Learning Discriminative Canonical Correlations for Object Recognition with Image Sets

Tae-Kyun Kim[1], Josef Kittler[2], and Roberto Cipolla[1]

[1] Department of Engineering, University of Cambridge,
Cambridge, CB2 1PZ, UK
{tkk22, cipolla}@eng.cam.ac.uk
[2] Centre for Vision, Speech and Signal Processing, University of Surrey,
Guildford, GU2 7XH, UK
J.Kittler@surrey.ac.uk

Abstract. We address the problem of comparing sets of images for object recognition, where the sets may represent arbitrary variations in an object's appearance due to changing camera pose and lighting conditions. The concept of Canonical Correlations (also known as principal angles) can be viewed as the angles between two subspaces. As a way of comparing sets of vectors or images, canonical correlations offer many benefits in accuracy, efficiency, and robustness compared to the classical parametric distribution-based and non-parametric sample-based methods. Here, this is demonstrated experimentally for reasonably sized data sets using existing methods exploiting canonical correlations. Motivated by their proven effectiveness, a novel discriminative learning over sets is proposed for object recognition. Specifically, inspired by classical Linear Discriminant Analysis (LDA), we develop a linear discriminant function that maximizes the canonical correlations of within-class sets and minimizes the canonical correlations of between-class sets. The proposed method significantly outperforms the state-of-the-art methods on two different object recognition problems using face image sets with arbitrary motion captured under different illuminations and image sets of five hundred general object categories taken at different views.

1 Introduction

Whereas most previous works for object recognition have focused on the problems of ***single-to-single*** or ***single-to-many*** vector matching, many tasks can be cast as matching problems of vector sets (i.e. ***many-to-many***) for robust object recognition. In object recognition, e.g., 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 efficiently classify a novel set of vectors to one of the training classes, each also represented by one or several vector sets. In this study, sets may be derived from sparse and unordered observations acquired by e.g. multiple still shots of a three dimensional object or a long term surveillance systems, where a subject would not face the camera all the time. Without temporal coherence,

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 251–262, 2006.

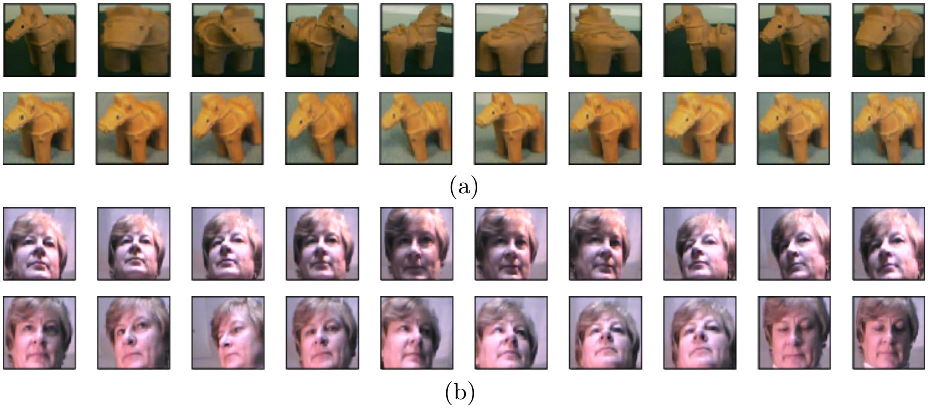


Fig. 1. Examples of image sets. (a) Two sets (top and bottom) contain images of an 3D object taken from different views but with a certain overlap in their views. (b) Face image sets collected from videos taken under different illumination settings. Face patterns of the two sets (top and bottom) vary in both lighting and pose.

training sets can be conveniently augmented. See Figure 1 for examples of pattern sets of objects. The previous works exploiting temporal coherence between consecutive images [1, 2] are irrelevant to this study. Furthermore, this work does not explicitly exploit any data-semantics in images, but is purely based on automatic learning of given labelled image sets. Therefore, we expect that the proposed method can be applied to many other problems requiring a set comparison.

Relevant previous approaches for set matching can be broadly partitioned into ***model-based*** and ***sample-based*** methods. In the parametric model-based approaches [3, 4], each set is represented by a parametric distribution function, typically Gaussian. The closeness of the two distributions is then measured by the Kullback-Leibler Divergence (KLD) [3]. Due to the difficulty of parameter estimation under limited training data, these methods easily fail when the training and test sets do not have strong statistical correlations.

More suitable methods for comparing sets are based on the matching of pairwise samples of sets, e.g. Nearest Neighbour (NN) or Hausdorff distance matching [5]. The methods are based on the premise that similarity of a pair of sets is reflected by the similarity of the modes (or NNs) of the two respective sets. This is useful in many computer vision applications, where the data acquisition conditions and the semantics of sets may change dramatically over time. However, they do not take into account the effect of outliers as well as the natural variability of the sensory data due to the 3D nature of the observed objects. Note also that such methods are very computationally expensive as they require a comparison of every pairwise samples of any two sets.

