On-line Learning of Mutually Orthogonal Subspaces for Face Recognition by Image Sets

Tae-Kyun Kim, Josef Kittler, *Member, IEEE,* and Roberto Cipolla, *Senior Member, IEEE*

Abstract—We address the problem of face recognition by matching image sets. Each set of face images is represented by a subspace (or linear manifold) and recognition is carried out by subspace-to-subspace matching. In this paper, 1) a new discriminative method that maximises *orthogonality between subspaces* is proposed. The method improves the discrimination power of the subspace angle based face recognition method by maximizing the angles between different classes. 2) We propose a method for on-line updating the discriminative subspaces as a mechanism for continuously improving recognition accuracy. 3) A further enhancement called locally orthogonal subspace method is presented to maximise the orthogonality between competing classes. Experiments using 700 face image sets have shown that the proposed method outperforms relevant prior-art and effectively boosts its accuracy by on-line learning. It is shown that the method for on-line learning delivers the same solution as the batch computation at far lower computational cost and the locally orthogonal method exhibits improved accuracy. We also demonstrate the merit of the proposed face recognition method on Portal scenarios of Multiple Biometric Grand Challenge.

Index Terms—On-line learning, manifold-to-manifold matching, subspace, face recognition, image sets, mutually orthogonal subspace

1 Introduction

Whereas considerable advances have been made in face recognition in controlled environments, recognition in unconstrained and changing environments still remains a challenging problem. Face recognition by image sets has been increasingly popular because of their greater accuracy and robustness as compared with the approaches exploiting a single image as input [2]-[8], [21]. Image set harvested in either a video or a set of multiple still-shots captures various facial appearance changes and thus provides more evidence on face identity than a single image alone. Prior video-based attempts [18], [19], [20] have shown that including a strong temporal constraint deteriorates recognition performance when persons move arbitrarily in a testing video sequence. Moreover, temporal continuity assumption between consecutive face images is often invalid when subjects do not face a camera and move abruptly. In this work, we consider a general scenario where an image set is a more pertinent input than video.

Of the methods that compare an image set to an image set, subspace (or manifold) matching based methods have been shown superior to other approaches such as aggregation of multiple nearest neighbour vector-matches [6] and probability-density based methods [4], [5] in many studies e.g. [7], [21], [1], [2]. Subspace representation of image sets allows interpolation of data vectors thus yielding a robust matching of new data in the subspaces. Conventionally, when a face image is given as a vector, distance of the face vector to each model subspace is measured and the nearest subspace is picked for its class. Now that we want to classify a subspace instead of a single vector (i.e. subspace-to-subspace matching), the distance is no longer valid but angles between subspaces (called canonical angles, principal angles or canonical correlations) become a reasonable measurement. The subspace angle method also yields an economical matching in time and memory compared to aggregation of all pairwise vector matches of two sets [6]. Methods beyond the subspace angles have also been explored: a generalised form called Grassmannian distance is proposed for face recognition in e.g. [3] where the principal angle has been shown as one of the Grassmannian distances. In [2], a nonlinear manifold is obtained as a set of subspaces and the angles between pairwise subspaces are exploited for manifold-to-manifold matching. Prior to [2], a mixture of subspaces for manifold principal angles have similarly been proposed in [31]. More traditionally, a kernel version of principal angles has also been proposed to deal with nonlinear manifolds e.g. in [8].

Since Hotelling [22], Canonical Correlation Analysis (CCA) has been a standard tool to inspect linear relations between two random variables. Goloub's formulation [14] for subspace angles is mathematically equivalent to Hotelling's. CCA has received increasing attention in related literature: Yamaguchi et al. have adopted the standard CCA for face recognition by image-sets [7]

- T-K. Kim is with the Sidney Sussex College, University of Cambridge, Cambridge, CB2 3HU, UK. E-mail: tkk22@cam.ac.uk, see http://mi.eng.cam.ac.uk/∼tkk22.
- J. Kittler is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. E-mail: J.Kittler@surrey.ac.uk.
- R. Cipolla is with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK. E-mail: rc10001@cam.ac.uk.

(called Mutual Subspace Method (MSM)) and subsequently proposed the constrained subspace which improves the discrimination power of the manifold-angle method [9], [10], [12] (called Constrained Mutual Subspace Method). Bach and Jordan [23] have proposed a probabilistic interpretation, and Wolf and Shashua [8] a kernel version to capture nonlinear manifolds. Kim, Kittler and Cipolla have proposed discriminative learning for CCA and have shown superior accuracy to other CCA-based methods [21].

In practice, a complete set of training images is not given in advance and the execution of the batch-computation [1] is required whenever new images are presented. This is too expensive in both time and space. An efficient model update would be greatly desired to accumulate the information conveyed by new data so that the method's future accuracy is enhanced. E.g. once a face image set is classified by matching subspaces, the image set could be exploited to update the existing subspaces, as shown in Figure 1. Time-efficient recognition and update by the method proposed in this paper facilitates interaction between users and a system.

Numerous algorithms have been developed to update eigenspace as more data samples arrive. The computational cost of an eigenproblem is cubic in the size of the respective scatter matrix. In [16], the size of the matrix to be eigendecomposed is reduced by using the sufficient spanning set [2], greatly speeding up the computation of the eigenproblem for update. The method also allows update over a set of new vectors at a time. Methods for incremental learning of discriminative subspaces have also been proposed. Gradient-based incremental learning of a modified LDA was proposed by Hiraoka et al. [25]. This, however, requires setting of a learning rate. Ye et al. [24] have proposed an incremental version of LDA, which can include a single new data point in each time step. An important limitation is the computational complexity of the method when the number of classes C is large. In [17], an incremental LDA has been achieved by updating the components of both the between class scatter and total class scatter matrix, thus remaining efficient regardless of the number of classes. At each update, the sufficient spanning set is exploited to reduce the size of the scatter matrix yielding a speed up similarly to [16]. While it is worth noting the existence of efficient algorithms for kernel PCA and LDA [26], [27], the computational cost of feature extraction of new samples in these methods is high for a large-size recognition problem.

