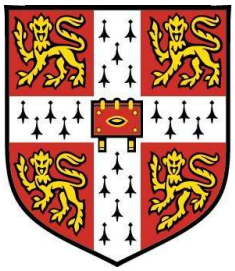
Discriminant Analysis of Patterns in Images, Image Ensembles, and Videos

TAE-KYUN KIM
Jesus College
MICHAELMAS TERM 2007



Department of Engineering
University of Cambridge

This dissertation is submitted for
the degree of *Doctor of Philosophy*

Abstract

This work addresses three visual classification tasks: face recognition from a single model image, object recognition by image sets (or ensembles) and action classification in videos. The work assumes that images and videos are given as 2D and 3D bounding boxes of patterns respectively, focusing on classification of isolated patterns. Whereas traditional classification problems have involved a single query image and a set of model images per class, the so called Single-to-Set matching task, the three tasks require different matching strategies: Single-to-Single, Set-to-Set, and Video-to-Video matching to each of the three tasks (in the afore-mentioned order) respectively. They are difficult to tackle in conventional ways due to extremely limited model data and lack of principles to exploit image sets or videos as inputs.

We propose novel methods of Discriminant Analysis (DA) for tackling the problems concerned. Discriminant Analysis (DA) is a well-established method of classification that approaches and often outperforms more complex modern methods. Owing to its simplicity and powerfulness as a statistical representation method, Discriminant Analysis (DA) could be best developed for the three problems.

To tackle the Single-to-Single matching task where both a query and a class model are single samples, we ought to exploit class priors for robust matching. Discriminant Analysis is performed for a set of prototype classes that have plenty of model samples and is applied to a set of classes of single model samples concerned. Note that the two sets do not involve same classes. The discriminative information learnt by DA from the prototype classes plays as priors on the classes to be recognized. Under this framework, two novel updates on Discriminant Analysis are proposed to 1) capture non-linearity of data and 2) to perform on-line learning for the Single-to-Single matching task.

We extend Discriminant Analysis to cope with image sets/videos as inputs for the Set-to-Set/Video-to-Video tasks where both a query and a model are image sets/videos respectively. Canonical Correlation Analysis (CCA), a standard tool for inspecting linear relations between two random variables, is set for measuring similarity between two sets of images. CCA yields a subspace-based matching which is effectively invariant to pattern variations on the subspaces. The CCA for multi-array data is also developed for similarity between two videos. Novel methods of Discriminant Analysis with the image-set and video similarity are then proposed for robust object recognition by image sets and action classification in videos respectively.

Finally, we integrate the concepts/methods of learning non-linearity and on-line learning developed for the Single-to-Single task to the methods for the Set-to-Set and Video-to-Video tasks. The Discriminant Analysis methods are conveniently integrated owing to their coherency. Moreover, the method proposed as a general meta-algorithm can be combined with other pre- or post-processing algorithms for further improvement, which

is exemplified.

Key words: Object Recognition, Action Classification, Face Recognition, Image Retrieval, Image Set, Ensemble, Video, Tensor, Statistical Feature Extraction, Discriminant Analysis, Bayesian Decision Rule, Bayes Classifier, Linear Discriminant Analysis, Kernel Classifier, Canonical Correlation Analysis, On-line Learning, Mixture of Experts.

This thesis is submitted to the Department of Engineering, University of Cambridge, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work and, except where otherwise stated, describes my own research.

Tae-Kyun Kim, Jesus College

Copyright ©2007
Tae-Kyun Kim
All Rights Reserved

Email: tkk22@cam.ac.uk
WWW: http://mi.eng.cam.ac.uk/∼tkk22

To my family

Acknowledgement

Firstly, I would like to express my heartfelt gratitude to my supervisors, Roberto Cipolla and Josef Kittler, for their immense advice, support and enthusiasm. I have been truly indebted to them for all their help during my PhD study. Working with them has been a great pleasure and this thesis has been a rewarding experience.

I also must thank to Zoubin Ghahramani and Jiri Matas for all useful comments and kind encouragement on my study. My former supervisors, Myung Jin Chung and In So Kweon, both of who have consistently given me their thoughtful consideration, are extremely appreciated.

Over the past years I have worked with a number of people. My study at Cambridge would have been far less fruitful and enjoyable without the fellows in the Machine Intelligence Laboratory. I am deeply grateful to Björn Stenger and Shu-Fai Wong for our close collaboration and friendship. I would like to thank Nanthan, Tom, Stefano, Ramanan, Fabio, Julia, Julien, Jamie, Neill, Gabriel, Matthew, George, Carlos, Ollie, Ognjen, Giovanni, Martin, Ben and Frank. They are the cleverest and friendliest people whom I have ever met. I also wish to offer my thanks to all my friends and colleagues, far and near, for their care and help. I have been especially fortunate for KwangSeok's coming and living nearby.

Financial support from the scholarships that I have had is gratefully acknowledged: Toshiba Research Europe, Chevening and Cambridge Oversea Trust. They have made my study at Cambridge possible. I also thank Jesus College and the Department of Engineering for supporting my participation in conferences and providing excellent work environment.

Last, but by no means least, my wholehearted thanks go to my family for their never-ending support, love and sacrifices. I would like to dedicate this work to my parents and wife. My son, who was born last year, has been indeed a great source of energy and happiness of my whole family.

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

1.1 Problems

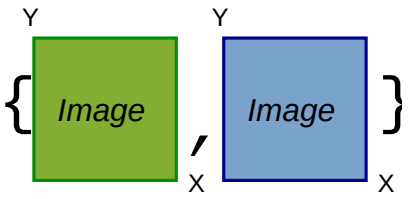
We aim to classify patterns for three visual recognition problems; face recognition from a single model image, object recognition by image sets (or ensembles) and action classification in videos. This work assumes that images and videos are given as 2D and 3D bounding boxes of patterns respectively, focusing on classification of isolated patterns in spatial or spatiotemporal domain. Pattern registration for the bounding boxes has been done simply by a state-of-the-art object detector or a known color model of simple background. While typical classification problems have involved a single query image and a set of model images of each class (so called Single-to-Set matching problem), the three tasks require different matching strategies: Single-to-Single, Set-to-Set, and Video-to-Video matching. Each of the three matching tasks is briefly explained as follows:

Single-to-Single:

The task of recognition with a single model image has received increasing attention because of important applications such as automatic passport control at airports, where a single photo in the passport is available as a model, and face image retrieval in Internet/or unknown database. In the retrieval task, a single arbitrary query image is supplied by users and every single image in the unknown database is matched with the single query, producing a *Single-to-Single* matching problem. The task has emerged as an active research area in Face Recognition Test (FRT) [154] and Moving Picture Experts Group (MPEG-7) Standardisation for face image retrieval [1, 134, 80, 100]. The example patterns used for evaluation of this study are shown in Figure 1.1.

Set-to-Set:

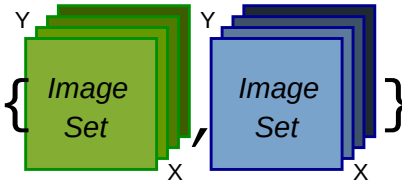
Rather than a single image input, more robust object recognition can be achieved classifying a set of images which represents a variation in object's appearance. Sets may be derived from sparse and unordered observations acquired by multiple still shots as well as from a video sequence. Thus, the objective of this task is to classify an unknown set of images to one of the training classes, each of which is also represented by image sets. The example sets of an isolated general object are shown in Figure 1.2.



(a) Single-to-Single matching strategy

(b) Face images captured from different views (XM2VTS).

Figure 1.1: Example Patterns For Single-to-Single Matching. Given a single model image per subject (for e.g. a frontal view), novel view images should be classified.



(a) Set-to-Set matching strategy

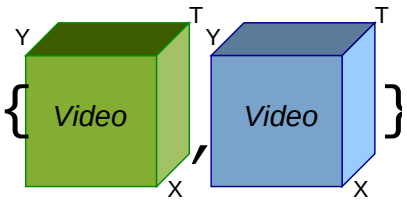
(b) A set of object images collected from a turntable sequence.

(c) A set of object images collected from a random-moving-camera sequence.

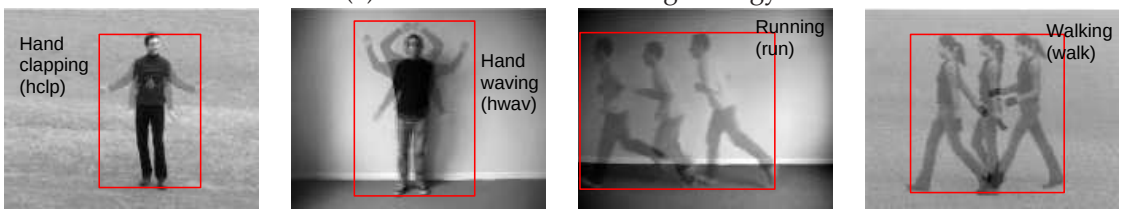
Figure 1.2: Examples Patterns For Set-to-Set Matching. The two sets containing different pattern variations resulting from different views and lighting.

Video-to-Video:

Over the last decades, human action/gesture classification has become an important topic in computer vision for a variety of tasks such as video surveillance, object-level video summarisation, video indexing, digital library organisation, etc [168, 143]. This task may be tackled by classifying spatiotemporal patterns in aligned videos. Action detection may first be performed to localise unit actions in input videos, followed by classification of the localised actions. Example video samples with indication of action alignments are shown in Figure 1.3.



(a) Video-to-Video matching strategy



(b) Example action classes (KTH). The bounding box and superimposition indicate the action alignment in a spatiotemporal space.

Figure 1.3: Examples Patterns For Video-to-Video Matching.

1.2 Challenges

Although numerous methods have been developed to tackle the above matching problems, given its importance for real world applications, the tasks are still challenging. The challenges of the problems are listed as follows:

- **Learning with Extremely Limited Training Samples:** Most state-of-the-art classification methods (Support Vector Machine [22], boosted classifiers [164, 193], conventional Discriminant Analysis (DA), probabilistic-Principal Component Analysis(PCA) method [140] etc.) fail to work when only a single training sample is available (i.e. for the Single-to-Single matching task). As there is no intra-class information available from the given classes, many studies have tackled the problem by unsupervised learning, e.g. the Principal Component Analysis (PCA)-based methods for greater robustness, by perturbing original samples [204, 224, 227], an image-as-matrix approach [215], or discarding some dimensions [195, 81, 82, 87, 204]. Part-based representations [136, 198] have been also developed. However, these unsupervised methods are inherently limited in delivering maximum discriminative information of classes. The methods of generating novel view images [191, 15] are highly dependent on dense correspondence of facial features, which is difficult to establish in practice.

- **Handling a Large Class Data Set:** The important applications of the Single-to-Single matching task such as passport control and face image retrieval handle a large number of classes which exhibit wide variation in object appearance. This raises a number of important issues both in learning and classification. With regard to learning, a method should learn complex non-linear manifolds of the data set. It is difficult to assume that all classes are simply linearly separable. On-line learning is also greatly needed to update a current algorithm over new training sets in a time-efficient manner. Practically, a complete set of training samples is not given in advance. For applications involving a large data set, computational complexity is particularly important at a classification stage. Existing kernel-based non-linear classification meth-

ods [188, 8] require high computational cost in classification.

- **Developing Image Set-to-Set Similarity:** Whereas many previous works on matching image sets for object recognition exploit temporal continuity between consecutive images [116, 232, 118, 125, 58], this study does not make any such assumption. *Set-to-Set* matching can be tackled in a straightforward way by assembling conventional Single-to-Set matching problems, but the combining rules are ad-hoc, leaving room for systematic exploration of set-property. The set-similarity achieved by comparing probability density functions [27, 167, 234] is often inappropriate for the classification task with image sets, as the image sets exhibit significantly different behavior from training to testing in distributions (See Figure 1.2 for the example of two sets of the same class). Computationally, the method using more than normal densities has to resort to numerical methods for computing the probabilistic distances. On the other hand, subspace-based set matching [38, 207, 199] seems effective in absorbing the large intra-class variation of patterns placed on the subspaces and is also computationally efficient. However, there are no previous studies on optimal classification of image sets either by density functions or subspaces.

- **Exploiting Full Spatiotemporal Information:** Many earlier methods for action/gesture recognition have mainly conveyed only partial data of the space-time information (mainly motion data) [168, 16, 202, 85, 14]. The part-based approaches [165, 32, 143], so called bag-of-words, are based purely on local appearance, ignoring global shape information. Despite recent progress in integrating global structural information to the local appearance [203, 163], their performance is highly dependent on empirical parameter settings of the space-time interest points and the code book. On the other hand, previous studies on tensors [189, 7, 210] may be useful in analysing videos as full spatiotemporal volumes (i.e. holistic methods). However, holistic methods are typically more sensitive to background changes and geometrical variations in human pose than part-based approaches.

- **Small Sample Size:** Action/gesture recognition may be seen as a spatiotemporal pattern classification problem. Learning with video inputs is generally more difficult than learning with images regarding Small Sample Size problems. Simple vectorization of a video achieved by concatenating all pixels in a three-dimensional volume causes a high dimension of N^3, which is much larger than N^2 of an image (where N is a dimension of each axis). Also, it is often difficult to collect a sufficient number of video samples for training. Most state-of-the-art classification methods (for e.g. Support Vector Machine [22], boosting classifiers [164, 193]) depend on a large number of and representative class samples in training.

1.3 Discriminant Analysis vs. Classifier

This study approaches the three visual pattern classification problems through a statistical learning method based on Discriminant Analysis (DA). Originally developed in 1936 by R.A. Fisher, Discriminant Analysis (DA) is a classic proven method of classification. It often produces models whose accuracy approaches and even exceeds more complex modern methods. For clarity, it may be worth comparing Discriminant Analysis with a classifier, another popular classification approach.

Traditionally, pattern classification studies are divided into two large parts: representation of inputs (or feature extraction) and classification. The former concerns learning low-dimensional representations of given high-dimensional inputs, s.t. $\mathcal{F} : \mathcal{X} \in \mathbb{R}^N \longrightarrow \mathcal{Y} \in \mathbb{R}^n, N > n$. Low-dimensional representation greatly helps subsequent class modelling and facilitates classification. The latter learning a classifier (or a discriminant function) which assigns a class to given inputs as $\mathcal{G} : \mathcal{X} \in \mathbb{R}^N \longrightarrow \mathcal{Y} \in \{1, 2\}$ in a typical binary case.

Discriminant Analysis learns optimal representation of input vectors for maximum class separation. With the learnt representation, a simple Nearest Neighboring classifier(NN) delivers good recognition accuracy. Discriminant Analysis has often been applied to a large number of classes with a few training images per class, e.g. face recognition tasks with hundreds of persons. On the other hand, classifiers directly learn a decision boundary rather than the representation of individual samples. Classifiers such as Support Vector Machines [22] and Boosting [164, 193], were originally designed for a binary class task. Typically, for given high-dimensional visual inputs, classifiers are heavily dependent on a large number of and representative training samples per class to learn a proper decision boundary. They have mostly been adopted for the case of binary or a small number of classes with plenty of training samples per class.

Discriminant Analysis is closely related to a discriminant function (or classifier). It is equivalent to optimal Bayesian discriminant function (or Bayes classifier) on the Gaussian equal-class covariance assumption [33]. This simple assumption greatly reduces a number of parameters to estimate for learning under small sample size and could be reasonable for e.g. a face recognition task, where the most appearance changes in facial images may be dominated by extrinsic factors such as lighting conditions and view-point changes rather than differences in faces themselves. Different face classes may therefore exhibit similar covariance structures. Once images are projected into a lower-dimensional space by the transformation function, class diversity in covariances can be still reflected in subsequent classification e.g. in a non-parametric way by Nearest Neighbor (NN) classifiers.

1.4 Contributions

We extend the classical Discriminant Analysis method to cope with a single model image and image sets/videos as inputs for the three matching tasks. Four major contributions and integration efforts are explained.

Conventional Discriminant Analysis (DA) can not directly handle the ***Single-to-Single*** matching task where each class has a single model sample and thus there is no intra-class variation available from a given model set. Assuming that human faces exhibit similar intra-class variation, we have proposed learning the transformation function of Discriminant Analysis from an independent training set which contains prototype face classes of multiple samples per class, and apply the leant function to compare any single query and the model set, as illustrated in Figure 1.4. The training and the model sets contain different face classes. The Discriminant Analysis method has well generalised the intra-class information learnt from the training set, delivering good retrieval accuracy in the model set. This method was adopted as a key function of the MPEG-7 (Moving Picture Expert Group) standard for face image retrieval [1, 134, 80, 100]. Motivated by our success, we have updated the method for the Single-to-Single matching task in two ways:

- **A Non-linear Extension of Discriminant Analysis [98]:** A single linear model is in-

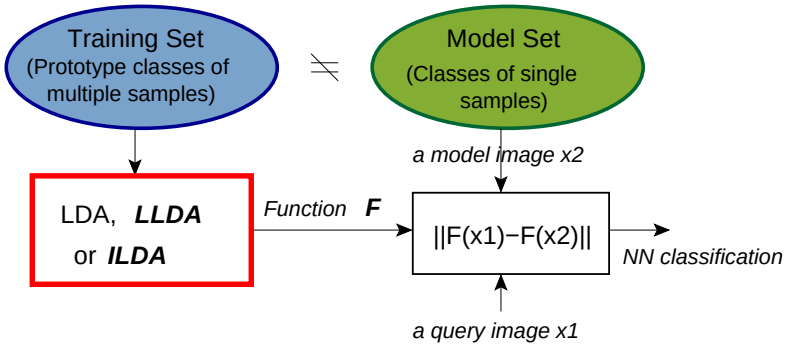


Figure 1.4: Block Diagram of the Proposed Methods for the Single-to-Single Matching Task. The method learns Discriminant Analysis, the non-linear DA (called Locally Linear Discriminant Analysis (LLDA)), or the on-line DA (called Incremental Linear Discriminant Analysis (ILDA)) from an independent training set and applies the learnt DA to compare a single query with any image in a model set.

sufficient to learning non-linear manifolds of a data set exhibiting large appearance changes. We propose a method of non-linear discriminant analysis for the recognition task with a single model image. A set of locally linear transformations, so-called ***Locally Linear Discriminant Analysis (LLDA)***, are learnt to capture non-linear manifolds of prototypes and are applied to novel faces. A novel key idea is that each local model is aligned with the others while being locally discriminative. This facilitates recognition of multi-modally distributed classes with single model samples. Compared with Kernel Discriminant Analysis [8], our method is highly time-efficient and avoids overfitting owing to its linear base structures. See Chapter 4 for details.

- **Incremental Learning of Discriminant Analysis [106]:** In practice, a complete set of prototypes or training samples is not given in advance. Execution of the batch-computation, whenever new data is presented, is too expensive in terms both of time and space. An efficient update algorithm of LDA is needed to accumulate the information conveyed by new data so that the method's future accuracy is enhanced. A new on-line solution yielding close agreement with the batch-mode, which is called ***Incremental Linear Discriminant Analysis (ILDA)***, is obtained by applying the concept of *sufficient spanning set* approximation in each update step of LDA. The proposed method is evaluated on a database merging scenario for face image retrieval, i.e. the Single-to-Single matching task. See Chapter 5.

For the ***Set-to-Set*** and ***Video-to-Video*** matching tasks, we propose novel Discriminant Analysis methods on top of Canonical Correlation Analysis [60, 5] (CCA) as a tool to measure pairwise set-similarity or video-similarity. CCA, since Hotelling (1936), has provided a standard tool for inspecting linear relations between two random variables or two sets of vectors. A key purpose in using CCA in this work is to obtain an efficient invariant subspace-based matching of two image sets (or videos) to pattern variations that resort to the subspaces (See Chapter 6 for details).

- **Discriminant Analysis of Image Set Classes [104]:** We propose novel discriminant analysis methods for optimal image-set classification based on the CCA-based image-set similarity. The proposed methods, called ***Discriminative Canonical Correlations (DCC)*** and ***Orthogonal Subspace Method (OSM)***, learn the transformation

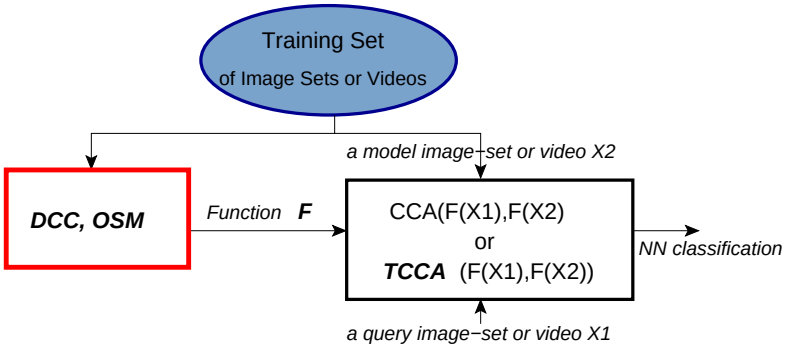


Figure 1.5: Block Diagram of the Proposed Methods for the Set-to-Set or Video-to-Video Matching Task. The proposed methods, called ***Discriminative Canonical Correlations (DCC)*** and ***Orthogonal Subspace Method (OSM)***, learn the transformation of input vectors so that the transformed image sets or videos are maximally separated according to the similarity yielded by CCA (for Set-to-Set) or ***Tensor CCA*** (for Video-to-Video).

of input vectors so that the transformed image sets are maximally class-wise separated according to the set-similarity yielded by Canonical Correlation Analysis. The proposed solution for object recognition with image sets is illustrated in Figure 1.5.

- **Generalisation of CCA into High-order Tensors [105]:** Conventional CCA is insufficient for action/gesture classification tasks as it does not encode temporal (ordering) information. CCA simply treats a video as a set of frames. We extend classical CCA into that of two high-order tensors (or multi-array data) for Video-to-Video matching. We call this ***Tensor Canonical Correlation Analysis (TCCA)***. The proposed extension of CCA for two videos is seen as the aggregation of many different sub-CCAs, one of which corresponds to the classical CCA of two image sets. Similarly, the discriminant analysis methods of image sets are applied with the new Video-to-Video similarity to robust action/gesture classification problems.

Finally, we explain integrations. The proposed methods are integrated within the Discriminant Analysis framework. The non-linear model and incremental learning method, which were proposed for Single-to-Single matching, are integrated into the methods for Set-to-Set and Video-to-Video matching [103]. The proposed method as a general meta-algorithm can, moreover, be combined with other pre- or post-processing for further improvement. As an example, the method is combined with the image representation based on the *Scale-Invariant-Feature-Transform (SIFT)*[1] rather than the raw-pixel representation of images in gesture recognition [107].

1.5 Structure of This Report

The following chapters are organised as follows:

- Chapter 2 reviews literature on the three recognition problems.

[1]SIFT is a histogram representation of gradient directions of local regions. It helps invariance to image scale, rotation and partially changing viewpoints and illumination [129].

- Chapter 3 provides background to Discriminant Analysis and Canonical Correlation Analysis and their bibliographic notes.
- Chapter 4 explains the non-linear extension of Discriminant Analysis for the Single-to-Single matching task. A set of locally linear models is proposed for view-invariant face recognition with a single model image.
- Chapter 5 presents a novel solution for incremental learning of Linear Discriminant Analysis which is useful for model reinforcement by new training data. The method is evaluated on a database merging scenario for face image retrieval, i.e. the Single-to-Single matching task.
- Chapter 6 addresses various object recognition problems with image sets (i.e. Set-to-Set matching). In the proposed method, Canonical Correlation Analysis (CCA) is exploited for image-set similarity and novel discriminant analysis methods are proposed to maximise image-set class separation in terms of CCA.
- Chapter 7 tackles the problem of Video-to-Video matching for human action/gesture classification by extending Canonical Correlation Analysis into multi-array data. The proposed method, so called Tensor Canonical Correlation Analysis (TCCA), is evaluated on a public data set for comparison with many state-of-the-art methods.
- Chapter 8 describes integration efforts. The Discriminant Analysis method developed for image sets and the SIFT-based representation are integrated to Tensor CCA for more robust gesture recognition.
- Chapter 9 explains the integration of the non-linear model and the incremental learning method which are proposed for Single-to-Single matching, into the methods for Set-to-Set and Video-to-Video matching.
- Chapter 10 concludes the dissertation with suggested directions for future research.

CHAPTER 2
Literature Review

In this chapter, we offer a comprehensive but not exhaustive literature review by categorising methods to address each of the three recognition problems.

2.1 Single-to-Single matching

'Human face' is a special topic in computer vision research because of its importance in many applications. In this subsection, literature on automatic recognition of faces with a single-per-class image is reviewed. The problem of so-called 'a single-per-class (or person) sample' has been intensively studied for important applications such as automatic passport control at airports (where a single photo is available in a passport as a model) and face image retrieval on Web or unknown face database. In the retrieval task, a single arbitrary query image is supplied by users and every single image in the unknown database is matched with the single query. Both tasks require robust matching between a single model and a single query image, i.e. *Single-to-Single* matching. This task is also an extreme case of the small sample size problem in general pattern classification studies [77, 156]. Most visual learning methods suffer from a paucity of training samples due to high-dimensional visual inputs. Extremely limited training samples, i.e. a single-per-class sample, makes this task very challenging. This problem has rapidly emerged as an active research sub-area of Face Recognition (FR) test protocol FERET [153] in recent years and MPEG-7 (Moving Picture Experts Group) Standardisation for face image retrieval [80, 99, 134].

First of all, most state-of-the-art classification methods fail to work, relying heavily on large size and representative of training sets. They include Support Vector Machine (SVM) [22], boosting classifiers [164], Linear Discriminant Analysis (LDA) [9, 226], and state-of-the-art face recognition methods, e.g. probabilistic eigenface [140], and Laplacianface methods [66]. Other feasible approaches are summarized as follows:

2.1.1 Unsupervised Learning Methods

The recognition task with a single-per-class image belongs by nature to an unsupervised learning problem, as there is no intra-class information available from given classes. Many

extensions of Principal Component Analysis (PCA), a representative method of unsupervised learning, have been developed for higher robustness [204, 224, 215]. Many decompose an original face space and discard some less stable dimensions for robust feature extraction [195, 81, 82, 87, 204]. For example, second-order PCA methods attempted to remove illumination effects by removing the first few eigenvectors [195, 81, 82, 87]. In another method, image matrices have been exploited as inputs rather than raster-scanned vectors for reliable estimation of a covariance matrix [215]. Unsupervised methods can be directly applied to this task, as they do not require class information. These are, however, inherently inferior to any discriminative approach which captures maximum discriminative information, addressed in Section 2.1.4.

2.1.2 Part-based Methods

Part-based representation has benefits for dimension and flexibility in learning under small sample size. Local region-based image descriptions return a lower-dimensional input space rather than holistic representations of images. Local methods are also flexible in weighting each region according to its importance. These properties of local methods facilitate robust learning and recognition under small sample size. Gabor wavelet-based algorithm on a deformable topology graph was developed [133, 198]. The locally-applied wavelet features are robust in illumination change, distortion and scaling. The correspondence solving required for the elastic grid is, however, difficult and computationally demanding in practice. The Hidden Markov Model (HMM) characterises face pattern as a dynamic random process [213, 89, 142, 141, 34]. The drawback of this method is that it resorts to a local minimum if sufficient training samples are not given. In [137], subspaces were locally computed. To deal with pose change under the one sample condition, a multi-subregion based probabilistic approach [84] was proposed similar to the Bayesian method [140].

2.1.3 Generating Virtual Samples

Another mainstream attempt to solve the one-sample problem is to enlarge an actual training set by e.g. perturbing original images [227, 136, 224] or by using several filters [227]. In [74], a Linear Discriminant Analysis (LDA)-based method was proposed by the spatial perturbation to handle the one-sample problem. However, the enlarged training set by perturbation contains highly correlated samples, which do not greatly help accuracy enhancement. Meanwhile, numerous approaches have tried to generate virtual face images by geometrical transformation such as rotation [191, 15] and by facial symmetry [212, 57]. They require dense correspondences of facial features for image normalization prior to transformation. Unfortunately, the correspondence-solving itself is difficult in practice and errors in correspondences seriously degrade recognition performance as shown in [15].

2.1.4 MPEG-7 Competition

In the protocol of MPEG-7 [134] and FERET [153], a prototype set, which is independent of both gallery and probe set, is additionally defined for training. The prototype set serves as a generic training set, consisting of multiple samples per class. The gallery and probe set do not include the classes of the prototype set. Motivated by the experimental protocol of

MPEG-7 and FERET, many studies have learnt class priors from the independent prototype set and applied them to the gallery and probe set for testing. Most methods [195, 81, 82, 87], however, have resorted to unsupervised learning partly due to difficulty in achieving good generalisation of discriminative information (as priors) across classes. During the MPEG-7 standard efforts, we have proposed Linear Discriminant Analysis (LDA) to learn the discriminative information from the prototype face classes and generalise it to new face classes, assuming that humans exhibit similar intra-class variation [99]; that is, the transformation matrix of LDA is learnt from the prototype set and applied to new face classes. A face image is further partitioned into several facial components to simplify image statistics for modelling and the components are then encoded by Linear Discriminant Analysis (LDA). The method has well delivered the discriminant information to novel classes achieving the best retrieval accuracy among proposals for the MPEG7 international standard. Competitive proposals include PCA-based methods (the second-order PCA methods [195] and the Fourier spectral PCA method [81]), part-based methods (The Pseudo 2D-HMM [34], the embedded HMM method [142, 141] and the eHMM with the second-order eigenvectors [89]) and Discriminant Analysis-based methods (the Generalised Discriminant Analysis [39] and component-based Linear Discriminant Analysis methods [99]). The method promoted to an international standard combines Fourier spectral space and our component-based LDA method [99, 83].

