on the example of sets of face appearance images.

2.2. Learning the subspace similarity function

In Section 1.1 it was argued that one of the weaknesses of previous approaches in the literature is their use of only the first few principal angles. While these do correspond to most similar modes of variation of two subspaces, they may be caused by extrinsic factors: in the case of face images these may be changed corresponding to extreme illumination conditions, see Fig. 3(a). Given a set of first N principal angles $\Theta = \{\theta_1, \ldots, \theta_N\}$, our aim is to learn the optimal similarity function $f(\Theta)$ between the two subspaces.

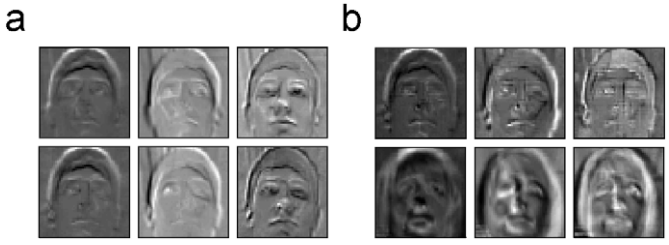

Fig. 2. *Principal vectors in MSM:* The first three pairs (top and bottom rows) of principal vectors for a comparison of two linear subspaces corresponding to the same (a) and different individuals (b). In the former case, the most similar modes of pattern variation, represented by principal vectors, are very much alike *in spite of different illumination conditions* used in data acquisition.

2.2.1. Boosted principal angles

In general, each principal angle θ_i carries some information for discrimination between the corresponding two subspaces. We use this to build simple weak classifiers $\mathscr{M}(\theta_i) = \mathrm{sign}[\cos(\theta_i) - C]$. In the proposed method, these are combined

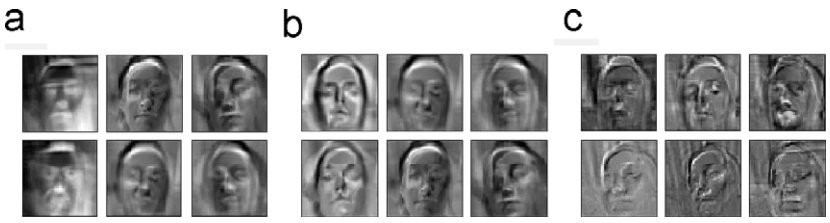

Fig. 3. *MSM, BPA and MPA*: (a) The first three principal vectors between two linear subspaces which MSM incorrectly classifies as corresponding to the same person (the two data sets are shown in Fig. 1). In spite of different identities, the most similar modes of variation are very much alike and can be seen to correspond to especially difficult illuminations. (b) Boosted principal angles (BPA), on the other hand, chooses different principal vectors as the most discriminating—these modes of variation are now less similar between the two sets. (c) Modelling of nonlinear manifolds corresponding to the two image sets produces a further improvement. Shown are the most similar modes of variation amongst all pairs of linear manifold patches. Local information is well captured and even these principal vectors are now very dissimilar.

using the now acclaimed AdaBoost algorithm [21]. In summary, AdaBoost learns a weighting $\{w_i\}$ of decisions cast by weak learners to form a classifier $\mathcal{M}(\Theta)$:

$$\mathcal{M}(\Theta) = \text{sign}\left[\sum_{i=1}^{N} w_i \mathcal{M}(\theta_i) - \frac{1}{2}\sum_{i=1}^{N} w_i\right]. \quad (3)$$

In an iterative update scheme classifier performance is optimized on training data which consists of in-class and out-of-class features (i.e. principal angles). Let the training database consist of sets $S_1, \ldots, S_K \equiv \{S_i\}$, corresponding to K classes. In the framework described, the $K(K-1)/2$ out-of-class principal angles are computed between pairs of linear subspaces corresponding to training data sets $\{S_i\}$, estimated using PCA. On the other hand, the K in-class principal angles are computed between a pair of randomly drawn subsets for each S_i. PCA was adopted to learn effective low-dimensional global subspaces of image sets. It has been known face image classes are well-characterized by eigen-subspaces, which have also been proven to be beneficial in low computational complexity of subsequent principal angle analysis (See Section 3.2).

