# Learning over Sets using Boosted Manifold Principal Angles (BoMPA)

Tae-Kyun Kim Ognjen Arandjelović Roberto Cipolla
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ, UK
{tkk22,oa214,cipolla}@eng.cam.ac.uk

**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.*

## 1 Introduction

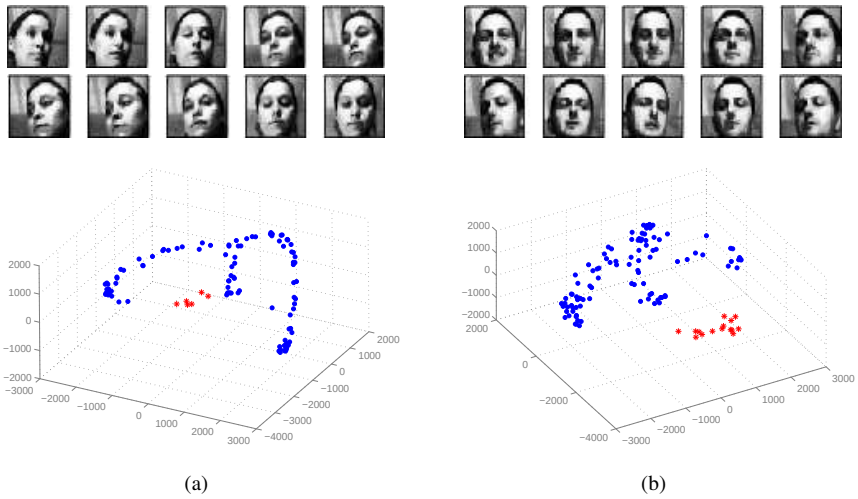
Many computer vision tasks can be cast as learning problems over vector *sets*. In object recognition, for example, a set of vectors may represent a variation in an object's appearance – be it due to camera pose changes, non-rigid deformations or variation in illumination conditions. The objective of this work is to classify a novel set of vectors to one of the training classes, each also represented by a vector set.

**Problem challenges** Pattern variations with a class are usually complex and nonlinear (see Figure 1 for example), often with greater intra then inter class differences, e.g. see [1]. This makes their modelling difficult, requiring models expressive enough to capture such complex behaviour, yet simple enough to allow for efficient estimation in the presence of missing data. The problem is further complicated by the sheer volume of data – practical limitations in terms of available storage space and time constraints on recognition frequently demand compact models that can be rapidly matched.

### 1.1 Previous Work

Most of the previous work on matching vector or image sets exploits their semantics to a certain degree, typically 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

BMVC 2005 doi:10.5244/C.19.58



(a) (b)

Figure 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).*

recent literature we recognize two groups of approaches to learning over sets of vectors: statistical and principal-angle based.

**Statistical methods** Statistical learning approaches rely on the assumption that vectors $\mathbf{x}$ of the $i$-th 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.* [19], densities $p^{(i)}(\mathbf{x})$ are modelled as multivariate Gaussians, estimated with Probabilistic PCA [21] and compared using the Kullback-Leibler (KL) divergence [6]. 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 [2]. 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 [13] on the task of AFR under mildly varying imaging conditions. In [3], Arandjelović *et al.* proposed a Gaussian Mixture Model 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 of increased difficulty of divergence computation, performed using a Monte-Carlo algorithm.

**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 [12], it has been applied to in various fields [10, 14, 17]. Of most relevance to the work addressed in this paper is the Mutual Subspace Method (MSM) of

Yamaguchi *et al.* [24]. 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 [24] and ship identification [16] (for evaluation results also see [2, 3]). In the related Constrained MSM [9], vector sets are projected to the linear Constraint 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 [2]. In [23] Wolf and Shashua show how principal angles between nonlinear subspaces can be computed using the "kernel trick" [18]. 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 [5], while the estimation of linear subspaces can be performed in an incremental manner [11, 20].

