Randomised Manifold Forests
for Principal Angle-Based Face Recognition

Ujwal D. Bonde[1], Tae-Kyun Kim[2], and K.R. Ramakrishnan[1]

[1] Department of Electrical Engg., Indian Institute of Science, Bangalore, India
[2] Department of Engineering, University of Cambridge, Cambridge, UK

Abstract. In set-based face recognition, each set of face images is often represented as a linear/nonlinear manifold and the Principal Angles (PA) or Kernel PAs are exploited to measure the (dis-)similarity between manifolds. This work systemically evaluates the effect of using different face image representations and different types of kernels in the KPA setup and presents a novel way of randomised learning of manifolds for set-based face recognition. First, our experiments show that sparse features such as Local Binary Patterns and Gabor wavelets significantly improve the accuracy of PA methods over 'pixel intensity'. Combining different features and types of kernels at their best hyper-parameters in a multiple classifier system has further yielded the improved accuracy. Based on the encouraging results, we propose a way of randomised learning of kernel types and hyper-parameters by the set-based Randomised Decision Forests. We observed that the proposed method with linear kernels efficiently competes with those of nonlinear kernels. Further incorporation of discriminative information by constrained subspaces in the proposed method has effectively improved the accuracy. In the experiments over the challenging data sets, the proposed methods improve the accuracy of the standard KPA method by about 35 percent and outperform the Support Vector Machine with the set-kernels manually tuned.

1 Introduction

Unlike traditional access control scenarios, face recognition in dynamic environments is yet extremely challenging due to uncontrolled lighting conditions, large pose variations, facial expressions and severe occlusions. For the past decade set-based face recognition has gained a huge interest in related fields. Over conventional single-shot based face recognition, the main benefits have been two folds: a) its ability to represent and match data over a combination of face exemplars and b) its natural extension to videos where a tracked object can be represented as a set of images. This has led to significant improvement in accuracy and efficiency for face recognition.

Among different methods for set-based face recognition, the most widespread one is the Principal Angle (PA) method. It represents a set of face images as a subspace and matches one set to another set using subspace angles. Despite the popularity of the Principal Angle based methods, it has not received much

attention on its efficacy using state-of-the-art face image representations (e.g. Local Binary Patterns, Gabor features) other than 'pixel intensity': one reason for this could be that their very sparse representations might be thought difficult to be constrained on linear subspaces. Nonlinear extension of the Principal Angle method by a kernel trick [26] or a set of linear subspaces [15,31] and discriminative versions of the PA technique e.g. Constrained Mutual Subspace Method [8] have been successfully developed. They have shown significantly improved accuracy over the standard method but their good performance is highly subjective to the settings in the methods. In the Kernel Principal Angle technique (KPA) [26], it is not a trivial problem to automatically set the best types of kernels and kernel hyper-parameters. This paper systematically evaluates the Principal Angle methods over a number of respective issues and proposes a novel way of randomised learning for the PA methods using Randomised Decision Forests [2,3]. In this work we look at the following key areas in the framework of Principal Angle based face recognition:

- Performance of features such as Local Binary Patterns (LBP) and Gabor wavelets, both are sparse representations, over the pixel intensity representation.
- Combining different features and kernels for the KPA method by a multiple classifier system.
- Proposing randomised manifold (or kernel) learning by Random Decision Forests.
- Using the idea of CMSM to incorporate discriminative information for the proposed method.

We demonstrate these for a video-based face recognition problem.

Rest of the paper is structured as follows. In Section 2 we briefly review related work. Section 3 details KPA using non-linear feature extraction techniques and LBPs, followed by the method for combining these features with different kernels as a multiple classifier system. The Random Manifold Forest is proposed in Section 4. Experimental setup and the results obtained are presented in Section 5. We conclude our work in Section 6.

2   Related Work

Use of the principal angles(PA) for matching sets of face images was initially proposed by Yamaguchi *et al.* [28]. This has become more widely applicable since the Kernalized version was proposed by Wolf *et al.* [26]. A large number of related methods including Boosted Manifold Principal Angles [15], Constrained Mutual Subspace Method(CMSM) [8] and Orthogonal Subspace Method(OSM) [9] have been proposed as an improvement over the original PA or KPA method. The PA-based methods have shown superior to other alternatives such as parametric distribution matching and simple aggregation of individual sample matching consistently in literature e.g. [32,31]. However all of the PA methods above have considered raw pixel intensity images for their input and have not paid much attention to representations.