We use the learnt weights $\{w_i\}$ for computing the following similarity measure between two linear subspaces:

$$f(\Theta) = \frac{1}{N}\frac{\sum_{i=1}^{N} w_i \cos(\theta_i)}{\sum_{i=1}^{N}.w_i}. \quad (4)$$

A typical set of weights $\{w_i\}$ we obtained for our AFR application is shown graphically in Fig. 4(a). The plot shows an interesting result: the weight corresponding to the first principal angle is not the greatest. Rather it is the second principal angle that is most discriminating, followed by the third one. This confirms our observation that the most similar mode of variation across two subspaces can indeed be due an extrinsic factor. Fig. 3(b) shows the three most discriminating principal vector pairs selected by our algorithm for data incorrectly classified by MSM—the most weighted principal vectors are now much less similar. The gain achieved with boosting is also apparent from Fig. 4(b). A significant improvement can be seen both for a small and a large number of principal angles. In the former case this is because our algorithm chooses not the first but the most discriminating set of angles. The latter case is practically more important—as more principal angles are added to MSM,

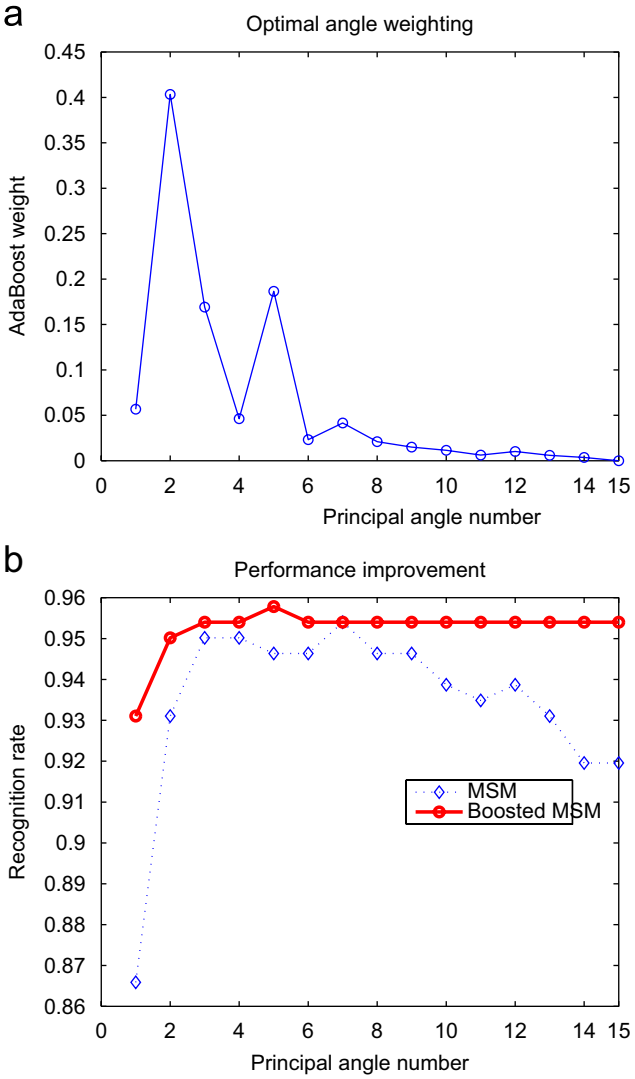

Fig. 4. *Boosted principal angles*: (a) A typical set of weights corresponding to weak principal angle-based classifiers, obtained using AdaBoost. This figure confirms our criticism of MSM-based methods for (i) their simplistic fusion of information from different principal angles and (ii) the use of only the first few angles, see Section 1.1. (b) The average performance of a simple MSM classifier and our boosted variant.

its performance first improves, but after a certain point it starts *worsening*. This highly undesirable behaviour is caused by effectively equal weighting of base classifiers in MSM. In contrast, the performance of our algorithm never decreases as more information is added. As a consequence, no further provision for choosing the optimal number of principal angles is needed by the weights learnt.

At this point it is worthwhile mentioning the work of Maeda et al. [22] in which the third principal angle was found to be useful for discriminating between sets of images of a face and its photograph. Much like the methods described in Section 1.1, the use of a single principal angle was motivated only empirically—the described framework can be used for a more principled feature selection in this setting as well.

Basically, the problem we tackled by Adaboost is to learn the best combination of principal angles and the individual importance of principal angles. The entire parameter space of this problem is certainly huge. We have proposed Adaboost as a reasonable efficient method, which is based on learning. Using the learnt weights we could eliminate special provisions to decide the best combinations and weights of principal angles.

2.3. Nonlinear subspaces

The assumption that patter variations within each class are well represented by a linear subspace is usually severely limiting, see Fig. 1. Our aim is to extend the described framework of boosted principal angles to being able to effectively capture nonlinear data behaviour. We propose a method that combines *global* manifold variations with more subtle, *local* ones.