**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.* [3], 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.

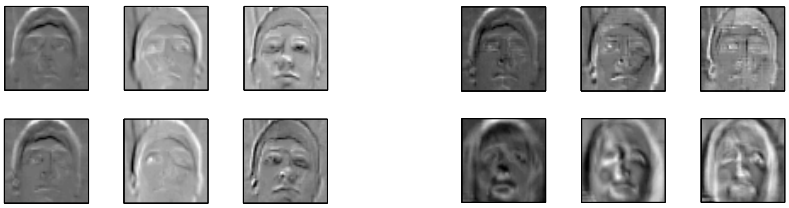
# 2 Boosted Manifold Principal Angles (BoMPA)

Principal, or canonical, angles $0 \leq \theta_1 \leq \ldots \leq \theta_D \leq (\pi/2)$ 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:

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

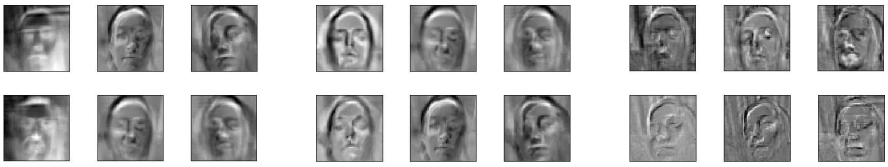
[1] In statistics, the cosines of canonical angles are termed canonical correlations.

[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.



(a) (b)

Figure 2: ***Principal vectors in MSM:*** *The first 3 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.*



(a) (b) (c)

Figure 3: ***MSM, BPA and MPA:*** *(a) The first 3 principal vectors between two linear subspaces which MSM incorrectly classifies as corresponding to the same person (the two data sets are shown in Figure 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.*

subject to:

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

We will refer to $\mathbf{u}_i$ and $\mathbf{v}_i$ as the $i$-th 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 Figure 2 on the example of sets of face appearance images.

## 2.1 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 changes corresponding to extreme illumination conditions, see Figure 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.

**Boosted Principal Angles** In general, each principal angle $\theta_i$ carries some information for discrimination between the corresponding two subspaces. We use this to build simple weak classifiers $\mathscr{M}(\theta_i) = \text{sign}\,[\cos(\theta_i) - C]$. In the proposed method, these are combined using the now acclaimed AdaBoost algorithm [8]. In summary, AdaBoost learns a weighting $\{w_i\}$ of decisions cast by weak learners to form a classifier $\mathscr{M}(\Theta)$:

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

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 Principal Component Analysis (PCA). On the other hand, the $K$ in-class principal angles are computed between a pair of randomly drawn subsets for each $S_i$.

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

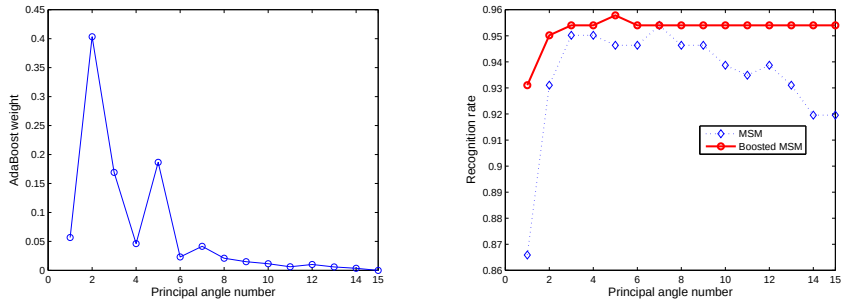
$$f(\Theta) = \frac{1}{N}\sum_{i=1}^{N} w_i \cos(\theta_i) / \sum_{i=1}^{N} w_i \tag{4}$$

A typical set of weights $\{w_i\}$ we obtained for our AFR application is shown graphically in Figure 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. Figure 3 (b) shows the 3 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 Figure 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, 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 special provision for choosing the optimal number of principal angles is needed.

At this point it is worthwhile mentioning the work of Maeda *et al.* [15] 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.

## 2.2 Nonlinear Subspaces

The assumption that patter variations within each class are well represented by a linear subspace is usually severely limiting, see Figure 1. Our aim is to extend the described




(a) (b)

Figure 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.*

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.