Existing discriminative methods for subspace-to-subspace matching are limited in the aspect of on-line learning: Constrained Mutual Subspace Method (CMSM) [9], [10] requires manual setting for the dimensionality of the constrained subspace and its accuracy is dependent on the dimension, which precludes automatic on-line learning. Iterative optimization required in Kim et al.'s method [21] is computationally costly, making on-line update difficult. Non-liner extensions of CMSM [12] comprise kernel parameters to set and require high computational cost in both matching and model learning.

This paper presents a method of on-line learning of discriminative subspaces for principal-angle based face recognition. The earlier version of this work [30] has been rewritten for clarification and conciseness. More comparisons and experiments have also been added. 1) The discriminative subspace was first proposed in the earlier version of this study [30] by generalisaing Oja and Kittler's formulation [13] (the proposed method has been later adopted in e.g. [28], [29]). The method enforces orthogonality between subspaces and hence improves the discrimination power of the subspace angle based classification method. 2) The mutually orthogonalised subspaces are incrementally learnt by updating components of the numerator and denominator of the objective function respectively. Each update is benefited in both time and space by the concept of the sufficient spanning set used for the incremental Principal Component Analysis (PCA) in [16]. The proposed method yields an identical solution to that of the batch-mode computation but at a far lower computational and space cost. The on-line method also allows multiple sets of vectors to be added in a single update, thus avoiding frequent updates. 3) Finally, recognition accuracy of the discriminative subspace method is improved by maximising the orthogonality between rival classes, which is seen as an extension to nonlinear manifolds in a sense (see Section 5). In this paper, we mainly explain our method for subspaces i.e. linear manifolds but the proposed method may be further generalised to nonlinear manifolds by representing a manifold as a set of linear manifolds similarly to [2], [31].

The next section reviews the subspace-angle method and the Oja and Kittler's formulation. The proposed orthogonalisation between subspaces is explained in Section 3. The on-line learning method of the orthogonal subspaces is proposed in Section 4 and the method to improve the discrimination power in Section 5. Section 6 and Section 7 provide comparative evaluations and conclusions respectively.

2 Background

2.1 Subspace Angles

Canonical correlations [14], which are cosines of principal angles between any two d-dimensional linear manifolds (or subspaces) $\mathcal{L}_1$ and $\mathcal{L}_2$, are uniquely defined as

$$\cos\theta_i = \max_{\mathbf{u}_i \in \mathcal{L}_1} \max_{\mathbf{v}_i \in \mathcal{L}_2} \mathbf{u}_i^T \mathbf{v}_i, \qquad i = 1, ..., d. \tag{1}$$

subject to $\mathbf{u}_i^T\mathbf{u}_i = \mathbf{v}_i^T\mathbf{v}_i = 1,\ \mathbf{u}_i^T\mathbf{u}_j = \mathbf{v}_i^T\mathbf{v}_j = 0,\ i \neq j$. If $\mathbf{P}_1, \mathbf{P}_2$ denote basis matrices of the two subspaces,

1. All existing images are re-used with new images for computing a new model.

2. It is a reduced set of basis vectors spanning the space of most data variation

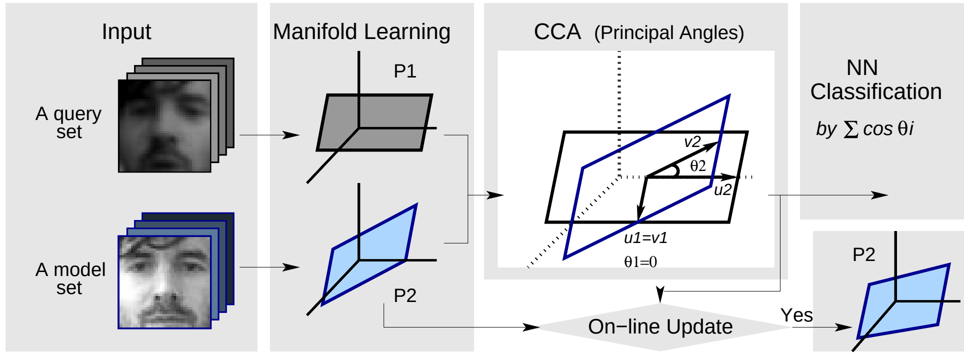


Fig. 1. **Proposed method.** *The similarity between linear manifolds (or subspaces) is computed as the sum of principal angles and is used for NN classification. Once a query set has been classified, it can be included in the model by on-line updating the existing manifold.*

canonical correlations are conveniently obtained as singular values of $\mathbf{P}_1^T\mathbf{P}_2 \in \mathcal{R}^{d\times d}$, only taking $O(d^3)$:

$$\mathbf{P}_1^T\mathbf{P}_2 = \mathbf{Q}_L\Lambda\mathbf{Q}_R^T, \qquad \Lambda = diag(\sigma_1, ..., \sigma_d) \tag{2}$$

where $\mathbf{Q}_L, \mathbf{Q}_R$ are orthogonal matrices [3]. Similarity of two subspaces is then defined as the average of the canonical correlations and Nearest Neighbor (NN) classification is performed based on the subspace similarity [7], [8], [21], [28], [29].