Without the loss of generality, let S_1 and S_2 be two vector sets and Θ the set of principal angles between two linear subspaces. We derive a measure of similarity ρ between S_1 and S_2 by comparing the corresponding linear subspaces $U_{1,2}$ and locally linear patches $L_{1,2}^{(i)}$ corresponding to piece-wise linear approximations of manifolds of S_1 and S_2:

$$\rho(S_1, S_2) = (1-\alpha) f_G[\Theta(U_1, U_2)] + \alpha \max_{i,j} f_L[\Theta(L_1^{(i)}, L_2^{(j)})], \quad (5)$$

where f_G and f_L have the same functional form as f in Eq. (4), but separately learnt base classifier weights $\{w_i\}$. Put in words, the proximity between two manifolds is computed as a weighted average of the similarity between global modes of data variation and the best matching local behaviour. The two terms complement each other: the former provides (i) robustness to noise, whereas the latter ensures (ii) graceful performance degradation with missing data and (iii) flexibility in modelling complex manifolds, see Fig. 3(c).

2.3.1. Finding stable locally linear patches

In the proposed framework, stable locally linear manifold patches are found using mixtures of probabilistic PCA (PPCA) [23]. The main difficulty in fitting of a PPCA mixture is the requirement for the local principal subspace dimensionality to be set *a priori*. We solve this problem by performing the fitting in two stages. In the first stage, a GMM constrained to diagonal covariance matrices is fitted first. This model is crude as it is insufficiently expressive to model local variable correlations, yet too complex (in terms of free parameters) as it does not encapsulate the notion of intrinsic manifold dimensionality and additive noise. However, what it is useful for is the *estimation* of the intrinsic manifold dimensionality d, from the eigenspectra of its covariance matrices, see Fig. 5(a). Once d is estimated (typically $d \ll D$), the fitting is repeated using a mixture of PPCA.

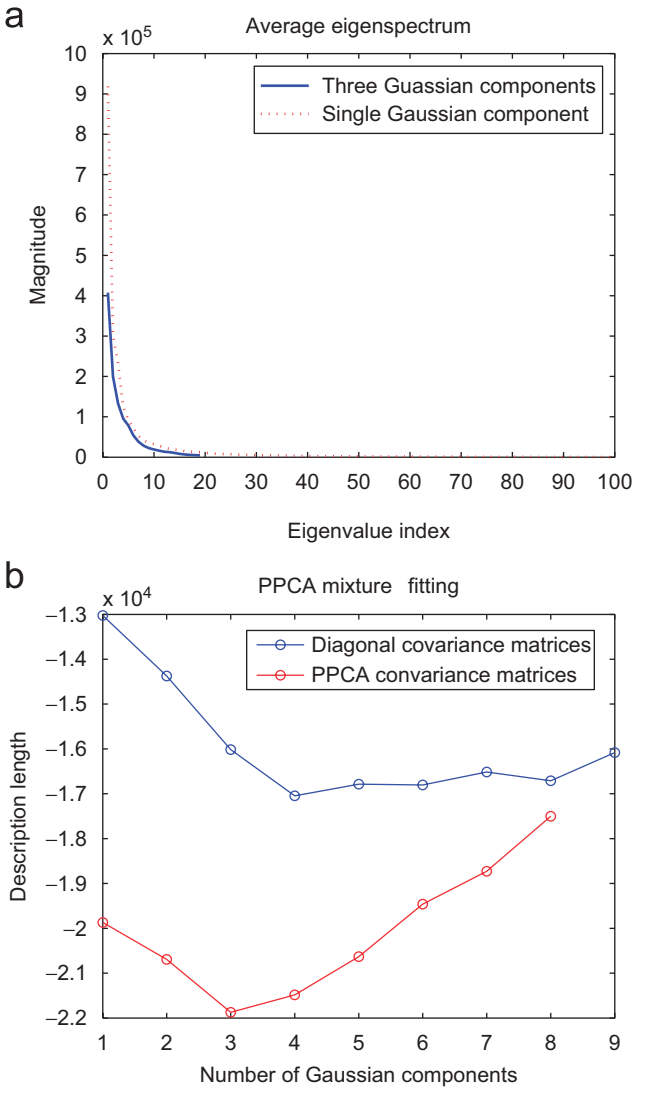

Fig. 5. *Piece-wise linear manifolds*: (a) Average eigenspectrum of diagonal covariance matrices in a typical intermediate GMM fit. The manifold dimension was chosen to represent more than 90% energy, which is around 10. (b) Description length as a function of the number of Gaussian components in the intermediate and final, PPCA-based GMM fitting on a typical data set. The latter results in fewer components and a significantly lower MDL.