2.2 Set-to-Set Matching

Many computer vision tasks may be cast as image-set matching problems (or generally vector-set). In object recognition, for example, a set of vectors may represent a variation in an object's appearance – be it due to camera pose changes, non-rigid deformations or variation in illumination conditions. Image sets may be derived not only from video but also sparse and unordered observations acquired by multiple still shots. The objective of this task is to classify an unknown set of vectors to one of the training classes, each of which is also represented by vector sets, thus requiring *Set-to-Set* matching. Relevant approaches to the set-to-set matching can be broadly partitioned into a probability density-based or manifold (or subspace)-based method, and a straightforward assembly approach.

2.2.1 Probabilistic Density-based Methods

In probability density-based approaches, each set is represented by a parametric distribution function, typically a Gaussian. The closeness of the two distributions is then measured by e.g. the Kullback-Leibler Divergence (KLD) [27, 167]. However, the Gaussian assumption is often invalid when modelling complex nonlinear manifolds. Mixture models [4] or non-parametric densities have been used to reflect nonlinearity. A drawback of these methods lies in the requirement of numerical methods to compute distances of mixtures or non-parametric densities. Nonlinearity has been also tackled by kernel methods in [233]. Various probabilistic distance measures include Chernoff distance [25], Matusita distance [138], Symmetric KL divergence [27], Patrick-Fisher distance [150], Lissack-Fu distance [122], Kolmogorov distance [2] and resistor-average distance (RAD) [3]. Overall, it is difficult to estimate parameters of densities with limited samples (typically the number of images in a single set is small) and probability density-based methods easily fail when the

training and novel test sets do not show strong statistical relations.

2.2.2 Manifold (or Subspace)-based Methods

A manifold can be effectively captured by a mean and a set of basis vectors, which are obtained by subspace analysis. A joint manifold distance to cluster appearances was proposed in [38] and piece-wise linear identity manifold in [119] where a video sequence corresponds to a trajectory traced out in the identity surface. Relatively recently, Canonical Correlation Analysis (CCA) (or Principal Angles), which is an established method of inspecting linear relations between two random variables [71, 79, 51, 11], has received increasing attention for image set matching [207, 199, 45, 144, 110, 46]. Each image set is represented by a linear subspace and the angles between two low-dimensional subspaces are exploited as a similarity measure between two image sets (See Chapter 3 and Chapter 6 for more details). The benefits of using CCA over other methods for object recognition with image sets have been noted, e.g. efficiency, accuracy and robustness [100, 4]. Computation of the principal angles has been extended into a nonlinear feature space called reproducing kernel Hilbert space (RKHS) induced by a positive definite kernel function in [199, 60]. An additional potential benefit of the kernel extension is that, given such kernel function, it can be readily plugged into a classification scheme such as support vector machine (SVM) [199]. Another non-linear extension method of CCA has been proposed by multiple subspaces in our work [100].

To summarise, manifold-based matching methods are less constrained than probabilistic density-based methods, yielding invariance to data variations on manifolds. Invariance up to the manifolds greatly helps classification of image sets which exhibit large intra-class variations. Previous studies have, however, not generally addressed optimal set classification by exploiting given class information: An exception lies in the Constrained Mutual Subspace Method (CMSM) [45, 144], which is that most closely related to this study. In CMSM, a constrained subspace is defined as the subspace in which the entire class population exhibits small variance. The authors have showed that the face image sets were class-wise more separated in the constrained subspace in terms of canonical correlations (See Chapter 3 and Chapter 6 for more details).

2.2.3 Simple Assembly Algorithms

The *Set-to-Set* matching problem can be tackled by straightforward assembly methods. These combine results of many Single-to-Single or Single-to-Set matchings and include Nearest Neighbour (NN) and Hausdorff distance matching, both of which are based on matching of paired individual samples of two sets [162, 33]. The methods are based on the premise that similarity of a pair sets is reflected by the similarity of the nearest samples of the two respective sets. Thus, the overall recognition performance of the assembly algorithms mainly depends on that of base algorithm, i.e. Single-to-Single matching. Some approaches might use several representative samples of sets for e.g. by clustering techniques. As mentioned, the assembly algorithms use multiple observations in a straightforward fashion, neglecting set properties. The combining rules are somewhat ad-hoc. Note also that such methods are very time consuming as they require comparison of every pair of samples drawn from the two sets. Instead, any model-based method afore-mentioned greatly speeds up the Set-to-Set matching task.

2.3 Video-to-Video Matching

The topic of discriminating patterns in a spatiotemporal space is addressed for action and gesture recognition problems.

2.3.1 Explicit Motion Estimation

Many methods for action categorisation have been suggested. Traditional approaches are based on the comparison of motion data requiring explicit motion estimation [12, 35]. The performance of such algorithms is highly dependent on the quality of the motion estimation, which is a hard problem in practice due to smooth surfaces, singularities, self-occlusions, appearance changes and the aperture problem.

2.3.2 Analysis of Space-Time Volumes

Some recent work has analysed human action directly in the space-time volume without explicit motion estimation [168, 16, 202, 85]. Motion history images and the space-time local gradients are used to represent video data in [16, 202] and [168] respectively, having the benefits of being able to analyse quite complex and low-resolution dynamic scenes. Both representations, however, only partially convey space-time information (mainly the motion data) and are unreliable in cases of motion discontinuities and motion aliasing. Additionally, the method in [168] involves manual setting of important parameters such as positions and scales of the local space-time patches.

Importantly, it has been reported that spatial information contains cues as important as dynamic information for human action classification [14]. In this study, actions are represented as space-time shapes by the silhouette images and the Poisson equation. It assumes, however, that silhouettes are extracted from video. Furthermore, as noted in [14], the silhouette images are insufficient to represent complex spatial information.

2.3.3 Approach by Bag-of-Words

There is another important line of action recognition methods which are based on space-time interest points and visual code words [143, 165, 32, 113]. Originally, this technique was widely applied to image-based object categorisation tasks. Local variations around the interest points in videos are quantized by a code book. Histogram representations are then combined with either a Support Vector Machine (SVM) [165] or a probabilistic generative model [143]. Although they have yielded good accuracy mainly due to the high discrimination power of individual local space-time descriptors, they exhibit ambiguity by ignoring global space-time shape information of action classes. Approaches by so called 'bag-of-words' do not exploit any global structural information since purely based on the local appearance information. In spite of recent attempts to combine the structural with local information [203, 163], methods based on the bag-of-words often suffer from difficulties in setting of the parameters of the space-time interest points and the code book, whose best setting is application or data dependent.

2.3.4 Learning over Tensors

Traditional classification approaches may be applied to action recognition by simple vectorization or tensor representations of videos.

Many previous studies [189, 7, 210] have dealt with tensor data in its original form to consider multi-dimensional relations of data and to avoid the *curse of dimensionality* when the multi-dimensional data array are simply vectorized. Whereas various learning methods associated with tensors have been proposed, they do not take video data as input for classification. A video sequence can be treated as a general 3rd-order tensor. The ensembles of multilinear classifiers and discriminant analysis method have been developed for the tensors obtained from color images [7] and filter banks applied to a single gray image [210] respectively.

Recently, the Support Vector Machine with general tensor representation has been proposed in [201], where a small experiment was carried out with videos for action classification. Holistic representations are, however, typically sensitive to background changes and geometrical variations in human pose, requiring an efficient feature extraction method.

2.4 Limits of Scope

We limit the scope of our study as above. Other topics on video-based recognition are briefly discussed in the following.

2.4.1 Object Recognition in Video with Temporal Constraint

Studies on object recognition in videos propose simultaneous tracking and recognition [231, 232] to incorporate temporal coherence of images on the model of state transition probability and observational likelihood. Video-based face recognition under pose variation has been studied by several pose appearance manifolds [116]. Temporal continuity is directly captured by the transition probabilities between pose manifolds. The adaptive Hidden Markov Model (HMM) has been proposed to capture the dynamics [125]. However, these temporal information may not directly help object classification because an object (face) or a camera often moves arbitrarily. It is difficult to impose a strong constraint on temporal appearance changes. In this thesis, the topic of object recognition in videos will be regarded as the recognition problems by image sets without assuming temporal continuity. For pattern discrimination in the full spatiotemporal domain, human action and gesture recognition will be tackled.

2.4.2 Dynamic Texture Recognition

Dynamic texture recognition has been studied in recent decades [161]. It deals with videos which exhibit some patterns of temporal stationarity such as ocean waves, smoke and waterfalls. Images of stationary process are represented as the output of a stochastic dynamical model. It would be of interest to see how our proposed method works for this type of repetitive video patterns in future.

2.4.3 Shape from Motion

If a set of images or a video is available as input, 3D depth information can in theory be obtained and exploited for recognition. In spite of extensive studies on 3D model reconstruction, few approaches perform the recognition task directly using the 3D model from Shape from Motion. Most methods are dependent on some special 3D input equipment (See Face Recognition Grand Challenge (FRGC) [37] where the challenge include recognition tasks by 3D face models obtained by a laser scanner). This may be partly because current techniques are not accurate enough to produce 3D models for recognition and there remain many problems such as registration and matching even after obtaining 3D information. Due to the difficulty in explicit 3D representation and matching in real environments, we omit this topic from this thesis.

CHAPTER 3

Discriminant Analysis: Backgrounds and Updates

This chapter offers an introductory explanation of Discriminant Analysis based on its relations to Bayesian decision theory, which is a fundamental probabilistic approach to pattern classification problems. Canonical Correlation Analysis (CCA) as a key method for Set-to-Set and Video-to-Video matching tasks is also explained. Interestingly, CCA itself can be seen as a general form of Discriminant Analysis, as will be explained in Section 3.5. A bibliographical note on Discriminant Analysis and Canonical Correlation Analysis follows.

3.1 Bayesian Decision Theory

Bayesian decision theory is a fundamental pattern classification approach based on statistical and probabilistic theories. It defines an optimal rule by which to assign correct classes to given a set of vectors $\mathbf{x}$ by minimising the expected error caused by the decision rule.

Let $\mathbf{x}$ be a feature vector in a N-dimensional feature space $\mathbb{R}^N$. In classification tasks, each x belongs to one of finite classes denoted by w_i. The set of total c classes is $\{w_1, ..., w_c\}$. Possible a actions are denoted by $\{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_a\}$. In classification problems, $\boldsymbol{\alpha}_i$ is the decision that the true class is w_i, therefore $a = c$. $P(w_i)$ describes the prior probability of class w_i and $P(\mathbf{x}|w_i)$ the class-conditional probability density function of $\mathbf{x}$ for given a class w_i (also called *likelihood*). By Bayes formula, the posterior probability of a class w_i for given $\mathbf{x}$, $P(w_i|\mathbf{x})$, can be computed from the likelihood and priors [53, 19, 33, 131] by

$$P(w_i|\mathbf{x}) = \frac{P(\mathbf{x}|w_i)P(w_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|w_i)P(w_i)}{\sum_{j=1}^{c} P(\mathbf{x}|w_j)P(w_j)}. \tag{3.1}$$

The loss function denoted by $E(\boldsymbol{\alpha}_i|w_j)$ represents the loss incurred by taking the decision $\boldsymbol{\alpha}_i$ for a given true class w_j. The conditional loss of the decision $\boldsymbol{\alpha}_i$ for a given $\mathbf{x}$ is defined as

$$E(\boldsymbol{\alpha}_i|\mathbf{x}) = \sum_{j=1}^{c} E(\boldsymbol{\alpha}_i|w_j)P(w_j|\mathbf{x}), \tag{3.2}$$

where $P(w_j|\mathbf{x})$ is the posterior probability [158, 33]. Thus, the conditional loss is the sum of losses over all classes. The overall risk can then be defined as the sum of the conditional losses over a whole given set of vectors $\mathbf{x}$. For a more comprehensive representation, the decision rule is now given as a functional form of $\mathbf{x}$ by $\alpha(\mathbf{x})$. The function takes one of possible decisions in the set $\{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_a\}$. The overall risk E is $E = \int E(\alpha(\mathbf{x})|\mathbf{x})P(\mathbf{x})d\mathbf{x}$. By finding a decision rule which minimises the overall risk, the Bayes decision process yields the optimal performance. In classification problems, the loss function $E(\boldsymbol{\alpha}_i|w_j)$ can be more precisely specified. If we make a decision $\boldsymbol{\alpha}_i$ when the true class is w_j, then the decision is correct if $i = j$ and is erroneous if $i \neq j$. We assign no loss to a correct decision and a unit loss to any erroneous decision, giving the symmetrical or zero-one loss function by

$$E(\boldsymbol{\alpha}_i|w_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

$i, j = 1, ..., c$ [158, 47]. Thus the conditional loss of the decision $\boldsymbol{\alpha}_i$ for a given $\mathbf{x}$ in (3.2) is

$$E(\boldsymbol{\alpha}_i|\mathbf{x}) = \sum_{j \neq i} P(w_j|\mathbf{x}) = 1 - P(w_i|\mathbf{x}). \tag{3.3}$$

The optimal decision rule is to select i which minimises errors. This is equivalent to choosing i which maximises the posterior probability $P(w_i|\mathbf{x})$. The Bayes decision rule to minimise the risk [47, 158] is therefore to

$$\text{Assign} \quad w_i \quad \text{if} \quad P(w_i|\mathbf{x}) \geq P(w_j|\mathbf{x}) \qquad \text{for all} \quad j \neq i. \tag{3.4}$$

3.2 Bayes Classifiers and Discriminant Functions

Classifiers may be represented by a set of discriminant functions. This section gives optimal discriminant functions drawn by the Bayesian decision theory, which are either linear or non-linear depending on assumptions on given distributions. Let $g_i(\mathbf{x})$, $i = 1, ..., c$ denote a set of discriminant functions of a classifier. The corresponding classifier is interpreted as a decision rule to assign a feature vector $\mathbf{x}$ to class w_i if

$$g_i(x) \geq g_j(x) \qquad \text{for all} \;\; j \neq i, \tag{3.5}$$

[33, 47, 158, 19]. Thus the classifier consists of the c discriminant functions. Bayes decision theory in the previous section is directly associated with the Bayes optimal discriminant function as $g_i(x) = -E(\alpha_i|\mathbf{x})$ in general. For the minimum error case in classification problems, the discriminant functions is

$$g_i(\mathbf{x}) = P(w_i|\mathbf{x}) = \frac{P(\mathbf{x}|w_i)P(w_i)}{\sum_{j=1}^{c} P(\mathbf{x}|w_j)P(w_j)}. \tag{3.6}$$

A vector $\mathbf{x}$ is assigned to the class which has the maximum discriminant function or the maximum posterior probability. If we have a monotonic increasing function $f(\cdot)$, we can

replace the discriminant function $g_i(x)$ with $f(g_i(x))$ [19, 47]. Consequently there are many equivalent discriminant functions yielding the optimal classification result. Some of them are simpler in analytical and computational aspects. For the minimum-error-rate classification, the discriminant function $g_i(x)$ can further be simplified as

$$g_i(\mathbf{x}) = P(\mathbf{x}|w_i)P(w_i). \tag{3.7}$$

The discriminant function can also take a logarithm of the above equation by

$$g_i(\mathbf{x}) = \ln P(\mathbf{x}|w_i) + \ln P(w_i). \tag{3.8}$$

The discriminant functions are further derived for the special cases of normal distributions $p(\mathbf{x}|w_i)$. A normal distribution $P(\mathbf{x}|w_i) \sim N(\mathbf{m}_i, \Sigma_i)$ is given by

$$P(\mathbf{x}|w_i) = \frac{1}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \Sigma_i^{-1}(\mathbf{x} - \mathbf{m}_i)]. \tag{3.9}$$

The discriminant function in (3.8) can then be simply given as

$$g_i(x) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \Sigma_i^{-1}(\mathbf{x} - \mathbf{m}_i) - \frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i). \tag{3.10}$$

3.2.1 Case 1. $\Sigma_i = \Sigma$

If the class conditional densities $p(x|w_i)$ are Gaussian functions with equal covariance matrices such that $\Sigma_i = \Sigma$, the resulting discriminant function is linear. A classifier that uses linear discriminant functions is called *a linear machine* and such a classifier has many benefits in computational efficiency and good generalisation performance for novel samples in pattern classification problems. From a geometrical point of view, all the samples of the ith class are in hyperellipsoidal clusters of equal size and shape being centered at the mean vector $\mathbf{m}_i$ [33, 47, 158].

In the case of $\Sigma_i = \Sigma$, the terms $|\Sigma_i|$ and $(N/2)\ln 2\pi$ in (3.10) are common to all classes and can thus be eliminated. Also, the quadratic term $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ in the form of $(\mathbf{x} - \mathbf{m}_i)^T\Sigma^{-1}(\mathbf{x} - \mathbf{m}_i)$ is independent of classes, and thus removed to yield the linear discriminant function as

$$g_i(\mathbf{x}) = t_i^T\mathbf{x} + t_{i0}, \tag{3.11}$$

where

$$t_i = \Sigma^{-1}\mathbf{m}_i, \quad t_{i0} = -\frac{1}{2}\mathbf{m}_i^T\Sigma^{-1}\mathbf{m}_i + \ln P(w_i). \tag{3.12}$$

The decision boundary between two neighboring decision regions $\mathcal{D}_i$ and $\mathcal{D}_j$ [47, 158, 19] is represented as

$$t^T(\mathbf{x} - \mathbf{x}_0) = 0, \tag{3.13}$$

where

$$t = \Sigma^{-1}(\mathbf{m}_i - \mathbf{m}_j) \tag{3.14}$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\mathbf{m}_i + \mathbf{m}_j) - \frac{\ln[P(w_i)/P(w_j)]}{(\mathbf{m}_i - \mathbf{m}_j)^T \Sigma^{-1} (\mathbf{m}_i - \mathbf{m}_j)}(\mathbf{m}_i - \mathbf{m}_j). \tag{3.15}$$

The vector $\mathbf{x}_0$ is generally called a threshold of a classifier.

3.2.2 Case 2. Σ_i = arbitrary

The Bays optimal discriminant function for the general multivariate normal distributions $P(x|w_i)$ is clearly quadratic. When Σ_i = arbitrary, the term $(N/2)\ln 2\pi$ in (3.10) can be eliminated, thus rendering the discriminant function [33, 47, 19] as

$$g_i(\mathbf{x}) = \mathbf{x}^T T_i \mathbf{x} + t_i^T \mathbf{x} + t_{i0}, \tag{3.16}$$

where

$$T_i = -\frac{1}{2}\Sigma_i^{-1}, \quad t_i = \Sigma_i^{-1}\mathbf{m}_i \tag{3.17}$$

and

$$t_{i0} = -\frac{1}{2}\mathbf{m}_i^T \Sigma_i^{-1}\mathbf{m}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i). \tag{3.18}$$

In this case, the discriminant function is non-linear and the decision regions are not simply connected, resulting in complex decision boundaries even for a small number of classes [131, 33, 47, 158].

3.3 Discriminant Analysis

Originally developed in 1936 by R.A. Fisher, Discriminant Analysis is a proven method of classification. Discriminant Analysis often produces models whose accuracy approaches and occasionally exceeds that of more complex modern methods. Discriminant analysis has proved a powerful method for dimensionality reduction and classification that projects high-dimensional data onto a low-dimensional space where the data achieves maximum class separation [33, 47]. After low-dimensional representation of the data, it may be combined with any parametric or nonparametric methods (for e.g. Nearest Neighboring classifier (NN)) for classification.

3.3.1 Two Class Case

Assume that a set of M N-dimensional samples $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$ is given. The set is partitioned into two subsets of vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, each of which corresponds to a class label w_1 or w_2 respectively. Each subset $\mathbf{X}_i$ consists of M_i sample vectors. A linear combination of the components of $\mathbf{x}$ [47, 158, 53] is defined with $\mathbf{t}$ such that $\mathbf{y} = \mathbf{t}^T\mathbf{x}$, where $\mathbf{y}$ is an output vector. The set of M output vectors $\mathbf{y}_1, ...\mathbf{y}_M$ is divided into the two subsets $\mathbf{Y}_1$ and $\mathbf{Y}_2$ which correspond to $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. The magnitude of the linear combination $\mathbf{t}$ is not useful for classification as it just scales the output vectors $\mathbf{y}$. Thus $\mathbf{t}$ is normalized s.t. $||\mathbf{t}|| = 1$. To find the optimal direction of t which yields accurate classification of the vectors, the separation measures of the transformed (or projected) vectors are defined. A

measure of the separation between the projected points of the classes, w_1, w_2, is defined as the difference of the class means by

$$(\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2)^2 = \mathbf{t}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{t} = \mathbf{t}^T\mathbf{B}\mathbf{t}, \tag{3.19}$$

where $\mathbf{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ is the between-class scatter matrix and $\widetilde{\mathbf{m}}_i, \mathbf{m}_i$ are the i-th class mean of the projected data and the input data vectors respectively. The difference in class means should be maximised relatively to a certain measure of the standard deviations of each class [158, 53]. The sum of scatters of the projected samples within each class [33, 47] is given as

$$\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2 = \sum_{i=1}^{2} \sum_{\mathbf{x} \in \mathbf{X}_i} \mathbf{t}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\mathbf{t} = \mathbf{t}^T\mathbf{W}\mathbf{t},$$

where the within-class scatter matrix $\mathbf{W}$ is defined as the sum of the scatter matrices by $\mathbf{W} = \sum_{i=1}^{2} \sum_{\mathbf{x} \in \mathbf{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$. The criterion function of the Fisher linear discriminant is given by

$$J(\mathbf{t}) = \frac{\mathbf{t}^T\mathbf{B}\mathbf{t}}{\mathbf{t}^T\mathbf{W}\mathbf{t}}. \tag{3.21}$$

The Fisher linear discriminant finds the direction of $\mathbf{t}$ which maximises $J(\mathbf{t})$ under the condition that $||\mathbf{t}|| = 1$ [53, 9, 226, 158, 47]. Solution to the simple two-class case problem can be directly obtained since the between-class scatter matrix has just one rank and thus $\mathbf{B}\mathbf{t}$ is always in the direction of $\mathbf{m}_1 - \mathbf{m}_2$ [33]. As the direction of the solution $\mathbf{t}$ alone is significant, the solution of the function $\mathbf{t}$ is

$$\mathbf{t} = \mathbf{W}^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{3.22}$$

Its Equivalence to Bayes Optimal Discriminant Function.

It is noteworthy that the Fisher's linear discriminant function is identical with the Bays's optimal discriminant function given in (3.14) when the classes have equal covariance matrices such that $\Sigma_i = \Sigma$. The threshold of the Fisher's linear classifier can be set at that of the Bayes classifier as given in (3.14).

3.3.2 Multiple Class Case

As derived so far, Fisher's linear discriminant function was originally developed for the two-class problem. For the c-class problem, the generalisation of Fisher's linear discriminant involves $c - 1$ discriminant functions. Thus the projection is performed from a N-dimensional space to a $(c-1)$-dimensional space, where $N \geq c$.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$ be a data set of given N-dimensional vectors. Each data point belongs to one of C classes $\{\mathbf{X}_1, ..., \mathbf{X}_c, ..., \mathbf{X}_C\}$. The between-class scatter matrix and the within-class scatter matrix are defined as

$$\mathbf{B} = \sum_{c=1}^{C} M_c(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T, \quad \mathbf{W} = \sum_{c=1}^{C} \sum_{\mathbf{x} \in \mathbf{X}_c} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^T,$$

where $\mathbf{m}_c$ denotes the class mean and $\mathbf{m}$ is the global mean of the entire sample. The number of vectors in class $\mathbf{X}_c$ is denoted by M_c. LDA finds a matrix, $\mathbf{U}$, maximising the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix as

$$\mathbf{U}_{opt} = \max_{arg\mathbf{U}} \frac{|\mathbf{U}^T\mathbf{B}\mathbf{U}|}{|\mathbf{U}^T\mathbf{W}\mathbf{U}|} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N].$$

This formula is the well-known generalised Rayleigh quotient in Mathematical Physics [172, 53, 47]. Clearly, the solution $\{\mathbf{u}_i | i = 1, 2, ..., N\}$ is a set of generalised eigenvectors of $\mathbf{B}$ and $\mathbf{W}$ i.e., $\mathbf{B}\mathbf{u}_i = \lambda_i \mathbf{W}\mathbf{u}_i$. If the within-class scatter matrix $\mathbf{W}$ is non-singular, the generalised eigenvalue problem can be

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{u}_i = \lambda \mathbf{u}_i. \tag{3.23}$$

The solution $\mathbf{u}_i$ is obtained by solving a conventional eigenvalue problem [172]. Usually PCA is performed first to avoid a singularity of the within-class scatter matrix, which is often encountered in recognition problems of high-dimensional inputs [9, 226].

3.4 Bibliographic Notes on Discriminant Analysis

There is indeed extensive literature on Discriminant Analysis (DA). This section categorises existing works and briefly explains some important studies.

3.4.1 Kernel or Generalised Discriminant Analysis

Numerous studies have been carried out on non-linear discriminant analysis by kernel methods, so-called Kernel Discriminant Analysis (KDA) or Generalised Discriminant Analysis (GDA). The underlying theory is close to that of support vector machines (SVM) insofar as this method provides a mapping of the input vectors into high-dimensional feature space. In the transformed space, linear properties make it easy to extend and generalise the classical Linear Discriminant Analysis (LDA) to non-linear discriminant analysis. Some representative works are found in [8, 139]. Prior to these works, a nonparametric version of discriminant analysis was proposed to yield richer nonlinear classification schemes by using a mapping function (similar to kernel methods) in [62] and a compromise between Linear and Quadratic Discriminant Analysis, so-called Regularized Discriminant Analysis (RDA) in [43]. The RDA method was recently updated by the optimal regularization parameter estimation [220].

The Generalised Discriminant Analysis (GDA) [8] is explained in detail: The method is designed for non-linear classification based on a kernel function Φ which transforms the original space $\mathbf{X}$ to a new high dimensional feature space $\mathbf{Z}$ s.t. $\Phi : \mathbf{X} \rightarrow \mathbf{Z}$. The within-class (or total) scatter and between-class scatter matrix of the non-linearly mapped data is

$$\mathbf{B}^{\Phi} = \sum_{c=1}^{C} M_c \mathbf{m}_c^{\Phi} (\mathbf{m}_c^{\Phi})^T, \quad \mathbf{W}^{\Phi} = \sum_{c=1}^{C} \sum_{\mathbf{x} \in \mathbf{X}_c} \Phi(\mathbf{x}) \Phi(\mathbf{x})^T,$$

where $\mathbf{m}_c^{\Phi}$ is the mean of class $\mathbf{X}_c$ in $\mathbf{Z}$ and M_c is the number of samples belonging to $\mathbf{X}_c$.