Nonlinear extension of the PA methods has been obtained largely either by a kernel technique [26] or by expressing a manifold as a set of linear subspaces [15, 31]. Although they have been shown better than the standard PA method, their good performance is highly dependent on how to form a nonlinear subspace or manifold. Despite the popularity of the KPA for face recognition, it has not received attention on its effectiveness using different kernels and hyperparameters. Based on the encouraging results by the LBP and Gabor features, we investigate a way to combine different features and different types of kernels in a multiple classifier system, firstly assuming the best hyper-parameters given, and later propose a novel method of randomised learning for both *kernel types* and their *hyper-parameters* by the KPA and random decision forests [2].

CMSM as a discriminative method has been shown to significantly improve standard PA and KPA techniques. It has since then been used for automatic character listing in videos [14], recognition in image sets [8]. The main drawback in CMSM is the choice of the *sum space* and its dimensionality. Methods such as multiple CMSM [29] and boosted CMSM [30] have been proposed to address this problem to a certain extent. But its efficacy is still restricted to a certain range that needs to be estimated. Here we propose a method that circumvents this question similar to the problem of choosing a kernel and its hyper-parameter in the PA or KPA methods.

Multiple classifier system refers to techniques to aggregate the evidences from multiple sources (or classifiers) and typically provides better performance than individual base classifiers. These techniques have been widely used in combining results obtained in biometric systems and also in face recognition example [11, 16]. A large number of methods have been developed for combining the classifier outputs at different levels: the simplest yet robust methods are the sum and the product rules which combine the classifiers at the measurement/or confidence level [16]. These methods assume that individual classifiers are uncorrelated. Some methods fuse classifiers at the classifier confidence level, establishing the classifier weights by their performances on a validation set [6]. Mixture of experts (MoE) [10] jointly learns multiple classifiers, their weights and data partitions for binary class problems. This was extended to multi-class problems in Chen *et al.* [5]. MoE provides an unified framework of multiple classifier learning and fusion, though it resorts to a local optimal solution due to the iterative algorithm, EM used for optimisation. We have formulated a novel closed-form solution for learning classifier weights for *multi-class* problems and have shown that this outperforms the sum, product, minimum score, maximum score and weighted sum rules, where the weights were set according to the classifier accuracies.

Random Decision Forests (RF) introduced by Breimen [2] and Geurts *et al.* [3] is a powerful ensemble learning technique and has been used in various applications such as image segmentation [23], classification [1] and tracking problems and has shown competitive results in these areas. It is inherently for multi-classes and shows fast learning and classification performance. Randomised learning is useful particularly when features to be learnt are difficult to be explicitly

represented due to a high dimensional space. Comparative studies have been performed on the accuracy of RF against Support Vector Machines (SVM) [24] and in many cases, for example in Gene selection [7], RF was shown superior to the traditional SVMs. In this paper, a novel method for randomised kernel learning is proposed by RF and in the process, Random Manifolds are defined. In the experiments over the challenging data sets, the proposed methods have been shown to outperform the Support Vector Machine with the set-kernels manually tuned.

3 Combining Features and Kernels for KPA

3.1 Base Classifier Design

Previous face recognition methods based on PA have used raw pixel values as features. We use LBP (and Gabor) features which have gained an increasing interest owing to its good performance for classification. Despite having sparse representations these features have shown to perform very well in our setup and have significantly improved the accuracy over the exisitng methods using raw pixels. In addition, we consider nonlinear feature extraction in KPA before computing the principal angles. In the previous KPA method [26] the dimension of the set subspace is fixed as the set cardinality. However, the intrinsic dimension of the subspace by the set is in general much lower than the number of data points for faces as shown by Kriegman *et al.* [20]. We apply the Kernel PCA technique [22] to get 'k' ($\ll$ cardinality of set) dimensional approximation of the original subspace before calculating the principal angles. We get the eigenvectors associated with the k largest eigenvalues as $Q_{k(\Phi(A))}, Q_{k(\Phi(B))}$ for the reduced dimensional subspaces. The principal angles between two reduced subspaces are then computed using a kernel trick [26]. Each base classifier is defined as Nearest Neighbor (NN) classifier in terms of the KPA similarity as

$$d(A, B) = \frac{1}{k} \sum_{i=1}^{k} \cos \theta_i \qquad (1)$$

where A, B are two sets of the LBP (or Gabor) vectors and $\cos \theta_i, i = 1, ..., k$ are the kernel principal angles for the reduced dimensional subspaces and the kernel used. Here all k principal angles are equally considered and feature selection for better recognition accuracy [15] is left as future work. The two features, LBP [27] and Gabor, and three different kernels are used: Gaussian kernel is defined as $K(x, y) = \exp(\frac{\|x-y\|_2^2}{\sigma^2})$, Fractional power kernel as $K(x, y) = (sign(x^T y) \times (x^T y))^a$, $0 < a < 1$, and Polynomial kernel as $K(x, y) = (x^T y)^b$, $b \geq 1$ respectively.