Both the intermediate diagonal and the final PPCA mixtures are estimated using the expectation maximization (EM) algorithm [24] which is initialized by K-means clustering. Automatic model order selection is performed using the well-known minimum description length (MDL) criterion [24], see Fig. 5(b). Typically, the optimal (in the MDL sense) number of components for face data sets used in Section 3 was three.

Table 1
Database: Age distribution for database used in the experiments

Age	18–25	26–35	36–45	46–55	65+
Percentage	29%	45%	15%	7%	4%

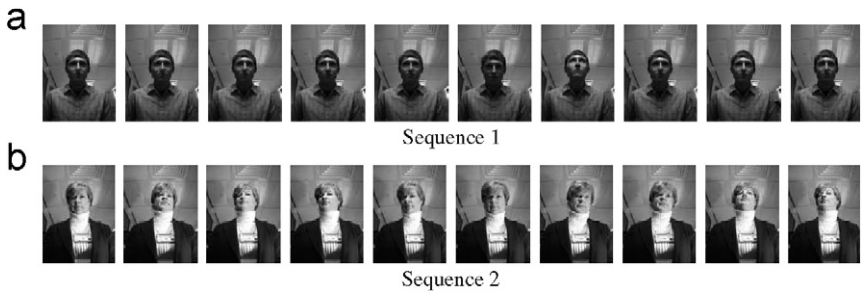

Fig. 6. *Raw data:* Frames from two typical video sequences from the database used for evaluation. The motion of the user was not controlled, leading to different motion patterns and assumed poses.

Fig. 7. *Illuminations:* Different illumination conditions in databases. Note that in spite of the same spatial arrangement of light sources for a particular illumination configuration, its effect on the appearance of faces changes significantly due to variations in people's heights and their *ad lib* chosen position relative to the camera.

Fig. 8. *Data preprocessing*: (a) Left to right—typical input frame from a video sequence of a person performing unconstrained head motion (320 × 240 pixels), output of the face detector (72 × 72 pixels) and the final image after resizing to uniform scale (50 × 50 pixels) and histogram equalization. (b) Typical outliers—face detector false positives—present in our data.

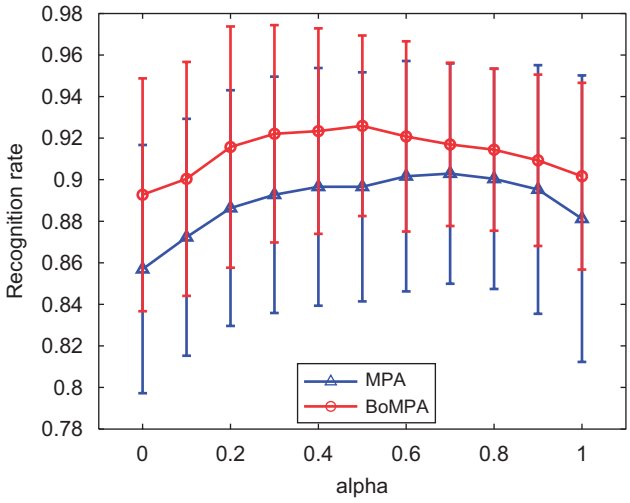

Fig. 9. *Accuracy of MPA and BoMPA for the different setting of* α: The best performance is observed at around 0.5 for both methods.

3. Empirical evaluation

The proposed algorithm was evaluated in the framework of automatic face recognition. We used a database with 100 individuals of varying age (see Table 1) and ethnicity, and equally represented genders. For each person in the database we collected seven video sequences of the person in arbitrary motion (significant translation, yaw and pitch, and negligible roll). The users were instructed not to perform extreme facial expressions but many users talked or smiled during the acquisition, see Fig. 1. Each sequence was recorded in a different illumination setting for 10 s at 10 fps and 320 × 240 pixel resolution, see Fig. 6.[3] After automatic localization using a cascaded detector [25] and cropping to the uniform scale of 50 × 50 pixels, images of faces were histogram equalized, see Fig. 8. Training of all algorithms was performed with data acquired in a single illumination setting and testing with a single other—we used 9 randomly selected training/test combinations, see Figs. 6–8.

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

Table 2
Evaluation results: The mean recognition rate and its standard deviation across different training/test illuminations (in %)

	KLD	NN-HD-PCA	NN-HD-LDA	NN-LDA	FaceIt	MSM	KPA	MPA	BoMPA
Mean	19.8	44.6	40.7	76.6	87.11	84.9	89.1	89.7	92.6
std	9.7	7.9	6.6	7.8	8.8	6.8	10.1	5.5	4.3
Time	7.8	11.8	11.8	11.8	—	0.8	45	7.0	7.0

The last row shows the average time in seconds for 100 set comparisons.