Another model-based approaches are based on the concept of ***canonical correlations***, which has attracted increasing attention for image set matching in [8]-[11], following the early works [12, 13]. Each set is represented by a linear subspace and the angles between two subspaces are exploited as a similarity measure of two sets. As a method of comparing sets, the benefits of canonical correlations, as compared with both, distribution based and sample based matching, have been noted in [4, 10]. A nonlinear extension of canonical correlation has been proposed in [9, 10]. The previous work called Constrained Mutual Subspace Method (CMSM) [11] is the most related with this paper. In CMSM, a constrained subspace is defined as the subspace in which the entire class population exhibits small variance. The authors showed that the sets of different classes in the constrained subspace had small canonical correlations. However, the principle of CMSM is rather heuristic, especially the process of selecting the dimensionality of the constrained subspace. If the dimensionality is too low, the subspace will be a null space. In the opposite case, the subspace simply passes all the energy of the original data and thus could not play a role as a discriminant function.

Given a similarity function of two sets, an important problem in set classification is how to learn discriminative information (or a discriminant function) from data associated with the given similarity function. To our knowledge, the topic of discriminative learning over sets has not been given a proper attention in literature. This paper presents a novel method for an optimal linear discriminant function of image sets based on canonical correlations. A linear discriminant function that maximizes the canonical correlations of within-class sets and minimizes the canonical correlations of between-class sets is devised, by analogy to the optimization concept of Linear Discriminant Analysis (LDA) [6]. The linear mapping is found by a novel iterative optimization algorithm. The discriminative capability of the proposed method is shown to be significantly better than the method [8] that simply aggregates canonical correlations and the k-NN methods in LDA subspace [5]. Compared with CMSM [11], the proposed method is more practical by easiness of feature selection as well as it is more theoretically appealing.

2 Discriminative Canonical Correlations (DCC)

2.1 Canonical Correlations

Canonical correlations, which are cosines of principal angles $0 \leq \theta_1 \leq \ldots \leq \theta_d \leq (\pi/2)$ between any two d-dimensional linear subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$ are uniquely defined as:

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

subject to $\mathbf{u}_i^T\mathbf{u}_i = \mathbf{v}_i^T\mathbf{v}_i = 1, \ \mathbf{u}_i^T\mathbf{u}_j = \mathbf{v}_i^T\mathbf{v}_j = 0, \ \text{for } i \neq j.$ Of the various ways to solve this problem, the Singular Value Decomposition (SVD) solution [13] is more numerically stable. The SVD solution is as follows: Assume that $\mathbf{P}_1 \in \mathbf{R}^{n \times d}$ and $\mathbf{P}_2 \in \mathbf{R}^{n \times d}$ form unitary orthogonal basis matrices for two linear subspaces, $\mathcal{L}_1$ and $\mathcal{L}_2$. Let the SVD of $\mathbf{P}_1^T\mathbf{P}_2$ be

$$\mathbf{P}_1^T\mathbf{P}_2 = \mathbf{Q}_{12}\mathbf{\Lambda}\mathbf{Q}_{21}^T \quad s.t. \quad \mathbf{\Lambda} = diag(\sigma_1, ..., \sigma_d) \tag{2}$$

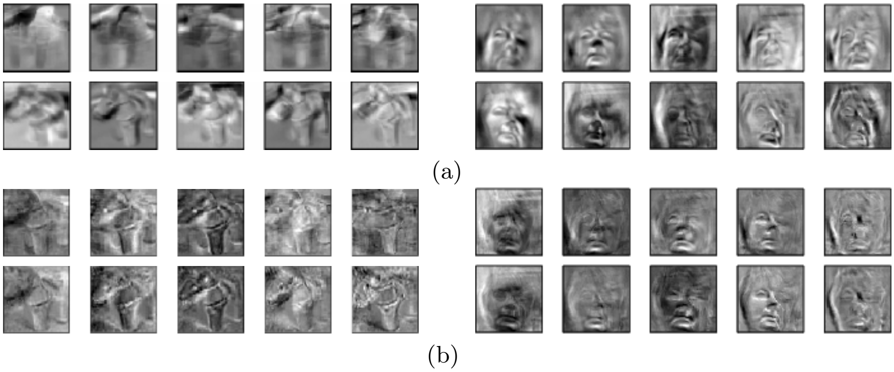


Fig. 2. Principal components vs. canonical vectors. (a) The first 5 principal components computed from the four image sets shown in Figure 1. The principal components of the different image sets (see each column) show significantly different variations even for the same objects. (b) The first 5 canonical vectors of the four image sets. Every pair of canonical vectors (each column) well captures the common modes (views and illuminations) of the two sets, i.e. the pairwise canonical vectors are almost similar. The canonical vectors of different dimensions represent different pattern variations e.g. in pose or lighting.