3.2 Combining Features and Kernels

We propose a novel way of learning classifier weights in multiple classifier system by a closed-form solution. We later show in experiments that this technique

outperforms some baseline methods. The classifiers obtained using two different features (LBP and Gabor) and three different kernels are considered giving a total of six base classifiers.

Let, $\hat{y}_i^c \in \mathbb{R}^Z$ be an indicator vector representing the predicted output of the c-th classifier for the i-th sample with one in the predicted class and zeros elsewhere, $y_i \in \mathbb{R}^Z$ is an indicator vector for the true class label where Z is the number of classes. The cost F is defined by:

$$F = \min_{w^c} \sum_i || \sum_c (w^c \hat{y}_i^c) - y_i ||_2^2. \tag{2}$$

Also, if $\hat{Y}_i = [\hat{y}_i^1, \hat{y}_i^2, \ldots, \hat{y}_i^C] \in \mathbb{R}^{Z \times C}$ and $w = [w^1, w^2, \ldots, w^C]^T \in \mathbb{R}^{C \times 1}$, where C is the number of classifiers (here six), then the cost is rewritten as:

$$F = \min_w \sum_i ||\hat{Y}_i w - y_i||_2^2.$$

Now if $\hat{Y} = [\hat{Y}_1^T, \hat{Y}_2^T, \ldots, \hat{Y}_N^T]^T \in \mathbb{R}^{NZ \times C}$ and $Y = [y_1^T, y_2^T, \ldots, y_N^T]^T \in \mathbb{R}^{NZ \times 1}$ where N is the number of data points (or data sets), then the cost function can be further rewritten as: $F = \min_w ||\hat{Y}w - Y||_2^2 \Rightarrow F = \min_w \left(w^T \hat{Y}^T \hat{Y} w - w^T \hat{Y}^T Y - Y^T \hat{Y} w - Y^T Y \right)$. Differentiating it with respect to w and equating it to zero gives:

$$\Rightarrow w_F = \left(\hat{Y}^T \hat{Y} \right)^{-1} \hat{Y}^T Y. \tag{3}$$

Thus a least square solution for w is obtained using the cost function F.
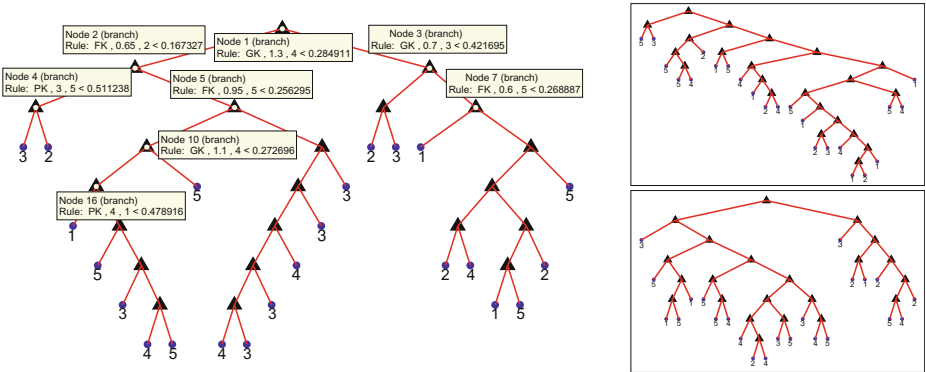

Fig. 1. Example trees in random manifold forests. *(left) Decision Rule: First two characters represent the type of kernel ('GK': Gaussian, 'PK': Polynomial, 'FK': Fractional) this is followed by the hyper-parameters of the kernel, next is the reference set followed by the threshold. (right) More example trees in the forests.*

4   Random Manifold Forests

Until this section we have not discussed how to obtain the hidden parameters like the hyper-parameters used by the kernels in KPA or the constraint subspace dimension in CMSM. In this section we propose the set based Random Forest method and define Random Manifolds for randomised kernel learning. We have also tested the proposed method with linear kernels and have observed it to efficiently compete with nonlinear kernels. Encouraged by this we go on to propose a method to incorporate discriminative information.