3.1. Methods

We compared the performance of our learning algorithm, without (MPA) and with (BoMPA) boosted feature selection, to that of:

- KL divergence algorithm (KLD) of Shakhnarovich et al. [1],[4]
- MSM of Yamaguchi et al. [11],[4]
- kernel principal angles (KPA) of Wolf and Shashua [14],[5] and
- nearest neighbour (NN) in the sense of the shortest and Hausdorff distance (HD),[6] in (i) LDA [26] and (ii) PCA [27] subspaces, estimated from data,
- NN by FaceIt (v.5.0), the commercial face recognition software of Identix, which ranked top overall in the Face Recognition Vendor test [34,35].

In KLD 90% of data energy was explained by the principal subspace used. In MSM, the dimensionality of PCA subspaces was set to 9 [13]. A sixth degree monomial expansion kernel was used for KPA [14]. In BoMPA, we set the value of parameter α in (5) to 0.5. See the performance of the method for the different setting of α in Fig. 9. All algorithms were preceded with PCA estimated from the entire training data set which, depending on the illumination setting used for training, resulted in dimensionality reduction to around 150 (while retaining 95% of data energy).

3.2. BoMPA implementation

From a practical stand, there are two key points in the implementation of the proposed method: (i) the computation of principal angles between linear subspaces and (ii) time efficiency. These are now briefly summarized for the implementation used in the evaluation reported in this paper. We compute the cosines of principal angles using the method of Björck and Golub [16], as singular values of the matrix $B_1^{\mathrm{T}} B_2$ where $B_{1,2}$ are orthonormal basis of two linear subspaces. This method is numerically more stable than the eigenvalue decomposition-based method used in Ref. [11] and with roughly the same computational demands, see Ref. [16] for a thorough discussion on numerical issues pertaining to the computation of principal angles.

[4]The algorithm was reimplemented through consultation with the authors.

[5]We used the original authors' implementation.

[6]It is defined as $\max_{x_1 \in S_1} \min_{x_2 \in S_2} d(x_1, x_2)$.

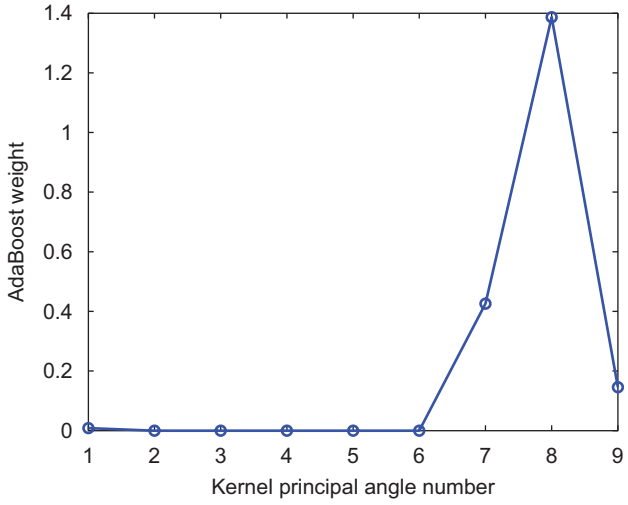

Fig. 10. *Boosted kernel principal angles*: A typical set of weights corresponding to weak kernel principal angle-based classifiers, obtained using AdaBoost. This shows that no discriminatory information was contained in the first few principal angles in the kernel space.

A computationally far more demanding stage of the proposed method is the PPCA mixture estimation. In our implementation, a significant improvement was achieved by dimensionality reduction using the incremental PCA algorithm of Hall et al. [18]. Finally, we note that the proposed model of pattern variation within a set inherently places low demands on storage space.

3.3. Results

The performance of evaluated recognition algorithms is summarized in Table 2. Firstly, the KL-Divergence (KLD) based method achieved by far the worst recognition rate. Seeing that the illumination conditions varied across data and that the face motion was largely unconstrained, the distribution of intra-class face patterns was significant making this result unsurprising. This is consistent with results reported in the literature [6]. Note the much poorer performance of the two NN methods in the Hausdorff distance (HD) sense in PCA and LDA subspaces than the NN method taking the shortest distance in LDA subspace. This can be similarly explained as for the poor performance of the KLD method. The intra-class sets have the significantly different overall distributions and contain only few similar face patterns. The NNs in the HD sense might fail to reflect such similar patterns due to the max operator defined in the HD measure. The NN method in the shortest distance sense in