where $\mathbf{Q}_{12}, \mathbf{Q}_{21}$ are orthogonal matrices, i.e. $\mathbf{Q}_{ij}^T\mathbf{Q}_{ij} = \mathbf{Q}_{ij}\mathbf{Q}_{ij}^T = \mathbf{I}_d$. Canonical correlations are $\{\sigma_1, ..., \sigma_d\}$ and the associated canonical vectors are $\mathbf{U} = \mathbf{P}_1\mathbf{Q}_{12} = [\mathbf{u}_1, ..., \mathbf{u}_d]$, $\mathbf{V} = \mathbf{P}_2\mathbf{Q}_{21} = [\mathbf{v}_1, ..., \mathbf{v}_d]$. The canonical correlations tell us how close are the closest vectors of two subspaces. Different canonical correlations tell about the proximity of vectors in other dimensions (perpendicular to the previous ones) of the two subspaces. See Figure 2 for the canonical vectors computed from the sample image sets given in Figure 1. Whereas the principal components vary for different imaging conditions of the sets, the canonical vectors well capture the common modes of the two different sets.

Compared with the parametric distribution-based matching, this concept is much more flexible as it effectively places a uniform prior over the subspace of possible pattern variations. Compared with the NN matching of samples, this approach is much more stable as patterns are confined to certain subspaces. The low computational complexity of matching by canonical correlations is also much favorable.

3 Learning a Discriminant Function of Canonical Correlations

3.1 Problem Formulation

Assume m sets of vectors are given as $\{\mathbf{X}_1, ..., \mathbf{X}_m\}$, where $\mathbf{X}_i$ describes a data matrix of the i th set containing observation vectors (or images) in its columns. Each set belongs to one of object classes denoted by C_i. A d-dimensional linear subspace of the i th set is represented by an orthonormal basis matrix $\mathbf{P}_i \in \mathbf{R}^{n \times d}$ s.t. $\mathbf{X}_i\mathbf{X}_i^T \simeq \mathbf{P}_i\mathbf{\Lambda}_i\mathbf{P}_i^T$, where $\mathbf{\Lambda}_i, \mathbf{P}_i$ are the eigenvalue and eigenvector matrices of the d largest eigenvalues respectively and n denotes the vector dimension. We define a transformation matrix $\mathbf{T}$ s.t. $\mathbf{T} : \mathbf{X}_i \rightarrow \mathbf{Y}_i = \mathbf{T}^T\mathbf{X}_i$. The matrix $\mathbf{T}$ is to transform images so that the transformed image sets are more class-wise discriminative using canonical correlations.

Representation. Orthonormal basis matrices of the subspaces for the transformed data are obtained from the previous matrix factorization of $\mathbf{X}_i\mathbf{X}_i^T$:

$$\mathbf{Y}_i\mathbf{Y}_i^T = (\mathbf{T}^T\mathbf{X}_i)(\mathbf{T}^T\mathbf{X}_i)^T \simeq (\mathbf{T}^T\mathbf{P}_i)\mathbf{\Lambda}_i(\mathbf{T}^T\mathbf{P}_i)^T \tag{3}$$

Except when $\mathbf{T}$ is an orthogonal matrix, $\mathbf{T}^T\mathbf{P}_i$ is not generally an orthonormal basis matrix. Note that canonical correlations are only defined for orthonormal basis matrices of subspaces. Any orthonormal components of $\mathbf{T}^T\mathbf{P}_i$ now defined by $\mathbf{T}^T\mathbf{P'}_i$ can represent an orthonormal basis matrix of the transformed data. See Section 3.2 for details.

Set Similarity. The similarity of any two transformed data sets are defined as the sum of canonical correlations by

$$F_{ij} = \max_{\mathbf{Q}_{ij},\mathbf{Q}_{ji}} \mathrm{tr}(M_{ij}), \tag{4}$$

$$M_{ij} = \mathbf{Q}_{ij}^T\mathbf{P'}_i^T\mathbf{T}\mathbf{T}^T\mathbf{P'}_j\mathbf{Q}_{ji} \quad \text{or} \quad \mathbf{T}^T\mathbf{P'}_j\mathbf{Q}_{ji}\mathbf{Q}_{ij}^T\mathbf{P'}_i^T\mathbf{T}, \tag{5}$$

as $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any matrix A, B. $\mathbf{Q}_{ij}, \mathbf{Q}_{ji}$ are the rotation matrices defined in the solution of canonical correlations (2).