2.2 Orthogonality Between Subspaces

We revisit Oja and Kittler's class-wise feature extraction method [13]. The method finds the class-specific components on which class data have maximum variance while those of all other classes have zero variance. Then, a new vector is classified by conventionally measuring the distance of the vector to the class-specific subspaces. The method is as follows, replacing their vector notations with matrices:

Denote the correlation matrices of C classes by $\mathbf{R}_i,\ i = 1, ..., C$, where $\mathbf{R}_i = 1/M_i \sum \mathbf{x}\mathbf{x}^T$ and M_i is the number of data points, $\mathbf{x}$, of i-th class. The total correlation matrix is defined as $\mathbf{R}_T = \sum_{i=1}^{C} w_i\mathbf{R}_i$ where $w_i\ \forall i$ denotes class priors. The total correlation matrix is eigen-decomposed s.t. $\mathbf{P}_T^T\mathbf{R}_T\mathbf{P}_T = \Lambda_T$. We then have $\mathbf{Z}^T\mathbf{R}_T\mathbf{Z} = \mathbf{I}$ by $\mathbf{Z} = \mathbf{P}_T\Lambda_T^{-1/2}$. This means that matrices $w_i\mathbf{Z}^T\mathbf{R}_i\mathbf{Z}$ and $\sum_{j\neq i} w_j\mathbf{Z}^T\mathbf{R}_j\mathbf{Z}$ have the same eigenvectors and the respective eigenvalue sum must be equal to one: let $\mathbf{U}_i$ be the eigenvector matrix of the i-th class having the eigenvalues equal to unity in the transformed space by $\mathbf{Z}$ s.t.

$$w_i\mathbf{U}_i^T\mathbf{Z}^T\mathbf{R}_i\mathbf{Z}\mathbf{U}_i = \mathbf{I}_i, \tag{3}$$

then

$$\sum_{j\neq i} w_j\mathbf{U}_i^T\mathbf{Z}^T\mathbf{R}_j\mathbf{Z}\mathbf{U}_i = \mathbf{O} \rightarrow \tag{4}$$

$$w_j\mathbf{U}_i^T\mathbf{Z}^T\mathbf{R}_j\mathbf{Z}\mathbf{U}_i = \mathbf{O},\ \text{for all } j \neq i$$

where $\mathbf{O}$ is a zero matrix and every matrix $w_j\mathbf{U}_i^T\mathbf{Z}^T\mathbf{R}_j\mathbf{Z}\mathbf{U}_i$ is positive semi-definite. If we have the eigenvector matrix of unity eigenvalues of the j-th class s.t. $w_j\mathbf{Z}^T\mathbf{R}_j\mathbf{Z} \simeq \mathbf{U}_j\mathbf{U}_j^T$, by (4),

$$w_j\mathbf{U}_i^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_i = \mathbf{O} \rightarrow \mathbf{U}_i^T\mathbf{U}_j = \mathbf{O}. \tag{5}$$

Two linear manifolds spanned by $\mathbf{U}_i, \mathbf{U}_j$ are mutually orthogonal since all the vectors of each space are orthogonal to those of the other space.

3 Generalised Mutually Orthogonal Subspaces

Clearly, canonical correlations of mutually orthogonal subspaces are zero by (2) and (5) (put $\mathbf{U}$ in the place of $\mathbf{P}$). The decision is simply made to label a query set as the same class if the canonical correlations are non-zero and a different class otherwise. However, in practice, the eigenvectors having the eigenvalues which are exactly equal to one in (3), do not often exist. We propose using the eigenvectors corresponding to the largest few eigenvalues. The mutual orthogonal subspace (3) is thus generalized into

$$w_i\mathbf{U}_i^T\mathbf{Z}^T\mathbf{R}_i\mathbf{Z}\mathbf{U}_i = \Delta_i, \quad \sum_{j\neq i} w_j\mathbf{U}_i^T\mathbf{Z}^T\mathbf{R}_j\mathbf{Z}\mathbf{U}_i = \mathbf{I} - \Delta_i, \tag{6}$$

where $\mathbf{Z} = \mathbf{P}_T\Lambda_T^{-1/2}$ (and other notation) as defined in the previous section and Δ_i is the diagonal matrix corresponding to the largest few eigenvalues. Clearly, the method seeks the most important basis vectors of each class that are at the same time the least significant basis vectors of the ensemble of the rest of the classes. If we write $\overline{\mathbf{U}}_i = \mathbf{Z}\mathbf{U}_i$, where the orthogonal basis matrix of i-th class model is denoted by $\mathbf{U}_i$, the problem can be written as

$$\max_{arg\overline{\mathbf{U}}_i} \frac{|\overline{\mathbf{U}}_i^T\mathbf{R}_i\overline{\mathbf{U}}_i|}{|\overline{\mathbf{U}}_i^T\mathbf{R}_T\overline{\mathbf{U}}_i|}, \quad i = 1, ..., C. \tag{7}$$

3. An orthogonal matrix is a square matrix $\mathbf{Q}$ whose transpose is its inverse: $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$. Note that this is different from the orthogonality between subspaces in Section 2.2.

From (6)

$$\max_{arg\overline{\mathbf{U}}_i} \frac{|\overline{\mathbf{U}}_i^T w_i \mathbf{R}_i \overline{\mathbf{U}}_i|}{|\overline{\mathbf{U}}_i^T \sum_{j\neq i} w_j \mathbf{R}_j \overline{\mathbf{U}}_i|}$$

$$= \max_{arg\overline{\mathbf{U}}_i} \frac{|\overline{\mathbf{U}}_i^T w_i \mathbf{R}_i \overline{\mathbf{U}}_i|}{|\overline{\mathbf{U}}_i^T (w_i\mathbf{R}_i + \sum_{j\neq i} w_j \mathbf{R}_j) \overline{\mathbf{U}}_i|},$$

the proposed orthogonalisation improves the discrimination power of the subspace-angle method (See Section 6). The solution is given by successively diagonalising matrices as in section 4. Interestingly, the principle of the Orthogonal Subspace Method (OSM) is very close to that of CMSM [9]. Both methods find the components which maximally represent the class data while minimizing the variances of all the other classes. However, OSM provides the optimal way to choose the number of such components based on the eigenvalues, while CMSM requires an empirical setting for the number of the components, which is practically unfavorable for online learning.