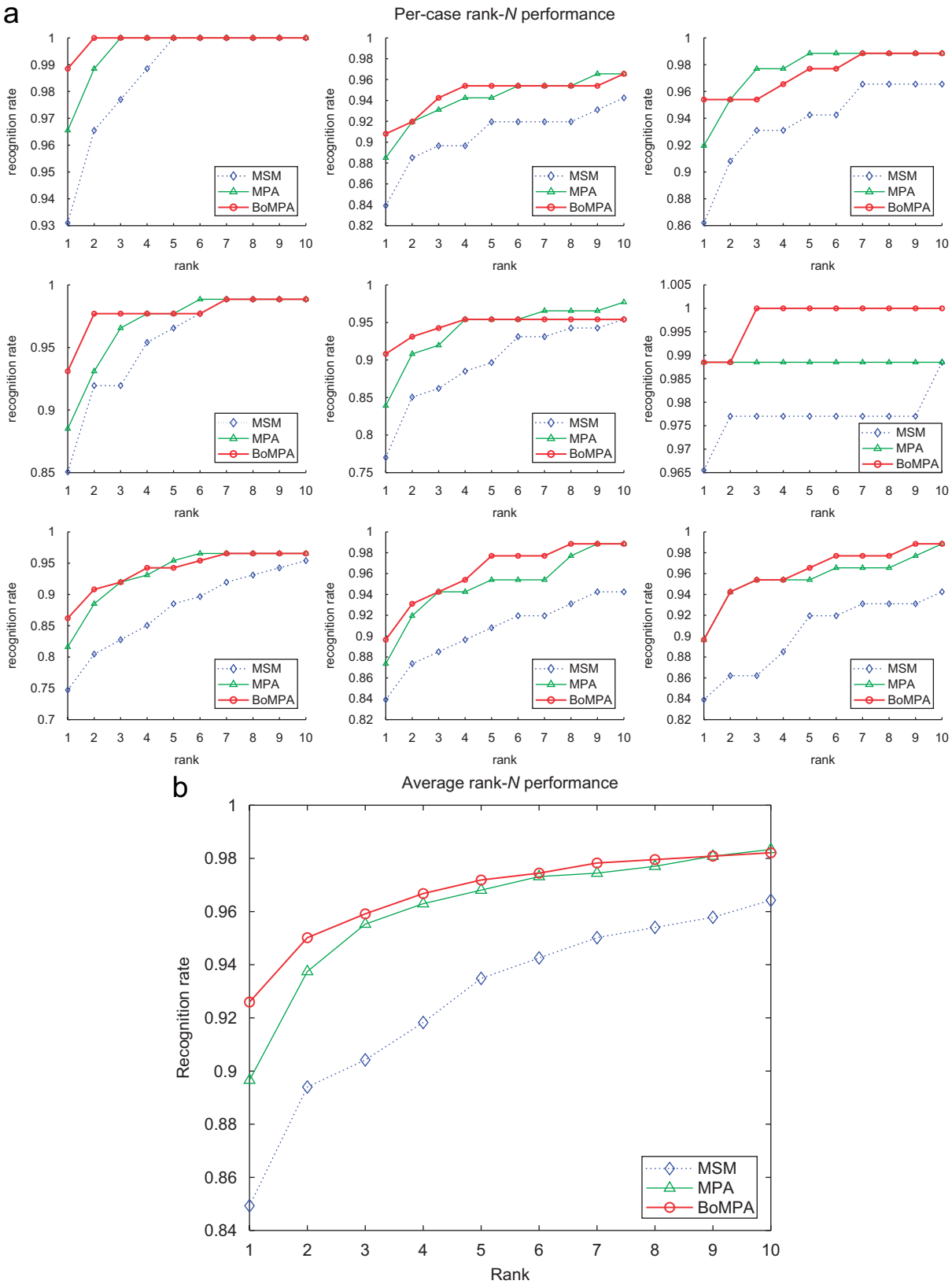

Fig. 11. *Rank-N recognition*: Shown is the improvement in rank-*N* recognition accuracy of the basic MSM, MPA and BoMPA algorithms for (a) each training/test combination and (b) on average. A consistent and significant improvement is seen with nonlinear manifold modelling, which is further increased using boosted principal angles.