Discriminant Function. The discriminative function $\mathbf{T}$ is found to maximize the similarities of any pairs of sets of within-classes while minimizing the similarities of pairwise sets of between-classes. Matrix $\mathbf{T}$ is defined by

$$\mathbf{T} = \arg\max_{\mathbf{T}} \frac{\sum_{i=1}^{m}\sum_{k \in W_i} F_{ik}}{\sum_{i=1}^{m}\sum_{l \in B_i} F_{il}} \tag{6}$$

where $W_i = \{j \,|\mathbf{X}_j \in C_i\}$ and $B_i = \{j \,|\mathbf{X}_j \notin C_i\}$. That is, the two sets W_i, B_i denote the within-class and between-class sets of a given set class i respectively, which are similarly defined with [7].

3.2 Iterative Optimization

The optimization problem of $\mathbf{T}$ involves the variables of $\mathbf{Q}, \mathbf{P'}$ as well as $\mathbf{T}$. As the other variables are not explicitly represented by $\mathbf{T}$, a closed form solution for $\mathbf{T}$ is hard to find. We propose an iterative optimization algorithm. Specifically, we compute an optimal solution for one of the three variables at a time by fixing the other two and repeating this for a certain number of iterations.

Algorithm 1. Discriminative Canonical Correlations (DCC)

Input: All $\mathbf{P}_i \in \mathcal{R}^{n\times d}$ **Output:** $\mathbf{T} \in \mathcal{R}^{n\times n}$

```
$\mathbf{T} \leftarrow \mathbf{I}_n$
Do iterate the followings:
   For all $i$, do QR-decomposition: $\mathbf{T}^T\mathbf{P}_i = \mathbf{\Phi}_i\mathbf{\Delta}_i \rightarrow \mathbf{P'}_i = \mathbf{P}_i\mathbf{\Delta}_i^{-1}$
   For every pair $i,j$, do SVD: $\mathbf{P'}_i^T\mathbf{T}\mathbf{T}^T\mathbf{P'}_j = \mathbf{Q}_{ij}\mathbf{\Lambda}\mathbf{Q}_{ji}^T$
   Compute $\mathbf{S'_b} = \sum_{i=1}^m \sum_{l\in B_i}(\mathbf{P'}_l\mathbf{Q}_{li} - \mathbf{P'}_i\mathbf{Q}_{il})(\mathbf{P'}_l\mathbf{Q}_{li} - \mathbf{P'}_i\mathbf{Q}_{il})^T$,
           $\mathbf{S'_w} = \sum_{i=1}^m \sum_{k\in W_i}(\mathbf{P'}_k\mathbf{Q}_{ki} - \mathbf{P'}_i\mathbf{Q}_{ik})(\mathbf{P'}_k\mathbf{Q}_{ki} - \mathbf{P'}_i\mathbf{Q}_{ik})^T$.
   Compute eigenvectors $\{\mathbf{t}_i\}_{i=1}^n$ of $(\mathbf{S'_w})^{-1}\mathbf{S'_b}$,   $\mathbf{T} \leftarrow [\mathbf{t}_1, ..., \mathbf{t}_n]$
End
```

Thus, the proposed iterative optimization is comprised of the three main steps: normalization of $\mathbf{P}$, optimization of matrices $\mathbf{Q}$, and $\mathbf{T}$. Each step is explained below:

Normalization. The matrix $\mathbf{P}_i$ is normalized to $\mathbf{P'}_i$ for a fixed $\mathbf{T}$ so that the columns of $\mathbf{T}^T\mathbf{P'}_i$ are orthonormal. QR-decomposition of $\mathbf{T}^T\mathbf{P}_i$ is performed s.t. $\mathbf{T}^T\mathbf{P}_i = \mathbf{\Phi}_i\mathbf{\Delta}_i$, where $\mathbf{\Phi}_i \in \mathbf{R}^{n\times d}$ is the orthonormal matrix with the first d columns and $\mathbf{\Delta}_i \in \mathbf{R}^{d\times d}$ is the invertible upper-triangular matrix with the first d rows. From (3), $\mathbf{Y}_i = \mathbf{T}^T\mathbf{P}_i\sqrt{\mathbf{\Lambda}}_i = \mathbf{\Phi}_i\mathbf{\Delta}_i\sqrt{\mathbf{\Lambda}}_i$. As $\mathbf{\Delta}_i\sqrt{\mathbf{\Lambda}}_i$ is still an upper-triangular matrix, $\mathbf{\Phi}_i$ can represent an orthonormal basis matrix of the transformed data $\mathbf{Y}_i$. As $\mathbf{\Delta}_i$ is invertible,

$$\mathbf{\Phi}_i = \mathbf{T}^T(\mathbf{P}_i\mathbf{\Delta}_i^{-1}) \quad \rightarrow \quad \mathbf{P'}_i = \mathbf{P}_i\mathbf{\Delta}_i^{-1}. \tag{7}$$