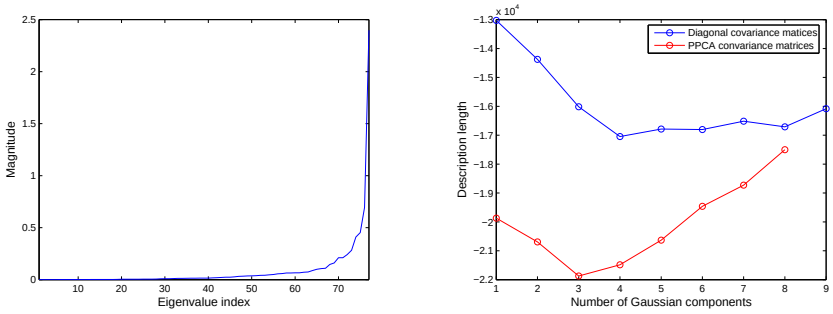
Without the loss of generality, let $S_1$ and $S_2$ be two vector sets and $\Theta$ the set of principal angles between two linear subspaces. We derive a measure of similarity $\rho$ between $S_1$ and $S_2$ by comparing the corresponding linear subspaces $U_{1,2}$ and locally linear patches $L_{1,2}^{(i)}$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

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

where $f_G$ and $f_L$ have the same functional form as $f$ in (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 Figure 3 (c)

**Finding stable locally linear patches** In the proposed framework, stable locally linear manifold patches are found using Mixtures of Probabilistic PCA (PPCA) [21]. 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 Gaussian Mixture Model (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 Figure 5 (a). Once $d$ is estimated (typically $d \ll D$), the fitting is repeated using a Mixture of PPCA.

Both the intermediate diagonal and the final PPCA mixtures are estimated using the




(a) (b)

Figure 5: ***Piece-wise Linear Manifolds:*** *(a) Average eigenspectrum of diagonal covariance matrices in a typical intermediate GMM fit. The approximate intrinsic manifold dimensionality can be seen to be 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.*

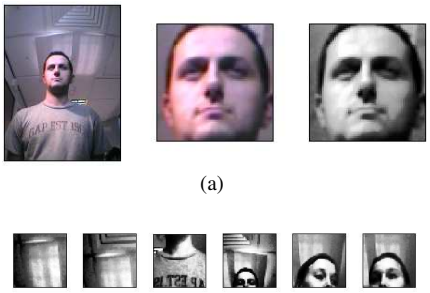
Expectation Maximization (EM) algorithm [7] which is initialized by K-means clustering. Automatic model order selection is performed using the well-known Minimum Description Length (MDL) criterion [7], see Figure 5 (b). Typically, the optimal (in the MDL sense) number of components for face data sets used in Section 3 was 3.

# 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 and ethnicity, and equally represented genders. For each person in the database we collected 7 video sequences of the person in arbitrary motion (significant translation, yaw and pitch, and negligible roll). The users were instructed not to perform extreme facial expressions but many users talked or smiled during the acquisition, see Figure 1. Each sequence was recorded in a different illumination setting for 10s at 10fps and $320 \times 240$ pixel resolution. After automatic localization using a cascaded detector [22] and cropping to the uniform scale of $50 \times 50$ pixels, images of faces were histogram equalized, see Figure 6. 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.

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

- KL divergence algorithm (KLD) [19] ,
- Mutual Subspace Method (MSM) [24],
- Kernel Principal Angles (KPA) [23], and
- Nearest Neighbour (NN) in the Hausdorff distance sense in (i) LDA [4] and (ii) PCA subspaces, estimated from data.



(a)

(b)

Figure 6: ***Data preprocessing:*** *(a) Left to right – typical input frame from a video sequence of a person performing unconstrained head motion ($320 \times 240 pixels$), output of the face detector ($72 \times 72$ pixels) and the final image after resizing to uniform scale ($50 \times 50$ pixels) and histogram equalization. (b) Typical outliers – face detector false positives – present in our data.*

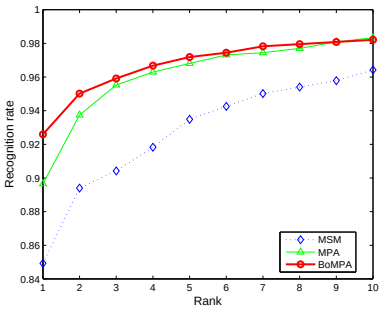



Figure 7: ***Rank-N Recognition:*** *Shown is the improvement in rank-N recognition accuracy of the basic MSM, MPA and BoMPA algorithms. A consistent and significant improvement is seen with nonlinear manifold modelling, which is further increased using boosted principal angles.*

In KLD 90% of data energy was explained by the principal subspace used. In MSM, the dimensionality of PCA subspaces was set to 9 [9]. A sixth degree monomial expansion kernel was used for KPA [23]. In BoMPA, we set the value of parameter $\alpha$ in (5) to 0.5. All algorithms were preceded with PCA estimated from the entire training dataset which, depending on the illumination setting used for training, resulted in dimensionality reduction to around 150 (while retaining 95% of data energy).