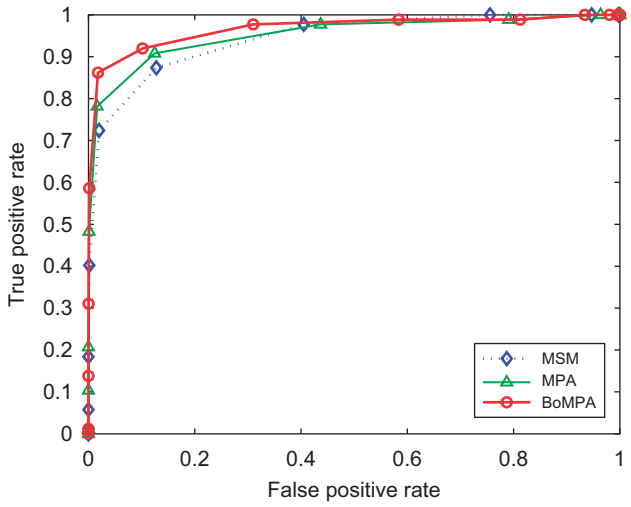

Fig. 12. Receiver–operator characteristic curves of MSM/MPA/BoMPA.

LDA subspace performed much better. The LDA-based methods have long been established in the single-shot face recognition literature, e.g. see Ref. [26,28–31]. FaceIt, which has been ranked at the top of the well-recognized face recognition vendor tests twice, delivered impressive accuracy but with large deviation for the different illumination settings. Note that the original 320×240 images with the localization results by [25] were input to the software, which involves its own training and preprocessing of face images [34,35]. This method can be considered as a proxy for gauging the difficulty of the recognition task.

The performance of the four principal angle-based methods confirms the premises of our work. Basic MSM performed well, but worst of the four. The inclusion of nonlinear manifold modelling, either by using the "kernel trick" or a mixture of linear subspaces, achieved an increase in the recognition rate of about 5%. While the difference in the average performance of MPA and the KPA methods is probably statistically insignificant, it is worth noting the greater robustness to specific imaging conditions of our MPA, as witnessed by a much lower standard deviation of the recognition rate. Further performance increase of 3% is seen with the use of boosted angles, the proposed BoMPA algorithm correctly recognizing 92.6% of the individuals with the lowest standard deviation of all methods compared. For completeness, we also combined the KPA method with the proposed boosting observing 2.2% accuracy gain over the KPA and 10% standard deviation. The typical weights learnt for the first few KPAs are shown in Fig. 10, where the first six angles appear to be ineffective for discrimination in kernel space. This is contrasted with the case of linear subspaces where only the first one is less important than the successive few angles as shown in Fig. 4(a). More flexibility of the first few KPAs in finding similar modes might result in the loss of discrimination powers. An illustration of the improvement provided by each novel step in the proposed algorithm is shown in Fig. 11. Moreover, the receiver–operator characteristic curve in Fig. 12 compares the three methods, MSM/MPA/BoMPA in terms of ratio of the true positive rate over the false positive rate. Finally, its computational superiority to the best performing method in the literature, Wolf and Shashua's KPA, is clear from a 7-fold difference in the average recognition time.

4. Conclusions and future work

In this paper we introduced a novel method for discrimination over vector sets. Our approach was based on modelling pattern variations within a set and comparing them using principal angles. We showed that principal angles provide an effective means of comparing only the most similar regions of two linear subspaces, while achieving numerical stability and robustness to noise. Our first contribution was to introduce a learning framework by which focus is put on the most discriminative regions of the subspaces. Next, we extended the method to more effectively model non-linear pattern variations within a set and proposed an extended similarity criterion. In an extensive empirical evaluation it was demonstrated to perform better than state-of-the-art algorithms in the literature on the task of face recognition from image sets, extracted from 700 face motion video sequences and including 70,000 detected faces.

The main research direction we intend to pursue in the future is the extension of the concept of principal angles to comparisons of probability densities. This would allow us to avoid the hard cut-off of higher dimensions of linear subspaces that are being compared. Additionally, it may prove beneficial to incorporate more specific domain knowledge, in particular illumination models, in guiding the mixture component estimation. Finally, an interesting application of our work could be to use an ensemble of BoMPA learners for object recognition using local image features.

Acknowledgements

The authors would like to express their gratitude to Josef Kittler whose valuable comments and suggestions helped this research. Funding was kindly provided by Chevening Scholarship, Toshiba Corporation and Trinity College, Cambridge.

References

[1] G. Shakhnarovich, J.W. Fisher, T. Darrel, Face recognition from long-term observations, in: Proceedings of the IEEE European Conference on Computer Vision, vol. 3, 2002, pp. 851–868.
[2] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Statist. Soc. 3 (61) (1999) 611–622.
[3] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[4] O. Arandjelović, R. Cipolla, An information-theoretic approach to face recognition from face motion manifolds, Image Vision Comput. 24 (5) (2006) in press.
[5] D.H. Johnson, S. Sinanović, Symmetrizing the Kullback–Leibler distance, Technical Report, Rice University, 2001.
[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, vol. 1, 2005, pp. 581–588.
[7] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–372.