Computation of Rotation Matrices Q. Rotation matrices $\mathbf{Q}_{ij}$ for every i,j are obtained for a fixed $\mathbf{T}$ and $\mathbf{P'}_i$. The correlation matrix M_{ij} in the left of (5) can be conveniently used for the optimization of $\mathbf{Q}_{ij}$, as it has $\mathbf{Q}_{ij}$ outside of the matrix product. Let the SVD of $\mathbf{P'}_i^T\mathbf{T}\mathbf{T}^T\mathbf{P'}_j$ be

$$\mathbf{P'}_i^T\mathbf{T}\mathbf{T}^T\mathbf{P'}_j = \mathbf{Q}_{ij}\mathbf{\Lambda}\mathbf{Q}_{ji}^T \tag{8}$$

where $\mathbf{\Lambda}$ is a singular matrix and $\mathbf{Q}_{ij}, \mathbf{Q}_{ji}$ are orthogonal rotation matrices.

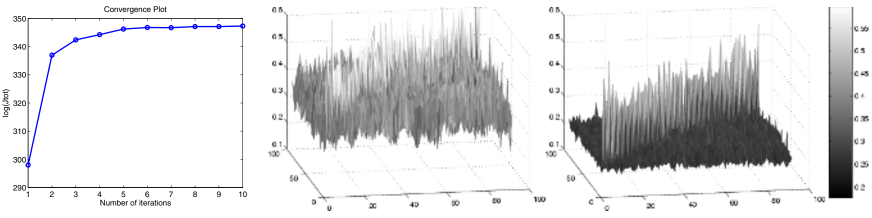
Computation of T. The optimal discriminant transformation $\mathbf{T}$ is computed for given $\mathbf{P'}_i$ and $\mathbf{Q}_{ij}$ by using the definition of M_{ij} in the right of (5) and (6). With $\mathbf{T}$ being on the outside of the matrix product, it is convenient to solve for. The discriminative function is found by

$$\mathbf{T} = \max_{arg\mathbf{T}} \ \mathrm{tr}(\mathbf{T^T S_w T})/\mathrm{tr}(\mathbf{T^T S_b T}) \tag{9}$$

$$\mathbf{S_w} = \sum_{i=1}^m \sum_{k\in W_i} \mathbf{P'}_k\mathbf{Q}_{ki}\mathbf{Q}_{ik}^T\mathbf{P'}_i^T, \quad \mathbf{S_b} = \sum_{i=1}^m \sum_{l\in B_i} \mathbf{P'}_l\mathbf{Q}_{li}\mathbf{Q}_{il}^T\mathbf{P'}_i^T \tag{10}$$

where $W_i = \{j \ |\mathbf{X}_j \in C_i\}$ and $B_i = \{j \ |\mathbf{X}_j \notin C_i\}$. For a more stable solution, an alternative optimization is finally proposed by

$$\mathbf{T} = \max_{arg\mathbf{T}} \ \mathrm{tr}(\mathbf{T}^T\mathbf{S'_b}\mathbf{T})/\mathrm{tr}(\mathbf{T}^T\mathbf{S'_w}\mathbf{T}) \tag{11}$$



(a) (b) (c)

Fig. 3. Example of learning. (a) The cost function for the number of iterations. Confusion matrices of the training set (b) before the learning ($\mathbf{T} = \mathbf{I}$) and (c) after the learning. The discriminability of canonical correlations was significantly improved by the proposed learning.

$$\mathbf{S'_b} = \textstyle\sum_{i=1}^{m} \sum_{l \in B_i} (\mathbf{P'}_l \mathbf{Q}_{li} - \mathbf{P'}_i \mathbf{Q}_{il})(\mathbf{P'}_l \mathbf{Q}_{li} - \mathbf{P'}_i \mathbf{Q}_{il})^T, \qquad (12)$$

$$\mathbf{S'_w} = \textstyle\sum_{i=1}^{m} \sum_{k \in W_i} (\mathbf{P'}_k \mathbf{Q}_{ki} - \mathbf{P'}_i \mathbf{Q}_{ik})(\mathbf{P'}_k \mathbf{Q}_{ki} - \mathbf{P'}_i \mathbf{Q}_{ik})^T. \qquad (13)$$

Note that no loss of generality is incurred by this modification of the objective function as

$$A^T B = \mathbf{I} - 1/2 \cdot (A - B)^T (A - B),$$

where $A = \mathbf{T}^T \mathbf{P'}_i \mathbf{Q}_{ij}, B = \mathbf{T}^T \mathbf{P'}_j \mathbf{Q}_{ji}$. The solution $\{\mathbf{t}_i\}_{i=1}^n$ is obtained by solving the following generalized eigenvalue problem: $\mathbf{S'_b t} = \lambda \mathbf{S'_w t}$. When $\mathbf{S'_w}$ is non singular, the optimal $\mathbf{T}$ is computed by eigen-decomposition on $(\mathbf{S'_w})^{-1}\mathbf{S'_b}$. Note also that the proposed learning can avoid a singular case of $\mathbf{S'_w}$ by pre-applying PCA to data similarly with the Fisherface method [6] and speed up by using a small number of nearest neighboring sets for B_i, W_i in (6) like [7].