The aim of the GDA is to find such projection matrix $\mathbf{U}^{\Phi}$ that maximises the ratio

$$\mathbf{U}^{\Phi}_{opt} = \underset{arg\mathbf{U}^{\Phi}}{\max} \frac{|(\mathbf{U}^{\Phi})^T\mathbf{B}^{\Phi}\mathbf{U}^{\Phi}|}{|(\mathbf{U}^{\Phi})^T\mathbf{W}^{\Phi}\mathbf{U}^{\Phi}|} = [\mathbf{u}^{\Phi}_1, ..., \mathbf{u}^{\Phi}_N].$$

The vectors, $\mathbf{u}^{\Phi}$ can be found as the solution of the generalised eigenvalue problem i.e. $\mathbf{B}^{\Phi}\mathbf{u}^{\Phi}_i = \lambda_i\mathbf{W}^{\Phi}\mathbf{u}^{\Phi}_i$. The training vectors are supposed to be centered (zero mean, unit variance) in the feature space $\mathbf{Z}$. From the theory of reproducing kernels, any solution $\mathbf{u}^{\Phi} \in \mathbf{Z}$ must lie in the span of all training samples in $\mathbf{Z}$, i.e.

$$\mathbf{u}^{\Phi} = \sum_{c=1}^{C}\sum_{i=1}^{M_c} \alpha_{ci}\Phi(\mathbf{x}_{ci}),$$

where α_{ci} are real weights and $\mathbf{x}_{ci}$ is the i-th sample of class c. The solution is obtained by solving

$$\lambda = \frac{\boldsymbol{\alpha}^T\mathbf{KDK}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{KK}\boldsymbol{\alpha}},$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_c), c = 1, ..., C$ is a vector of weights with $\boldsymbol{\alpha}_c = (\alpha_{ci}), i = 1, , M_c$. The kernel matrix $K(M \times M)$ is composed of the dot products of non-linearly mapped data, i.e.

$$\mathbf{K} = (\mathbf{K}_{kl})_{k=1,...,C,\ l=1,...,C},$$

where $\mathbf{K}_{kl} = (k(\mathbf{x}_{ki}, \mathbf{x}_{lj}))_{i=1,...,M_k,\ j=1,...,M_l}$. The matrix $D(M \times M)$ is a block diagonal matrix such that

$$\mathbf{D} = (\mathbf{D}_c)_{c=1,...,C},$$

where c-th matrix $\mathbf{D}_c$ on the diagonal has all elements equal to $1/M_c$. Solving the eigenvalue problem yields the coefficient vectors $\boldsymbol{\alpha}$ that define the projection vectors $\mathbf{u}^{\Phi} \in \mathbf{Z}$. A projection of a testing vector $\mathbf{x}_{test}$ is computed as

$$(\mathbf{u}^{\Phi})^T\Phi(\mathbf{X}_{test}) = \sum_{c=1}^{C}\sum_{i=1}^{M_c} \boldsymbol{\alpha}_{ci}k(\mathbf{x}_{ci}, \mathbf{x}_{test}).$$

3.4.2 Multiple Local Analysers

This category of study aims at solving nonlinear classification problems using a set of local discriminant analysers. It may be considered a locally linear yet globally nonlinear discriminant analyser; a special Kernel Discriminant Analysis (KDA) with a geometry-adaptive-kernel, in contrast to traditional KDA whose kernel is independent of samples. Its computation and memory cost are reduced a great deal compared with traditional KDA, owing to its linear base structure, which is important for cases with a large number of samples. The linear property also helps to soften overfitting of KDA whose hyper-parameters are hard to set. Since the work on Discriminant Analysis by Gaussian mixtures [64], many studies have been conducted especially in the past few years. Some recent works have been motivated by the concept of preserving local structures in a supervised

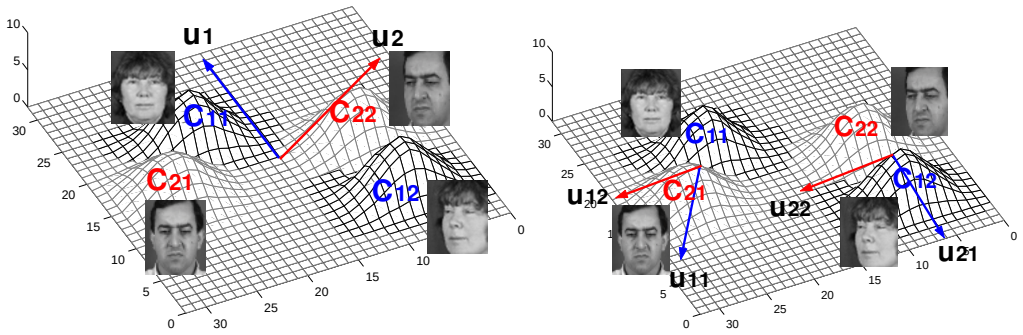


Figure 3.1: Comparison between LDA and LDA mixture for non-linear classification problem. Classical LDA (left) cannot dichotomise the classes exhibiting mixtures of Gaussian, whereas the LDA mixture model (right) solves the problem locally.

manner [173, 23, 128, 209]. Multi-modal class distributions of varying covariance are considered in [183] and the criterion for the best clustering of each class into a set of subclasses is derived in [236]. Figure 3.1 illustrates a comparison between classical LDA and the Discriminant Analysis using Gaussian mixtures for the non-linear classification problem where each class exhibits multi-modal distributions.

3.4.3 Discriminant Analysis for Small Sample Size

One of main difficulties in applying classical LDA is that it needs to take an inverse of a (within-class) scatter matrix which is singular under small sample size. Accordingly, lots of variations of LDA have been developed to address this problem. One popular approach is to perform Principal Component Analysis (PCA) prior to LDA [9, 226]. By reducing an input dimension by PCA, the singular problem of LDA could be corrected. This PCA/LDA approach has shown great success in face recognition tasks. Its theoretical justification has also been studied [214]. Another two-stage approach exploits QR-decomposition instead of PCA [218], so-called LDA/QR which achieves efficiency while overcoming the singularity problem of classical LDA. For a similar purpose, the generalised Singular Value Decomposition (SVD) is adopted to solve the generalised DA criterion in [72, 217]. More traditionally, methods exploiting null-space of the with-class scatter, which are often called Direct LDA or Null-space LDA, have been widely adopted for small sample size [24, 223].

3.4.4 Tensor Discriminant Analysis

A single image is originally given as a matrix rather than a concatenated vector. Rather than vectoring an image in classical LDA, methods of discriminant analysis for handling matrices as input have been developed. They consider multi-dimensional relations of data by representing the data in its original form, i.e. matrices, and help to avoid the small-sample size problem. When the multi-dimensional data array is simply vectorised in classical LDA, the input dimension is huge. The methods of two-dimensional LDA are proposed in [216, 109]. Beyond 2D LDA, a more general framework of discriminant analysis for high-order tensors (cf. a matrix is a second-order tensor) has been developed in for e.g.

[210] where an image is encoded as a general tensor by applying Gabor wavelets to the image, and in [7] where a single color image is a third-order tensor.

The 2D LDA [216] is briefly described: Let $A_i \in \mathbb{R}^{r\times c}$, for $i = 1, ..., n$, be the n images in the data set, clustered into classes $\Pi_1, ..., \Pi_k$, where Π_i has n_i images. Let $M_i = 1/n_i \sum_{X\in\Pi_i} X$ be the mean of the $i-$th class, $1 \leq i \leq k$, and $M = 1/n \sum_{i=1}^{k} \sum_{X\in\Pi_i} X$ be the global mean. In 2D LDA, images are regarded as two-dimensional signals. It aims to find two transformation matrices $L \in \mathbb{R}^{r\times l_1}$ and $R \in \mathbb{R}^{c\times l_2}$ that map each $A_i \in \mathbb{R}^{r\times c}$, for $1 \leq i \leq n$, to a matrix $B_i \in \mathbb{R}^{l_1\times l_2}$ such that $B_i = L^T A_i R$. The within-class and between-class distances D_w and D_b can be computed as follows:

$$D_w = tr(\sum_{i=1}^{k} \sum_{X\in\Pi_i} (X - M_i)(X - M_i)^T), D_b = tr(\sum_{i=1}^{k} n_i(M_i - M)(M_i - M)^T).$$

In the transformed space by the transformations L, R, the within-class and the between-class distances become

$$\widetilde{D}_w = tr(\sum_{i=1}^{k} \sum_{X\in\Pi_i} L^T(X - M_i)RR^T(X - M_i)^T L),$$

$$\widetilde{D}_b = tr(\sum_{i=1}^{k} n_i L^T(M_i - M)RR^T(M_i - M)^T L).$$

The optimal transformations L and R are found to maximise $\widetilde{D}_b$ and minimise $\widetilde{D}_w$. An iterative algorithm which solves either L or R fixing the other at each iteration is derived.

3.4.5 Robust Fisher Discriminant Analysis

In computer vision applications, outliers are often incurred within a sample (image) due to corruption pixels by noise, alignment errors or occlusion. Similarly to the robust Principal Component Analysis method [182], which makes PCA robust to the outliers, the method of discriminant analysis has been developed for the pixel outliers. In [90], sensitivity of classical Fisher Linear Discriminant (FLD) is alleviated by explicitly incorporating a model of data uncertainty and optimizing for the worst-case scenario.

3.4.6 Nonparametric Discriminant Analysis

In Discriminant Analysis studies, the terminology of *nonparametric* has meant different things. Generally Nonparametric Discriminant Analysis defines another version of scatter matrices whose parameters (mean and covariance in the classical LDA) are estimated from the data, not assuming the Gaussian class distributions. In pattern classification, a nonparametric method has not received great attention. Nearest neighbour classification (NN) instead, has a well-established position among other classification techniques due to its practical and theoretical properties. A discriminant analysis framework in the sense that a linear transformation is sought to improve NN performance was developed in [20].

In [40], the Support Vector Machine (SVM)-based Nonparametric Discriminant Analysis, whose scatter matrix is associated with the decision boundary of SVM, was proposed. The semi-parametric discriminant analysis in the sense that class distributions are multivariate normal after unspecified univariate monotone transformations, was proposed in [121].

The nonparametric discriminant analysis method developed in the Nearest Neighbour sense [20] is further explained: Assume that a data matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\} \in \mathbb{R}^{N \times M}$ is given, where $\mathbf{x}_i \in \mathbb{R}^N$ is a N-dimensional column vector obtained by raster-scanning an image. Each vector belongs to one of the classes denoted by C_i. Linear discriminant analysis (LDA) finds a transformation $\mathbf{T} \in \mathbb{R}^{N \times n}$ $(n \leq N)$ which maps a vector $\mathbf{x}$ to $\widetilde{\mathbf{x}} = \mathbf{T}^T\mathbf{x} \in \mathbb{R}^n$ such that the transformed data have maximum separation between classes and minimum separation within classes with respect to the defined between-class and within-class scatter measures. A nonparametric form of these scatter matrices is proposed in [20] with the definition of the between-class and within-class neighbours of a sample $\mathbf{x}_i \in C_c$ given by

$$\mathbf{B} = \frac{1}{M}\sum_{i=1}^{M} w_i(\mathbf{\Delta}_i^B)(\mathbf{\Delta}_i^B)^T, \quad \mathbf{W} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{\Delta}_i^W)(\mathbf{\Delta}_i^W)^T \tag{3.24}$$

where $\mathbf{\Delta}_i^B = \mathbf{x}_i - \mathbf{x}_i^B$, $\mathbf{\Delta}_i^W = \mathbf{x}_i - \mathbf{x}_i^W$, $\mathbf{x}^B = \{\mathbf{x}' \in \overline{C}_c \,|\, \|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{z} - \mathbf{x}\|, \forall \mathbf{z} \in \overline{C}_c\}$ and $\mathbf{x}^W = \{\mathbf{x}' \in C_c \,|\, \|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{z} - \mathbf{x}\|, \forall \mathbf{z} \in C_c\}$. The sample weight w_i is to "deemphasise" samples away from class boundaries. Nonparametric Discriminant Analysis (NDA) finds the optimal $\mathbf{T}$ which maximises $trace(\tilde{\mathbf{B}})$ and minimises $trace(\tilde{\mathbf{W}})$, where $\tilde{\mathbf{B}}, \tilde{\mathbf{W}}$ are the scatter matrices of the transformed data. As these are explicitly represented with $\mathbf{T}$ by $\tilde{\mathbf{B}} = \mathbf{T}^T\mathbf{B}\mathbf{T}$, $\tilde{\mathbf{W}} = \mathbf{T}^T\mathbf{W}\mathbf{T}$, the solution $\mathbf{T}$ can easily be obtained by solving the generalised eigen-problem, $\mathbf{B}\mathbf{T} = \mathbf{W}\mathbf{T}\mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is the eigenvalue matrix.

3.4.7 Incremental LDA

Visual learning is supposed to be a continuous process, motivating a recent research trend in online learning approaches. Classical machine learning performed in an isolation is limited [31]. Inspiration for incremental LDA can also be drawn from numerous works on incremental Principal Component Analysis (PCA) [59, 169]. However, extension of it to LDA is not straightforward due to the difficulty of updating a discriminative model rather than a generative model like PCA.

A number of incremental versions of LDA have been proposed [68, 120, 149, 219]. An incremental version of LDA was proposed by assuming that the number of classes C is negligible compared to the number of images in [219]. In [149] updating of the between-class and within-class scatter matrices is mainly concerned, without consideration of the subsequent LDA steps. Modified LDA criteria have been exploited for incremental learning [68, 208]. The work [120] deals with online updating of discriminative models for the purpose of object tracking. Note that their use is just for binary classification. Efficient algorithms for Kernel PCA and Kernel Discriminant Analysis have also been developed [26, 178].

3.4.8 Probabilistic Linear Discriminant Analysis

LDA is often used for feature extraction in object recognition, but does not address the problem of how to use these features for recognition. In Probabilistic PCA [179], it is demonstrated that how the principal components of data can be determined by maximum-likelihood estimation of parameters in a latent variable model. An EM algorithm has been adopted for estimating the principal components iteratively. Similarly, the probabilistic LDA method has been developed [76].

3.4.9 Heteroscedastic LDA (HLDA)

Different class covariances are considered in the method of discriminant analysis called Heteroscedastic LDA. On the other hand, classical LDA is often called homoscedastic because classes are assumed to have a common covariance structure. The method for different class covariances has often been applied in a pairwise fashion, i.e. class separation is measured as the sum of all pairwise class distances each of which reflects different class covariances [155, 127]. Weighting to reduce the role of the least stable sample classes was proposed in [155, 126, 63].

3.4.10 Other Issues

It has been reported that the performance of LDA depends on feature selection. Automatic determination of the optimal number of features for quadratic discriminant analysis has been proposed via the normal approximation to the discriminant distribution [73]. Selecting the principal components in a two-stage PCA/LDA algorithm has also been studied [235].

Independence or orthogonality of the discriminant components has been discussed for e.g. in [206] and the Least Square solution for LDA proposed in [221]. Fast LDA has been motivated by the study of binary PCA [177] using binary bases [176], which are linear combinations of Haar-like box functions. This facilitates computational efficient subspace projection. In a recent work [194], Discriminant Analysis with maximum margin criterion has been proposed for reducing structural risk and promoting better generalisation for novel data similarity with Support Vector Machine (SVM). The weighted maximum margin discriminant analysis method has been shown to be better than SVM in [229].

3.5 Relations Between Discriminant Analysis(DA), Canonical Correlation Analysis(CCA) and Regression

Canonical Correlation Analysis (CCA) (since Hotelling (1936) [71]) for the analysis of relations between two sets of variables, embodies Discriminant Analysis and multiple regression as its special cases. CCA is equivalent to Discriminant Analysis or multiple regression under certain conditions. Canonical Correlation Analysis involves two sets of variables: it supposes that there exist n observations of each of p variables x_i $(i = 1, ..., p)$ and of q variables y_i $(i = 1, ..., q)$. On the other hand, Discriminant Analysis includes n

observations of each of p variables x_i $(i = 1, ..., p)$, which are tagged with their respective categories or classes. Thus the Discriminant Analysis associates a set of *independent variables* with a categorical *target* or *dependent variable*. Multiple linear regression described by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \tag{3.25}$$

where β_is are the respective parameters of independent variables and ε is a random term, is concerned with the estimation of one dependent variable by a linear combination of independent variables. This is a special case of Canonical Correlation Analysis with $q = 1$.

Canonical Correlation Analysis vs. Multiple Regression: As described above, whereas multiple regression is used for many-to-one relations, canonical correlation analysis is used for many-to-many. It has been reported that the two things are mathematically equivalent under the condition that $q = 1$.

Discriminant Analysis vs. Linear Regression: Discriminant Analysis can be used only for classification (i.e., with a categorical target variable), not for regression. The target variable in the Discriminant Analysis is a discrete variable of two or more categories, whereas it is a continuous variable in Linear Regression in general.

Discriminant Analysis vs. Logistic Regression: Both Logistic Regression and Discriminant Analysis allow one to predict a discrete outcome such as group membership from a set of independent variables. The logistic regression may be better suited to cases where the dependant variable is dichotomous, while the independent variables may be nominal, ordinal, ratio or interval. Discriminant analysis might be more appropriate when the dependant variable has more than two groups/categories. Logistic regression is much more relaxed and flexible in its assumptions than discriminant analysis. Unlike discriminant analysis, logistic regression does not have the requirements of the independent variables to be normally distributed, linearly related, nor of equal variance within each group [175]. Freedom of being from the assumption of the discriminant analysis posits the logistic regression as a tool to be used in many situations. However, when the assumptions regarding the distribution of predictors are met, discriminant function analysis may be a more powerful and efficient analytic strategy [175].

3.6 Canonical Correlation Analysis (CCA)

As seen above, Canonical Correlation Analysis (CCA) is a general method for inspecting linear relations between two sets of variables, which will be exploited as a useful tool to measure set-similarity or video-similarity in the following chapters. This section gives an overview of CCA and its brief bibliographic notes. CCA has been adopted in various fields, from Economics, Medicine, Meteorology and recently computer vision and pattern recognition studies.

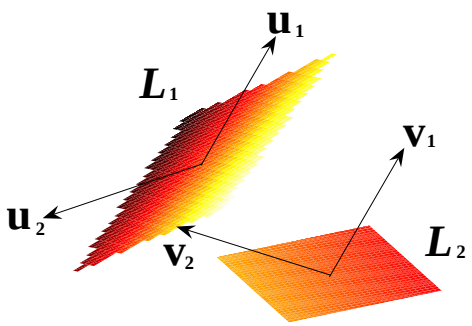


Figure 3.2: Representation of Canonical Correlation Analysis (CCA). CCA measures principal angles between two linear subspaces. Two sets of samples are represented as linear subspaces which are here planes (denoted by $\mathcal{L}_1$ and $\mathcal{L}_2$). Canonical vectors $\mathbf{u}_1, \mathbf{v}_1$ on the planes are found to yield maximum correlations. The second canonical vectors $\mathbf{u}_2, \mathbf{v}_2$ are determined to be perpendicular to the first.

3.6.1 Standard Formulation of CCA

A standard interpretation and solution of CCA is given as a method of correlating linear relationships between two multidimensional variables. It finds basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximised [71, 60]. Given two random vectors $\mathbf{x} \in \mathbb{R}^{m_1}, \mathbf{y} \in \mathbb{R}^{m_2}$, a pair of transformations $\mathbf{u}, \mathbf{v}$ is found to maximise a correlation of the two vectors $\mathbf{x}' = \mathbf{u}^T\mathbf{x}, \mathbf{y}' = \mathbf{v}^T\mathbf{y}$ as

$$\rho = \max_{\mathbf{u},\mathbf{v}} \frac{E[\mathbf{x}'\mathbf{y}'^T]}{\sqrt{E[\mathbf{x}'\mathbf{x}'^T]E[\mathbf{y}'\mathbf{y}'^T]}} = \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^T\mathbf{C}_{\mathbf{xy}}\mathbf{v}}{(\mathbf{u}^T\mathbf{C}_{\mathbf{xx}}\mathbf{u}\mathbf{v}^T\mathbf{C}_{\mathbf{yy}}\mathbf{v})^{1/2}} \tag{3.26}$$

where ρ is called the canonical correlation and multiple canonical correlations $\rho_1, ...\rho_d$ where $d < min(m_1, m_2)$ are defined by the next pairs of $\mathbf{u}, \mathbf{v}$ which are orthogonal to the previous ones.

Affine invariance of CCA.

Canonical correlations are invariant to affine transformations w.r.t. inputs, i.e. $\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{C}\mathbf{y} + \mathbf{d}$ for arbitrary $\mathbf{A} \in \mathbb{R}^{m_1 \times m_1}, \mathbf{b} \in \mathbb{R}^{m_1}, \mathbf{C} \in \mathbb{R}^{m_2 \times m_2}, \mathbf{d} \in \mathbb{R}^{m_2}$ [17]. This proof is straightforward from (3.26) as $\mathbf{C}_{\mathbf{xy}}, \mathbf{C}_{\mathbf{xx}}, \mathbf{C}_{\mathbf{yy}}$ are covariance matrices and are multiplied by canonical transformations $\mathbf{u}, \mathbf{v}$.

3.6.2 Principal Angles of Linear Subspaces

An alternative (and geometrical) formulation, which is equivalent to the standard formulation of CCA (See [11] for details), is given by subspace concepts. Canonical correlations are often referred to as principal angles (precisely cosine of principal angles) of linear subspaces. 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

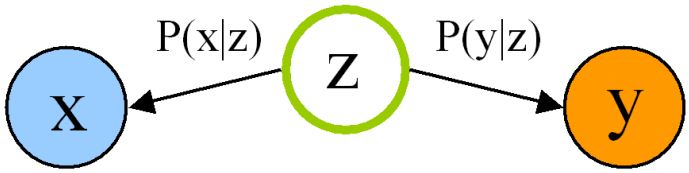


Figure 3.3: Probabilistic Canonical Correlation Analysis tells how well two random variables $\mathbf{x}, \mathbf{y}$ are represented by a common source variable $\mathbf{z}$ [5].

as:

$$\cos\theta_i = \max_{\mathbf{u}_i \in \mathcal{L}_1} \max_{\mathbf{v}_i \in \mathcal{L}_2} \mathbf{u}_i^T \mathbf{v}_i \tag{3.27}$$

subject to $\mathbf{u}_i^T \mathbf{u}_i = \mathbf{v}_i^T \mathbf{v}_i = 1,\ \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0,\ i \neq j$. The concept of principal angles between two linear subspaces is shown in Figure 3.2.

There are various ways to solve this problem. They are all *equivalent*, but the Singular Value Decomposition (SVD) solution [11] is known to be more numerically stable than the standard solutions [71, 60], as the number of free parameters to be estimated is smaller. The solutions offered by neural networks have also been available through e.g. [48].

3.6.3 Probabilistic Interpretation of CCA

Several studies have been conducted in probabilistic interpretation of Canonical Correlation Analysis. Two graphical models have been proposed in a recent work [5] and other probabilistic derivations of CCA by neural networks in [112, 111]. The probabilistic interpretation seems useful in deepening understanding of CCA.

A probabilistic model of CCA [5] is further detailed: as shown in Figure 3.3, the model reveals how well two random variables $\mathbf{x}, \mathbf{y}$ are represented by a common source variable $\mathbf{z} \in \mathbb{R}^d$ by the two likelihoods $p(\mathbf{x}|\mathbf{z}), p(\mathbf{y}|\mathbf{z})$, comprised of affine transformations. The posterior expectations of $\mathbf{z}$ given $\mathbf{x}, \mathbf{y}$ are

$$E(\mathbf{z}|\mathbf{x}) = \mathbf{M}_1^T \mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}_x),\ E(\mathbf{z}|\mathbf{y}) = \mathbf{M}_2^T \mathbf{V}^T (\mathbf{y} - \boldsymbol{\mu}_y), \tag{3.28}$$

where $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_d], \mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_d]$, $\mathbf{M}_1, \mathbf{M}_2$ are arbitrary matrices s.t. $\mathbf{M}_1 \mathbf{M}_2^T = diag(\rho_1, ..., \rho_d)$ and $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y$ are the data means. Whatever $\mathbf{M}_1, \mathbf{M}_2$ are, the projections $E(\mathbf{z}|\mathbf{x}), E(\mathbf{z}|\mathbf{y})$ lie in the d-dimensional subspaces that are identical to those obtained from standard CCA by the canonical transformations $\mathbf{U}, \mathbf{V}$.

3.6.4 CCA by Mutual Information

A new method based on information theory for Canonical Correlation Analysis has been proposed [222]. The method finds canonical coefficient vectors by maximising mutual information of two random variables. Relations of a standard CCA to mutual information and relevant information theories are also found in [17].

3.6.5 Generalisation of CCA with Kernels

The CCA may be generalised by standard kernel tricks. Studies on Kernel CCA for non-linear relations between two sets of vectors have been conducted in [199, 60, 197]. An alternative model for non-linear CCA by alignment of local models has been proposed in [190].

3.6.6 Generalisation of CCA for Multiple Sets

Standard Canonical Correlation Analysis has been designed for two sets of vectors. It has been extended into that for relating several sets of vectors [86, 192, 60]. See [60] for some relevant discussion.

3.6.7 Applications of CCA

There is much literature where CCA is successfully applied for various visual classification tasks. As noted in Chapter 2, CCA has been successful in comparing two sets of images for object (face) recognition. Relevant works include the simple application of principal angles [207], efforts to improve discriminatory powers of principal angle features [45] (See Section 3.6.8 for details), the kernel version of discriminative principal angles [46] and multiple expert design by bagging or boosting [144]. Object recognition problems have been tackled by a standard Kernel CCA [199, 200]. Moreover, CCA has been adopted for classification of temporal stationary images, so-called *dynamic texture recognition* problems in [161].

Canonical Correlation Analysis (CCA) has also been examined for a set-similarity measure of local image descriptions for general object recognition [36]. Each object image is represented as a collection of local image descriptions and compared with other images by CCA. The output of CCA is then supplied into Support Vector Machine for classification. Many other kernels have been compared with the CCA in [36], where CCA was found to be poorer. This might be partly because the sets of local descriptors, as inputs of CCA, might be not well characterised by their low-dimensional subspaces. Also, the output of the CCA for the input of the SVM, is an already compressed form of data learnt in a discriminative way so that the subsequent SVM does learn properly.

CCA has also been applied to problems of estimation of face depth maps from color textures [157], Robot localization [170], detection of neural activity in functional MRI [44], seasonal climate forecasts [166], underwater target classification [152], face recognition by generalised CCA [174], facial expression recognition by kernel CCA [230], and differentiating presence of a real face from that of its photograph [132].

3.6.8 Constrained Mutual Subspace Method

A method for further improvement of discrimination power of CCA has been proposed in [45]. In the method called Constrained Mutual Subspace Method (CMSM), a constrained subspace is defined as that on which the entire class population exhibits small variance. It has been shown that the sets of inter-classes have smaller canonical correlations than those of intr-classes in the constrained subspace, thus facilitating better classification. The

constrained subspace $\mathbf{D}$ is spanned by N_d eigenvectors $\mathbf{d}$ of the matrix $\mathbf{G} = \sum_{i=1}^{C} \mathbf{P}_i \mathbf{P}_i^T$ s.t.

$$\mathbf{G}\mathbf{d} = \lambda \mathbf{d}$$

where C is the number of training classes, $\mathbf{P}_i$ is a basis matrix of the original i-th class data and eigenvector $\mathbf{d}$ corresponds to the N_d smallest eigenvalues. The optimal dimension N_d of the constrained subspace is set experimentally. The subspace $\mathbf{P}_i$ is projected onto $\mathbf{D}$ and the orthogonal components of the projected subspace, normalised to unit length, are obtained as inputs for computing canonical correlations by MSM [207].

Their success has encouraged us to develop an optimal discriminative transformation which ensures that the transformed image sets are class-wise optimally separated by CCA.

Part I

Discriminant Analysis for Single-to-Single Matching

CHAPTER 4

Locally Linear Discriminant Analysis for Recognizing Multi-modally Distributed Classes with a Single Model Image

In light of our success in MPEG-7 (Moving Picture Expert Group) standard and general interest in non-linear discriminant analysis, this chapter presents a novel method of non-linear discriminant analysis for the object recognition task with a single-per-class sample, i.e. a ***Single-to-Single Matching*** task. Linear Discriminant Analysis (LDA) learns the optimal Bayesian discriminant function under the equal-class covariance assumption, which enables convenient generalisation across different object classes. That is, it can learn covariance structures from an independent prototype set of multiple-images-per-class and apply them to a new set of a single-image-per-class. However, a single linear model would be insufficient to capture complex non-linear manifolds of the data set. Here we propose a method called "Locally Linear Discriminant Analysis (LLDA)" for a non-linear classification problem. Note that the proposed method embodies the conventional LDA as a special case.

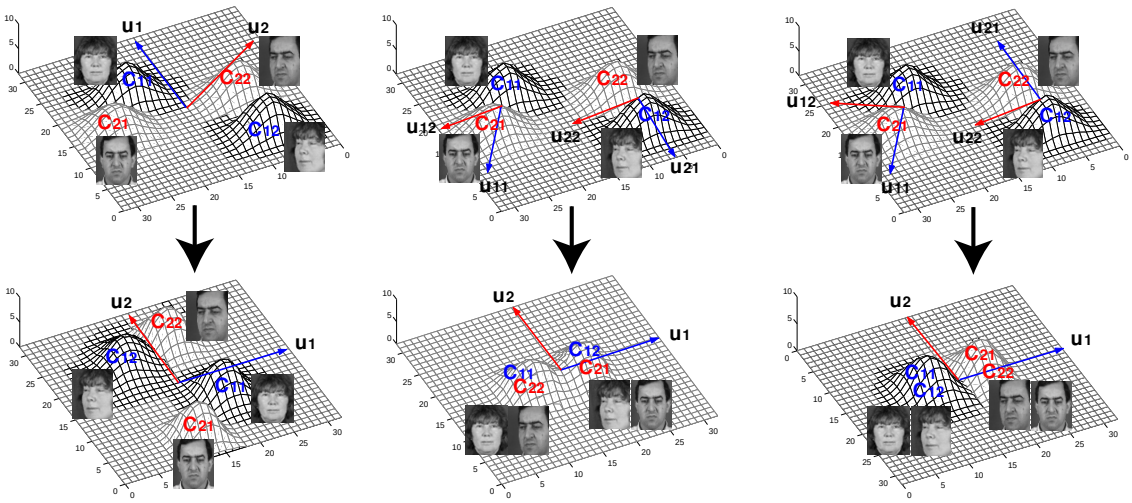
The key novel idea of the proposed method is to maximise the separability of classes locally while promoting consistency between the local representations of a single object class. The method assumes that global nonlinear data structures are locally linear and local structures can be linearly aligned. Input vectors are projected into each local feature space by linear transformations found to yield locally linearly transformed classes that maximise between-class covariance while minimising within-class covariance in the aligned output space. The learnt model can then be applied to recognition of any new class which has a model image in one local cluster and a query image in one of the other clusters. Compared with conventional Generalised Discriminant Analysis (GDA) [8], our method is highly time efficient and avoids overfitting owing to its linear base structures. A novel gradient-based learning algorithm is proposed for finding the optimal set of local linear bases. Experiments have been done for synthetic data and for pose-invariant face recognition with a single model image.