The similarity of two orthogonal subspaces $\mathbf{U}_1, \mathbf{U}_2$ is given as $tr(\Lambda)$ where Λ is the singular value matrix in (2) (Put $\mathbf{U}$ in the place of $\mathbf{P}$). Nearest Neighbor (NN) classification is then performed based on the similarity measure.

4 Incremental Learning of Orthogonal Subspaces

There are many previous studies for incremental PCA, but the involvement of matrix inverse and product in $\mathbf{R}_T^{-1}\mathbf{R}_i$ in the Orthogonal Subspace Method (OSM) makes incremental learning not straightforward from prior methods. Among the existing methods for online discriminative subspaces discussed in Section 1, the framework of [17] is the most appropriate for the OSM that needs an efficient update for both numerator and denominator of the OSM criterion. Following the three step framework of [17], we define new sufficient spanning sets and a new online method for the OSM.

The incremental OSM solution we propose involves the following three steps: update the principal components of each class correlation matrix $\mathbf{R}_i$, update the principal components of the total correlation matrix $\mathbf{R}_T$ and compute the orthogonal components using the updated sets of principal components. The method using a sufficient spanning set for incremental PCA [16] is conveniently applied to each step to reduce the size of the matrices to be eigendecomposed. The proposed method provides the same solution as the batch-mode OSM with far lower computational cost. When a new data point or set is added to an existing data set, existing orthogonal subspaces $\mathbf{U}_i,\ i = 1, ..., C$ are updated to $\mathbf{U}'_i$ as follows:

a) Updating principal components of class correlation matrix. The update is defined as

$$\mathcal{F} : (M_i, \mathbf{P}_i, \Lambda_i, M_i^n, \mathbf{P}_i^n, \Lambda_i^n) \longrightarrow (M'_i, \mathbf{P}'_i, \Lambda'_i), \quad (8)$$

where the number of samples, eigenvector and eigenvalue matrices corresponding to the first few eigenvalues of the i-th class correlation matrix R_i in an existing data set are $(M_i, \mathbf{P}_i, \Lambda_i)$ respectively. The set $(M_i^n, \mathbf{P}_i^n, \Lambda_i^n)$ denotes those of a new data set. This update is applied only to the classes i that have new data points. For all other classes, $(M'_i, \mathbf{P}'_i, \Lambda'_i) = (M_i, \mathbf{P}_i, \Lambda_i)$. The proposed update is similar to [16] except that correlation matrices are used instead of covariance matrices. The updated class correlation matrix is $\mathbf{R}'_i \simeq \frac{M_i}{M'_i}\mathbf{P}_i\Lambda_i\mathbf{P}_i^T + \frac{M_i^n}{M'_i}\mathbf{P}_i^n\Lambda_i^n\mathbf{P}_i^{nT}$ where $M'_i = M_i + M_i^n$. The sufficient spanning set of $\mathbf{R}'_i$ is given as $\Upsilon_i = \mathcal{H}([\mathbf{P}_i,\ \mathbf{P}_i^n])$, where $\mathcal{H}$ is an orthonormalisation function of column vectors (e.g. QR decomposition) followed by removing zeros vectors. The updated principal components are then written as $\mathbf{P}'_i = \Upsilon_i\mathbf{Q}_i$, where $\mathbf{Q}_i$ is a rotation matrix. By this representation, the eigenproblem of the updated class correlation matrix is changed into a new low dimensional eigenproblem as

$$\mathbf{R}'_i \simeq \mathbf{P}'_i\Lambda'_i\mathbf{P}_i'^T \quad \longrightarrow \quad \Upsilon_i^T\mathbf{R}'_i\Upsilon_i \simeq \mathbf{Q}_i\Lambda'_i\mathbf{Q}_i^T. \quad (9)$$

where $\mathbf{Q}_i, \Lambda'_i$ are eivenvector and eigenvalue matrices of $\Upsilon_i^T\mathbf{R}'_i\Upsilon_i$. Note that the new eigenvalue problem requires only $O({d_i}^3)$ computations, where d_i is the number of columns of Υ_i. The total computational cost of this stage takes $O(C^n \times (d_i^3 + min(N, M_i^n)^3))$, where N is the dimension of input space and C^n is the number of classes in the new data set given. The latter term is for computing $(M_i^n, \mathbf{P}_i^n, \Lambda_i^n)$ from the new data set in order to perform the update.