With the identity matrix $\mathbf{I} \in \mathbf{R}^{n \times n}$ as the initial value of $\mathbf{T}$, the algorithm is iterated until it converges to a stable point. A Pseudo-code for the learning is given in **Algorithm 1**. See Figure 3 for an example of learning. It converges fast and stably and dramatically improves the discriminability of the simple aggregation method of canonical correlations (i.e. $\mathbf{T} = \mathbf{I}$). After $\mathbf{T}$ is found to maximize the canonical correlations of within-class sets and minimize those of between-class sets, a comparisons of any two sets is achieved using the similarity value defined in (4).

4 Experimental Results and Discussion

4.1 Experimental Setting for Face Recognition

Database and Protocol. We have acquired a database called the ***Cambridge-Toshiba Face Video Database*** with 100 individuals of varying age and ethnicity, and equally represented genders. For each person, 7 video sequences of the person 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 4). Following automatic localization using a cascaded face detector [14]

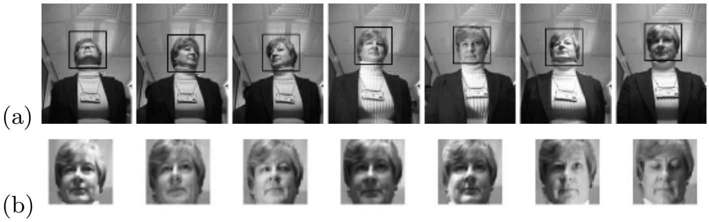


Fig. 4. Face data sets. (a) Frames of a typical face video sequence. (b) Face prototypes of 7 different lighting sequences.

and cropping to the uniform scale of 20×20 pixels, images of faces were histogram equalized. Training of all the algorithms was performed with data acquired in a single illumination setting and testing with a single other setting. We used 18 randomly selected training/test combinations for reporting identification rates.

Comparative Methods. We compared the performance of our learning algorithm (DCC) to that of:

- K-L Divergence algorithm (KLD) [3],
- k-Nearest Neighbours (k-NN) and Hausdorff distance[1] in (i) PCA, and (ii) LDA [6] subspaces estimated from training data [5],
- Mutual Subspace Method (MSM) [8], which is equivalent to the simple aggregation of canonical correlations,
- Constrained MSM (CMSM) [11] used in a state-of-the-art commercial system FacePass [16].

Dimensionality Selection. In KLD, 96% of data energy was explained by the principal subspace of training data used. In NN-PCA, the optimal number of principal components was 150 without the first three. In NN-LDA, PCA with 150 dimensions (removal of the first 3 principal components did not improve the LDA performance) was applied first to avoid singularity problems and the best dimension of LDA subspace was 150 again. In both MSM and CMSM, the PCA dimension of each image set was fixed to 10, which represents more than 98% of data energy of the set. All 10 canonical correlations were exploited. In CMSM, the best dimension of the constrained subspace was found to be 360 in terms of the test identification rates as shown in Figure 5. The CMSM exhibits a peaking and does not have a principled way of choosing dimensionality of the constrained subspace in practice. By contrast, the proposed method provided constant identification rates regardless of dimensionality of $\mathbf{T}$ beyond a certain point, as shown in Figure 5. Thus we could fix the dimensionality at 400 for all experiments. This behaviour is highly beneficial from the practical point of view. The PCA dimension of image sets was also fixed to 10 for the proposed method.