4.1 Discriminant Analysis for Non-linear Problems

The effectiveness of pattern classification methods can seriously be compromised by various factors which often affect sensory information on an object. Frequently observations from a single object class are multi-modally distributed and samples of objects from different classes in the original data space are more closely located to each other than to those of the same class. The data set of face images taken from a certain number of viewing angles is a typical example of such problems. It is because the change in appearance of face images due to pose changes is usually larger than that due to identities of faces. Generally, the face manifold is known to be continuous with respect to continuous pose changes [52]. The proposed method for multi-modally distributed face classes may be useful generally, as a continuous pose set can be divided into many subsets of multi-modal distributions.

Linear Discriminant Analysis (LDA) [47, 9, 226] is a powerful method for face recognition yielding an effective representation that linearly transforms the original data space into a low dimensional feature space where the data is as well separated as possible on the assumption that the data classes are gaussian with equal co-variance structure. However, the method fails to solve non-linear problems, as illustrated in Figure 4.1 (a), because LDA only considers a single linear transformation in a global coordinate system. The transformed face classes are still multi-modally distributed. The multiple LDA system [88, 180, 151] which adopts several independent local transformations attempts to overcome the shortcomings of LDA but it fails to learn any global data structure, as shown in Figure 4.1 (b). In the LDA mixture model [88, 180], it is assumed that single class objects are distributed normally with an identity covariance matrix structure. It then just focuses on maximising the discriminability of the local structures and it does not make any effort to achieve consistency in the local representations of any single object class. In the upper picture of Figure 4.1 (b), the two data sets C_{11} and C_{12} corresponding to the different modes of a class are unfortunately positioned in different directions of the corresponding local components, u_{11} and u_{21}, thus having different representations in a global coordinates as illustrated below. Classes are mixed up in the transformed space. The view-based method for face recognition proposed by Pentland [151] would exhibit the same difficulty in these circumstances. Following their idea, we could divide images into distinct pose groups and then train LDA separately for each group, which is similar to using the LDA mixture. Because these LDA bases do not encode any relations of the different pose groups, it is not guaranteed that this 'view-based LDA' would yield a consistent representation of different pose images of a single identity. In many conventional face recognition systems [88, 91, 9, 226, 151] which adopt a linear machine such as LDA or LDA mixture model, as many gallery samples as possible are required to capture all the modes of class distributions. However, it is often difficult to obtain various mode (or pose) images of one person.

Support vector machine (SVM) based on kernels has been successfully applied for non-linear classification problems such as face detection [188, 148]. This is, however, inefficient for multi-class recognition and inappropriate when a single sample per class is available to build a class model. By design generalised discriminant analysis (GDA) [8, 124, 139, 211] is suitable for multi-class face recognition problems where the original data is mapped into a high-dimensional feature space via a kernel function. The GDA representation learnt from training face classes of various pose images can be exploited to achieve pose robust representation of novel face classes. Recognition with a single model image of the novel classes is thus facilitated. However, GDA generally suffers from the drawback of high



(a) LDA (b) LDA mixture (c) LLDA

Figure 4.1: Comparison of LDA, LDA mixture and LLDA for the non-linear classification problem. Only LLDA guarantees that the multi-modally distributed face classes in the input space are transformed into the class-wise single-modal distributions in the output space. Each upper plot shows the simulated data distributions and the components found by LDA, LDA mixture and LLDA. In the lower graphs the transformed class distributions in the global output coordinate system are drawn. The data are generated by $C_{11} = \{X \sim N(21.6, 2), Y \sim N(21.6, 1)\}$, $C_{12} = \{X \sim N(7.5, 2), Y \sim N(7.5, 0.8)\}$, $C_{21} = \{X \sim N(26, 2), Y \sim N(16, 2)\}$, and $C_{22} = \{X \sim N(8, 2), Y \sim N(16, 1.2)\}$, where $N(a, b)$ is a normal variable with the mean a and standard deviation b. 200 data points are drawn for each mode. C_{ij} is the j-th cluster of the i-th class, u_{ij} is the j-th component of the i-th cluster and u_i denotes the i-th component of the output coordinate system.

computational cost in classification and overfitting. In applications such as classification of large data sets on the Internet or video, computational complexity is particularly important. The global structure of nonlinear manifolds was represented by a locally linear structure in [159, 65]. These methods perform unsupervised learning for locally linear dimensionality reduction but not a supervised learning for discrimination.

4.2 Locally Linear Discriminant Analysis (LLDA)

In this study, several locally linear transformations are concurrently sought so that the class structures manifest by the locally transformed data are well separated in the output space. The proposed method is called "Locally Linear Discriminant Analysis (LLDA)". The underlying idea of this approach is that global nonlinear data structures are locally linear and local structures can be linearly aligned. Single-class training objects, even if multi-modally distributed, are transformed into a cluster that is as small as possible with maximum distance to the different class training objects, by a set of locally linear functions, as illustrated in Figure 4.1 (c). The linear functions learnt from training face classes of various pose images can be efficiently generalised to novel classes. Even when a single

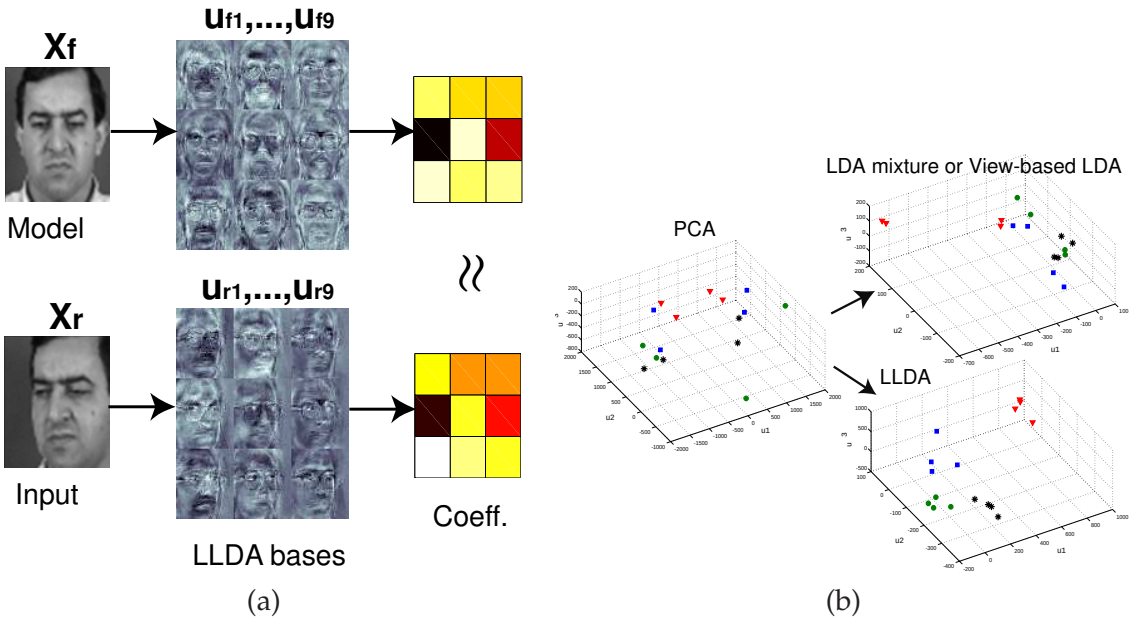


Figure 4.2: LLDA Representation. (a) Locally discriminative and aligned LLDA bases yield similar representations of posed face images. u_{ij} denotes the j-th component of the i-th cluster. (b) Face-image distributions in the first three dimensions of PCA, view-based LDA and LLDA. Whereas LDA and view-based LDA have shuffled class samples, LLDA achieves class-distinctive distributions. Classes are marked with different symbols.

model image per class is provided, it is much easier to recognize a novel view image in the aligned output space.

The method advocated maximises the separability of classes locally while promoting consistency between the multiple local representations of single class objects. Compared with the conventional nonlinear methods based on kernels, the proposed method is much more computationally efficient because it only involves linear transformations. By virtue of its linear base structure, it also reduces overfitting normally exhibited by conventional non-linear methods. The transformation functions (or bases) learned from the face images of two views are visualised in Figure 4.2 (a). The functions can be exploited as the bases of a low dimensional subspace for robust face recognition. The basis functions of each cluster are specific to a particular facial pose. We note two interesting points in this Figure. Firstly the bases of each cluster are similar to those of classical LDA and this ensures that face images of different individuals in the same pose are discriminative. Secondly, the corresponding components of the two clusters, for example, u_{f1} and u_{r1} are aligned. They are characterised by a certain rotation and scaling with similar intensity variation. In consequence, face images of the same individual in different poses have quasi-invariant representation as shown in Figure 4.2 (a) and (b). For conciseness, only four face classes are plotted in the subspaces of Principal Component Analysis (PCA) [187], view-based LDA (or LDA mixture) and LLDA in Figure 4.2 (b). Each class has the four samples of two poses and two time sessions. While LDA and view-based LDA have shuffled class samples, LLDA achieves class-distinctive distributions of samples.

The remainder of this chapter is organized as follows: The proposed LLDA method is formulated in Section 4.3 and a solution of the optimization problem involved is presented in Section 4.4. Section 4.5 further simplifies the proposed method by replacing the

Gaussian mixture model with K-means clustering. Section 4.6 is devoted to the analysis of computational complexity. Section 4.7.1 presents the results of experiments performed to demonstrate the beneficial properties of the proposed method on synthetic data. In Section 4.7.2, the method is applied to the problem of face recognition. Summary is given in Section 4.8.

4.3 LLDA Formulation

The proposed method, LLDA is applicable to multi-class nonlinear classification problems by using a set of locally linear transformations. Consider a data set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$ of N-dimensional vectors of face images of multiple poses and C classes $\{\mathbf{X}_1, ..., \mathbf{X}_c, ..., \mathbf{X}_C\}$. The input vectors are clustered into K subsets denoted by $k,\ k = 1, ..., K$ and each subset k represents a cluster to which a different transformation function is applied. As the multi-modality of the face data distribution is typically caused by the variety of poses, the clusters may correspond to facial poses. Note that each pose set is assumed to contain data of every class. Clusters are obtained by K-means clustering, Gaussian mixture modelling or pose labels (if available) of the input vectors. The number of clusters K is chosen to maximise an objective function defined on the training set. Because K is usually a small positive integer, we can make the best choice of K empirically. It is also pertinent to select K as the number of pose sets, if it is known. However, general model order selection for a high dimensional data set remains an open problem. The basic LLDA approach draws on the notion of 'soft clustering', in which each data point belongs to each of the clusters with a posterior probability $P(k|\mathbf{x})$. The algorithm that is combined with 'hard' K-means clustering will be discussed in Section 4.5. We define the locally linear transformation $\mathbf{U}_k = [\mathbf{u}_{k1}, \mathbf{u}_{k2}, ..., \mathbf{u}_{kN}],\ k = 1, ..., K$ such that

$$\mathbf{y}_i = \sum_{k=1}^{K} P(k|\mathbf{x}_i)\mathbf{U}_k^T(\mathbf{x}_i - \boldsymbol{\mu}_k), \tag{4.1}$$

where N is the dimension of the transformed space. The mean vector of the k-th cluster $\boldsymbol{\mu}_k$ is described by

$$\boldsymbol{\mu}_k = \left(\sum_{i=1}^{M} P(k|\mathbf{x}_i)\mathbf{x}_i\right) / \left(\sum_{i=1}^{M} P(k|\mathbf{x}_i)\right). \tag{4.2}$$

The locally linear transformation matrices $\mathbf{U}_k$ are concurrently found so as to maximise the criterion function, J. Two objective functions are considered,

$$J_1 = \log(|\widetilde{\mathbf{B}}|/|\widetilde{\mathbf{W}}|), \text{ and } J_2 = (1-\alpha)|\widetilde{\mathbf{B}}| - \alpha|\widetilde{\mathbf{W}}|, \tag{4.3}$$

where $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{W}}$ are the between-class and within-class scatter matrices in the locally linear transformed feature space respectively. The constant α takes values from the interval [0 1]. The objective functions maximise the between-class scatter while minimising the within-class scatter in the locally transformed feature space. One of the differences between the two defined objective functions is manifest in the efficiency of "learning". The log objective function J_1 has the benefit of not requiring a free parameter α but it is more

costly computationally. The function J_2 can efficiently be optimised iteratively, once α is selected. This is exemplified in the subsequent section. In terms of performance, the two approaches are similar, as reported in the experimental section 4.7.1. The global mean $\widetilde{\mathbf{m}}$ of all the transformed samples is

$$\widetilde{\mathbf{m}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{y}_i = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} P(k|\mathbf{x}_i)\mathbf{U}_k^T(\mathbf{x}_i - \boldsymbol{\mu}_k), \tag{4.4}$$

where M is the total number of the samples. By substituting for $\boldsymbol{\mu}_i$ from equation (4.2), we get $\widetilde{\mathbf{m}} = \overrightarrow{0}$. The sample mean for class c which consists of M_c samples is given by

$$\widetilde{\mathbf{m}}_c = \frac{1}{M_c}\sum_{\mathbf{x}\in\mathbf{X}_c}\mathbf{y} = \sum_{k=1}^{K}\mathbf{U}_k^T\mathbf{m}_{ck}, \tag{4.5}$$

where $\mathbf{m}_{ck} = \frac{1}{M_c}\sum_{\mathbf{x}\in\mathbf{X}_c} P(k|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_k)$.

The term $\mathbf{m}_{ck}$ denotes the sample mean of a class c in the k-th cluster. Because the transformation is defined with respect to the original cluster mean $\boldsymbol{\mu}_k$, the total mean $\widetilde{\mathbf{m}}_k$ of the transformed data in every cluster becomes zero. Using equations (4.4) and (4.5), the transformed between-class scatter matrix is given as:

$$\begin{aligned}
\widetilde{\mathbf{B}} &= \sum_{c=1}^{C} M_c(\widetilde{\mathbf{m}}_c - \widetilde{\mathbf{m}})(\widetilde{\mathbf{m}}_c - \widetilde{\mathbf{m}})^T \\
&= \sum_{c=1}^{C} M_c\left(\sum_{k=1}^{K}\mathbf{U}_k^T\mathbf{m}_{ck}\right)\left(\sum_{k=1}^{K}\mathbf{U}_k^T\mathbf{m}_{ck}\right)^T \\
&= \sum_{k=1}^{K}\mathbf{U}_k^T\mathbf{B}_k\mathbf{U}_k + \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^T\mathbf{B}_{ij}\mathbf{U}_j + \left(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^T\mathbf{B}_{ij}\mathbf{U}_j\right)^T
\end{aligned} \tag{4.6}$$

where

$$\mathbf{B}_k = \sum_{c=1}^{C} M_c\mathbf{m}_{ck}\mathbf{m}_{ck}^T, \text{ and } \mathbf{B}_{ij} = \sum_{c=1}^{C} M_c\mathbf{m}_{ci}\mathbf{m}_{cj}^T.$$

The between-class scatter matrix consists of the scatter matrices associated with the respective clusters and the correlation matrix of the data samples belonging to two discrete clusters. The correlation matrix encodes the relations of the two local structures for alignment. Similarly, the within-class scatter is defined by

$$\begin{aligned}
\widetilde{\mathbf{W}} &= \sum_{c=1}^{C}\sum_{\mathbf{x}\in\mathbf{X}_c}(\mathbf{y} - \widetilde{\mathbf{m}}_c)(\mathbf{y} - \widetilde{\mathbf{m}}_c)^T \\
&= \sum_{k=1}^{K}\mathbf{U}_k^T\mathbf{W}_k\mathbf{U}_k + \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^T\mathbf{W}_{ij}\mathbf{U}_j + \left(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\mathbf{U}_i^T\mathbf{W}_{ij}\mathbf{U}_j\right)^T,
\end{aligned} \tag{4.7}$$

$$\mathbf{W}_k = \sum_{c=1}^{C} \sum_{\mathbf{x} \in \mathbf{X}_c} \left(P(k|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_k) - \mathbf{m}_{ck}\right) \left(P(k|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_k) - \mathbf{m}_{ck}\right)^T$$

$$\mathbf{W}_{ij} = \sum_{c=1}^{C} \sum_{\mathbf{x} \in \mathbf{X}_c} \left(P(i|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_i) - \mathbf{m}_{ci}\right) \left(P(j|\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_j) - \mathbf{m}_{cj}\right)^T .$$

Matrix $\mathbf{W}_k$ describes a local cluster and $\mathbf{W}_{ij}$ is the cross-term of two local clusters.

Relations to classical LDA and LDA mixture model. Please note that the defined criterion with $K = 1$ is identical with that of the conventional LDA. Moreover, the proposed algorithm without the cross terms $\mathbf{B}_{ij}$ and $\mathbf{W}_{ij}$ would adhere to the same concept as that of the LDA mixture model by focusing just on the local separability. In this circumstance, clusters are treated independently of each other, thus not requiring alignment of the local representations. Any subset of classes (except only one class) is allowed in clusters whereas every class is assumed to be presented in each cluster (pose) for LLDA. In general, data may have heterogeneous class distributions and local structures making our assumption in LLDA invalid. Interesting follow-up study is found in [29] where this issue is handled by locating local discriminant analysers at their optimal places.

4.4 Gradient-based Solution for LLDA

In this section, we provide an efficient iterative optimisation method based on a gradient learning algorithm for an optimal set of locally linear transformation functions. While it is hard to find good parameters of a kernel function for new data in the conventional GDA, the proposed learning has only parameters which reduce or eliminate overfitting. The discriminant based on such a piecewise linear structure has the benefit of optimising a convex function with respect to the set of basis vectors of the local coordinates, yielding a unique maximum.

The method is based on a one-basis vector solution for $\mathbf{u}_{k1}$, $k = 1, ..., K$. Other methods based on incremental one-basis at a time solution can be found in [75, 147, 228] for discriminant or independent component analysis criteria. The proposed gradient method yields a global maximum solution by virtue of the criterion function's being 2nd-order convex with respect to all the variables $\mathbf{u}_{k1}$, $k = 1, ..., K$. We need to run the one-basis algorithm several times to obtain a multidimensional solution $\mathbf{U}_k = [\mathbf{u}_{k1}, \mathbf{u}_{k2}, ..., \mathbf{u}_{kN}]$, $k = 1, ..., K$. The vector orthogonalization is performed to prevent different vectors from converging to the same maxima in every iteration. We seek the vectors $\mathbf{u}$ which maximise the criterion function under the constraint of their being unit norm:

$$\begin{aligned} &\text{Max } J_1 \text{ or } J_2, \\ &\text{for } ||\mathbf{u}_{kn}|| = 1, k = 1, ..., K \text{ and } n = 1, ..., N. \end{aligned} \tag{4.8}$$

This constrained optimization problem is solved by the method of projection on the constraint set [75]. A vector normalization imposing a unit norm is executed after every update of the vector. The learning rules are as follows:

Take the following steps with an index n starting from 1 to N for $\mathbf{u}_{kn}$, $k = 1, ..., K$.

1. Randomly initialize K unit vectors $\mathbf{u}_{kn}$.

2. Calculate the gradient of the objective function with respect to the variables $\mathbf{u}_{kn}$ by

$$\frac{\partial J_1}{\partial \mathbf{u}_{kn}} = \left(2\widetilde{\mathbf{B}}^{-1}\mathbf{B}_k - 2\widetilde{\mathbf{W}}^{-1}\mathbf{W}_k\right)\mathbf{u}_{kn} + \sum_{i=1, i\neq k}^{K}\left(2\widetilde{\mathbf{B}}^{-1}\mathbf{B}_{ki} - 2\widetilde{\mathbf{W}}^{-1}\mathbf{W}_{ki}\right)\mathbf{u}_{in}, \text{or}$$

$$\frac{\partial J_2}{\partial \mathbf{u}_{kn}} = \left(2(1-\alpha)\mathbf{B}_k - 2\alpha\mathbf{W}_k\right)\mathbf{u}_{kn} + \sum_{i=1, i\neq k}^{K}\left(2(1-\alpha)\mathbf{B}_{ki} - 2\alpha\mathbf{W}_{ki}\right)\mathbf{u}_{in}. \tag{4.9}$$

3. Update with an appropriate step size η as

$$\Delta\mathbf{u}_{kn} \leftarrow \eta\frac{\partial J}{\partial \mathbf{u}_{kn}}. \tag{4.10}$$

4. Carry out the deflationary orthogonalization by

$$\mathbf{u}_{kn} \leftarrow \mathbf{u}_{kn} - \sum_{i=1}^{n-1}\left(\mathbf{u}_{kn}^T\mathbf{u}_{ki}\right)\mathbf{u}_{ki}. \tag{4.11}$$

5. Normalize the vectors $\mathbf{u}_{kn}$ by

$$\mathbf{u}_{kn} \leftarrow \mathbf{u}_{kn}/||\mathbf{u}_{kn}||. \tag{4.12}$$

Repeat the processes $2 \sim 5$ until the algorithm converges to a stable point, set $n := n + 1$ and then go to step 1.

Note that the two objective functions have different costs in learning process. When calculating the gradients of J_2 in (4.9), all the matrices, here scalar values, are previously given but the two matrices $\widetilde{\mathbf{B}}^{-1}, \widetilde{\mathbf{W}}^{-1}$ in the learning of J_1 should be iteratively updated. For the synthetic data example given in Figure 4.1, the optimization of J_1 takes about 15 times longer than that of J_2. While the learning of J_1 has a benefit of avoiding a free parameter α, J_2 has a simpler optimization cost when the parameter α is fixed. By changing α, one can control the importance of the variance of the between-class to that of the within-class data distributions. Orthogonalization (4.11) ensures that the proposed discriminant is defined by orthonormal basis vectors in each local coordinate system. The orthonormalisation of the bases yields more robust performance in the presence of estimation error (please refer to [147, 228] for the details). The benefits of orthonormal bases in discriminant analysis over classical LDA have also been explained in previous studies. Although we do not provide a proof of convergence or uniqueness of the gradient-based iterative learning method, its convergence to a global maximum can be expected by virtue of the criterion's being a 2nd-order convex function with respect to a basis vector, $\mathbf{u}_{kn}$, of each

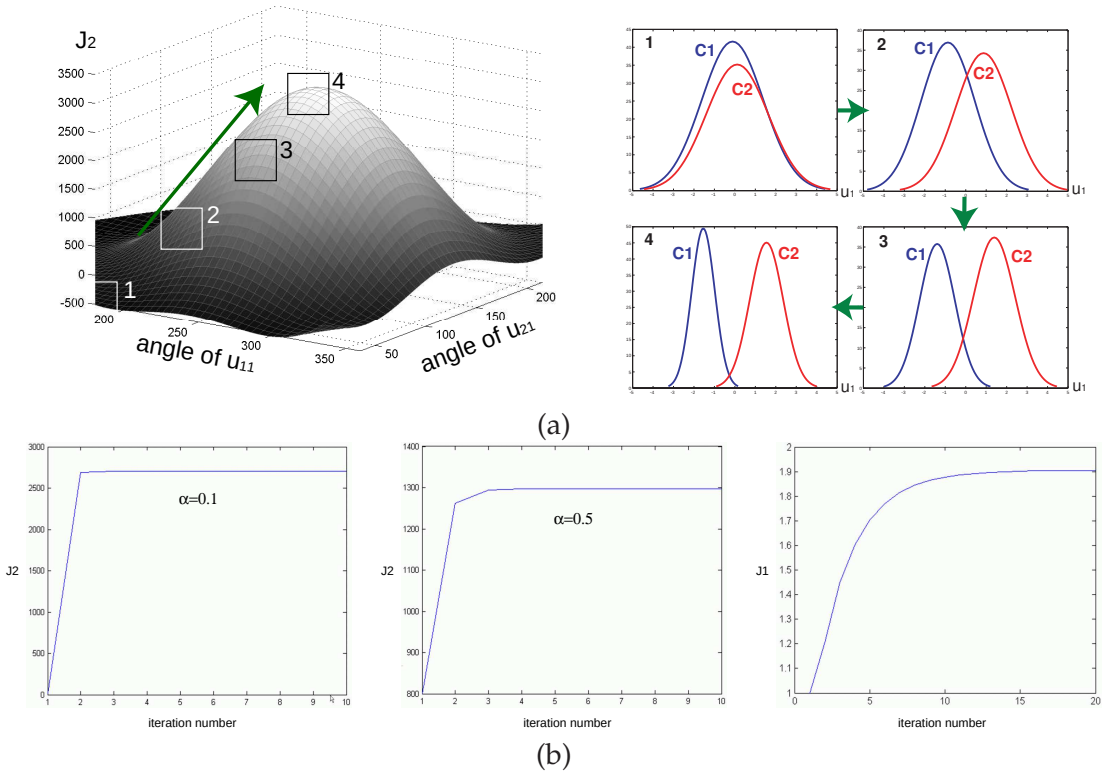


Figure 4.3: Convex optimization in LLDA learning. The proposed gradient-based learning is performed for the data distribution shown in Figure 4.1, where K is set to 2 and step size η is fixed to 0.1. (a) Value of the criterion J_2 (left) as a function of orientation of $\mathbf{u}_{11}, \mathbf{u}_{21}$ with $\alpha = 0.1$. The distributions of the two classes $C_1 = C_{11} \bigcup C_{12}, C_2 = C_{21} \bigcup C_{22}$ on the first major component $\mathbf{u}_1$, are drawn (right) as a series while J_2 is maximised. (b) Convergence graphs of J_2 with $\alpha = 0.1, 0.5$ and J_1.

local coordinate system, and the joint set of the basis vectors $\mathbf{u}_{kn}$, $k = 1, ..., K$, as explained in [115, 114]. Figure 4.3 shows the convergence characteristics of the learning process for the synthetic data presented in Figure 4.1. The constant α was explored in steps of 0.1 for the best classification rate of the training data. The value of J_2 according to the angles of basis vectors has a unique global maximum. It is also noted that the gradient optimization method of the objective function quickly converges regardless of constant α. The learning which deploys the objective function J_1 also stably approaches a unique maximum.

Lagrangian Method for Constrained Optimization. A solution to the constrained optimization problem can also be obtained by using the method of Lagrangian multipliers as

$$L = (1 - \alpha)|\widetilde{\mathbf{B}}| - \alpha|\widetilde{\mathbf{W}}| - \sum_{k=1}^{K} \Lambda_k (\mathbf{U}_k^T \mathbf{U}_k - \mathbf{I}), \tag{4.13}$$

where $\mathbf{I}$ is the identity matrix and the diagonal matrix of eigen-values is

$$\Lambda_k = \begin{bmatrix} \lambda_{k1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_{kN} \end{bmatrix}$$

The third term in (4.13) represents the constraints that the local transformations have orthonormal bases. The gradient of the Lagrangian function with respect to the basis vectors is

$$\frac{\partial L}{\partial \mathbf{u}_{kn}} = (2(1-\alpha)\mathbf{B}_k - 2\alpha\mathbf{W}_k - 2\lambda_{kn}\mathbf{I})\,\mathbf{u}_{kn} + \sum_{i=1, i\neq k}^{K} (2(1-\alpha)\mathbf{B}_{ki} - 2\alpha\mathbf{W}_{ki})\,\mathbf{u}_{in} = 0 \tag{4.14}$$

The solution can be found by numerical optimization of the Lagrangian function. However, in practice, a numerical optimization can only be used in low dimensional data spaces. As a reference, we used the numerical optimization "solve" function in Matlab to solve the two-dimensional problem shown in Figure 4.1. The constraint optimization took 600 times longer than the gradient-based of J_2. The two proposed methods of gradient based learning are much favoured for their efficiency.

4.5 LLDA with K-means Clustering

Let us revisit the basic model derived in Section 4.3 by considering the special case involving a discrete posterior probability. K-means clustering divides a data set into disjoint subsets. If the data point $\mathbf{x}$ belongs to the k^*-th cluster, $P(k^*|\mathbf{x}) = 1$ and $P(k|\mathbf{x}) = 0$ for all the other k's. The mean vector of the k-th cluster $\boldsymbol{\mu}_k$ in (4.2) can be rendered by

$$\boldsymbol{\mu}_k = \left(\sum_{\mathbf{x}} P(k|\mathbf{x})\mathbf{x}\right) / \left(\sum_{\mathbf{x}} P(k|\mathbf{x})\right) = \left(\sum_{\mathbf{x}\in k} \mathbf{x}\right) / M_k', \tag{4.15}$$

where M_k' is the sample number of the cluster k. The defined transformation in (4.1) becomes

$$\mathbf{y} = \mathbf{U}_k^T(\mathbf{x} - \boldsymbol{\mu}_k) \text{ for } \mathbf{x} \in k. \tag{4.16}$$

The definition of the global mean (4.4) and the class mean (4.5) changes as follows:

$$\widetilde{\mathbf{m}} = \frac{1}{M}\sum_{k=1}^{K} \mathbf{U}_k^T \sum_{\mathbf{x}\in k}(\mathbf{x} - \boldsymbol{\mu}_k) = \overrightarrow{0}, \qquad \widetilde{\mathbf{m}}_c = \sum_{k=1}^{K} \mathbf{U}_k^T \mathbf{m}_{ck}, \tag{4.17}$$

where

$$\mathbf{m}_{ck} = \frac{1}{M_c} \sum_{\mathbf{x}\in \mathbf{X}_c \bigcap k} (\mathbf{x} - \boldsymbol{\mu}_k).$$

The transformed between-class matrix (4.6) and the within-class scatter matrix (4.7) can similarly be expressed by changing the notation from $P(k|\mathbf{x})$ to $\mathbf{x} \in k$. The learning algorithm in Section 4.4 finds the optimal set of locally linear transformation $\mathbf{U}_k,\ k = 1, ..., K$.