**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 [5], as singular values of the matrix $B_1^T B_2$ 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 [24] and equally efficient, see [5] for details. 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.* [11]. Finally, we note that the proposed model of pattern variation within a set inherently places low demands on storage space.

## 3.1 Results

The performance of evaluated recognition algorithms is summarized in Table 1. Firstly, note the relatively poor performance of the two nearest neighbour-type methods – the Hausdorff NN in PCA and LDA subspaces. These can be considered as proxies for gauging the difficulty of the recognition task, seeing that both can be expected to perform

| | KLD | NN-LDA | NN-PCA | MSM | KPA | MPA | BoMPA |
|---|---|---|---|---|---|---|---|
| **mean** | 19.8 | 40.7 | 44.6 | 84.9 | 89.1 | 89.7 | 92.6 |
| **std** | 9.7 | 6.6 | 7.9 | 6.8 | 10.1 | 5.5 | 4.3 |
| **time** | 7.8 | 11.8 | 11.8 | 0.8 | 45 | 7.0 | 7.0 |

Table 1: ***Evaluation results:*** *The mean recognition rate and its standard deviation across different training/test illuminations (in %). The last row shows the average time in seconds for 100 set comparisons.*

relatively well if the imaging conditions do not greatly differ between training and test data sets. The KL-divergence 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 [3].

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. An illustration of the improvement provided by each novel step in the proposed algorithm is shown in Figure 7. 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

BoMPA, a novel method for discrimination over vector sets was introduced. 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 video.

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. Another interesting direction could be to use an ensemble of BoMPA learners for object recognition using local image features.

# Acknowledgements

The authors would like to express their deep gratitude to Josef Kittler for his valuable comments and suggestions. Funding of this research was kindly provided by Toshiba Corporation and Trinity College, Cambridge.

# References


[1] Y Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[2] O. Arandjelović and R. Cipolla. Face recognition from face motion manifolds using robust kernel resistor-average distance. *IEEE Workshop on Face Processing in Video*, 5:88, 2004.

[3] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[5] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wily & Sons, Inc., New York, 2nd edition, 2000.

[8] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Proceedings of the 2nd European Conference on Computational Learning Theory*, pages 23–37, 1995.

[9] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *Int'l Symp. of Robotics Research*, 2003.

[10] R. Gittins. Canonical analysis: A review with applications in ecology. *Biomathematics*, 12, 1985.

[11] P. Hall, D. Marshall, and R. Martin. Merging and splitting eigenspace models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–372, 1936.

[13] D. H. Johnson and S. Sinanović. Symmetrizing the Kullback-Leibler distance. *Technical report, Rice University*, 2001.

[14] T. Kailath. A view of three decades of linear filtering theory. *IEEE Transactions on Information Theory*, 20(2):146–181, 1974.

[15] K. Maeda, O. Yamaguchi, and K. Fukui. Towards 3-dimensional pattern recognition. *Statistical Pattern Recognition*, 3138:1061–1068, 2004.

[16] A. Maki and K. Fukui. Ship identification in sequential isar imagery. *Machine Vision and Applications*, 15(3), 2004.

[17] E. Oja. *Subspace Methods of Pattern Recognition.* Research Studies Press and J. Wiley, 1983.

[18] B. Schölkopf, A. Smola, and K. Müller. Kernel principal component analysis. *Advances in Kernel Methods - SV Learning*, pages 327–352, 1999.

[19] G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *In Proc. IEEE European Conference on Computer Vision*, pages 851–868, 2002.

[20] D. Skocaj and A. Leonardis. Weighted and robust incremental method for subspace learning. *In Proc. IEEE International Conference on Computer Vision*, pages 1494–1501, 2003.

[21] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[22] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[23] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4(10):913–931, 2003.

[24] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *IEEE International Conference on Automatic Face and Gesture Recognition*, (10):318–323, 1998.