4.1   Random Manifold Forests for Randomised Kernel Learning

We begin by considering a reference set R for each class. At every node we randomly select the following: 1) a kernel type, 2) kernel hyper-parameters and 3) a reference set. The split function at a node for a data set X is defined by

$$f(X, R) = d(X, R) - t = \frac{1}{k} \sum_{i=1}^{k} \cos \theta_i - t, \tag{4}$$

where k is the reduced subspace dimension, R is the reference set randomly chosen at that node and d is the sum of kernel principal angles for the chosen type of kernel and its hyper-parameters. Note that a set of vectors is taken as input of the spit function, whereas a single vector is the input in traditional RFs This choice is repeated m times from which we select the one that gives us the best split in terms of the information gain [3] [2]. Figure 1 shows the example trees built using this method. The decision taken at few of the nodes for one of the trees is also displayed. At every node we observe that the tree is projecting the data sets to a different feature space depending upon the choice of the kernel and its hyper-parameters. This feature space is split into two regions based on the choice of the reference set R and the threshold t calculated at that node. A decision is taken based on the region in which the subspace spanned by the test data set X lies. The decision surface is given by

$$\mathbb{L} : f(X, R) = 0 \tag{5}$$

It is the separating region at that node. Based on this region the sets X will either go to the left or the right child of the node, i.e:

$$\begin{aligned} I_l &= \{i | f(X_i, R) < 0\} \\ I_r &= I_n \setminus I_l \end{aligned} \tag{6}$$

where I_n is the total data sets arriving at the node n.

For a better intuition let us consider that at a particular node in a tree the sets are projected into a three dimensional feature space and the reference set spans a line as shown in Figure 2 then, the split function and the separating region are given by

$$f(X, R) = \cos \theta_1 - t.$$

$$\mathbb{L} : f(X, R) = 0 \Rightarrow \cos \theta_1 = t. \tag{7}$$

i.e, the threshold t parameterizes a cone as a separating surface. All the sets which span a line (plane) that lies in (passes through) this cone go to the left child and the rest to the right child of this node. This goes on until we reach a leaf node. Thus in essence a leaf node signifies a group of such discriminating regions lying in different spaces. As a result we are no more concerned about choosing a kernel and obtaining its best hyper-parameters since this method allows us to obtain a combination of discriminating regions lying in different spaces. We call this region a *Random Manifold*. Based on this node split strategy, we choose best split functions that maximize the information gain by Shannon entropy [2].
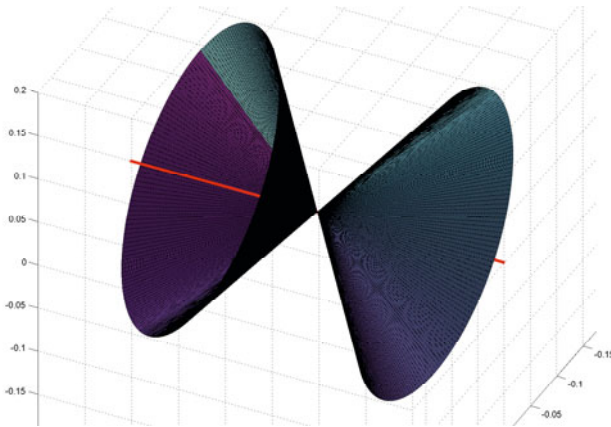

Fig. 2. A three dimensional example for a separating surface. Note that, unlike standard Random Decision Trees, we use set-similarity for splitting nodes.

More Randomness. Inspired from the work on random sampling [25] we also consider using random face subspace for the reference sets, i.e instead of choosing the best k rank approximation, as explained in Section 3.1, we choose the best $k/2$ rank approximates and randomly choose the other $k/2$ bases from the rest of the columns in $Q_{\Phi(X)}$.

In order to further increase the diversity of individual trees in a forests we consider another setting for RF. In this, the threshold t in equation (4) is set randomly rather than being chosen optimally in terms of the information gain. Thus at every node the following need to be randomly chosen: a kernel type, its hyper-parameter, a reference set, its random subspace and the threshold. At each node from all the possible combinations we choose m combinations and retain the one that gives us the best split. This setting is denoted as the Randomised Threshold Random Forest (RtRF).

4.2   Constrained Random Manifold Forests

We have observed that the linear kernels in tree structures were able to capture the inherent non-linearity of the data and gave competitive results compared

with non-linear kernels This motivated us to furtherincorporate discriminative information obtained from Constrained Mutual subspace matching. CMSM uses only that information which is essential for recognition. However the main problem in CMSM is obtaining the optimal constraint space and its dimension. Similar to our approach in randomised kernel learning we use Random Manifold Forest to learn this. At every node we randomly choose the following parameters: 1) a fixed number of sets per subject from the training data to build the constraint space 2) the dimension of the constraint subspace. As a result of this, similar to the previous setup, during classification at every node, the test set is projected onto a different constraint subspace. The combination of discriminating regions at each leaf node that form the Random Manifolds thus lie in these constrained subspaces.

Fig. 3. Large variations in pose, illumination, expression and scale for a subject in the database. *Red outlines show the detected/tracked faces.*

More Flexibility. In order to increase the flexibility of random manifolds, instead of a single reference set per subject, we use multiple reference sets. Each of these are obtained by randomly choosing different combinations of the training data.