When a new pattern $\mathbf{x}_{test}$ is presented, it is first assigned to one of the clusters by

$$\mathbf{x}_{test} \in k^* = \min_{argk} ||\mathbf{x}_{test} - \boldsymbol{\mu}_k|| \tag{4.18}$$

and transformed by using the corresponding function

$$\mathbf{y}_{test} = \mathbf{U}_{k^*}^T(\mathbf{x}_{test} - \boldsymbol{\mu}_{k^*}). \tag{4.19}$$

4.6 Computational Complexity

The complexity of the algorithms depends on the computational costs associated with extracting the features and with matching.

Feature extraction. For the linear subspace methods such as PCA and LDA, the cost of feature extraction is determined by the dimensionality N of the input vector, $\mathbf{x}$, and the number of components of the subspace S. The cost of extracting features using linear methods is roughly proportional to $N \times S$. In nonlinear subspace methods like the GDA, the n-th component of the projection of vector $\mathbf{x}$ is computed as

$$\mathbf{y}_n = \sum_{i=1}^{M} \alpha_{ni} k(\mathbf{x}_i, \mathbf{x}), \tag{4.20}$$

where M is the total number of training patterns, α_{ni} is a real weight and k denotes a kernel function. The cost of extracting features of the GDA is about $N \times S \times M$. The proposed method, LLDA has a similar cost with that of PCA or LDA depending on the preceding clustering algorithm. When a hard clustering such as K-means is applied, the cost of extracting features is $N \times (S + K)$, where the additional term $N \times K$ is for assigning a cluster to the input. When a soft clustering is applied, the cost is multiplied by the number of clusters, i.e., $N \times S \times K$. Note that usually $K \ll M$.

Matching. When the data points are represented as the S dimensional feature vectors and C gallery samples are given for the C class categories, the matching cost for recognition is $C \times S$. This applies to all, the linear, nonlinear and proposed subspace methods.

4.7 Experiments

4.7.1 Results on Synthetic Data

Two sets of 2-dimensional synthetic data, i.e. $\mathbf{x} \in \mathbb{R}^2$, underwent experiment. The distribution of c-th class data was generated as Gaussian mixtures by

$$p(\mathbf{x}|c) = \sum_{k=1}^{K} \frac{1}{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{ck}, \Sigma_{ck})$$

where K is the number of clusters and $\boldsymbol{\mu}_{ck}, \Sigma_{ck}$ are the mean and covariance matrix of k-th cluster data of c-th class. 200 data points were drawn from each Gaussian mode. Set 1 has three classes which have two distinct modes while Set 2 has two classes with three distinct modes, as shown in Figure 4.4. Conventional LDA, mixture of LDA, and GDA with the radial basis function (RBF) as a kernel are compared with LLDA in terms of classification error. Euclidean distance(E.D.), normalized correlation(N.C.) and Mahalanobis distance(M.D.) were used as similarity functions for the nearest neighbor (N.N.) classification. It is noted that all the transformed data points were compared with the sample mean of each class in (4.5).

In the LLDA method, the number of clusters, K, was selected to maximise the value of the objective function. For the example of the data of Set 1, the peak values of J_1 changed with K as follows: -7.14, 2.97, 0.85 for $K = 1, 2, 3$ respectively, so the number $K = 2$ was chosen. This is much simpler than the parameter selection of RBF as a kernel function in GDA, because the standard deviation of RBF is hard to initialize and it is a real (non-integer) value. The axes of LDA, LDA mixture, LLDA are drawn in Figure 4.4. Table 4.1 shows the average number of classification errors with their standard deviation and the relative costs of feature extraction. It is apparent that the proposed discriminant can well solve the non-linear classification problem on which the conventional linear methods fail and it is much more profitable than GDA in terms of computational efficiency. The feature extraction complexity of the proposed method is about 1/270 of that of GDA in this example. Although the GDA was slightly more accurate, it is noted that the kernel parameter of RBF in GDA was exhaustively searched to find the best performance for the given data. In contrast, the proposed algorithm based on the log objective function has only a small integer K to be adjusted and the learning process is also much faster. Note that additionally, when the class distributions have a single mode, LLDA with $K = 1$ yields a successful separation by behaving like the conventional LDA. LLDA with $K = 1$ is identical with the conventional LDA, with the exception of the orthonormal constraint imposed on the axes by LLDA.

4.7.2 View-invariant Face Recognition with One Sample Image

The proposed algorithm has been validated on the problem of free-pose face recognition in the scenario when only a single frontal image of each class is available as a gallery image.

In our experiments, the proposed algorithm, LLDA, is compared with PCA, LDA and GDA as the benchmark subspace methods that have been successfully applied to face recognition in the past and FaceIt(v.5.0), the commercial face recognition system from Identix. FaceIt ranked top overall in the Face Recognition Vendor Test 2000 and 2002 [154, 13].

Database: We used the XM2VTS data set annotated with pose labels of the face. The face

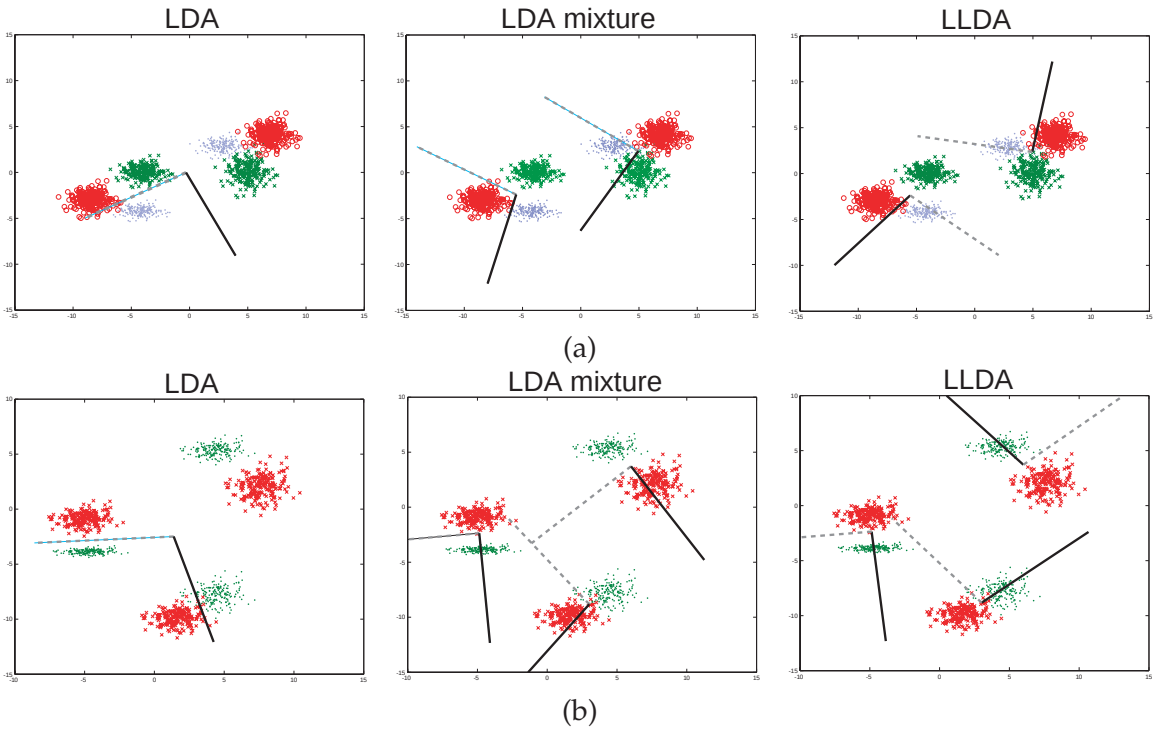


Figure 4.4: Simulated data distributions and the components found. Colors (or symbols) indicate different classes. Black solid lines represent the first major components and gray dashed lines the second components. (a) For Set 1. (b) For Set 2.

	E.D.	N.C.	M.D.	Cost
Set1 (400 samples/class)				
LDA	266±115	266±115	81±61	1
LDA mixture	254±27	255±23	169±45	1+ω
GDA	4.3±1.1	4.3±1.1	4.4±0.5	270
LLDA J_1 + km	7.6±3.5	7.6±3.5	7±3.4	1+ω
LLDA J_2 + km	7.6±3.5	8±3.6	7.3±3.7	1+ω
LLDA J_1 + GMM	7.6±3.5	8±3.6	7.3±3.7	2+ω
Lagran. J_2	7.6±3.2	8±2.6	7.3±2.8	1+ω
Set2 (600 samples/class)				
LDA	308±129	308±129	207±272	1
LDA mixture	205±1.4	205±1.4	206±7	1+ω
GDA	4±1.4	4±1.4	4±0	278
LLDA J_1 + km	9.5±3.5	9.5±3.5	7.5±3.5	1+ω
LLDA J_2 + km	8±1.4	8±1.4	7±2.8	1+ω

Table 4.1: Classification Results (number of errors). ω indicates the computational cost of deciding to which cluster a new pattern belongs. It is usually less than 1. 'LLDA J_1 +km' is the LLDA of the objective function J_1 with K-means clustering algorithm. 'LLDA J_1 + GMM' indicates the LLDA of the objective function J_1 with Gaussian mixture modelling. 'Lagrangian J_2' denotes a numerical solution of the Lagrangian formulation.

database consists of 2950 facial images of 295 persons with 5 pose variations and 2 different time sessions which have 5 months time elapse. The data set consists of 5 pose groups

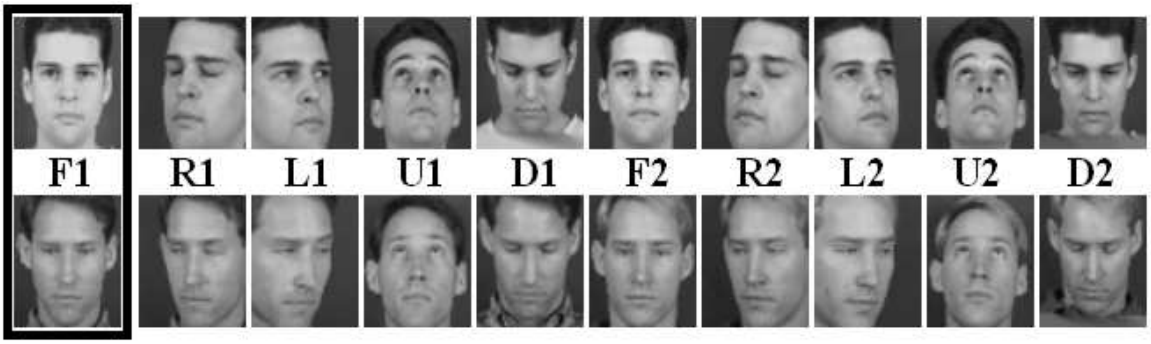


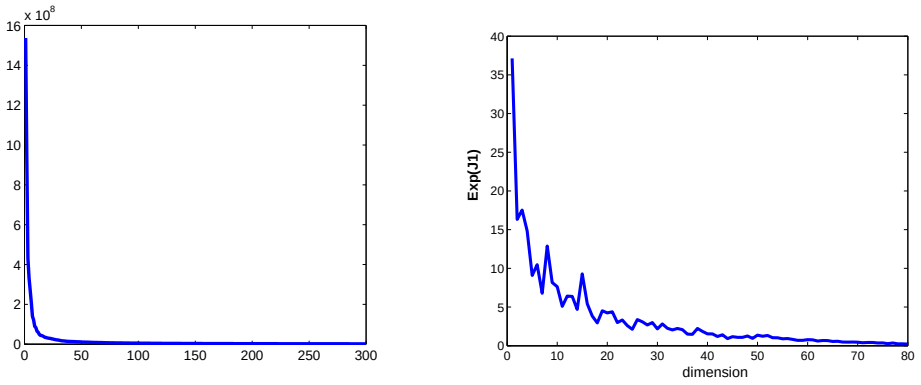
Figure 4.5: Normalized data samples. The leftmost image is given as the gallery image and other rotated face images are used as testing images.

(F,R,L,U,D) which are captured in frontal view, about ± 30 horizontal rotations and ± 20 vertical rotations. The two images of a pose group 'F' captured at different times are denoted by F1 and F2. This may be the largest public database that contains images of faces taken from different viewpoints. The images were normalized to 46*56 pixel resolution with a fixed eye position and some normalized data samples are shown in Figure 4.5. The face set is partitioned into the three subsets: 1250 images of 125 persons, 450 images of 45 persons and 1250 face images of 125 persons for the training(Tr), evaluation(Ev) and test(Te) respectively. Please note that the three sets have different face identities. For the test of the commercial FaceIt system, the original images were applied to the system with the manual eye positions.

Protocol and Setting: The training set is utilised to learn the subspace representation of the conventional PCA/LDA/GDA methods and LLDA with K-means. For efficiency of learning, all of the algorithms were applied to the first 80 ($\lambda_{80}/\lambda_1 = 0.004$) eigenfeatures of the face images. Figure 4.6 shows the plots of eigenvalues and J_1 of LLDA as a function of dimensionality. The evaluation set is used to adjust the kernel parameter of GDA(an RBF kernel with an adjustable width) and the dimensionality of the output vectors for all methods. The parameters are properly quantized and all combinations of the discrete values of the quantized parameters are examined to get the best recognition rate on the evaluation set. In LLDA, the number of clusters corresponded to the number of pose groups and K-means algorithm was applied. The log objective function J_1 was utilised to learn the set of transformation functions and the learning rate was controlled to achieve faster convergence. The learning typically took 2 or 3 minutes in Pentium IV 2GHz PC.

In the test the frontal face images of the test set, which are the leftmost images in Figure 4.5, are registered as a gallery and all the other images of the test set are exploited as queries. All the test images are projected into the learned subspace and Nearest-Neighbor based classification is performed based on the projection coefficients. $\%$ recognition rates are measured. In LLDA, test face images were assigned to one of the clusters by equation (4.18) and projected onto the corresponding subspace by (4.19).

Results : Table 4.2 presents the recognition rates on the evaluation and test set and Figure 4.7 shows the performance curves of the test set as a function of dimensionality. The recognition rate of the evaluation and test set was much enhanced by the proposed algorithm. FaceIt exhibited the best recognition performance for the frontal images F2 but quite



(a) (b)

Figure 4.6: (a) Eigenvalues of the face data. (b) Plot of J1 as a function of dimensionality.

low recognition rates for the rotated faces especially those involving up/down rotations. More results showing the effects of the elapsed time and the size of test population are given in Figure 4.8.

In LLDA, the number of clusters was chosen as the number of pose groups as previously mentioned by assuming that the multi-modality of the face class distributions is caused by different poses. In each cluster, classes are assumed to be linearly separable. Although this assumption may not be true, as other factors such as time lapse can make a class distribute multi-modally and not linearly separable, we found that LLDA performed much better than LDA/GDA/FaceIt. A performance degradation as a function of time was observed for all methods but a relative performance gain exhibited by LLDA was still preserved as shown in Figure 4.8. As mentioned above, the results of the test set were obtained by using the output dimensionality found to be the best for the evaluation set. The establishment of a proper evaluation set is important because the test results are sensitive to the output dimensionality, as shown in Figure 4.7. This may be because pose variation is so large that the methods find only few meaningful axes. We can see that the evaluation set used proved adequate to solving this peaking problem as the recognition results on the test set using the best dimensionality indicated by the evaluation set in Table 4.2 agreed with the best results of the graph in Figure 4.7. GDA had the tendency highly to overfit on the training set so that a separate evaluation set was needed to suppress this behaviour.

Regarding the complexity of the feature extraction, PCA, LDA and the LLDA are roughly identical and GDA about 40 times worse than the linear methods. Please note that the complexity of GDA depends on the size of the training set. The proposed method is not expensive in terms of computational costs and provides more robust and accurate performance for all the dimensionalities than other methods.

4.8 Summary

A novel non-linear discriminant analysis method has been proposed for a challenging task, a non-linear classification problem with a single model image. Object classes, for e.g. face images exhibiting large pose variations, often have nonlinear manifolds (or multi-modal distributions) and are not linearly separable. In addition, it is required to learn relations

	PCA		LDA		GDA		LLDA		FaceIt	
	Ev	Te	Ev	Te	Ev	Te	Ev	Te	Ev	Te
R1	13	4	55	43	66	49	66	56	73	**64**
L1	8	8	55	45	77	57	73	**64**	66	52
U1	28	16	53	43	73	52	71	**66**	46	36
D1	33	29	68	55	84	**66**	75	60	37	24
F2	75	70	73	63	82	71	75	66	95	**83**
R2	8	3	42	22	46	29	40	35	46	**36**
L2	4	4	33	27	44	36	48	**47**	46	30
U2	17	15	28	28	35	35	40	**44**	24	23
D2	20	10	31	32	42	32	35	**40**	33	9
Avg.	23	18	49	40	61	47	58	**53**	51	39

Table 4.2: Face Recognition Rates (%).

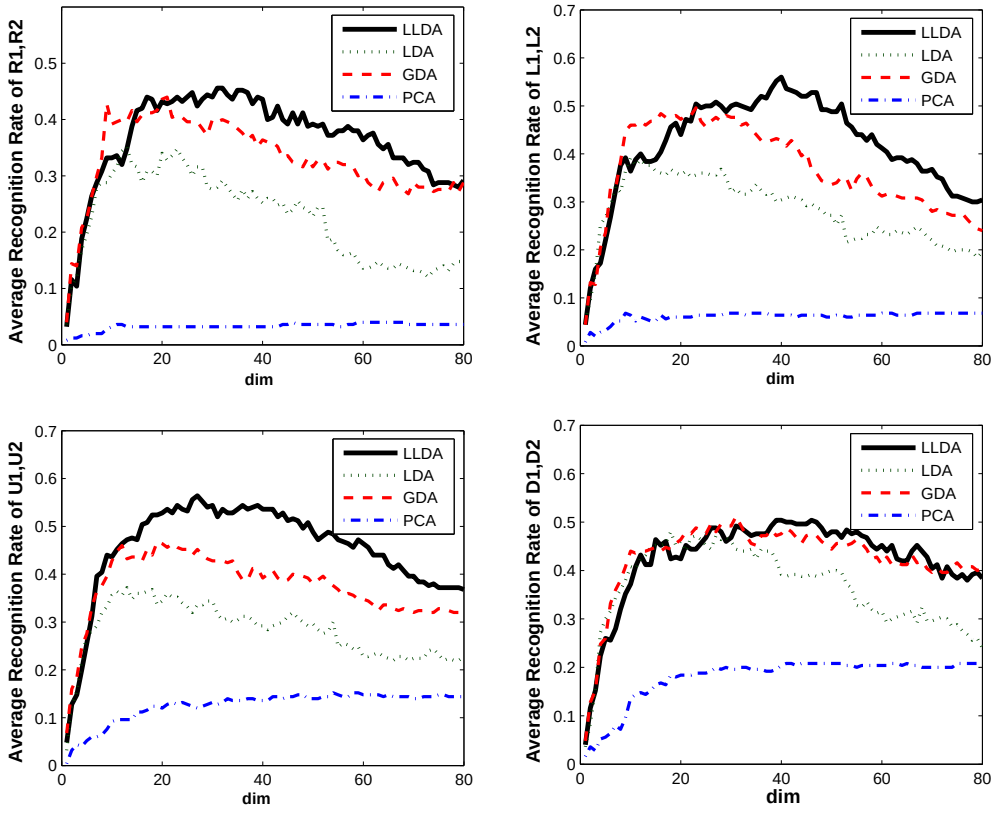
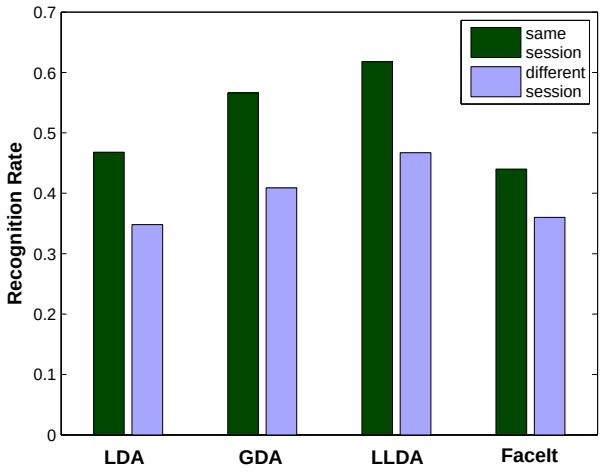
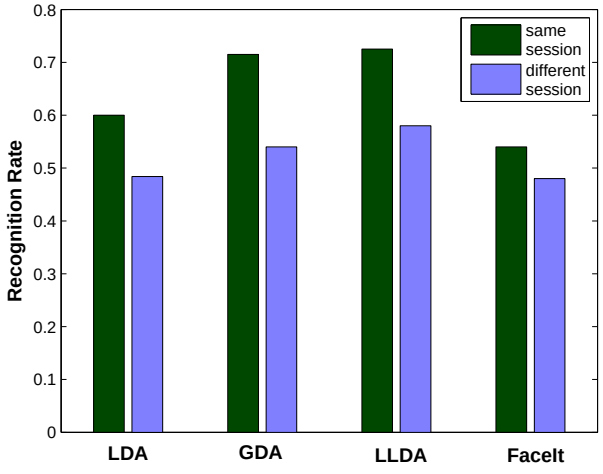


Figure 4.7: The test performance curves (in %) as a function of dimensionality.

between pose groups to recognize a novel view face image for given a single-model face image. The highly non-linear manifolds of data and the minimal training information make this task extremely challenging. In the view-invariant face recognition task with a single model image, the proposed method considerably outperformed all comparative methods: conventional PCA, LDA, LDA mixture, GDA methods and a commercial face recognition system. The recognition accuracy of the proposed method was about 70% in the best scenario. Although this might be not sufficient for a strict security system such



(a)



(b)

Figure 4.8: Recognition rates under aging for different sizes of test population. (a) Recognition rates on the test set consisting of 125 identities . (b) Recognition rates on the test set consisting of randomly chosen 50 identities.

as an automatic access control, this could still facilitate the retrieval task which does not have to get the correct images at the first rank all the time. Note that the proposed method guarantees the winning performance of the LDA-based method in the MPEG-7 protocol as it embodies the LDA as a special case. In addition to its improved accuracy, the proposed method does not suffer from the local-maxima problem and stably converges to a global maximum point and is computationally highly efficient as compared with the conventional non-linear discriminant analysis based on the kernel approach.

For further accuracy improvement, one may exploit more priors on object classes. For example, in view-invariant face recognition, more elaborate regularization using dense facial feature correspondences is expected to promote face class structures that are better separated; similarly to the results of [56]. Correspondence solving, however, is difficult

itself and errors in correspondences seriously degrade the performance of the subsequent recognition methods, as shown in [15]. It may be worth seeking an iterative algorithm of the two steps, the proposed solution and the correspondence-solving so that they help each other. Collecting more representative training samples, i.e. increasing the prototype set here, could be an another way to boost accuracy of the method. As observed in Figure 4.8, greater accuracy could be obtained with more training classes. As it is rash to assume that all sufficient training sets are given initially, and it is time-consuming to learn whenever a new set of training data is given, an efficient update method is required. This is explained in the next chapter.

CHAPTER 5

Incremental Linear Discriminant Analysis Using Sufficient Spanning Set Approximations

A new incremental (or on-line) learning solution for Linear Discriminant Analysis (LDA) is proposed in this chapter. It is often beneficial to learn the LDA bases from a large training set. We have observed that a larger training set delivers better retrieval accuracy by LDA. A complete set of training data may not be practically available initially. The execution of the batch-computation, whenever new training data is presented, is too expensive in terms of both time and space. An efficient update algorithm of LDA is greatly needed to accumulate the information conveyed by new data so that the method's future accuracy is enhanced.

We apply the concept of the *sufficient spanning set* approximation in each update step, i.e. for the between-class scatter, the total scatter and projected data matrices. The algorithm yields a more general and efficient solution for incremental LDA than previous attempts. It also significantly reduces the computational complexity while providing a solution which closely agrees with the batch LDA result. We show two applications of incremental LDA: Firstly, the method is applied to semi-supervised learning by integrating it into an EM framework. Secondly, we apply it to the task of merging large databases which were collected during MPEG-7 standardization for face image retrieval (i.e. the ***Single-to-Single Matching*** task). Note that the proposed framework for on-line learning is also shown useful for recognition tasks with image sets and videos in Chapter 9.

5.1 Drawbacks of Existing Methods

A number of incremental versions of LDA have been suggested which can be applied to on-line learning tasks [68, 120, 149, 219]. Ye et al. [219] proposed an incremental version of LDA which can include only a single new data point in each time step. A further limitation is the computational complexity of the method when the number of classes C is large, as the method involves an eigendecomposition of $C \times C$-dimensional scatter matrices. Pang et al. [149] introduced a scheme for updating the between-class and within-class scatter matrices. However, no incremental method is used for the subsequent LDA

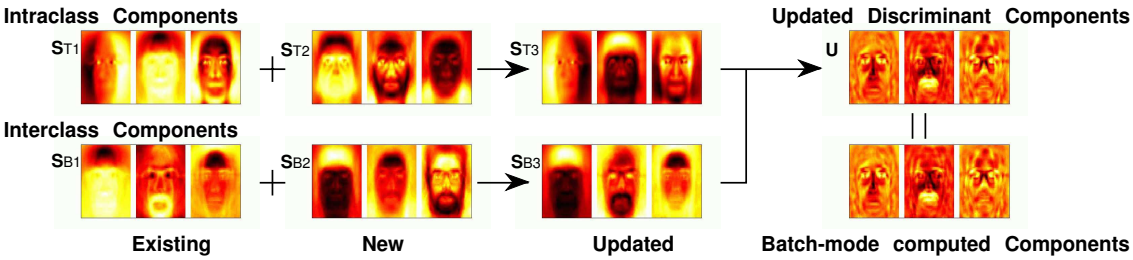


Figure 5.1: On-line update of an LDA basis. The basis computed by the new incremental LDA algorithm (top right) closely agrees with that computed by batch LDA (bottom right). Shown for each scatter matrix $\mathbf{S}_{\mathbf{T},\mathbf{i}}$ and $\mathbf{S}_{\mathbf{B},\mathbf{i}}$ are the first three principal components, which are combined by merging eigenspaces.

steps, i.e. eigenanalysis of the scatter matrices, which remains computationally expensive. Gradient-based incremental learning of a modified LDA was proposed by Hiraoka et al. [68]. Limitations of the method are that it includes only a single new data point at each time step and that it requires the setting of a learning rate. To circumvent the difficulty of incrementally updating the product of scatter matrices, Yan et al. [208] used a modified criterion by computing the difference of the between-class and within-class scatter matrices. This may lead, however, to regularization problems of the two scatter matrices. Lin et al. [120] dealt with the online update of discriminative models for the purpose of object tracking. As their task is binary classification, the discriminative model and the update method are limited to the two-class case.

Inspiration for incremental LDA can be drawn from work on incremental Principal Component Analysis (PCA). Numerous algorithms have been developed to update the eigenbasis as more data samples arrive. Most methods assume, however, a zero mean in updating the eigenbasis except [59, 169] where the update of the mean is handled correctly. The dimension of the eigenproblem can be reduced by using the *sufficient spanning set* (a reduced set of basis vectors spanning the space of most data variation). As the computation of the eigenproblem is cubic in the subspace dimension of the respective scatter matrix, this update scheme is highly efficient.

It is also worth noting the existence of efficient algorithms for kernel PCA and LDA [26, 178]. While studying the incremental learning of such non-linear models is worthwhile, when considering retrieval from large data sets, the computational cost of feature extraction of new samples is as demanding as updating the models [80, 99, 134]. Note also that the LDA method in [178] assumes a small number of classes for the update, similar to [219].

5.2 Overview of the Proposed Method

This study proposes a new solution for incremental LDA, which is accurate as well as efficient in both time and memory. The benefit of the proposed algorithm over other LDA update algorithms [120, 219] lies in its ability efficiently to handle large data sets with many classes. This is particularly important for the face image retrieval task, where hundreds of face classes have to be merged. An example of an LDA basis of face images is shown in Figure 5.1. The result obtained with the incremental algorithm closely agrees with the

batch LDA solution. Note that previous studies have not shown close agreement between incremental and batch LDA solutions [178, 219].

In the proposed method an LDA criterion which is equivalent to the Fisher criterion, namely maximising the ratio of the between-class and the total scatter matrix, is used to keep the discriminative information during the update. First the principal components of the two scatter matrices are efficiently updated and then the discriminant components are efficiently computed from these two sets of principal components. The concept of sufficient spanning sets is applied in each step, making the eigenproblem computation efficient.The algorithm is also memory efficient as it only needs to store the two sets of principal components to avoid losing discriminatory data.

The remainder of this chapter is structured as follows: Section 5.3 presents the new incremental LDA algorithm. In section 5.4 we show how it can be applied to semi-supervised learning within an EM-framework. We discuss the application to other discriminant models in section 5.5. Experimental results for the task of merging face databases are presented in section 5.6. Summary is presented in section 5.7.