Construction of Within-Class Sets for the Proposed Method. In the face image set experiment, the images drawn from a single video sequence of arbitrary

[1] $d(X_1, X_2) = \min_{x_1 \in X_1} \max_{x_2 \in X_2} d(x_1, x_2)$

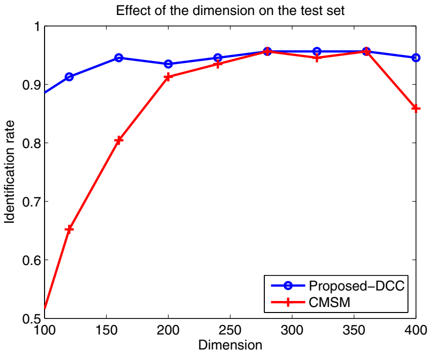


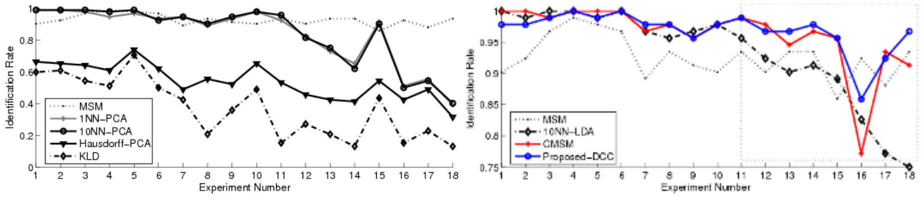
Fig. 5. Dimensionality selection for the proposed method and CMSM. The proposed method is more favorable than CMSM in dimensionality selection. CMSM shows a high peaking. The accuracy of CMSM at 400 is just equivalent to that of simple aggregation of canonical correlations.

head movement were randomly divided into the two within-class sets. The test recognition rates changed by less than 1-2 % for the different trials of random partitioning. In the experiment of general object recognition in Section 4.3, the two sets defined according to different viewing scopes comprised the within class sets.

4.2 Accuracy Comparison for Face Experiments

The 18 experiments were arranged in the order of increasing K-L Divergence between the training and test data. Lower K-L Divergence indicates more similar conditions. The identification rates of the evaluated algorithms is shown in Figure 6.

First, different methods of measuring set similarity were compared in Figure 6 (a). Most of the methods generally had lower recognition rates for experiments having larger KL-Divergence. The KLD 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 within-class face



(a) (b)

Fig. 6. Identification rates for the 18 experiments. (a) Methods of set matching. (b) Methods of set matching combined with discriminative transformations. (The variation between the training and test data of the experiments increases along the horizontal axis. Note that (a) and (b) have different scales for vertical axis.)

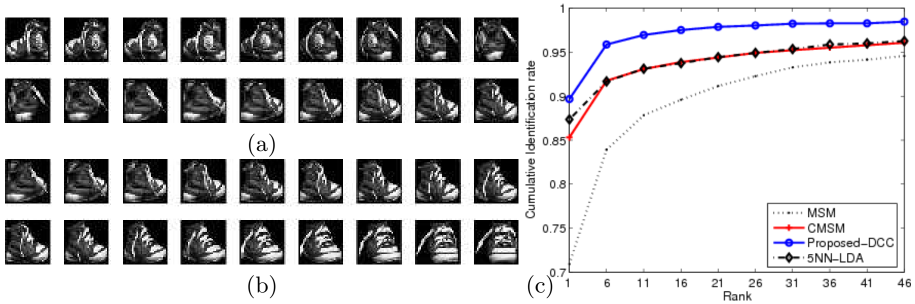
patterns was very broad, making this result unsurprising. As representatives of non-parametric sample-based matching, the 1-NN, 10-NN, and Hausdorff-distance methods defined in the PCA subspace were evaluated. It was observed that the Hausdorff-distance measure provided consistent but far poorer results than the NN methods. 10-NN yielded the best accuracy of the three, which is worse than MSM by 8.6% on average. Its performance greatly varied across the experiments while MSM showed robust performance under the different experimental conditions.

Second, methods combined with any discriminant function were compared in Figure 6 (b). Note that Figure 6 (a) and (b) have different scales. By taking MSM as a gauging proxy, 1-NN, 10-NN, and Hausdorff distance in the LDA subspace and CMSM were compared with the proposed algorithm. Here again, 10-NN was the best of the three LDA methods. For better visualization of comparative results, the performance of 1-NN and Hausdorff in LDA was removed from the figure. 10-NN-LDA yielded a big improvement over 10-NN-PCA but the accuracy of the method again greatly varied across the experiments. Note that 10-NN-LDA outperformed MSM for similar conditions between the training and test sets, but it became noticeably inferior as the conditions changed. The recognition rate of NN-LDA was considerably inferior to our method for the more difficult experiments (experiments 11 to 18 in Figure 6 (b)). NN-LDA yielded just 75% recognition rate for exp.18 where two very different illumination settings (see last two of Figure 4 (b)) were used for the training and test data. The accuracy of our method remained high at 97%. Note that the experiments 11 to 18 in Figure 6 are more realistic than the first half because they have greater variation in lighting conditions between training and testing. The proposed method also constantly provided a significant improvement over MSM. Just one exception for the proposed method due to overfitting were noted. Except this single case, the proposed method improved MSM by 5-10 % reaching almost more than 97% recognition rate.