5   Experimental Results

Experiments were performed in a video based face recognition framework. We have built our own database[1]. This database contains 10 subjects having 35 tracks (face sets) which were taken from two sitcoms 'Coupling' and 'Two and a half Men'. These tracks contain anywhere between 10-350 face images in them. Tracks were obtained using a slightly modified version of Anoop *et al's* tracker [21]. Detector used in this is the Viola Jones detector which detects frontal, left and right profile faces, if the detection is missed then a tracker is initiated to locate the face. Thus apart from the cropped location of the face, the detector/tracker also outputs the state of the face, i.e pose of the detected face (left profile, frontal, right profile) or a tracked face. This additional information was used to build separate manifolds for each detected pose and matching is done only winthin each these poses. Final similarity is given as the average of the measures obtained from each pose. Large variations in facial poses, illuminations, expressions and backgrounds contained in the database are shown for one of the subjects in Figure 3. The detected/tracked faces are shown by rectangles. Detected faces were resized to 100×100 and passed through the Multi-Scale Retinex filters [12] for compensating illumination changes. See Figure 4.

---

[1] Database will be made available on requesting the author.

Table 1. Performance of different features using Gaussian kernel and Polynomial kernel and CMSM. *Results of Fractional Powered kernel are not given due to space constraint. N* is the number of images in the set.*

No. of Training Images	Feature	Gaussian Kernel						Polynomial Kernel						CMSM
		k = N*	k = 50	k = 25	k = 15	k = 10	k = 5	k = N*	k = 50	k = 25	k = 15	k = 10	k = 5	k = 30
750	LBP	91.6	94.4	**94.8**	94.3	93.2	88.5	94.0	95.4	95.3	**95.4**	95.4	94.6	98.7
	Gabor	91.4	**93.8**	93.7	93.5	93.3	92.9	93.9	93.4	94.0	**94.2**	94.1	93.2	97.7
	Raw	72.3	**86.2**	85.1	82.9	80.8	71.6	74.2	**90.7**	90.6	88.9	86.7	78.2	95.7
500	LBP	87.6	91.6	**91.8**	91.5	90.3	85.2	91.7	93.6	94.2	**94.3**	93.8	92.7	95.1
	Gabor	88.9	90.9	**91.1**	91.0	90.9	90.2	91.2	**91.8**	91.6	91.2	90.8	89.9	88.1
	Raw	65.1	**82.3**	81.2	79.5	76.9	68.9	70.0	85.7	**85.8**	84.7	82.6	74.7	87.8
250	LBP	76.6	80.6	82.9	83.4	**83.5**	80.9	81.7	84.4	85.4	86.3	**86.4**	85.4	86.7
	Gabor	80.3	81.9	82.6	**82.9**	82.7	81.5	85.3	85.5	85.9	**86.0**	85.3	84.8	79.7
	Raw	56.9	71.4	75.6	**75.5**	74.2	70.0	59.0	73.6	**78.0**	77.9	77.5	72.8	75.3

To further validate our claims we perform experiments using the Oxford database. This database contains detected face of characters from the movies 'Player' and 'Groundhog Day'. We considered 6 subjects that had at least 100 images. We divided these images equally into 10 sets. We used only raw intensity features with a single manifold for all poses.

Fig. 4. Example of normalised face images in the sets

5.1 Performance of KPA and CMSM with Different Features and Kernels

For this experiment only 5 subjects were used from the sitcom database. Three different features, raw pixel intensity, Gabor and LBP were used. As in [27], the LBP feature vectors were set to have the length of 9440 and the Gabor feature vectors the length 9000. For raw pixel we resized the image to 15×15 and raster-scanned it to form a vector of size 225. Results are reported for the three kernels i.e Gaussian, Polynomial and Fractional powered kernels whose best performing kernel hyper-parameters were set with respect to the test set. For the training images (reference sets), tracks (face sets) of each subject were randomly selected so as to have a fixed number of images (750,500 and 250). These images were considered as a single set and were matched against the remaining sets. Each feature and the corresponding kernel projects the faces to a different feature space. The face subspace dimension (described in Section 3.1) at which it performs best needs not be the same. For this reason, we examined

various k values to get the best result. Table 1 shows the accuracies in percentage averaged over 15 different training/testing splits. For CMSM a fixed subspace dimesnion(30) was considered. LBP and Gabor significantly outperformed the raw pixel features in both KPA and CMSM setup. LBP performed slightly better than Gabor. Note also that the accuracy of the methods by the best k is much better than that of (k = the set cardinality as in [26]). The polynomial kernel delivered the best accuracy among the three kernels for all the cases.