5.3 Incremental LDA

As noted by Fukunaga [47], there are equivalent variants of Fisher's criterion used to find the projection matrix $\mathbf{U}$ to maximise class separability of the data set:

$$\max_{\arg \mathbf{U}} \frac{\mathbf{U}^T \mathbf{S}_B \mathbf{U}}{\mathbf{U}^T \mathbf{S}_W \mathbf{U}} = \max_{\arg \mathbf{U}} \frac{\mathbf{U}^T \mathbf{S}_T \mathbf{U}}{\mathbf{U}^T \mathbf{S}_W \mathbf{U}} = \max_{\arg \mathbf{U}} \frac{\mathbf{U}^T \mathbf{S}_B \mathbf{U}}{\mathbf{U}^T \mathbf{S}_T \mathbf{U}}, \tag{5.1}$$

where

$$\mathbf{S}_B = \Sigma_{i=1}^{C} n_i (\mathbf{m}_i - \boldsymbol{\mu})(\mathbf{m}_i - \boldsymbol{\mu})^T \tag{5.2}$$

is the between-class scatter matrix,

$$\mathbf{S}_W = \Sigma_{i=1}^{C} \Sigma_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{5.3}$$

is the within-class scatter matrix,

$$\mathbf{S}_T = \Sigma_{all\ \mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T = \mathbf{S}_B + \mathbf{S}_W \tag{5.4}$$

the total scatter matrix, C the total number of classes, n_i the sample number of class i, $\mathbf{m}_i$ the mean of class i, and $\boldsymbol{\mu}$ the global mean. The algorithm in this paper uses the third criterion in equation 5.1 and separately updates the principal components as the minimal sufficient spanning sets[1] of $\mathbf{S}_B$ and $\mathbf{S}_T$. The scatter matrix approximation with a small number of principal components (corresponding to significant eigenvalues) allows an efficient update of the discriminant components. The $\mathbf{S}_T$ matrix rather than $\mathbf{S}_W$ is used to avoid losing discriminatory data during the update. If we only kept track of the significant principal components of $\mathbf{S}_B$ and $\mathbf{S}_W$, any discriminatory information contained in the null

[1] It is defined as a minimal set of basis vectors spanning the space of most data variation.

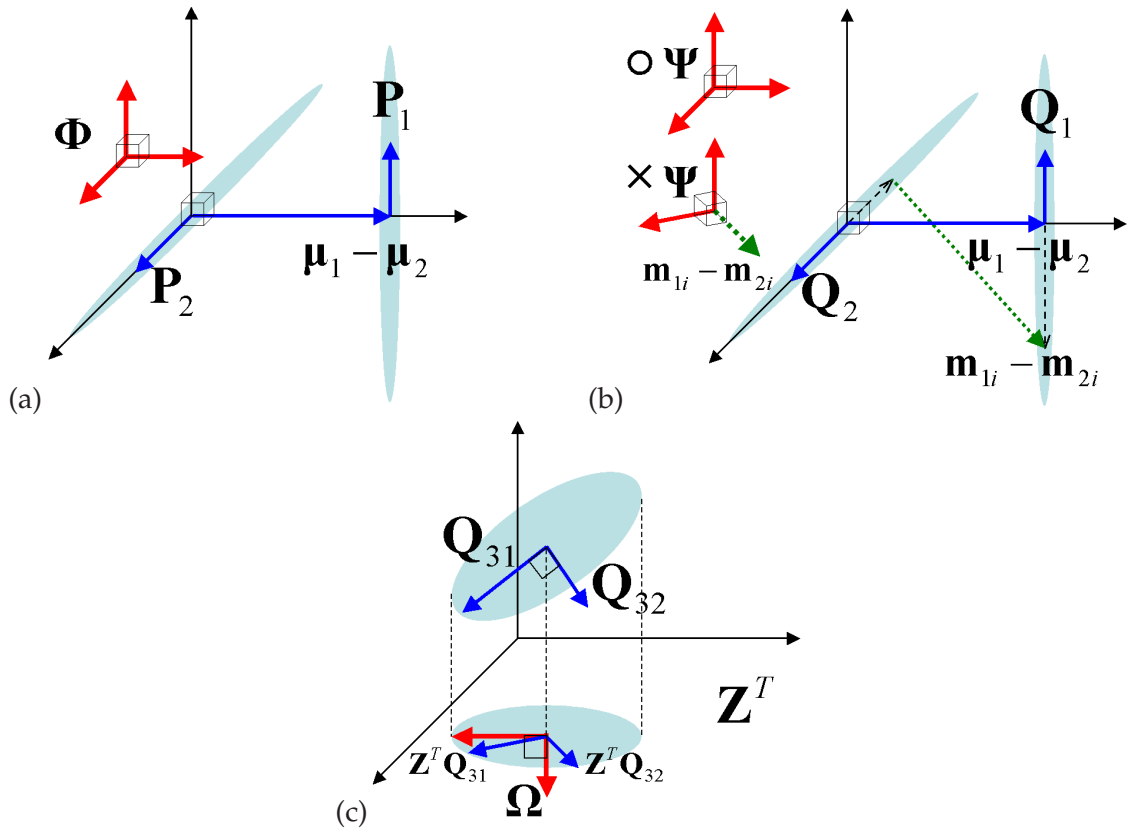


Figure 5.2: Concept of sufficient spanning sets of the total scatter matrix (a), the between-class scatter matrix (b) and the projected matrix (c). The union set of the principal components $\mathbf{P}_1, \mathbf{P}_2$ or $\mathbf{Q}_1, \mathbf{Q}_2$ of the two data sets and the mean difference vector $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ can span the respective total or between-class scatter data space **(left and middle)**. The dimension for the component $\mathbf{m}_{1i} - \mathbf{m}_{2i}$ should not be removed (cross=incorrect) from the sufficient set of the between-class scatter data but retained in the set (circle=correct) **(middle)**. The projection and orthogonalization of the original components $\mathbf{Q}_{31}, \mathbf{Q}_{32}$ yields the principal components of the projected data up to rotation **(right)**. See the corresponding sections for detailed explanations.

space of $\mathbf{S}_W$ would be lost (note that any component in the null space maximises the LDA criterion). However, as $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$ and both $\mathbf{S}_B$ and $\mathbf{S}_W$ are positive semi-definite, vectors in the null space of $\mathbf{S}_T$ are also in the null space of $\mathbf{S}_B$, and are thus ignored in the update.

The three steps of the algorithm are: (1) Update the total scatter matrix $\mathbf{S}_T$, (2) Update the between-class scatter matrix $\mathbf{S}_B$ and (3) from these compute the discriminant components $\mathbf{U}$. These steps are explained in more detail in the following sections.

5.3.1 Updating the total scatter matrix

The total scatter matrix is approximated with a set of orthogonal vectors that span the subspace occupied by the data and represent it with sufficient accuracy. The eigenspace merging algorithm of Hall et al. [59] can be used with the slight modifications ([59] considered merging covariances) in order incrementally to compute the principal components of

the total scatter matrix: Given two sets of data represented by eigenspace models

$$\{\boldsymbol{\mu}_i, M_i, \mathbf{P}_i, \boldsymbol{\Lambda}_i\}_{i=1,2}, \tag{5.5}$$

where $\boldsymbol{\mu}_i$ is the mean, M_i the number of samples, $\mathbf{P}_i$ the matrix of eigenvectors and $\boldsymbol{\Lambda}_i$ the eigenvalue matrix of the i-th data set, the combined eigenspace model $\{\boldsymbol{\mu}_3, M_3, \mathbf{P}_3, \boldsymbol{\Lambda}_3\}$ is computed. Generally only a subset of $d_{T,i}$ eigenvectors have significant eigenvalues and thus only these are stored in $\boldsymbol{\Lambda}_i$ and the corresponding eigenvectors in $\mathbf{P}_i$. We wish to compute the eigenvectors and eigenvalues of the new eigenspace model that satisfy $\mathbf{S}_{T,3} \simeq \mathbf{P}_3\boldsymbol{\Lambda}_3\mathbf{P}_3^T$. The eigenvector matrix $\mathbf{P}_3$ can be represented by a sufficient spanning set (see below for discussion) and a rotation matrix as

$$\mathbf{P}_3 = \boldsymbol{\Phi}\mathbf{R} = h([\mathbf{P}_1, \mathbf{P}_2, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2])\mathbf{R}, \tag{5.6}$$

where $\boldsymbol{\Phi}$ is the orthonormal column matrix spanning the combined scatter matrix, $\mathbf{R}$ is a rotation matrix, and h is an orthonormalization function (e.g. QR decomposition).

Using the sufficient spanning set, the eigenproblem is converted into a smaller eigenproblem as

$$\mathbf{S}_{T,3} = \mathbf{P}_3\boldsymbol{\Lambda}_3\mathbf{P}_3^T \;\Rightarrow\; \boldsymbol{\Phi}^T\mathbf{S}_{T,3}\boldsymbol{\Phi} = \mathbf{R}\boldsymbol{\Lambda}_3\mathbf{R}^T. \tag{5.7}$$

By computing the eigendecomposition on the r.h.s. one obtains $\boldsymbol{\Lambda}_3$ and $\mathbf{R}$ as the respective eigenvalue and eigenvector matrices. After removing nonsignificant components in $\mathbf{R}$ according to the eigenvalues in $\boldsymbol{\Lambda}_3$, the minimal sufficient spanning set is obtained as $\mathbf{P}_3 = \boldsymbol{\Phi}\mathbf{R}$. Note the matrix $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{S}_{T,3}\boldsymbol{\Phi}$ has the reduced size $d_{T,1} + d_{T,2} + 1$ and the dimension of the approximated combined total scatter matrix is $d_{T,3} \leq d_{T,1} + d_{T,2} + 1$, where $d_{T,1}, d_{T,2}$ are the number of the eigenvectors in $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively. Thus the eigenanalysis here only takes $O((d_{T,1} + d_{T,2} + 1)^3)$ computations, whereas the eigenanalysis in batch mode (on the l.h.s. of (5.7)) requires $O(\min(N, M_3)^3)$, where N is the dimension of the input data [2]. See Section 5.3.4 for a more detailed discussion of the time and space complexity.

Discussion. We conclude this section by giving more insight into the sufficient spanning set concept. Generally, given a data matrix $\mathbf{A}$, the sufficient spanning set $\boldsymbol{\Phi}$ can be defined as any set of vectors s.t.

$$\mathbf{B} = \boldsymbol{\Phi}^T\mathbf{A}, \quad \mathbf{A}' = \boldsymbol{\Phi}\mathbf{B} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T\mathbf{A} \simeq \mathbf{A}. \tag{5.8}$$

That is, the reconstruction $\mathbf{A}'$ of the data matrix by the sufficient spanning set should approximate the original data matrix. Let $\mathbf{A} \simeq \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$ where $\mathbf{P}, \boldsymbol{\Lambda}$ are the eigenvector and eigenvalue matrix corresponding to most energy. Then $\mathbf{PR}$ where $\mathbf{R}$ is an arbitrary rota-

[2] When $N \gg M$, the batch mode complexity can effectively be $O(M^3)$ as follows: $\mathbf{S}_T = \mathbf{Y}\mathbf{Y}^T$, where $\mathbf{Y} = [..., \mathbf{x}_i - \boldsymbol{\mu}, ...]$. SVD of $\mathbf{Y}$ s.t. $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T$ yields the eigenspace model of $\mathbf{S}_T$ by $\mathbf{U}$ and $\Sigma\Sigma^T$ as the eigenvector and eigenvalue matrix respectively. $\mathbf{Y}^T\mathbf{Y} = \mathbf{V}\Sigma^T\Sigma\mathbf{V}^T$ as $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. That is, by SVD of the low-dimensional matrix $\mathbf{Y}^T\mathbf{Y}$, the eigenvector matrix is efficiently obtained as $\mathbf{Y}\mathbf{V}\Sigma^{-1}$ and the eigenvalue matrix as $\Sigma^T\Sigma$. This greatly reduces the complexity when obtaining the eigenspace model of a small new data set in batch mode prior to combining.

tion matrix can be a sufficient spanning set:

$$\mathbf{A}' = \mathbf{\Phi}\mathbf{\Phi}^T\mathbf{A} \simeq \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T \simeq \mathbf{A} \tag{5.9}$$

as $\mathbf{R}\mathbf{R}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}$. This also applies to the sufficient spanning set in equation (5.6).

As visualised in Figure 5.2 (a), the union of the two principal components and the mean difference vector can span all data points of the combined set in the three-dimensional space. The principal components of the combined set are found by rotating this sufficient spanning set.

Note that this use of the sufficient spanning set is only possible in the case of merging generative models where the scatter matrix of the union set is represented as the sum of the scatter matrices of the two sets explicitly as

$$\mathbf{S}_{T,3} = \mathbf{S}_{T,1} + \mathbf{S}_{T,2} + M_1M_2/M_3 \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T, \tag{5.10}$$

where $\{\mathbf{S}_{T,i}\}_{i=1,2}$ are the scatter matrices of the first two sets. The method can therefore not be used directly to merge the discriminant components of LDA models.

5.3.2 Updating the between-class scatter matrix

In the update of the total scatter matrix, a set of new vectors is added to a set of existing vectors. The between-class scatter matrix, however, is the scatter matrix of the class mean vectors, see equation (5.12). Not only is a set of new class means added, but the existing class means also change when new samples belong to existing classes. Interestingly, the proposed update can be interpreted as simultaneous incremental (adding new data points) and decremental (removing existing data points) learning (see below).

The principal components of the combined between-class scatter matrix can be efficiently computed from the two sets of between-class data, represented by

$$\{\boldsymbol{\mu}_i, M_i, \mathbf{Q}_i, \mathbf{\Delta}_i, n_{ij}, \boldsymbol{\alpha}_{ij} \,|j = 1, ..., C_i\}_{i=1,2}, \tag{5.11}$$

where $\boldsymbol{\mu}_i$ is the mean vector of the data set i, M_i is the total number of samples in each set, $\mathbf{Q}_i$ are the eigenvector matrices, $\mathbf{\Delta}_i$ are the eigenvalue matrices of $\mathbf{S}_{B,i}$, n_{ij} the number of samples in class j of set i, and C_i the number of classes in set i. The $\boldsymbol{\alpha}_{ij}$ are the coefficient vectors of the j-th class mean vector $\mathbf{m}_{ij}$ of set i with respect to the subspace spanned by $\mathbf{Q}_i$, i.e. $\mathbf{m}_{ij} \simeq \boldsymbol{\mu}_i + \mathbf{Q}_i\boldsymbol{\alpha}_{ij}$. The task is to compute the eigenmodel $\{\boldsymbol{\mu}_3, M_3, \mathbf{Q}_3, \mathbf{\Delta}_3, n_{3j}, \boldsymbol{\alpha}_{3j} \,|j = 1, ..., C_3\}$ for the combined between-class scatter matrix. To obtain the sufficient spanning set for efficient eigen-computation, the combined between-class scatter matrix is represented by the sum of the between-class scatter matrices of the

first two data sets, similar to (5.10). The between-class scatter matrix $\mathbf{S}_{B,i}$ is rendered as

$$\mathbf{S}_{B,i} = \sum_{j=1}^{C_i} n_{ij}(\mathbf{m}_{ij} - \boldsymbol{\mu}_i)(\mathbf{m}_{ij} - \boldsymbol{\mu}_i)^T \tag{5.12}$$

$$= \sum_{j=1}^{C_i} n_{ij}\mathbf{m}_{ij}\mathbf{m}_{ij}^T - M_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T. \tag{5.13}$$

The combined between-class scatter matrix can further be written w.r.t. the original between-class scatter matrices and an auxiliary matrix $\mathbf{A}$ as

$$\mathbf{S}_{B,3} = \mathbf{S}_{B,1} + \mathbf{S}_{B,2} + \mathbf{A} + M_1M_2/M_3 \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T, \tag{5.14}$$

where

$$\mathbf{A} = \sum_{k\in s} \frac{-n_{1k}n_{2k}}{n_{1k} + n_{2k}}(\mathbf{m}_{2k} - \mathbf{m}_{1k})(\mathbf{m}_{2k} - \mathbf{m}_{1k})^T. \tag{5.15}$$

The set $s = \{k|k = 1, ..., c\}$ contains the indices of the common classes of both data sets. The matrix $\mathbf{A}$ needs to be computed only when the two sets have common classes, otherwise it is simply set to zero. If we assume that each between-class scatter matrix is represented by the first few eigenvectors such that $\mathbf{S}_{B,1} \simeq \mathbf{Q}_1\boldsymbol{\Delta}_1\mathbf{Q}_1^T$, $\mathbf{S}_{B,2} \simeq \mathbf{Q}_2\boldsymbol{\Delta}_2\mathbf{Q}_2^T$, the sufficient spanning set for the combined between-class scatter matrix can be similarly set as

$$\boldsymbol{\Psi} = h([\mathbf{Q}_1, \mathbf{Q}_2, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]), \tag{5.16}$$

where the function h is the orthonormalization function used in section 5.3.1. Note that the matrix $\mathbf{A}$ is negative semi-definite and does not add any more dimensions to $\boldsymbol{\Psi}$. As illustrated in Figure 5.2 (b), the sufficient spanning set can be a union set of the two eigencomponents and the mean difference vector. The negative semi-definite matrix $\mathbf{A}$ can conceptually be seen as the scatter matrix of the components to be removed from the combined data. When ignoring the scale factors, the decremental elements are $\mathbf{m}_{2i} - \mathbf{m}_{1i}$. This decreases the data variance along the direction of $\mathbf{m}_{2i} - \mathbf{m}_{1i}$ but the respective dimension should not be removed from the sufficient spanning set. The resulting variance reduction along this direction is taken into account when removing eigencomponents with nonsignificant eigenvalues in the subsequent eigenanalysis.

Let $d_{B,i}$ and N be the dimension of $\mathbf{Q}_i$ and input vectors, respectively. Whereas the eigenanalysis of the combined between-class scatter in batch mode [3] requires $O(\min(N, C_3)^3)$, the proposed incremental scheme requires only $O((d_{B,1} + d_{B,2} + 1)^3)$ computation for solving

$$\mathbf{S}_{B,3} = \boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Delta}_3\mathbf{R}^T\boldsymbol{\Psi}^T \;\Rightarrow\; \boldsymbol{\Psi}^T\mathbf{S}_{B,3}\boldsymbol{\Psi} = \mathbf{R}\boldsymbol{\Delta}_3\mathbf{R}^T, \tag{5.17}$$

where $\mathbf{R}$ is a rotation matrix. Finally, the eigenvectors of the combined between-class scatter matrix, which are memorized for the next update, are obtained by $\mathbf{Q}_3 = \boldsymbol{\Psi}\mathbf{R}$ after the components having zero eigenvalues in $\mathbf{R}$ are removed, i.e. $d_{B,3} \leq d_{B,1} + d_{B,2} + 1$. All

[3] The batch solution of the between-class scatter matrix can be computed using the low-dimensional matrix similarly to the total scatter matrix when $N \gg C$. Note $\mathbf{S}_{B,i} = \mathbf{Y}\mathbf{Y}^T$, $\mathbf{Y} = [..., \sqrt{n_{ij}}(\mathbf{m}_{ij} - \boldsymbol{\mu}_i), ...]$.

Algorithm 1. Incremental LDA (ILDA)

Input: The total and between-class eigenmodels of an existing data set, $\{\mathbf{P}_1, ...\}$, $\{\mathbf{Q}_1, ...\}$ and a new data set
Output: Updated LDA components $\mathbf{U}$

1. Compute $\{\mathbf{P}_2, ...\}, \{\mathbf{Q}_2, ...\}$ from the new data set in batch mode.
2. Update the total scatter matrix for $\{\mathbf{P}_3, ...\}$:
   Compute $\mathbf{S}_{T,3}$ by (5.10) and $\{\mathbf{S}_{T,i}\}_{i=1,2} \simeq \mathbf{P}_i \Lambda_i \mathbf{P}_i^T$.
   Set Φ by (5.6) and compute the principal components $\mathbf{R}$ of $\Phi^T \mathbf{S}_{T,3} \Phi$. $\mathbf{P}_3 = \Phi \mathbf{R}$.
3. Update the between-class scatter for $\{\mathbf{Q}_3, ...\}$:
   Obtain $\mathbf{S}_{B,3}$ from (5.14), $\{\mathbf{S}_{B,i}\}_{i=1,2} \simeq \mathbf{Q}_i \Delta_i \mathbf{Q}_i^T$ and $\mathbf{m}_{ij} \simeq \boldsymbol{\mu}_i + \mathbf{Q}_i \boldsymbol{\alpha}_{ij}$.
   Set Ψ by (5.16) and eigendecompose $\Psi^T \mathbf{S}_{B,3} \Psi$ for the eigenvector matrix $\mathbf{R}$. $\mathbf{Q}_3 = \Psi \mathbf{R}$.
4. Update the discriminant components:
   Compute $\mathbf{Z} = \mathbf{P}_3 \Lambda_3^{-1/2}$ and $\Omega = h([\mathbf{Z}^T \mathbf{Q}_3])$.
   Eigendecompose $\Omega^T \mathbf{Z}^T \mathbf{Q}_3 \Delta_3 \mathbf{Q}_3^T \mathbf{Z} \Omega$ for the eigenvector matrix $\mathbf{R}$. $\mathbf{U} = \mathbf{Z} \Omega \mathbf{R}$.

Table 5.1: Pseudocode of Incremental LDA.

remaining parameters of the updated model are obtained as follows: $\boldsymbol{\mu}_3$ is the global mean updated in Section 5.3.1, $M_3 = M_1 + M_2$, $n_{3j} = n_{1j} + n_{2j}$, $\boldsymbol{\alpha}_{3j} = \mathbf{Q}_3^T(\mathbf{m}_{3j} - \boldsymbol{\mu}_3)$, where $\mathbf{m}_{3j} = (n_{1j}\mathbf{m}_{1j} + n_{2j}\mathbf{m}_{2j})/n_{3j}$.

5.3.3 Updating discriminant components

After updating the principal components of the total scatter matrix and the between-class scatter matrix, the discriminative components are found using the updated total data $\{\boldsymbol{\mu}_3, M_3, \mathbf{P}_3, \boldsymbol{\Lambda}_3\}$ and the updated between-class data $\{\boldsymbol{\mu}_3, M_3, \mathbf{Q}_3, \boldsymbol{\Delta}_3, n_{3j}, \boldsymbol{\alpha}_{3j} \mid j = 1, ..., C_3\}$ using the new sufficient spanning set. Let $\mathbf{Z} = \mathbf{P}_3 \boldsymbol{\Lambda}_3^{-1/2}$, then $\mathbf{Z}^T \mathbf{S}_{T,3} \mathbf{Z} = \mathbf{I}$. As the denominator of the LDA criterion is the identity matrix in the projected space, the optimization problem is to find the components that maximise $\mathbf{Z}^T \mathbf{S}_{B,3} \mathbf{Z}$ s.t. $\mathbf{W}^T \mathbf{Z}^T \mathbf{S}_{B,3} \mathbf{Z} \mathbf{W} = \boldsymbol{\Lambda}$ and the final LDA components are obtained by $\mathbf{U} = \mathbf{Z}\mathbf{W}$. This eigenproblem of the projected data can be solved using the sufficient spanning set defined by

$$\boldsymbol{\Omega} = h([\mathbf{Z}^{\mathrm{T}} \mathbf{Q}_3]). \tag{5.18}$$

See Figure 5.2 (c). The original components are projected and orthogonalized to construct the sufficient spanning set. The principal components of the projected data can be found by rotating the sufficient spanning set. By this sufficient spanning set, the eigenvalue problem

	Batch LDA	Inc LDA
time	$O(NM_3^2$	$O(d_{T,3}^3 + d_{B,3}^3$
	$+min(N, M_3)^3)$	$+Nd_{T,3}d_{B,3})$
space	$O(NM_3 + NC_3)$	$O(Nd_{T,3} + Nd_{B,3})$

Table 5.2: Comparison of time and space complexity. The savings of incremental LDA are significant as usually $M_3 \gg d_{T,3} \geq d_{B,3}$. N is the data dimension and M_3, C_3 are the total number of data points and classes, respectively, $d_{T,i}, d_{B,i}$ are the dimensions of the total and between-class scatter subspaces.

changes into a smaller dimensional eigenvalue problem by

$$\mathbf{Z}^{\mathrm{T}}\mathbf{S}_{B,3}\mathbf{Z} = \mathbf{\Omega}\mathbf{R}\mathbf{\Lambda}\mathbf{R}^T\mathbf{\Omega}^T \;\Rightarrow\; \mathbf{\Omega}^T\mathbf{Z}^T\mathbf{S}_{B,3}\mathbf{Z}\mathbf{\Omega} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^T. \tag{5.19}$$

The final discriminant component is given as

$$\mathbf{Z}\mathbf{W} = \mathbf{Z}\mathbf{\Omega}\mathbf{R}. \tag{5.20}$$

This eigenproblem takes $O(d^3)$ time, where d is the dimension of $\mathbf{\Omega}$, which is equivalent to $d_{B,3}$, the dimension of $\mathbf{Q}_3$. Note that in LDA, $d_{T,3}$, the dimension of $\mathbf{P}_3$ is usually larger than $d_{B,3}$ and therefore the use of the sufficient spanning set further reduces the time complexity of the eigenanalysis: $O(d_{T,3}^3) \rightarrow O(d_{B,3}^3)$. The pseudocode of the complete incremental LDA algorithm is given in Table 5.1.

5.3.4 Time and space complexity

So far we have mainly considered the computational complexity of solving the eigenproblem for merging two data sets. Table 5.2 provides a comparison of the batch and the proposed incremental LDA in total time complexity (considering the necessary matrix products e.g. those in (5.7)) and space complexity, when the additional set is relatively small compared to the existing set, i.e. $M_2 \ll M_1$. The computational saving of the incremental solution compared to the batch version is large as normally $M_3 \gg d_{T,3} \geq d_{B,3}$. Both time and space complexity of the proposed incremental LDA are independent of the size of the total sample set and the total number of classes. The important observation from the face data base merging experiments (see Table 5.3) is that the intermediate dimensions $d_{T,3}$ and $d_{B,3}$ do not increase significantly when new data is successively added.

A more detailed analysis of the total complexity of the method summarized in Table 5.2 is as follows: Clearly, Batch LDA has a space complexity of $O(NM_3 + NC_3)$ and a time complexity of $O(NM_3^2 + \min(N, M_3)^3)$.

In the proposed incremental LDA, for the update of the principal components of the total scatter matrix, we only need to keep track of the data associated with $\{\boldsymbol{\mu}_3, M_3, \mathbf{P}_3, \mathbf{\Lambda}_3\}$ taking $O(Nd_{T,3})$ space. The total process can be partitioned into the merging and the eigenproblem of the new data set. Note that the computation cost of the orthonormalization in (5.6) and the necessary matrix products in (5.7) can be efficiently reduced by exploiting the orthogonality of the eigenvectors [59]. This cost is bounded by $O(Nd_{T,1}d_{T,2})$ and the

eigendecomposition takes $O(d_{T,3}^3)$. The eigenanalysis of the new data set is computed in $O(NM_2^2 + \min(N, M_2)^3)$.

Similarly only $\{\boldsymbol{\mu}_3, M_3, \mathbf{Q}_3, \boldsymbol{\Delta}_3, n_{3j}, \boldsymbol{\alpha}_{3j} \,|j = 1, ..., C_3\}$ is required to be stored for the update of the between-class scatter matrix, taking $O(Nd_{B,3})$. The computational complexity of this update is $O(Nd_{B,1}d_{B,2} + d_{B,3}^3)$, and $O(NC_2^2 + \min(N, C_2)^3)$ for the merging and the eigenanalysis of the new set.

The final LDA components are computed only from the two sets of data above in time $O(Nd_{T,3}d_{B,3})$.

5.4 Semi-supervised incremental learning

This section deals with the LDA update when the class labels of new samples are not given. Unlike incremental learning of generative models [59, 169], discriminative models such as LDA, require the class labels of additional samples for the model update. The proposed incremental LDA can be incorporated into a semi-supervised learning algorithm so that the LDA update can be computed efficiently without the class labels of the additional data set being known. For an overview of semi-supervised learning, including an explanation of the role of unlabeled data, see [237]. Although graph-based methods have been widely adopted for semi-supervised learning [237], the classic mixture model has long been recognized as a natural approach to modelling unlabelled data. A mixture model makes predictions for arbitrary new test points and typically has a relatively small number of parameters. Additionally mixture models are compatible with the proposed incremental LDA model assuming multiple Gaussian-distributed classes [47]. Here, classic EM-type learning is employed to generate the probabilistic labels of the new samples. Running EM in the updated LDA subspaces allows for more accurate estimation of the class labels. We iterate the E-step and M-step with all data vectors projected into the LDA subspaces (similarly to [205]), which are incrementally updated in an intermediate step. The class posterior probabilities of the new samples are set to the probabilistic labels.