b) Updating principal components of total correlation matrix. The subsequent update is described as

$$\mathcal{G} : (M_T, \mathbf{P}_T, \Lambda_T,\ M_i^n, \mathbf{P}_i^n, \Lambda_i^n) \longrightarrow (M'_T, \mathbf{P}'_T, \Lambda'_T), \quad (10)$$

where $i = 1, ..., C^n$ and C^n represents the number of classes of the new data. $M_T = \sum_{i=1}^{C} M_i$ and $\mathbf{P}_T, \Lambda_T$ are the first few eigenvector and eigenvalue matrices of the total correlation matrix of the existing data. The updated total correlation matrix is

$$\mathbf{R}'_T \simeq \frac{M_T}{M'_T}\mathbf{P}_T\Lambda_T\mathbf{P}_T^T + \frac{M_T^n}{M'_T}\sum_{i=1}^{C^n} w_i\mathbf{P}_i^n\Lambda_i^n(\mathbf{P}_i^n)^T \quad (11)$$

where $M'_T = M_T + M_T^n$, $M_T^n = \sum_i M_i^n$. The sufficient spanning set of $\mathbf{R}'_T$ is obtained as

$$\Upsilon_T = \mathcal{H}([\mathbf{P}_T, \mathbf{P}_1^n, ..., \mathbf{P}_{C^n}^n]) \quad (12)$$

and $\mathbf{P}'_T = \Upsilon_T\mathbf{Q}_T$, where $\mathbf{Q}_T$ is a rotation matrix. Note that the sufficient spanning set is independent of class prior w_i. Accordingly, the new low dimensional eigenproblem to solve is

$$\mathbf{R}'_T \simeq \mathbf{P}'_T\Lambda'_T\mathbf{P}_T'^T \quad \longrightarrow \quad \Upsilon_T^T\mathbf{R}'_T\Upsilon_T \simeq \mathbf{Q}_T\Lambda'_T\mathbf{Q}_T^T \quad (13)$$

The computation requires $O(d_T^3)$, where d_T^3 is the number of components of Υ_T. Note that all $\mathbf{P}_i^n$ have already been produced at the previous step.

Batch	$O(min(N,M'_T)^3 + C \times min(N,M'_i)^3)$
Incremental	$O(C^n \times (d_i^3 + min(N,M_i^n)^3))$ $+O(d_T^3) + O(C \times {d_i}^3)$

Fig. 2. **Computation cost for update.** *N is the dimension of input vectors, M'_T, M'_i are the number of vectors in total and i-th class of the combined data. M_i^n is the number of vectors of i-th class of the new set. The number of classes of the combined and the new set are denoted by C, C^n. d_i, d_T are the number of components of the sufficient spanning set of i-th class and the total set.*

c) Updating orthogonal components. The final step exploits the updated principal components of the previous steps, which are defined as

$$\mathcal{H} : (\mathbf{P}'_i, \Lambda'_i,\ \mathbf{P}'_T, \Lambda'_T) \longrightarrow \mathbf{U}'_i, \quad i = 1, ..., C. \tag{14}$$

Let $\mathbf{Z} = \mathbf{P}'_T {\Lambda'_T}^{-1/2}$, then the denominator term in (7) $\mathbf{Z}^T\mathbf{R}'_T\mathbf{Z} = \mathbf{I}$. The remaining problem is to find the components which maximise the variance of the numerator term in the projected subspace i.e. $\mathbf{Z}^T\mathbf{R}'_i\mathbf{Z}$. The sufficient spanning set of the projected data is given by $\Phi_i = \mathcal{H}({\mathbf{P}'_T}^T\mathbf{P}'_i)$. Then, the eigenproblem to solve is

$$\mathbf{Z}^T\mathbf{R}'_i\mathbf{Z} \simeq \mathbf{U}'_i\Delta_i{\mathbf{U}'_i}^T \quad \longrightarrow \quad \Phi_i^T\mathbf{Z}^T\mathbf{R}'_i\mathbf{Z}\Phi_i = \widetilde{\mathbf{Q}}_i\Delta_i\widetilde{\mathbf{Q}}_i^T, \tag{15}$$

where $\widetilde{\mathbf{Q}}_i, \Delta_i$ are eigenvector and eigenvalue matrix respectively. The final orthogonal components are given as $\mathbf{U}'_i = \Phi_i\widetilde{\mathbf{Q}}_i,\ i = 1, ..., C$. This computation only takes $O({d_i}^3)$, where d_i is the number of columns of $\mathbf{P}'_i$. Note usually $d_i < d_T$, where d_T is the number of columns of $\mathbf{P}'_T$.

Batch OSM vs. incremental OSM for time and space complexity. See Figure 2 for the computational cost. The batch computation of OSM for the combined data costs $O(min(N,M'_T)^3 + C \times min(N,M'_i)^3)$, where the former term is for the diagonalization of the total correlation matrix and the latter for the projected data of the C classes (Refer to Section 2 for the batch-mode computation). The batch computation also requires all data vectors or $N \times N$ correlation matrices to be updated. By contrast, the proposed incremental solution is much more time-efficient with the costs of $O(C^n \times (d_i^3 + min(N,M_i^n)^3))$, $O(d_T^3)$ and $O(C \times {d_i}^3)$ for the three steps respectively. Note $d_i \ll M'_i$, $d_T \ll M'_T$, $M_i^n \ll M'_i$. The proposed incremental algorithm is also very economical in memory costs, which corresponds to the data $(\mathbf{P}_i, \Lambda_i, \mathbf{P}_T, \Lambda_T),\ i = 1, ..., C$.

5 Locally Orthogonal Subspaces

The pairwise class prior w_j^i is proposed to improve the discriminatory power of the method. The locally orthogonal method defines

$$\max_{arg\overline{\mathbf{U}}_i} \frac{|\overline{\mathbf{U}}_i^T\mathbf{R}_i\overline{\mathbf{U}}_i|}{|\overline{\mathbf{U}}_i^T\mathbf{R}_T^i\overline{\mathbf{U}}_i|}, \quad \text{where} \quad \mathbf{R}_T^i = \sum_{j=1}^{C} w_j^i\mathbf{R}_j. \tag{16}$$