Table 2. Performance of individual classifiers against the sum, product, weighted sum(W-Sum), minimum score(Min), maximum score(max) least square methods. N^* is the number of images in the set.

Classifier	k = N*	k = 50	k = 25	k = 15	k = 10	k = 5
C1	89.0	90.9	91.1	91.1	90.9	90.2
C2	21.0	71.5	85.6	89.1	90.7	90.9
C3	91.2	91.8	91.6	91.2	90.8	89.9
C4	87.6	91.6	91.8	91.5	90.4	85.2
C5	80.3	91.4	92.1	92.5	92.0	87.1
C6	91.7	93.6	94.2	94.3	93.8	92.7
Min	91.2	91.8	91.6	91.2	90.8	89.9
Max	22.1	81.2	90.9	92.2	92.0	86.9
W-Sum	92.1	93.7	93.8	93.6	93.5	93.0
Product	91.1	93.6	93.8	93.8	93.5	93.2
Sum	88.7	93.3	94.2	93.8	94.1	93.6
LS	**92.9**	**94.5**	**95.0**	**95.2**	**95.2**	**94.0**

5.2   Multiple Classifier System for KPA

For the multiple classifier system the six base classifiers (2 features * 3 kernels) were considered: C1: Gaussian kernel with Gabor feature, C2: Fractional power kernel with Gabor feature, C3: Polynomial kernel with Gabor feature, C4: Gaussian kernel with LBP feature, C5: Fractional power kernel with LBP feature, C6: Polynomial kernel with LBP feature. The kernel hyper-parameters were set to perform best for the test set. The product, sum, weighted sum, minimum score and maximum score rules along with the proposed least square solution (LS) were compared in the multiple classifier systems. Table 2 shows the accuracies when the number of training images was 500. As shown our Least square formulation outperforms the individual classifiers and the baseline fusion methods for all k's.

5.3   Random Manifold Forests

For convenience, only LBP was exploited in the experiment for Random Manifold techniques. However, by using the type of features (i.e. LBP or Gabor) as one of the random choices at a split node, better recognition accuracy may be achieved. In this experiment, we considered the following choices: a kernel type, a kernel

Table 3. Performance of RF and RtRF with LBP features for four choices of training data. *'Raw' represents the best performance using raw pixel images with different kernels and k = No. of images in the set (as in [26]), 'Max-R' represents the best performance using raw pixels with different kernels across k, 'Max-LBP' represents the best performance using LBP features with different kernels across k, 'sum' stands for sum rule (LBP only) and 'LS' is for the least square method (LBP only).*

Training	Without Random Face Subspace(RFS)							With RFS	
Data	Raw	Max-R	Max-LBP	Sum	LS	RF	RtRF	RF	RtRF
1)	61.11	76.98	92.86	91.27	95.24	94.44	96.03	95.24	**96.84**
2)	47.20	75.20	86.40	81.60	87.20	91.20	92.00	92.80	**92.80**
3)	55.22	82.09	88.81	88.81	90.30	87.31	89.55	88.06	**90.30**
4)	59.50	90.08	94.21	93.39	**95.04**	92.56	94.21	92.56	94.21

hyper-parameter, a subject. For the reference set we randomly chose 500 images as explained earlier. For this experiment only 5 subjects were used from the sitcom database. Due to time and space complexity, we restricted the number of choices for kernel hyper-parameters to the following values: for Gaussian kernel: σ varies from $0.5 - 1.4$ in steps of 0.1, for Fractional Power Kernel: a varies from $0.5 - 0.95$ in steps of 0.05, for Polynomial Kernel: b varies from $1 - 5$ in steps of 1. The total number of choices (M) is thus given by: number of kernels and its hyper-parameter choices $(10 + 10 + 5) \times$ reference sets $(5) = 125$. We also considered the random face subspace dimension and have shown the results separately for this. To set the same number of random choices for different kernels, we considered ten different random face subspace dimensions for each hyper-parameter choice from the Gaussian and Fractional kernels and twenty different random face subspace dimensions for each hyper-parameter in Polynomial kernel. In this case the total number of choices is: $(10 * 10 + 10 * 10 + 5 * 20) * 5 = 1500$.

As mentioned in Section 4.1, we considered another setting to increase the diversity among the trees and denoted the method RtRF. The following threshold values were experimented: $0.1 - 0.95$ in steps of 0.05, i.e a total of 18 choices. Thus the total number of choices in this case is $125 * 18 = 2250$ without the random face subspace and $1500 * 18 = 27000$ with the random face subspace. Table 3 shows the performance of these two settings for different choices of training data as the reference set.