[8] R. Gittins, Canonical analysis: a review with applications in ecology, Biomathematics 12 (1985).
[9] T. Kailath, A view of three decades of linear filtering theory, IEEE Trans. Inform. Theory 20 (2) (1974) 146–181.
[10] E. Oja, Subspace Methods of Pattern Recognition, Research Studies Press, Wiley, New York, 1983.
[11] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, vol. 10, 1998, pp. 318–323.
[12] A. Maki, K. Fukui, Ship identification in sequential isar imagery, Mach. Vision Appl. 15 (3) (2004).
[13] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, Int. Symp. Robot. Res. (2003).
[14] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, J. Mach. Learn. Res. 4 (10) (2003) 913–931.
[15] B. Schölkopf, A. Smola, K. Müller, Kernel principal component analysis, Adv. Kernel Methods (1999) 327–352.
[16] Å. Björck, G.H. Golub, Numerical methods for computing angles between linear subspaces, Math. Comput. 27 (123) (1973) 579–594.
[17] O. Arandjelović, R. Cipolla, Incremental learning of temporally coherent Gaussian mixture models, in: Proceedings of the IAPR British Machine Vision Conference, vol. 2, 2005, pp. 759–768.
[18] P. Hall, D. Marshall, R. Martin, Merging and splitting eigenspace models, IEEE Trans. Pattern Anal. Mach. Intell. (2000).
[19] Y. Li, J. Xu, L. amd Morphett, R. Jacobs, An integrated algorithm of incremental and robust PCA, Proceedings of IEEE International Conference on Image Processing (2003).
[20] D. Skocaj, A. Leonardis, Weighted and robust incremental method for subspace learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 1494–1501.
[21] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: Proceedings of the Second European Conference on Computational Learning Theory, 1995, pp. 23–37.
[22] K. Maeda, O. Yamaguchi, K. Fukui, Towards 3-dimensional pattern recognition, Statist. Pattern Recognition 3138 (2004) 1061–1068.
[23] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Comput. 11 (2) (1999) 443–482.
[24] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley, New York, 2000.
[25] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vision 57 (2) (2004) 137–154.
[26] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[27] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.
[28] W. Zhao, R. Chellappa, A. Krishnaswamy, Discriminant analysis of principal components for face recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 336–341.
[29] M.T. Sadeghi, J.V. Kittler, Decision making in the lda space: generalised gradient direction metric, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 248–253.
[30] X. Wang, X. Tang, Random sampling lda for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 259–265.
[31] T.-K. Kim, J.V. Kittler, Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 318–327.
[32] T.-K. Kim, J. Kittler, R. Cipolla, Learning discriminative canonical correlations for object recognition with image sets, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 251–262.
[33] T.-K. Kim, O. Arandjelović, R. Cipolla, Learning over sets using boosted manifold principal angles (BoMPA), in: Proceedings of the British Machine Vision Conference, 2005, pp. 779–788.
[34] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E. Tabassi, J.M. Bone, FRVT 2002: Evaluation Report, Mar. 2003. ⟨http://www.frvt.org/FRVT2002/⟩.
[35] D.M. Blackburn, M. Bone, P.J. Phillips, Facial Recognition Vendor Test 2000: Evaluation Report, 2000.

About the Author—TAE-KYUN KIM is a Ph.D. student of Machine Intelligence Laboratory at the University of Cambridge. He received the B.Sc. and M.Sc. degrees from the Dept. of EECS at the Korea Advanced Institute of Science and Technology (KAIST) in 1998 and 2000, respectively. He worked as a research staff member at Samsung Advanced Institute of Technology, Korea during 2000–2004. His research interests include computer vision, statistical pattern classification and machine learning. He regularly reviews for the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) and was the Korea delegate of MPEG-7. The joint proposal of face descriptor of Samsung and NEC, for which he developed the main algorithms, was accepted as the international standard of ISO/IEC JTC1/SC29/WG11.

About the Author—OGNJEN ARANDJELOVIĆ is a Ph.D. student in the Machine Intelligence Laboratory at the University of Cambridge. He graduated top of his class from the Department of Engineering at the University of Oxford in 2003. His research interests include computer vision and machine learning, and their application in other scientific disciplines. He is a research scholar of Trinity College, Cambridge and a Fellow of the Cambridge Overseas Trust.

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