Although the proposed method achieved a comparable accuracy with CMSM in the face recognition experiment, the latter had to be optimised aposteriori by dimensionality selection. By contrast, DCC does not need any feature selection. The underlying concept of CMSM is to orthogonalize different class subspaces [17], i.e. to make ${\mathbf{P}_i}^T\mathbf{P}_j = \mathbf{O}$ if $C_i \neq C_j$, where $\mathbf{O}$ is a zero matrix. Then, canonical correlations (2) of the orthogonal subspaces become zeros as $\mathrm{tr}(\mathbf{Q}_{ij}^T{\mathbf{P}_i}^T\mathbf{P}_j\mathbf{Q}_{ji}) = 0$. However, subspaces can not always be orthogonal to all the other subspaces. Then, a direct optimization of canonical correlations in the proposed method would be preferred.

4.3 Experiment on Large Scale General Object Classes

The ALOI database [15] with 500 general object categories of different viewing angles provides another experimental data set for the comparison. Object images were segmented from the simple background and scaled to 20×20 pixel size. The training and five test sets were set up with different viewing angles of the objects as shown in Figure 7 (a) and (b). All images in the test sets had at least


(a)

(b)

(c)

Fig. 7. ALOI experiment. (a) The training set consists of 18 images taken at every 10 degree. (b) Two test sets are shown. Each test set contains 9 images at 10 degree intervals, different from the training set. (c) Cumulative identification plots of several methods.

5 degree pose difference from every sample of the training set. The methods of MSM, NN-LDA and CMSM were compared with the proposed method in terms of identification rate. The PCA dimensionality of each set was fixed to 5 and thus 5 canonical correlations were exploited for MSM, CMSM and the proposed method. Similarly, 5 nearest neighbours were used in LDA. See Figure 7 (c) for the cumulative identification rates. Unlike the face experiment, NN-LDA yielded better accuracy than MSM. This might be due to the nearest neighbours of the training and test set differed only slightly by the five degree pose difference (The two sets had no changes in lighting and they had accurate localization of the objects.). Here again, the proposed method were substantially superior to both MSM and NN-LDA. The proposed method outperformed even the best behaviour of CMSM in this scenario.

4.4 Computational Complexity

The matching complexity of the methods using canonical correlations, $O(d^3)$, is far lower than that of the sample-based matching methods such as k-NN, $O(c^2n)$, where d is the subspace dimension of each set, c is the number of samples of each set and n is the dimensionality of feature space, since $d \ll c, n$.

5 Conclusions

A novel discriminative learning framework has been proposed for object recognition using canonical correlations of image sets. The proposed method has been evaluated on both face image sets obtained from videos and image sets of five hundred general object categories. The new technique facilitates effective discriminative learning over sets, thus providing an impressive set classification accuracy. It significantly outperformed the KLD method representing a parametric distribution-based matching and NN in both PCA/LDA subspaces as examples of non-parametric sample-based matching. It also largely outperformed MSM and achieved a comparable accuracy with the best behavior of CMSM but, more pertinently, without the need for feature selection. The proposed method is also more theoretically appealing than CMSM.

Acknowledgements

The authors acknowledge the support of the Toshiba Corporation and the Chevening Scholarship. They are grateful to Osamu Yamaguchi for motivating this study; to Ognjen Arandjelović for helpful discussions and to Gabriel Brostow for help with proof-reading.

References

1. K. Lee, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *CVPR*, pages 313–320, 2003.
2. S. Zhou, V. Krueger, and R.Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(1):214–245, 2003.
3. G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *ECCV*, pages 851–868, 2002.
4. O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *CVPR*, 2005.
5. S. Satoh. Comparative Evaluation of Face Sequence Matching for Content-based Video Access. *Int'l Conf. on Automatic Face and Gesture Recognition*, 2000.
6. P.N.Belhumeur, J.P.Hespanha, and D.J.Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *PAMI*, 19(7):711–720, 1997.
7. M. Bressan, J. Vitria Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24(2003):2743–2749, 2003.
8. O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *AFG*, (10):318–323, 1998.
9. L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *JMLR*, 4(10):913–931, 2003.
10. T.K. Kim, O. Arandjelović and R. Cipolla, Learning over Sets using Boosted Manifold Principal Angles (BoMPA). *BMVC*, 2005.
11. K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *Int'l Symp. of Robotics Research*, 2003.
12. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–372, 1936.
13. Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.
14. P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.
15. J.M. Geusebroek, G.J. Burghouts, and A.W.M. Smeulders. The Amsterdam library of object images. *IJCV*, 61(1):103–112, January, 2005.
16. Toshiba. Facepass. http://www.toshiba.co.jp/rdc/mmlab/tech/w31e.htm.
17. E.Oja, Subspace Methods of Pattern Recognition. Research Studies Press, 1983.