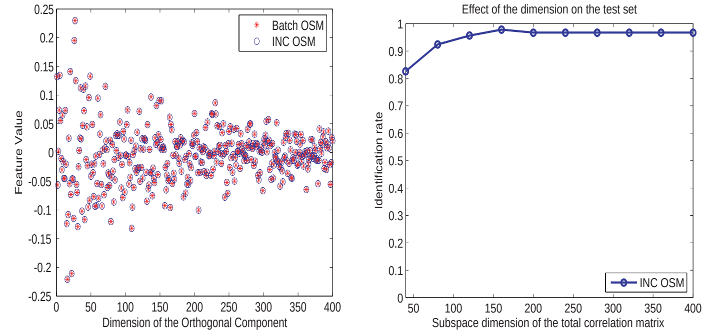
The pairwise class prior $w_j^i = 1$, if j-th class subspace is close to i-th class subspace in terms of the subspace similarity, $w_j^i = 0$ otherwise. That is, the method finds the component that maximises the variance of i-th class and minimises the variance of neighboring classes. The use of a set of total correlation matrices $\mathbf{R}_T^i$, which are locally defined, instead of a single total correlation matrix $\mathbf{R}_T$, is more appropriate to capture nonlinear manifolds of entire data vectors. Note, however, that in the proposed method each class data is still modeled as a single subspace. This may be further extended to a set of subspaces when each class exhibits highly nonlinear manifolds. The similar ideas have appeared in [2], [31].

Normalization. When classifying a query set, the locally orthogonal components of the query set are computed with respect to i-th model class using $\mathbf{R}_T^i$ for $i = 1, ..., C$. NN recognition is then performed in terms of the normalised subspace similarity as $(s_i - m_i)/\sigma_i$ where s_i is the subspace similarity between the query and i-th model and m_i, σ_i are the mean and standard deviation of subspace similarities of validation image sets with the i-th class model. As each class exploits a different total correlation matrix, the score normalization process is required for classification.

Time-efficient classification. Batch computation of the C locally orthogonal subspaces of a query set for classification is time-consuming, i.e. taking $O(C \times min(N, M_q)^3)$, where M_q is the number of vectors in the query set. This computational cost is reduced using the update function $\mathcal{H}(\mathbf{P}_q, \Lambda_q, \mathbf{P}_T^i, \Lambda_T^i)$ in Section 4, where $\mathbf{P}_q, \Lambda_q$ are the eigenvector and eigenvalue matrices of the correlation matrix of the query set and $\mathbf{P}_T^i, \Lambda_T^i$ of the class specific total correlation matrix $\mathbf{R}_T^i$ respectively. Note that this only requires $O(C \times d_q^3)$, where d_q is the number of columns of $\mathbf{P}_q$. The subsequent canonical correlation matching with C models is not computationally expensive. It only costs $O(C \times d^3)$ (Refer to Section 2.1), where d is the dimension of the orthogonal subspaces.

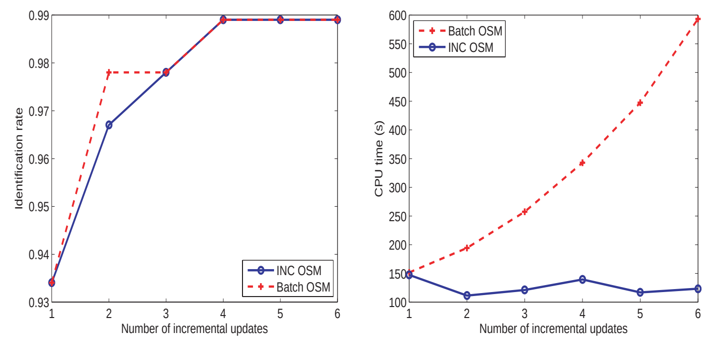
Incremental update of LOSM. Incremental update of the locally OSM may be similarly done as described in previous sections. The three steps in Section 4 are remained as the same except that the total correlation matrix is replaced with the class specific total correlation matrices defined above with w_j^i. Thus, when a new data set is added to the j-th class, the total correlation matrices that have nonzero w_j^i need to be updated, which increases the time complexity of the previous step up to C fold. In each update, the sufficient spanning set of the total correlation matrix remains the same as (12), since it is independent of the weight terms.

Fig. 3. **Data set.** *(top) Frames from a typical video sequence from the database used for evaluation. The motion of the user was not controlled, leading to different poses. (bottom) The 7 different illumination conditions in the database.*



(a) (b)

Fig. 4. **Batch vs. Incremental OSM-1.** *(a) Example orthogonal components, which are computed by the incremental and the batch-mode, are very alike. (b) Insensitivity of the incremental OSM to the dimensionality of the subspace of the total correlation matrix. The incremental solution yields the same solution as the batch-mode, provided the dimensionality of the subspace is high enough.*



(a) (b)

Fig. 5. **Batch vs. Incremental OSM-2.** *(a) Accuracy improvement of the incremental OSM for the number of updates. (b) Computational costs of the batch and incremental OSM.*

6 Evaluation

6.1 Data set

We used the face video database of 100 subjects. For each person, 7 video sequences of the individual in arbitrary motion were collected. Each sequence was recorded in a different illumination setting for 10s at 10fps and 320×240 pixel resolution (see Figure 3). Following automatic localization using a cascaded face detector [15] and cropping to the uniform scale, images of faces were histogram equalized. Each sequence is then represented by a set of raster-scanned vectors of the normalized images.