5.4   Constrained Random Manifold Forests

In the previous subsection we compared RF results using out-of-bag Error. But for a fair comparison with SVM techniques we have split the data into two sets training/testing. We have used 25 face sets per subject for training and 10 for testing. For the kernel gram matrix in SVM we used the KPA measure between two sets as given in equation 1. i.e KPA is taken as a kernel between two sets. The best kernel parameters were manually set for the SVM in the experiments. We compare this with the RtRF setup which gave the best result in Table 3. We

Table 4. Comparison on Sitcom database. CRMF-1 refers to a single reference set and CMRF-4 referes to 4 reference sets CS and CSD denote the number of constraint subspaces and the number of constraint subspace dimensions used, respectively.

Training Data	CMSM	SVM		RtRF with RFS		CRMF - 1		CRMF - 4	
		Linear	Non-L	Linear	Non-L	1CS,1CSD	5CS,5CSD	1CS,1CSD	5CS,5CSD
1)	81	95	96	89	92	97	**100**	99	99
				85.2	90	96.9	99.1	99	99
2)	89	91	96	92	92	93	97	98	**99**
				88.88	90.1	92.1	96.6	97.5	99
3)	90	93	94	87	90	90	93	96	**96**
				82.27	88.5	87.9	92.4	94.6	95.5
4)	86	94	96	91	93	93	95	98	**98**
				87.54	90.7	91.4	94.2	98	98
Avg	86.71	93.71	96.14	86.6	89.28	92.43	95.74	97.31	**97.73**

also compare this while using linear kernels where the random choice at every node are: 1) subject(10) 2) random Face subspace(20) 3) threshold(18). Thus the total number of choices at every node is $10*20*18 = 3600$. As mentioned in Section 4.2 we use discriminative information from CMSM. The constraint spaces are constructed using randomly selected 10 sets per subject. For space and time constraints we restrict the choices for constraint spaces to five possible subspace which are initially computed. We also restrict the dimension of these subspaces to $50\% - 90\%$ (in steps of 10%) of the total (least)significant eigenvalues. To increase flexibility we consider multiple references sets per subject each of which is obtained by randomly choosing 500 images from the training data as explained earlier. Thus the choices at every node are: 1) subject(10) 2) reference set(4) 3) constraint subspace(5) 4) constraint subspace dimension(5). Thus the total choices are $10*4*5*5 = 1000$. Results are shown in Table 4 and compared with SVM(1-vs-1) and original CMSM. For CMSM, from the training set we randomly choose 500 images as reference and rest are used to build the constraint space. Face subspace dimension(k) is kept as 30 and constraint subspace dimension as 90% of the (least)significant eigenvalues. 10 different trails are considered for RMF and the best along with the average performance is quoted.

For validation purpose, we also show results of the constraint random manifold forests on the Oxford dataset. Here we use raw pixel intensities, face subspace dimension(k) is kept at 10 and constraint sapce dimension is 90% of the (least)significant eigenvalues.

We have successfully included discriminative information(CMSM) in the linear version of random manifold forest and further improved its accuracy. Taking Non-Linear SVM as the baseline we get an average increase of 4% on the sitcom dataset and 32% on the Oxford database. Note that the best kernels were manually set for the SVM whereas they were automatically learnt in the proposed method.

Table 5. Comparison on Oxford database. CRMF-1 refers to a single reference set and CMRF-4 referes to 4 reference sets. CS and CSD denote the number of constraint subspaces and the number of constraint subspace dimensions used respectively.

Training Data	CMSM	SVM Non-L	CRMF - 1		CRMF - 4	
			1CS,1CSD	5CS,5CSD	1CS,1CSD	5CS,5CSD
1)	87.5	75	83.3	91.67	87.5	**100**
			81.67	88.75	86.67	99.58
2)	91.67	66.67	83.3	95.83	100	**100**
			82.08	91.67	99.17	99.58
3)	87.5	70.83	91.67	100	95.83	**100**
			87.92	97.08	92.92	100
4)	91.67	66.67	87.5	95.83	100	**100**
			84.17	93.33	96.67	100
Avg	85.12	67.26	82.74	92.92	93.04	**99.7**

6   Conclusions

In this paper we have explored the use of different features/kernels for KPA-based face recognition and have shown that the accuracy of the KPA method is significantly improved by using the sparse representations such as LBP and Gabor features. A novel least square formulation has been proposed for combining multiple classifiers and it has been shown to outperform some of the existing combining techniques. Both, the proposed and previous combining methods require setting the kernel hyper-parameters a priori, which is difficult in practice. To address this, we propose Random Manifold Forests that is learnt to combine discriminating regions obtained from different spaces (parameterized by the kernel type and its hyper-parameters).Hence, the method automatically learns the kernels and hyper-parameters. Taking the KPA with the raw-pixel representation as a base line, we have achieved the accuracy improvement by about *35 percent* on the challenging sitcom data set. We have successfully included discriminative information(CMSM) in the linear version of random manifold forests and further improved its accuracy.Compared to the non-linear SVM, whose kernels were manually tuned, we obtained an average increase of 4% on the sitcom dataset and 32% on the Oxford dataset.