Incremental LDA with EM. The proposed EM algorithm employs a generative model with the most recent LDA transformation $\mathbf{U}$ by

$$P(\mathbf{U}^T\mathbf{x}|\Theta) = \sum_{k=1}^{C} P(\mathbf{U}^T\mathbf{x}|C_k;\Theta_k)P(C_k|\Theta_k), \tag{5.21}$$

where class $C_k, k = 1, ..., C$ is parameterized by $\Theta_k, k = 1, ..., C$, and $\mathbf{x}$ is a sample of the initial labeled set $\mathcal{L}$ and the new unlabeled set $\mathcal{U}$. The E-step and M-step are iterated to estimate the MAP model over the projected samples $\mathbf{U}^T\mathbf{x}$ of the labeled and unlabeled sets. The proposed incremental LDA is performed every few iterations on the data sets $\{\mathbf{x}_j, y_j|\mathbf{x}_j \in \mathcal{L}\}$ and $\{\mathbf{x}_j, y'_{jk}|\mathbf{x}_j \in \mathcal{U}, k = 1, ..., C\}$, where y_j is the class label and y'_{jk} is the probabilistic class label given as the class posterior probability

$$y'_{jk} = P(C_k|\mathbf{U}^T\mathbf{x}_j). \tag{5.22}$$

We set

$$\mathbf{m}_{2i} = \frac{\sum_j \mathbf{x}_j y'_{ji}}{\sum_j y'_{ji}}, \quad n_{2i} = \sum_{j=1}^{M_2} y'_{ji}. \tag{5.23}$$

for the update of the between-class scatter matrix. All other steps for incremental LDA are identical to the description in Section 5.3 as they are independent of class label information.

Discussion. Using a common covariance matrix Θ_k for all class models rather than C class covariance matrices is more consistent with the assumption of LDA [47] and can additionally save space and computation time during the M-step. The common covariance matrix can conveniently be updated by $\mathbf{U}^T(\mathbf{S}_{T,3} - \mathbf{S}_{B,3})\mathbf{U}/M_3$, where $\mathbf{S}_{T,3}, \mathbf{S}_{B,3}$ are the combined total and between-class scatter matrices, which are kept track of in the incremental LDA as the associated first few eigenvector and eigenvalue matrices. The other parameters of Θ are also obtained from the output of the incremental LDA algorithm.

So far it is assumed that the new data points are in one of the existing classes, but this may not always be the case. Samples with new class labels can be screened out so that the LDA update is not biased to those samples by

$$y'_{jk} = P(C_k|\mathbf{U}^T\mathbf{x}_j) \cdot P(\mathbf{C}|\mathbf{U}^T\mathbf{x}_j), \tag{5.24}$$

where $P(\mathbf{C}|\mathbf{U}^T\mathbf{x}_j)$ denotes a probability of a hyper class. We can set this probability as close to zero for samples with new class labels.

5.5 Incrementally updating LDA-like discriminant models

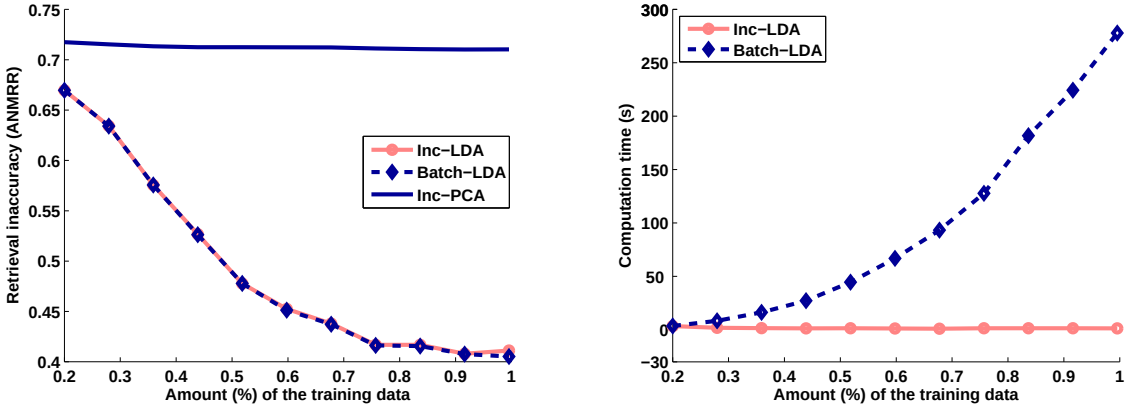
The proposed algorithm is general and can be applied to other incremental learning problems that seek to find discriminative components by maximising the ratio of two covariance or correlation matrices [6, 184, 145]. The method of using the sufficient spanning set for the three steps, the component analysis of the two matrices in the numerator and the denominator, respectively, and for the discriminant component computations, allows for efficient incremental learning. Note that the number of input vectors for the numerator matrices in the methods [184, 145] is often as large as those for the denominator matrices. In these cases the proposed incremental algorithm is still efficient, whereas the incremental LDA algorithm by Ye et al. [219] is no longer suitable as it assumes that the number of input vectors for the scatter matrix in the numerator, i.e. the number of classes, is small.

5.6 Experimental results

The algorithm was applied to the task of face image retrieval from a large database. All experiments were performed on a 3 GHz Pentium 4 PC with 1GB RAM.

LDA update	M_3 [# images]	C_3 [# classes]	$d_{T,3}$ [$\dim(S_{t,3})$]	$d_{B,3}$ [$\dim(S_{b,3})$]
1[first] – 10[final]	465–2315	93–463	158–147	85–85

Table 5.3: Efficient LDA update. Despite the large increase in the number of images and classes, the number of required principal components, $d_{T,3}$ and $d_{B,3}$, remains small during the update process implying that computation time remains low.



(a) (b)

Figure 5.3: Database merging experiments for the *MPEG+XM2VTS* data set. The solution of incremental LDA closely agrees to the batch solution while requiring much lower computation time. (a) Retrieval inaccuracy, ANMRR is 0 when all ground truth images are ranked on top, and 1 when none of the ground truth images are ranked among the first m images. (b) Computational cost.

5.6.1 Database and protocol

In the experiments we followed the protocols of evaluating face descriptors for MPEG-7 standardization [99]. Many MPEG-7 proposals, including the winning method, have adopted LDA features as their descriptors [80, 99]. A descriptor vector is extracted without knowledge of the test subject's identity, i.e. its statistical basis should be generated from images of subjects other than those in the test set. As it is necessary to learn the LDA basis from a very large training set, which may not be available initially, the proposed algorithm can be used to successively update the LDA basis as more data becomes available. An experimental face database was obtained consisting of the version 1 MPEG data set (635 persons, 5 images per person), the *Altkom* database (80 persons, 15 images per person), the *XM2VTS* database (295 persons, 5 images per person), and the *BANCA* database (52 persons, 10 images per person). The version 1 MPEG data set itself consists of several public face sets (e.g. *AR*, *ORL*). All 6370 images in the database were normalized to 46×56 pixels using manually labeled eye positions. The images for the experiments were strictly divided into training and test sets. All basis vectors were extracted from the training set. All test images were used as query images to retrieve other images of the corresponding persons (called ground truth images) in the test data set. As a measure of retrieval performance, we used the average normalized modified retrieval rate (ANMRR) [134]. The ANMRR is 0 when images of the same person (ground truth labeled) are ranked on top, and it is 1 when all images are ranked outside the first m images ($m = 2N_G$, where N_G is

the number of ground truth images in the test data set).

5.6.2 Results

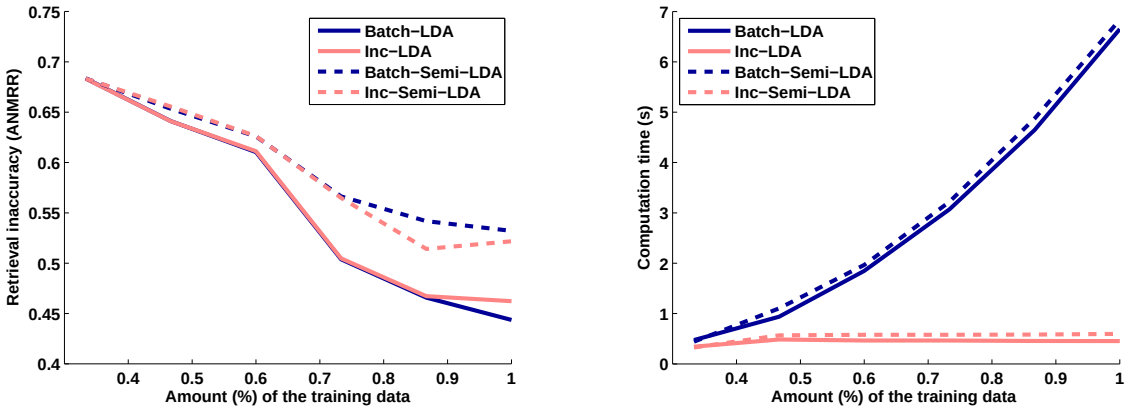
The training set was further partitioned into an initial training set and several new sets which are added successively for re-training. We used the combined set of MPEG and *XM2VTS* database (the total number of classes is 930) for the experiment where the new sets contain the images of new classes. We also performed the experiments for the *Altkom* and *BANCA* database separately where the additional sets contain new images of the existing classes of the initial training set. The proposed incremental LDA yielded nearly the same solution as batch LDA for both scenarios. The basis images of LDA of the incremental and batch versions are compared in Figure 5.1. The accuracy of the incremental solution can be seen in Figure 5.3 (a). Incremental LDA yields essentially the same accuracy as batch LDA, provided enough components are stored of the total and between-class scatter matrices. This is an accuracy vs. speed trade-off: using fewer components is clearly beneficial in terms of computational cost. The subspace dimensions for incremental learning were chosen from the eigenvalue plots by setting a fixed threshold on the variance of each component (similar results were obtained by choosing the first components that contain a specified fraction of the total variance)[4]. Table 5.3 shows the number of components selected during the experiment using the *MPEG+XM2VTS* data set. Even if the total number of images or classes increases, the number of components does not increase significantly, thus saving time and space (See section 5.3.4). The computational costs of the batch and the incremental version are compared in Figure 5.3 (b). Whereas the computational cost of the batch version increases significantly as data is successively added, the cost of the incremental solution remains low. Note that the cost of incremental LDA is comparable to that of incremental PCA while giving a much higher retrieval accuracy, as shown in Figure 5.3 (a). Incremental PCA did not significantly increase retrieval accuracy.

Figure 5.4 shows the result of comparing the proposed semi-supervised incremental LDA solution with the LDA solution when the correct class labels are provided. For this experiment the projection of all data points into the LDA subspace was performed once with the most recent LDA components before the EM iteration, and the incremental LDA with the probabilistic labels was carried out after EM converged, typically after ten iterations. The solution boosted the retrieval accuracy even without the class labels and its incremental solution yielding the same solution as the batch version. The cost of semi-supervised LDA is slightly higher than that of incremental LDA, but still far lower than any batch-mode computation.

5.7 Summary

The proposed incremental LDA method allows highly efficient learning to adapt to new data sets in a face image retrieval task (i.e. ***Single-to-Single*** matching problem), where thousands of face classes are merged in training. The solution exploits the low-rank matrix approximation. That face images are well confined to low-dimensional subspaces makes

[4]Note that accuracy of LDA is dependent on dimensionality of intermediate components (total scatter matrix) and final components (discriminant components). These dimensions of ILDA were set to be the same as those of batch LDA.



(a) (b)

Figure 5.4: Performance of semi-supervised incremental LDA. Semi-supervised incremental LDA decreases the error rate without the class labels of new training data being available, while being as time-efficient as incremental LDA with given labels. (a) Retrieval inaccuracy (ANMRR), (b) computational costs for the *Altkom* database. Similar results were obtained for the *BANCA* database.

this approximation effective. When provided the sufficient components spanning most energy of the data space, our method yields an accurate LDA solution. In the experiments on 11,845 face images, the solution closely agreed with the batch LDA result with far lower complexity in both time and space. Compared with previous attempts, this is a more general approach. We provided a conceptual comparison with related methods, arguing why none of them is suited to the on-line learning problem with a large number classes. Owing to its generality, the incremental LDA algorithm could be incorporated into a classic semi-supervised learning framework and to many other problems in which LDA-like discriminant components are required.

Directions for future research include the extension to the non-linear case, adaptive learning with a time-decaying function and using temporal information for more efficient semi-supervised learning. In particular, further studies in semi-supervised learning would be very interesting. The proposed semi-supervised incremental learning method may be useful for gradual pattern changes in steps. An ideal update method should be robust about outliers in a set of new samples and should exploit discriminative information at maximum in semi-supervised fashion. The concept of Active Learning may be also useful for robust update while minimising user intervention in the update. In the recognition task involving image sets, time-efficient learning is important over a large volume of image sets, which may be increasing over time. This will be addressed in Chapter 9.

Part II

Discriminant Analysis for Set-to-Set and Video-to-Video Matching

CHAPTER 6

Discriminant Analysis of Image Set Classes Using Canonical Correlations

This chapter addresses object recognition problems with image sets (or ensembles). The image sets may be collected not only from video but also sparse and unordered observations representing variations in an object's appearance. A key matter in robust recognition is how to represent and match image sets, i.e. ***Set-to-Set Matching***. Rather than conventional probability density based set-matching and ad-hoc assembly methods, the proposed method is based on a subspace-based set similarity, which facilitates robust set classification about pattern variations.

The benefits of using Canonical Correlation Analysis (CCA) as an image ensemble similarity is demonstrated and a novel discriminant analysis method of image sets based on CCA is proposed. We develop an optimal linear discriminant function which transforms input images so that the transformed image sets are best separated class-wise in terms of canonical correlations. The optimal transformation is found by a novel iterative optimization. An alternative and simpler method for discriminant analysis of image sets is also proposed by a classic method called orthogonal subspace method (OSM) [145]. To our knowledge, the close relationship of the orthogonal subspace method and canonical correlations has not been explored before. The proposed two methods are evaluated on various object recognition problems using face image sets with arbitrary motion captured under different illuminations and image sets of five hundred general objects taken at different views. The methods are also applied to object category recognition using ETH-80 database [117]. The proposed methods are shown to outperform the state-of-the-art methods not only in accuracy but also in time-efficiency. Note that the proposed discriminant analysis of image sets is readily plugged into the task of ***Video-to-Video Matching*** for action classification in the following chapters.

The chapter is organized as follows. Canonical Correlation Analysis is explained for image-set similarity in Section 6.1. Section 6.2 highlights the problem of discriminant analysis over sets and presents a novel optimal solution. In Section 6.3, the orthogonal subspace method is explained and related both to the proposed method and the prior art. The experimental results and their discussion are presented in Section 6.4 and the summary in Section 6.5.

(a) Two sets (top and bottom) contain images of a 3D object taken from different viewpoints but with a certain overlap in their views.

(b) Two face image sets (top and bottom) collected from videos taken under different illumination settings. Face patterns of the two sets vary in both lighting and pose.

Figure 6.1: Examples of image sets. The sets contain different pattern variations caused by different views and lighting.

6.1 Canonical Correlation Analysis as Image-Ensemble Similarity

The objective of this work is to classify an unknown set of images (or generally vectors) to one of the training classes, each also represented by image sets. The set of images may represent a variation in an object's appearance, for e.g. caused by object or camera viewpoint change, deformations, lighting variations, as shown in Figure 6.1.

When two sets contain images of an object taken from different viewpoints (but with a certain overlap in views) as shown in Figure 6.1 (a), statistical characteristics such as mean or variances of the two sets differ significantly. If each set was modelled by a probability density function and matched with that of the other set by for e.g. Kullback-Leibler Divergence (KLD) [27, 167], the two sets would return a low-similarity despite the same object contained. Due to the difficulty of parameter estimation under limited training data (typically a small number of images is in each set), the probabilistic density-based methods easily fail. Rather, a less constrained (i.e. more flexible) matching is needed to absorb large intra-class variations of image sets. Of course, there is a compromise issue between minimising intra-class variation and maximising inter-class variation for optimal classification.

Canonical Correlation Analysis, which is a classical method of inspecting linear relations between two random variables [71, 79, 51, 11], can yield flexible and yet descriptive set matching. A manifold of each image set can be effectively captured by a low-dimensional subspace and measuring angles between two low-dimensional subspaces gives an affine-invariant matching.

As explained in Chapter 3, 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

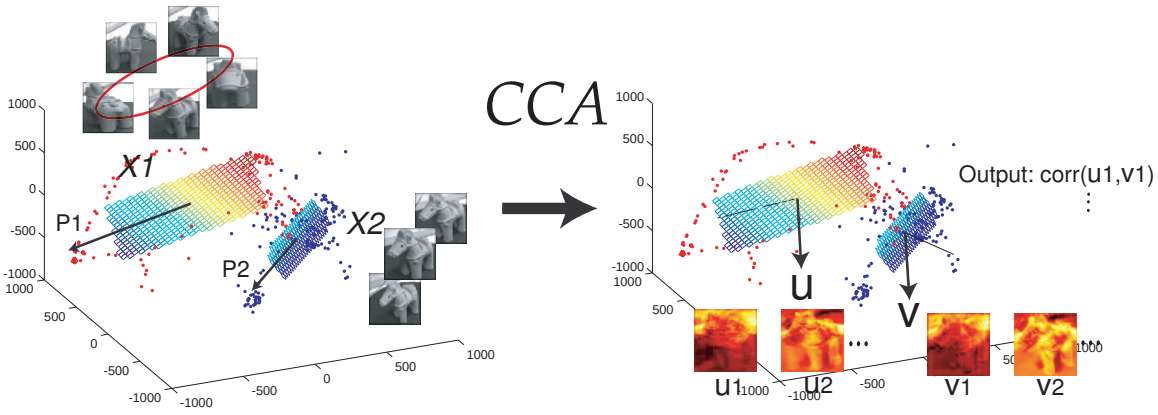


Figure 6.2: Conceptual illustration of canonical correlations. Two sets are represented as linear subspaces which are planes here. The principal components of the subspaces are P_1, P_2. Canonical vectors (u, v) on the planes are found to yield maximum correlation.

uniquely defined as:

$$\cos\theta_i = \max_{\mathbf{u}_i \in \mathcal{L}_1} \max_{\mathbf{v}_i \in \mathcal{L}_2} \mathbf{u}_i^T \mathbf{v}_i \tag{6.1}$$

subject to $\mathbf{u}_i^T \mathbf{u}_i = \mathbf{v}_i^T \mathbf{v}_i = 1,\ \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0,\ i \neq j$.

There are many equivalent ways to solve this problem. The Singular Value Decomposition (SVD) solution [11] is as follows: Assume that $\mathbf{P}_1 \in \mathbb{R}^{N \times d}$ and $\mathbf{P}_2 \in \mathbb{R}^{N \times d}$ $(N \gg d)$ form unitary orthogonal bases (eigenvectors in our study) for two linear subspaces, $\mathcal{L}_1$ and $\mathcal{L}_2$. Let the SVD of $\mathbf{P}_1^T \mathbf{P}_2 \in \mathbb{R}^{d \times d}$ be

$$\mathbf{P}_1^T \mathbf{P}_2 = \mathbf{Q}_{12} \mathbf{\Lambda} \mathbf{Q}_{21}^T \quad s.t. \quad \mathbf{\Lambda} = diag(\sigma_1, ..., \sigma_d) \tag{6.2}$$

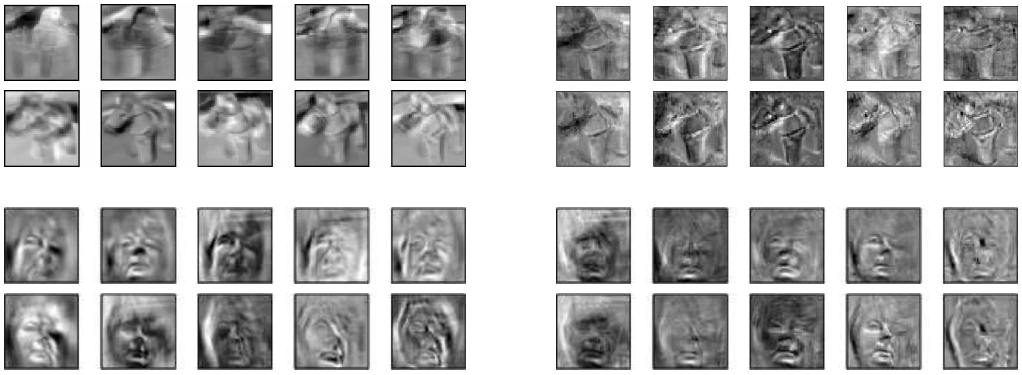
where $\mathbf{Q}_{12}^T \mathbf{Q}_{12} = \mathbf{Q}_{21}^T \mathbf{Q}_{21} = \mathbf{Q}_{12} \mathbf{Q}_{12}^T = \mathbf{Q}_{21} \mathbf{Q}_{21}^T = \mathbf{I}_d$. Canonical correlations are the singular values and the associated canonical vectors, whose correlations are defined as canonical correlations, are given by

$$\mathbf{U} = \mathbf{P}_1 \mathbf{Q}_{12} = [\mathbf{u}_1, ..., \mathbf{u}_d],\ \mathbf{V} = \mathbf{P}_2 \mathbf{Q}_{21} = [\mathbf{v}_1, ..., \mathbf{v}_d] \tag{6.3}$$

Canonical vectors are orthonormal in each subspace and $\mathbf{Q}_{12}, \mathbf{Q}_{21}$ can be seen as rotation matrices of $\mathbf{P}_1, \mathbf{P}_2$. An equivalence between the SVD solution and the previous method called MSM [207] is given in Appendix C. The complexity of SVD of a $d \times d$ dimensional matrix is very low. The concept is represented in Figure 6.2.

Affine invariance of CCA.

A set of object images is generally well confined to a low-dimensional subspace which retains most of the energy of the set. See Figure 6.3 for the principal components of the sets shown in Figure 6.1. The first few components mainly correspond to view changes or illumination changes of the objects in the image sets. The canonical vectors computed from the pairwise image sets are visualised in Figure 6.3. Note that the canonical vectors well capture the mutual information (for e.g. a particular view and illumination of the objects)



(a) (b)

Figure 6.3: Principal components vs. canonical vectors. (a) The first 5 principal components computed from the four image sets shown in Figure 6.1. The principal components of the different image sets are significantly different. (b) The first 5 canonical vectors of the four image sets, which are computed for each pair of the two image sets of the same object. Every pair of canonical vectors (each column) $\mathbf{U}, \mathbf{V}$ well captures the common modes (views and illuminations) of the two sets containing the same object. The pairwise canonical vectors are quite similar. The canonical vectors of different dimensions $\mathbf{u}_1, ...\mathbf{u}_5$ and $\mathbf{v}_1, ..., \mathbf{v}_5$ represent different pattern variations e.g. in pose or lighting.

between the pairwise sets yielding high correlations. Canonical vectors are very much pairwise alike despite the data changes across the sets. Intuitively, the first pair of canonical correlation tells us how close are the closest vectors from two subspaces. Similarly, the higher canonical correlations tell us about the proximity of vectors of the two subspaces in other dimensions (perpendicular to the previous ones) of the embedding space. Note that the vectors are represented by any linear combinations of basis vectors of the subspaces.

A key function in using CCA in matching high-dimensional vector sets is its affine invariance, which allows great flexibility and yet keeps sufficient discriminative information (See Chapter 3 for the proof on the affine-invariance of CCA). As observed above, images are well-constrained to lie on low-dimensional subspaces. The CCA effectively places a uniform prior over the subspaces and provides invariant matching of the image sets to the pattern variations subject to the subspaces. Note that the canonical vectors obtained were visually similar in each pair despite the large changes of pose and illuminations across the sets in Figure 6.3.

On the other hand, the canonical vectors computed from the sets of two distinct classes are not alike as shown in Figure 6.4 (b). Although the two sets were captured under the same lighting condition, CCA returned low correlations. Again, the canonical vectors in Figure 6.4 (a) computed from the two images sets of the same person are closely similar. The inter-/intra-class examples show that the canonical correlations can be good discriminative features for classification.

6.2 Discriminant analysis for Canonical Correlation analysis (DCC)

As shown above, the canonical correlation could be a promising measure of image-set similarity for object recognition. It helps a robust object recognition solution to variations

(a) (b)

Figure 6.4: Canonical Vectors of Same Class and Different Classes. The first 3 pairs (top and bottom rows) of canonical vectors for a comparison of two linear subspaces corresponding to the same (a) and different individuals (b). In the former case, the most similar modes of pattern variation, represented by canonical vectors, are closely similar *in spite of different illumination conditions* used in data acquisition. On the other hand, the canonical vectors in the latter case are clearly dissimilar despite the sequences captured in the same environment.

in observation data. Note, however, that the classical canonical correlation analysis does not exploit image set class information and thus is not optimal in view of classification. In this section, the discriminant analysis method is developed with the canonical correlation measure for optimal set classification. The optimal discriminant function is learnt so that the transformed image sets are best separated by CCA.

6.2.1 Nonparametric Discriminant Analysis

As explained in Chapter 3, the nonparametric discriminant analysis developed in Nearest Neighbor sense [20] defines the two scatter matrices as

$$\mathbf{B} = \frac{1}{M}\sum_{i=1}^{M} w_i(\mathbf{\Delta}_i^B)(\mathbf{\Delta}_i^B)^T, \quad \mathbf{W} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{\Delta}_i^W)(\mathbf{\Delta}_i^W)^T \tag{6.4}$$

where $\mathbf{\Delta}_i^B = \mathbf{x}_i - \mathbf{x}_i^B$, $\mathbf{\Delta}_i^W = \mathbf{x}_i - \mathbf{x}_i^W$, $\mathbf{x}^B = \{\mathbf{x}' \in \overline{C}_c \,|\, \|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{z} - \mathbf{x}\|, \forall \mathbf{z} \in \overline{C}_c\}$ and $\mathbf{x}^W = \{\mathbf{x}' \in C_c \,|\, \|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{z} - \mathbf{x}\|, \forall \mathbf{z} \in C_c\}$. w_i is a sample weight in order to deemphasise samples away from class boundaries. The magnitude of a data vector is often normalized so that $|\mathbf{x}| = 1$. As $trace(AB) = trace(BA)$ for any matrix A, B and $|\mathbf{x}| = 1$, $trace(\mathbf{W})$ in (6.4) equals $\frac{1}{M} trace(\sum_i 2(1 - \mathbf{x}_i^T \mathbf{x}_i^W))$. The problem of minimising $trace(\mathbf{W})$ can be changed into the maximisation of $trace(\mathbf{W}')$ and similarly the maximisation of $trace(\mathbf{B})$ into the minimisation of $trace(\mathbf{B}')$, where

$$\mathbf{B}' = \sum_i \mathbf{x}_i^T \mathbf{x}_i^B, \quad \mathbf{W}' = \sum_i \mathbf{x}_i^T \mathbf{x}_i^W \tag{6.5}$$

and $\mathbf{x}_i^B, \mathbf{x}_i^W$ indicate the closest between-class and within-class vectors of a given vector $\mathbf{x}_i$. Note the weight w_i is omitted for simplicity and the total number of training sets M does not change the direction of the desired components. We now see the optimization problem of classical NDA defined by correlations of individual pairwise vectors.

Rather than dealing with *correlations* of every pair of *vectors*, in the proposed method we exploit *canonical correlations* of pairwise *vector sets*.

6.2.2 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 $\mathbf{P}_i \in \mathbb{R}^{N\times d}$ 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 $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_n] \in \mathbb{R}^{N\times n}$, where $n \leq N$, $|\mathbf{t}_i| = 1$ s.t. $\mathbf{T} : \mathbf{X}_i \rightarrow \mathbf{Y}_i = \mathbf{T}^T\mathbf{X}_i$. The matrix $\mathbf{T}$ transforms images so that the transformed image sets are class-wise more discriminative using canonical correlations.

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

$$\mathbf{Y}_i\mathbf{Y}_i^T = (\mathbf{T}^T\mathbf{X}_i)(\mathbf{T}^T\mathbf{X}_i)^T \simeq (\mathbf{T}^T\mathbf{P}_i)\mathbf{\Lambda}_i(\mathbf{T}^T\mathbf{P}_i)^T \tag{6.6}$$

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 6.2.3 for details.

Set Similarity. The similarity of any two transformed data sets represented by $\mathbf{T}^T\mathbf{P'}_i$, $\mathbf{T}^T\mathbf{P'}_j$ is defined as the sum of canonical correlations by

$$F_{ij} = \max_{\mathbf{Q}_{ij}, \mathbf{Q}_{ji}} \text{tr}(M_{ij}), \tag{6.7}$$

$$M_{ij} = \mathbf{Q}_{ij}^T\mathbf{P'}_i^T\mathbf{T}\mathbf{T}^T\mathbf{P'}_j\mathbf{Q}_{ji} \quad \text{or} \quad \mathbf{T}^T\mathbf{P'}_j\mathbf{Q}_{ji}\mathbf{Q}_{ij}^T\mathbf{P'}_i^T\mathbf{T}, \tag{6.8}$$

as $\text{tr}(AB) = \text{tr}(BA)$ for any matrix A, B. $\mathbf{Q}_{ij}, \mathbf{Q}_{ji}$ are the rotation matrices similarly defined in the SVD solution of canonical correlations (6.2) with the two transformed subspaces.