6.2 Batch OSM vs Incremental OSM in accuracy and time complexity

The incremental OSM yielded the same solution as the batch-mode OSM for the data merging scenario, where the 100 sequences of 100 face classes of a single illumination setting were initially used for learning the orthogonal subspaces. Then, the sets of the 100 face classes of other illumination settings were additionally given for the update. We set the total number of updates including the initial batch computation to be 6 and the number of images to add at each iteration around 10,000. The dimension of the uniformly scaled images was 2,500 and the number of orthogonal components was around 10. The latter was set to capture more than 99% of the energy from the eigenvalue plot. We used all 10 canonical correlations for classification. See Figure 4 (a) for the example orthogonal component computed by the proposed incremental algorithm and the batch-mode. The figure shows the element values (y-axis) of the 400-dimensional (x-axis) basis vectors. The errors compare favorably with the working precision of our machine. Figure 4 (b) shows the insensitivity of the incremental OSM to the dimension of the subspace of the total correlation matrix. The incremental OSM yields the same accuracy as the batch-mode OSM, provided the retained dimensionality of the subspace is sufficient. The subspace dimensionality was automatically chosen from the eigenvalues plots of the correlation matrices at each update. Figure 5 (a) shows the accuracy improvement of the incremental OSM according to the number of updates. It efficiently updates the existing orthogonal subspace models with new evidence contained in the additional data sets, giving increasing accuracy. The computational costs of the batch OSM and the incremental OSM are compared in Figure 5 (b). Whereas the computational cost of the bath-mode is largely increased as the data is repeatedly added, the incremental OSM keeps the cost of the update low.

6.3 Accuracy comparison with prior arts

Another experiment was designed for comparing the accuracy of several other methods with the proposed orthogonal and locally orthogonal subspace methods. The training of all the algorithms was performed with the data acquired in a single illumination setting and testing with a single other setting. An independent illumination set comprising both training and test sets was used for the validation. We compared the performance of Mutual Subspace Method (MSM) [7], where the dimension of each subspace is 10, representing more than 99% energy of the data, CMSM [10] used in a state-of-the-art commercial system FacePass [11], where the dimension

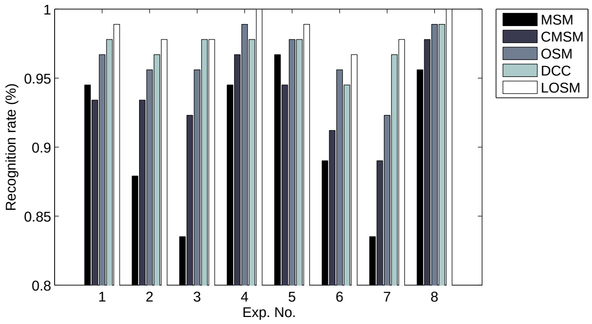


Fig. 6. **Accuracy comparison.**

of the constrained subspace was set to be 360, which yielded the best accuracy for the validation set, Discriminative Canonical Correlations (DCC) [21], Orthogonal Subspace Method (OSM), and Locally Orthogonal Subspace Method (LOSM), where the class priors were set by a threshold returning a half of the total classes as the neighboring classes. The component numbers of the total correlation matrix and the orthogonal subspaces of OSM and LOSM were 200 and 10 respectively. Figure 6 compares the recognition accuracy of all methods, where the experiment numbers correspond to the combinations of the training/test lighting sets. OSM was superior to CMSM and similar/or inferior to DCC except in experiment 4 and 6. The proposed locally orthogonal subspace method (LOSM) outperformed all the other methods.

6.4 Portal scenario of Multiple Biometric Grand Challenge

We have participated in the portal challenge of Multiple Biometric Grand Challenge [32]. The task is to match a query video captured at portal with still gallery images like passport photos i.e. a single image per person for face verification. The data set has in total 110 still images in gallery set and 140 videos in query set. The still images were captured in a studio quality (i.e. a good lighting, frontal facial pose and high resolution condition) and the videos in a poor indoor lighting including various head poses, scales and illumination changes. The challenge involves two experiments (called mask 1 and mask 2) taking different combinations of gallery and query subjects. See [32] for details. We have augmented 50 face gallery images by random affine transformations obtaining a set of face images per person. The other set of face images, typically composed of 150-200 images, was extracted from each query video, thus comprising set-to-set matching. Each face image is represented by multi-scale local binary pattern (LBP) histograms [33] (See Figure 7). For cropped face images of 142$\times$120 pixels, 10 LBP operators of the radius from one to ten were used. The number of non-overlapped regions was 81 or 100. We have proposed the three methods: the first is to compute the similarity score of each query image with a gallery image in the PCA+LDA space (leant by the augment gallery images) and to combine the similarity scores over the images of a query video. The second method is to match a query set to a gallery set by OSM based on LBPs: OSM is applied to each local component and the subspace similarities are summed over components. The third one is obtained by fusion of the two methods. Figure 8 shows the accuracy comparison by equal error rate (%). Increasing the number of components in LBPs improved the accuracy. The image frame based method with 100 components exhibited better accuracy than the image set based method OSM with 81 components in Mask1 but poorer in Mask2. The fusion method improved the best single method for both Masks. Note that the proposed subspace method worked well on the sparse representation of LBP.

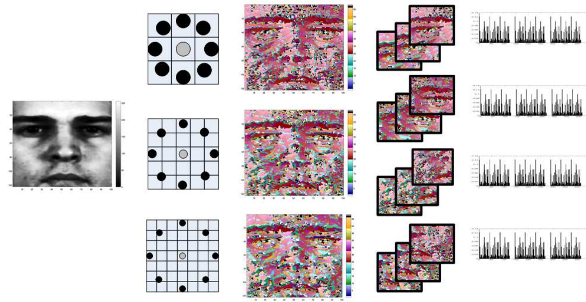


Fig. 7. **Multi-scale Local Binary Pattern Histograms.** *Multi-scale LBP images are divided into several non-overlap regions. Each region is represented as a histogram.*

Equal error rate(%)	K*	Mask 1	Mask 2
Method 1 (Image Frame based)	100	0.88%	1.36%
Method 1 (Image Frame based)	81	0.83%	1.50%
Method 2 (Image Set based)	81	1.93%	1.17%
Method 3 (Fusion)	81	0.70%	0.98%

Fig. 8. **MBGC Face verification results.** **denotes number of components in LBPs.*

7 Conclusions

In the object recognition task involving image sets, the development of an efficient incremental learning method for handling increasing volumes of image sets is important. Image data emanating from environments dramatically changing from time to time is continuously accumulated. The proposed incremental solution of the orthogonal subspaces and the locally orthogonal subspaces facilitates a highly efficient learning to adapt to new data sets. The same solution as the batch-computation is obtained with far lower complexity in both time and space. In the recognition experiments using 700 face image sets, the proposed LOSM delivered the best accuracy over all other relevant methods.