References

1. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: ICCV (2007)
2. Breiman, L.: Random forests. Journal of Machine learning 45(1), 5–32 (2001)
3. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Journal of Machile Learning 63(1) (2006)

4. Chan, C., Kittler, J., Messer, K.: Multi-scale local binary pattern histograms for face recognition. In: ICBA (2007)
5. Chen, K., Xu, L., Chi, H.: Improved learning algorithms for mixture of experts in multiclass classification. Neural Networks 12(9), 1229–1252 (1999)
6. Czyz, J., Vandendorpe, L., Kittler, J.: Combining face verification experts. In: ICPR (2002)
7. Díaz-Uriarte, R., de Andrés, A.: Gene selection and classification of microarray data using random forest. BMC bioinformatics 7(1), 3 (2006)
8. Fukui, K., Yamaguchi, O.: Face Recognition Using Multi-viewpoint Patterns for Robot Vision. In: Int. Sym. of Robotics Research (2003)
9. Fukui, K., Yamaguchi, O.: The kernel orthogonal mutual subspace method and its application to 3D object recognition. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 467–476. Springer, Heidelberg (2007)
10. Jacobs, R., Jordan, M., Nowlan, S., Hinton, G.: Adaptive mixtures of local experts. Neural computation 3(1), 79–87 (1991)
11. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition 38(12), 2270–2285 (2005)
12. Jobson, D., Rahman, Z., Woodell, G.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE TIP 6(7), 965–976 (1997)
13. Kapoor, A., Qi, Y., Ahn, H., Picard, R.: Hyperparameter and kernel learning for graph based semi-supervised classification. In: NIPS (2006)
14. Arandjelovic, O., Cipolla, R.: Automatic Cast Listing in Feature-Length Films with Anisotropic Manifold Space. In: CVPR (2006)
15. Kim, T.-K., Arandjelović, O., Cipolla, R.: Boosted manifold principal angles for image set-based recognition. Pattern Recogn 40(9), 2475–2484 (2007)
16. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Trans. on PAMI 20(3), 226–239 (1998)
17. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., Jordan, M.: Learning the kernel matrix with semidefinite programming. JMLR 5, 27–72 (2004)
18. Sonnenburg, S., Ratsch, G., Schafer, C., Scholkopf, B.: Large Scale Multiple Kernel Learning. Journal of Machine Learning Research 7, 1531–1565 (2006)
19. Zien, A., Ong, C.S.: Multiclass Multiple Kernel Learning. In: ICML (2007)
20. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. on PAMI, 684–698 (2005)
21. Rajagopal, A., Anandathirtha, P., Ramakrishnan, K., Kankanhalli, M.: Integrated Detect-Track Framework for Multi-view Face Detection in Video. In: ICVGIP (2008)
22. Scholkopf, B., Smola, A., Muller, K.: Nonlinear component analysis as a kernel eigenvalue problem. Neural computation 10(5), 1299–1319 (1998)
23. Shotton, J., Johnson, M., Cipolla, R., Center, T., Kawasaki, J.: Semantic texton forests for image categorization and segmentation. In: CVPR (2008)
24. Statnikov, A., Wang, L., Aliferis, C.: A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC bioinformatics 9(1), 319 (2008)
25. Wang, X., Tang, X.: Random sampling for subspace face recognition. IJCV 70(1), 91–104 (2006)
26. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. Journal of Machine Learning Research 4, 913–931 (2003)

27. Xiaoyu Wang, S.Y., Han, T.X.: An HOG-LBP Human Detector with Partial Occlusion Handling. In: ICCV (2009)
28. Yamaguchi, O., Fukui, K., Maeda, K.-i.: Face recognition using temporal image sequence. In: AFG (1998)
29. Nishiyama, M., Yamaguchi, O., Fukui, K.: Face recognition with the multiple constrained mutual subspace method. In: ACCV (2005)
30. Li, X., Fukui, K., Zheng, N.: Boosting Constrained Mutual Subspace Method for Robust Image-Set Based Object Recognition. IJCAI, 1132–1137 (2009)
31. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: CVPR (2008)
32. Kim, T.-K., Kittler, J., Cipolla, R.: Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. IEEE TPAMI 29(6) (2007)
33. Cevikalp, H., Triggs, B.: Face Recognition Based on Image Sets. In: Computer Vision and Pattern Recognition (2010)