Acknowledgment

The authors would like to thank Ken-ichi Maeda for motivating this study and giving his valuable comments. Chi-Ho Chan and Norman Poh are greatly appreciated for their help with MBGC. Multi-scale local binary patterns used in the experiment were received from Chi-Ho. This study was funded by the Toshiba-Cambridge Scholarship and the Chevening Scholarship. T-K. Kim is presently supported by the research fellowship of the Sidney Sussex College of the University of Cambridge. J. Kittler was partially supported by EU Projects VidiVideo and Mobio.

References

[1] O. Arandjelovic and R. Cipolla. A Pose-Wise Linear Illumination Manifold Model for Face Recognition using Video. *Computer Vision and Image Understanding*, 113(1):113–125, 2009.

[2] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, *Proc. of CVPR*, 2008.

[3] T. Wang and P. Shi, Kernel Grassmannian distances and discriminant analysis for face recognition from image sets, *Pattern Recognition Letters*, 30(13):1161–1165, 2009.

[4] G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *Proc. of ECCV*, pages 851–868, 2002.

[5] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *Proc. of CVPR*, 2005.

[6] S. Satoh. Comparative Evaluation of Face Sequence Matching for Content-based Video Access. *Int'l Conf. on Automatic Face and Gesture Recognition*, 2000.

[7] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *Int'l Conf. on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

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

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

[10] M. Nishiyama, O. Yamaguchi, and K. Fukui, Face Recognition with the Multiple Constrained Mutual Subspace Method. *Proc. of Audio Video Based Person Authentication*, 2005.

[11] Toshiba. Facepass. http://www.toshiba.co.jp/rdc/mmlab/tech/w31e.htm.

[12] K. Fukui, B. Stenger and O. Yamaguchi, A Framework for 3D Object Recognition Using the Kernel Constrained Mutual Subspace Method. *Proc. of ACCV*, 2006.

[13] E.Oja, Subspace Methods of Pattern Recognition. Research Studies Press, 1983.

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

[15] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[16] P. Hall, D. Marshall and R. Martin. Merging and Splitting Eigenspace Models. *IEEE Trans. on PAMI*, 22(9), 2000.

[17] T-K. Kim, S. Wong, B. Stenger, J. Kittler, R. Cipolla, Incremental Linear Discriminant Analysis Using Sufficient Spanning Set Approximations, Proc. of CVPR, Minneapolis, USA, 2007.

[18] S. Zhou, V. Krueger, and R. Chellappa, Probabilistic recognition of human faces from video, *Computer Vision and Image Understanding*, 91(1-2):214–245, 2003.

[19] X. Liu and T. Chen. Video-Based Face Recognition Using Adaptive Hidden Markov Models. *Proc. of CVPR*, pages 340–345, Madison, USA, 2003.

[20] K. Lee, M. Yang, and D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, *Proc. of CVPR*, volume 1, pages 313–320, Madison, USA, 2003.

[21] T-K. Kim, J. V. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. on PAMI*, 29(6), 2007.

[22] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(34):321–372, 1936.

[23] F.R. Bach and M.I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[24] J. Ye, Q. Li, H. Xiong, H. Park, V. Janardan, and V. Kumar. IDR/QR: An incremental dimension reduction algorithm via QR decomposition. *IEEE Trans. on Knowledge and Data Engineering*, 17(9):1208–1222, 2005.

[25] K. Hiraoka, K. Hidai, M. Hamahira, H. Mizoguchi, T. Mishima, and S. Yoshizawa. Successive learning of linear discriminant analysis: Sanger-type algorithm. *Proc. of ICPR*, 2000.

[26] T.-J. Chin and D. Suter. Incremental Kernel PCA for Efficient Non-linear Feature Extraction. *Proc. of BMVC*, 2006.

[27] X. Tao, J. Ye, Q. Li, R. Janardan, and V. Cherkassky. Efficient Kernel Discriminant Analysis via QR Decomposition. *In Advances in Neural Information Processing Systems*, 2004.

[28] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, O. Yamaguchi, Recognizing Faces of Moving People by Hierarchical Image-Set Matching. *Proc. of CVPR*, 2007.

[29] K. Fukui and O. Yamaguchi, The Kernel Orthogonal Mutual Subspace Method and Its Application to 3D Object Recognition. *Proc. of ACCV*, (2):467–476, 2007.

[30] T-K. Kim, J. Kittler and R. Cipolla, Incremental Learning of Locally Orthogonal Subspaces for Set-based Object Recognition, *Proc. of IAPR British Machine Vision Conference*, pages 559-568, Edinburgh, UK, Sept 2006.

[31] T-K. Kim, O. Arandjelović, R. Cipolla, Learning over sets using boosted manifold principle angles (BoMPA), *Proc. of BMVC*, Oxford, UK, 2005.

[32] Multiple Biometric Grand Challenge, http://face.nist.gov/mbgc/

[33] C-H. Chan, J. Kittler, K. Messer, Multi-scale Local Binary Pattern Histograms for Face Recognition. *Proc. of ICB* pages 809–818, 2007.