Discriminant Function. The discriminative function (or matrix) $\mathbf{T}$ is found to maximise the similarities of any pairs of within-class sets while minimising the similarities of pairwise sets of different classes. Matrix $\mathbf{T}$ is defined with the objective function J by

$$\mathbf{T} = \arg\max_{\mathbf{T}} J = \arg\max_{\mathbf{T}} \frac{\sum_{i=1}^m \sum_{k\in W_i} F_{ik}}{\sum_{i=1}^m \sum_{l\in B_i} F_{il}} \tag{6.9}$$

where the indices are defined as $W_i = \{j \,|\mathbf{X}_j \in C_i\}$ and $B_i = \{j \,|\mathbf{X}_j \notin C_i\}$. That is, the two index sets W_i, B_i denote, respectively, the within-class and between-class sets for a given set of class i, by analogy to [20]. See Figure 6.5 for conceptual illustration of the

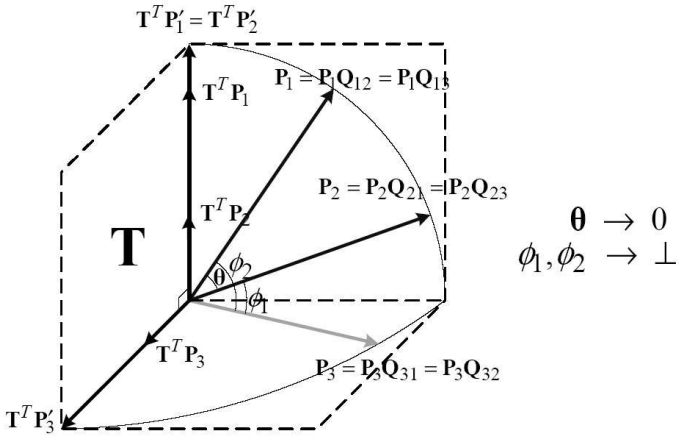


Figure 6.5: Conceptual illustration of the proposed method. Here are drawn the three sets represented by the basis vector matrices $\mathbf{P}_i,\ i = 1, ..., 3$. We assume that the two sets $\mathbf{P}_1, \mathbf{P}_2$ are within-class sets and the third one is coming from the other class. Canonical vectors $\mathbf{P}_i\mathbf{Q}_{ij},\ i = 1, ..., 3, j \neq i$ are equivalent to basis vectors $\mathbf{P}_i$ in this simple drawing where each set occupies a one-dimensional space. Basis vectors are projected on the discriminative subspace by $\mathbf{T}$ and normalized such that $|\mathbf{T}^T\mathbf{P}'| = 1$. Then, the principal angle of within-class sets, θ becomes zero and the angles of between-class sets, ϕ_1, ϕ_2 are maximised.

problem indicated. In the discriminative subspace represented by $\mathbf{T}$, canonical correlations of within-class sets are to be maximised and canonical correlations of between-class sets minimised.

6.2.3 Iterative Learning

The optimization problem of $\mathbf{T}$ involves the variables $\mathbf{Q}, \mathbf{P}'$ 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. 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 $\mathbf{\Phi}_i \in \mathbb{R}^{N\times d}$ is the orthonormal matrix composed by the first d columns and $\mathbf{\Delta}_i \in \mathbb{R}^{d\times d}$ is the $d \times d$ invertible upper-triangular matrix. From (6.6), $\mathbf{Y}_i = \mathbf{T}^T\mathbf{P}_i\sqrt{\mathbf{\Lambda}_i} = \mathbf{\Phi}_i\mathbf{\Delta}_i\sqrt{\mathbf{\Lambda}_i}$. 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,

$$\mathbf{\Phi}_i = \mathbf{T}^T(\mathbf{P}_i\mathbf{\Delta}_i^{-1}) \quad \rightarrow \quad \mathbf{P}'_i = \mathbf{P}_i\mathbf{\Delta}_i^{-1}. \tag{6.10}$$

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} defined in the left of (6.8) can conveniently

be 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

$$\mathbf{P'}_i^T \mathbf{T}\mathbf{T}^T \mathbf{P'}_j = \mathbf{Q}_{ij} \mathbf{\Lambda} \mathbf{Q}_{ji}^T \tag{6.11}$$

where $\mathbf{\Lambda}$ is a singular matrix and $\mathbf{Q}_{ij}, \mathbf{Q}_{ji}$ are orthogonal rotation matrices. Note that the matrices which are Singular-Value decomposed have only d^2 elements.

Computation of T. The optimal discriminant transformation matrix $\mathbf{T}$ is computed for given $\mathbf{P'}_i$ and $\mathbf{Q}_{ij}$ by using the definition of M_{ij} in the right of (6.8) and (6.9). With $\mathbf{T}$ being on the outside of the matrix product M_{ij}, it is convenient to solve for. The discriminative function is found by

$$\mathbf{T} = \max_{arg\mathbf{T}} \text{tr}(\mathbf{T}^T \mathbf{S_b} \mathbf{T}) / \text{tr}(\mathbf{T}^T \mathbf{S_w} \mathbf{T}) \tag{6.12}$$

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

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

where $B_i = \{j \,|\mathbf{X}_j \notin C_i\}$ and $W_i = \{j \,|\mathbf{X}_j \in C_i\}$. Note that no loss of generality is incurred from (6.9) 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 generalised 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 of $(\mathbf{S_w})^{-1} \mathbf{S_b}$. Note also that the proposed learning can avoid a singular case of $\mathbf{S_w}$ by pre-applying PCA to data in a way similar to that of the Fisherface method [9] and it can be speeded up by using a small number of nearest neighboring sets in B_i, W_i much as in [20]. Canonical correlation analysis for multiple sets [192] is also noteworthy here with regard to fast learning. It may be speeded up by reformulating the between-class and within-class scatter matrices in (6.12) by the canonical correlation analysis of multiple sets, thus avoiding the computation of the rotation matrices of every pair of image sets in the iterations.

With the identity matrix $\mathbf{I} \in \mathbb{R}^{N \times N}$ 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**. Once $\mathbf{T}$ maximising the canonical correlations of within-class sets and minimising those of between-class sets in the training data is found, a comparison of any two novel sets is achieved by transforming them by $\mathbf{T}$, and then computing canonical correlations (See (6.7)).

6.2.4 Discussion of Convergence

Although we do not provide a proof of convergence or uniqueness of the proposed optimization process, its convergence to a global maximum was confirmed experimentally. See Figure 6.6 for examples of the iterative learning. Each example is for the learning using a different training data set. The value of the objective function J for all cases becomes

Algorithm 1. Discriminant-analysis of Canonical Correlations (DCC)

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

1. $\mathbf{T} \leftarrow \mathbf{I}_N$
2. Do iterate the following:
3. For all i, do QR-decomposition: $\mathbf{T}^T\mathbf{P}_i = \mathbf{\Phi}_i\mathbf{\Delta}_i \rightarrow \mathbf{P'}_i = \mathbf{P}_i\mathbf{\Delta}_i^{-1}$
4. For every pair i, j, do SVD: $\mathbf{P'}_i^T\mathbf{T}\mathbf{T}^T\mathbf{P'}_j = \mathbf{Q}_{ij}\mathbf{\Lambda}\mathbf{Q}_{ji}^T$
5. Compute $\mathbf{S_b} = \sum_{i=1}^{m}\sum_{l\in B_i}(\mathbf{P'}_l\mathbf{Q}_{li} - \mathbf{P'}_i\mathbf{Q}_{il})(\mathbf{P'}_l\mathbf{Q}_{li} - \mathbf{P'}_i\mathbf{Q}_{il})^T$,
$\mathbf{S_w} = \sum_{i=1}^{m}\sum_{k\in W_i}(\mathbf{P'}_k\mathbf{Q}_{ki} - \mathbf{P'}_i\mathbf{Q}_{ik})(\mathbf{P'}_k\mathbf{Q}_{ki} - \mathbf{P'}_i\mathbf{Q}_{ik})^T$.
6. Compute eigenvectors $\{\mathbf{t}_i\}_{i=1}^{N}$ of $(\mathbf{S_w})^{-1}\mathbf{S_b}$, $\mathbf{T} \leftarrow [\mathbf{t}_1, ..., \mathbf{t}_N]$
7. End
8. $\mathbf{T} \leftarrow [\mathbf{t}_1, ..., \mathbf{t}_n]$

Table 6.1: Proposed iterative algorithm for finding $\mathbf{T}$, which maximises class separation in terms of canonical correlations.

stable after first few iterations, starting with the initial value $\mathbf{T} = \mathbf{I}$. This fast and stable convergence is highly efficient in keeping the learning cost low. Furthermore, as shown at bottom right in Figure 6.6, it was observed that the proposed algorithm converged to the same point irrespective of the initial value of $\mathbf{T}$. These results are indicative of the defined criterion's being a quadratic convex function with respect to the joint set of variables as well as each individual variable as argued in [115, 114].

For all experiments in Section 6.4, the number of iterations was fixed at 5. The proposed learning took about 50 seconds for the face experiments on a Pentium IV PC using non-optimized Matlab code, while the OSM/CMSM methods took around 5 seconds. Note the learning is performed once in an off-liner manner. On-line matching by the three recognition methods is time-efficient. See the experimental section for more information about the time complexity of the methods.

6.3 Orthogonal Subspace Method (OSM)

Orthogonality of two subspaces means that any vector of one subspace is orthogonal to any vector of the other subspace [145]. This requirement is equivalent to that of each basis vector of one subspace being orthogonal to each basis vector of the other. Recalling that canonical correlations are defined as maximal correlations between any two vectors of two subspaces as given in (6.1), it is very clear that canonical correlations of any two orthogonal subspaces are zeros. Measuring canonical correlations (or principal angles) of class specific orthogonal subspaces might thus be a basis for classifying image sets.

Let us assume that the subspaces of the between-class sets $B_i = \{j|\mathbf{X}_j \notin C_i\}$ of a given data set $\mathbf{X}_i$ are orthogonal to the subspace of the set $\mathbf{X}_i$. If the subspaces are orthogonal, all canonical correlations of those subspaces would also be zero as

$$\mathbf{P}_i{}^T\mathbf{P}_{l\in B_i} = \mathbf{O} \in \mathbb{R}^{d\times d} \quad \rightarrow \quad trace(\mathbf{Q}_{il}^T\mathbf{P}_i{}^T\mathbf{P}_l\mathbf{Q}_{li}) = 0 \tag{6.13}$$

where $\mathbf{O}$ is a zero matrix and $\mathbf{P}_i$ is a basis matrix of the set $\mathbf{X}_i$. The classical orthogonal subspace method (OSM) [145] has been developed as a method designed to obtain class-specific orthogonal subspaces. The OSM finds the common projection for all classes, which

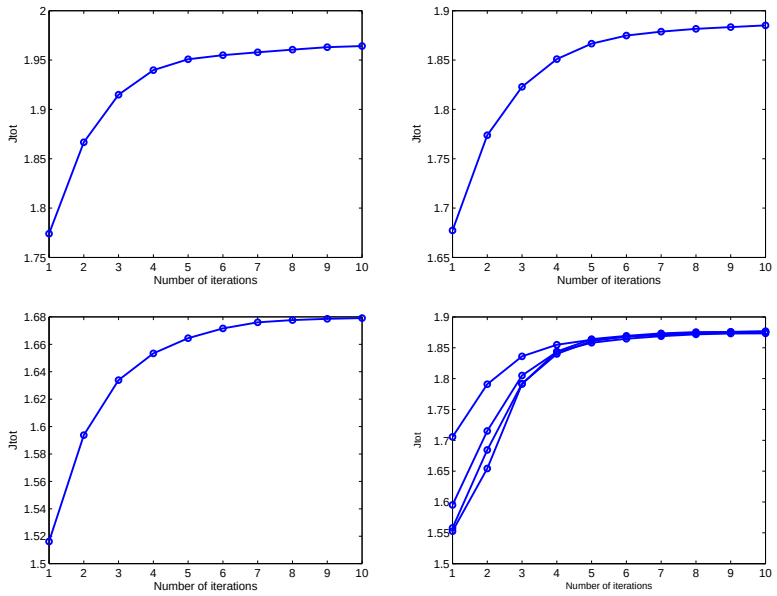


Figure 6.6: Convergence characteristics of the optimization. The cost of J of a given training set is shown as a function of the number of iterations.The bottom right shows the convergence to a unique maximum with different random initials of $\mathbf{T}$.

is represented by the matrix $\mathbf{P}_0$. In the space projected by $\mathbf{P}_0$, it computes subspaces, each of which maximally represents one class while minimally representing all the other classes. The details are explained in the following paragraph. Classification of a new image set is achieved by computing a subspace for the new set in the projected space by $\mathbf{P}_0$ and then by measuring canonical correlations between the subspace and the class orthogonal subspaces.

Denote the correlation matrices of the C classes by $\mathbf{C}^1, ...\mathbf{C}^C$ and the respective a priori probabilities by $\pi^1, ..., \pi^C$ [145]. Then matrix $\mathbf{C}_0 = \sum_{i=1}^{C} \pi^i \mathbf{C}^i$ is the correlation matrix of the mixture of all the classes. Matrix $\mathbf{C}_0$ can be diagonalized by $\mathbf{B}\mathbf{C}_0\mathbf{B}^T = \mathbf{\Lambda}$. Denoting $\mathbf{P}_0 = \mathbf{\Lambda}^{-1/2}\mathbf{B}$, we have $\mathbf{P}_0\mathbf{C}_0\mathbf{P}_0^T = \mathbf{I}$. Then,

$$\pi^1\mathbf{P}_0\mathbf{C}^1\mathbf{P}_0^T + ...\pi^C\mathbf{P}_0\mathbf{C}^C\mathbf{P}_0^T = \mathbf{I}$$

This means that matrices $\pi^i\mathbf{P}_0\mathbf{C}^i\mathbf{P}_0^T$ and $\Sigma_{j\neq i}\pi^j\mathbf{P}_0\mathbf{C}^j\mathbf{P}_0^T$ have the same eigenvectors but the eigenvalues λ_k^i of $\pi^i\mathbf{P}_0\mathbf{C}^i\mathbf{P}_0^T$ and $\overline{\lambda_k^i}$ of $\Sigma_{j\neq i}\pi^j\mathbf{P}_0\mathbf{C}^j\mathbf{P}_0^T$ are related by $\lambda_k^i = 1 - \overline{\lambda_k^i}$. That is, in the space rotated by matrix $\mathbf{P}_0$, the most important basis vectors of class i, which are the eigenvectors of $\pi^i\mathbf{P}_0\mathbf{C}^i\mathbf{P}_0^T$ corresponding to largest eigenvalues, are at the same time the least significant basis vectors for the ensemble of the rest of the classes. Let $\mathbf{P}_i$ be such an eigenvector matrix so that

$$\pi^i{\mathbf{P}_i}^T\mathbf{P}_0\mathbf{C}^i\mathbf{P}_0^T\mathbf{P}_i = \mathbf{\Lambda}^i$$

Then,

$$\Sigma_{j\neq i}\pi^j{\mathbf{P}_i}^T\mathbf{P}_0\mathbf{C}^j\mathbf{P}_0^T\mathbf{P}_i = \mathbf{I} - \mathbf{\Lambda}^i$$

Since every matrix $\pi^j\mathbf{P}_0\mathbf{C}^j\mathbf{P}_0^T$ for all $j \neq i$ is positive semidefinite, $\pi^j{\mathbf{P}_i}^T\mathbf{P}_0\mathbf{C}^j\mathbf{P}_0^T\mathbf{P}_i$

should be a diagonal matrix having smaller elements than $1 - \lambda^i$. If we let $\mathbf{P}_j$ denote the eigenvectors of j-th class by $\pi^j \mathbf{P}_0 \mathbf{C}^j \mathbf{P}_0^T \approx \mathbf{P}_j \mathbf{\Lambda}^j {\mathbf{P}_j}^T$, the matrix ${\mathbf{P}_i}^T \mathbf{P}_j \mathbf{\Lambda}^j {\mathbf{P}_j}^T \mathbf{P}_i$ now has small diagonal elements. ${\mathbf{P}_i}^T \mathbf{P}_j$ has, accordingly, small elements. In the ideal case when $\pi^i \mathbf{P}_0 \mathbf{C}^i \mathbf{P}_0^T$ has the eigenvalues which are exactly equal to one, the matrix ${\mathbf{P}_i}^T \mathbf{P}_j$ would be a zero matrix for all $j \neq i$. The two subspaces defined by $\mathbf{P}_i, \mathbf{P}_j$ are called orthogonal subspaces. That is, every column of $\mathbf{P}_i$ is perpendicular to every column of $\mathbf{P}_j$.

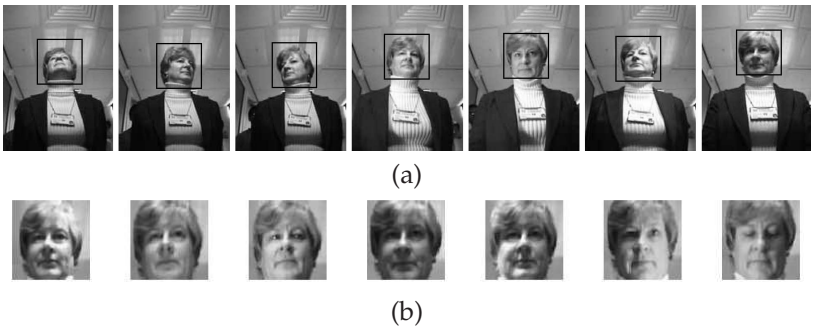
Note that the OSM method does not exploit the concept of multiple sets in a single class (or within-class sets). The method assumes that all data vectors of a single class i are represented by a single set $\mathbf{P}_i$. From the above, the matrix $\mathbf{P}_0$ could represent an alternative discriminative transformation by which the canonical correlations of between-class sets are minimised. Note that the matrix $\mathbf{P}_0$ is a square matrix.

Comparison with the iterative Solution, DCC. Note that the orthogonality of subspaces is a restrictive condition, at least when the number of classes is large. It is often the case that the subspaces of OSM represented by $\mathbf{P}_i$ and $\mathbf{P}_{l \in B_i}$ are correlated. If ${\mathbf{P}_i}^T \mathbf{P}_l$ has non-zero values, canonical correlations could be much greater than zero as

$$\mathbf{q}_{il}^T {\mathbf{P}_i}^T \mathbf{P}_l \mathbf{q}_{li} \gg 0 \tag{6.14}$$

where $\mathbf{q}$ is a column of the rotation matrix $\mathbf{Q}$ in the definition of canonical correlations. Generally, the problem of minimising correlations of basis matrices ${\mathbf{P}_i}^T \mathbf{P}_l$ in OSM is not equivalent to the proposed problem formulation where the canonical correlations $\mathbf{q}_{il}^T {\mathbf{P}_i}^T \mathbf{P}_l \mathbf{q}_{li}$ are minimised. That is, OSM tries orthogonalizatin for all axes of subspaces with equal importance but DCC does this for canonical axes with different importance, revealed by the canonical correlation analysis. Moreover, the orthogonal subspace method does not explicitly attempt to maximise canonical correlations of the within-class sets. It combines all examples of a class together. However, the OSM method is simpler in terms of learning and computationally economical as it does not require iterations in learning like DCC. It is shown that this simpler method also improves the simple canonical correlation method effectively in the experiments.

Comparison with Constrained Mutual Subspace Method (CMSM) It is worth noting that the previous method CMSM [45][144] can be seen to be closely related to the orthogonal subspace method. For the details of CMSM, refer to Section 3.6.8. CMSM finds the constrained subspace where the total projection operators have small variances. Each class is represented by a subspace which maximally represents the class data variances. The class subspace is then projected into the constrained subspace. The projected data subspace compromises the maximum representation of each class and the minimum representation of a mixture of all the other classes. This is similar in concept to the orthogonal subspace method explained above. Both methods try to minimise the correlation of between-class subspaces defined by ${\mathbf{P}_i}^T \mathbf{P}_{l \in B_i}$. However, dimensionality of the constrained subspace of CMSM should be optimised for each application. If dimensionality is too low, the constrained subspace will be a null space. In the opposite case, the constrained subspace simply retains all the energy of the original data and thus can not play a role as a discriminant function. This dependence of CMSM on the parameter (dimensionality) selection makes it empirical. By contrast, there is no need to choose any subspace from the discriminative space represented by the matrix $\mathbf{P}_0$ in the orthogonal subspace method. A full dimension



(a) (b)

Figure 6.7: Example images of the face data sets. (a) Frames of a typical face video sequence with automatic face detection (b) Face prototypes of the 7 illumination settings

of the matrix can simply be adopted. Note the proposed optimal solution, DCC, also exhibited insensitivity to dimensionality, thus practically, as well as theoretically appealing (See the experimental section).

6.4 Experimental Results and Discussion

The proposed method (the code is available at http://mi.eng.cam.ac.uk/∼tkk22) is evaluated on various object or object category recognition problems: using face image sets with arbitrary motion captured under different illuminations, image sets of five hundred general objects taken at different views and the 8 general object categories, each of which has several different objects. The task of all of the experiments is to classify an unknown set of vectors to one of the training classes, each also represented by vector sets.

6.4.1 Database of Face Image Sets

We have collected a database called the *Cambridge-Toshiba Face Video Database* with 100 individuals of varying age and ethnicity and equally represented genders. For each person, 14 (7 illuminations $\times$ two recordings) video sequences of the person in arbitrary motion were collected. Each sequence was recorded in a different illumination setting for 10s at 10fps and at 320$\times$240 pixel resolution. See Figure 6.7 for samples from an original image sequence and seven different lightings. Following automatic localization using a cascaded face detector [193] and cropping to a uniform scale of 20$\times$20 pixels, images of faces were histogram equalized. See Appendix A for more details on the data set. Note that face localization was performed automatically on the images of uncontrolled quality. Thus it was not as accurate as any conventional face registration with either manual or automatic eye positions performed on high-quality face images. Our experimental conditions are closer to those given for typical surveillance systems.

6.4.2 Comparative Methods and Parameter Setting

We compared the performance of :

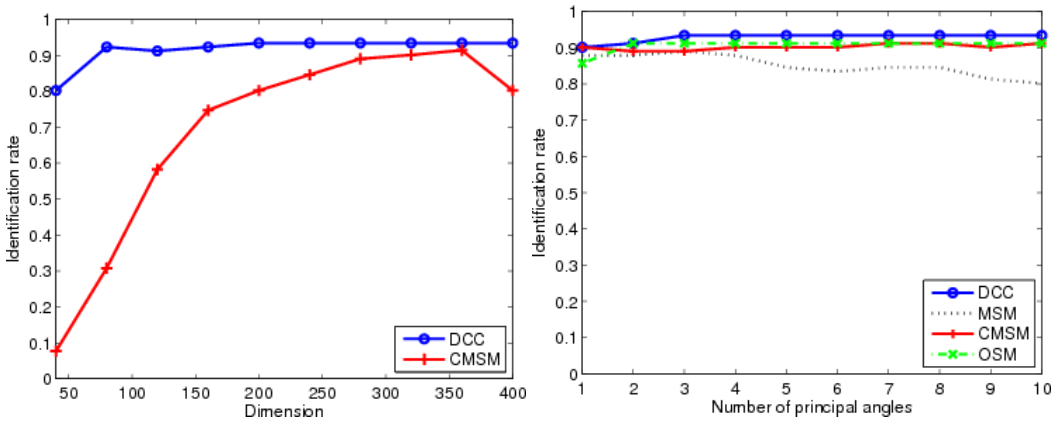
- KL-Divergence algorithm (KLD) [167] as a representative probability density-based method,

- Simple assembly methods such as Hausdorff Distance ($d(S_1, S_2) = \min_{x_1 \in S_1} \max_{x_2 \in S_2} d(x_1, x_2)$) and k-Nearest Neighbours (kNN) (in the sense of $\min_{x_1 \in S_1} \min_{x_2 \in S_2} d(x_1, x_2)$) [33] of images transformed by (i) PCA, and (ii) LDA [9] subspaces, which are estimated from training data similarly to [162],
- Nearest Neighbour (NN) by FaceIt (v.5.0), the commercial face recognition system from Identix, which ranked top overall in the Face Recognition Vendor Test 2000 and 2002 [154, 13],
- Mutual Subspace Method (MSM) [207], which is equivalent to a simple aggregation of canonical correlations (See Appendix C),
- Constrained MSM (CMSM) [45, 144] used in a state-of-the-art commercial system called FacePass [186] (See Section 3.6.8),
- Orthogonal Subspace Method (OSM) [145],
- and the proposed iterative discriminative learning, DCC.

To compare algorithms, important parameters of each method were adjusted and those optimal in terms of test identification rates were selected. In KLD, 96% of data energy was explained by the principal subspace of training data used [167]. In kNN methods, the dimension of PCA subspace was chosen to be 150, which represents more than 98% of training data energy (Note that removing the first 3 components improved accuracy in the face recognition experiment as similarly observed in [9]). The best dimension of LDA subspace was also found to be around 150. The number of nearest neighbors used was chosen from one to ten. In MSM/CMSM/OSM/DCC, the dimension of the linear subspace of each image set represented 98% of data energy of the set, which was around 10. PCA was performed for computing the subspace of each set in the MSM/CMSM/DCC methods.

Dimension Selection of the Discriminative Subspaces in CMSM/OSM/DCC. As shown in Figure 6.8 (a), CMSM exhibited a high peaking in the the relationship between accuracy and dimensionality of the constrained subspace, whereas the proposed method, DCC, provided constant identification rates regardless of dimensionality of $\mathbf{T}$ beyond a certain point. The best dimension of the constrained subspace of CMSM was found to be at around 360 and was fixed. For DCC, we fixed the dimension at 150 for all experiments (the full dimension can also be conveniently exploited without any feature selection). The full dimension was also used for the matrix $\mathbf{P}_0$ in OSM. Note that the proposed methods DCC and OSM do not require any elaborate feature selection and this behaviour of DCC/OSM is highly attractive from the practical point of view, compared with CMSM. Without feature selection, the accuracy of CMSM in the full space drops dramatically to the level equivalent to that of MSM, which is a simple aggregation of canonical correlations without any discriminative transformation.

Number of Canonical Correlations. Figure 6.8 (b) shows the accuracy of DCC, MSM, CMSM and OSM according to the number of canonical correlations used. This parameter does not affect the accuracy of the methods as much as the dimension of the discriminative subspace, as shown in Figure 6.8 (a). Overall, the methods DCC, CMSM and OSM were



(a) (b)

Figure 6.8: (a) The effect of the dimensionality of the discriminative subspace on the proposed iterative method (DCC) and CMSM. The accuracy of CMSM at 400 is equivalent to that of MSM, a simple aggregation of canonical correlations. (b) The effect of the number of canonical correlations on DCC/MSM/CMSM/OSM.

shown to be less sensitive to this parameter than MSM as they exploit their own discriminative transformations. More precisely, DCC and OSM showed desirably stable curves over this parameter whereas CMSM exhibited more or less fluctuating performance. For simplicity, the number of canonical correlations was fixed at the same (i.e. set as the dimension of linear subspaces of image sets) for all methods, MSM, CMSM, OSM and DCC.

6.4.3 Face-Recognition Experiments

Training of all algorithms was performed with data sequences acquired in a single illumination setting and testing with a single other setting. We used 18 randomly selected training/test combinations of the sequences for reporting identification rates. The performance of the evaluated recognition algorithms is shown in Figure 6.9 and Table 6.2. The 18 experiments were divided into two parts according to the degree of difference between the training and the test data of the experiments, which was measured by KL-Divergence between the training and test data. Figure 6.9 shows the cumulative recognition rates for the averaged results of all 18 experiments and Table 6.2 shows the results separately for the first (easier) and the second parts (more difficult) of the experiments.

In this experiment, all training samples of a class were drawn from a single video sequence of arbitrary head movement, so they were randomly divided into two sets for the within-class sets in the proposed learning. Note that the proposed method with this random partition still worked well regardless of the number of partitions as exemplified in Table 6.3. The test recognition rates changed by less than 1% in all of the various trials of random partitioning. This may be because no explicit discriminatory information is contained in the randomly partitioned intra-class sets. Rather, DCC is mostly concerned with achieving the maximum possible separation of inter-class sets as are CMSM/OSM. In this case, the numerator in the objective function (6.9) may just help finding the meaningful solution that minimises the denominator. Table 1 shows the averaged accuracy of 18 random trials of the 18 experiments. If samples of a class can be partitioned according to the data semantics, the concept of within-class sets would be more useful and realistic, which is the

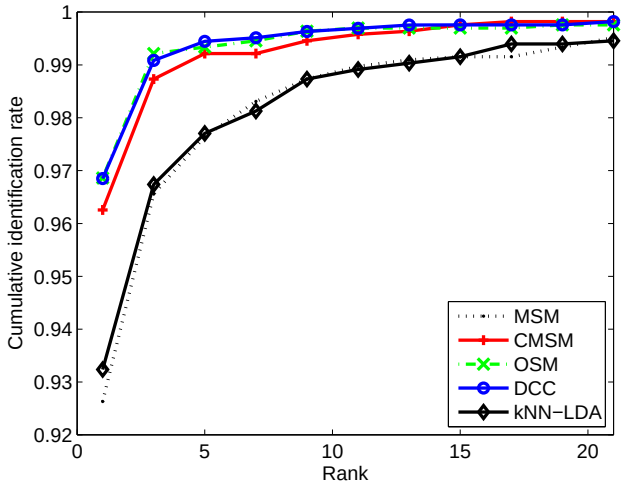


Figure 6.9: Cumulative recognition plot for the MSM/kNN-LDA/CMSM/OSM/DCC methods.

	KLD	HD-PCA	1NN-PCA	10NN-PCA	FaceIt S/W
1st half	0.49±0.14	0.60±0.07	0.95±0.03	0.96±0.03	0.90±0.09
2nd half	0.24±0.13	0.47±0.09	0.71±0.20	0.71±0.21	0.86±0.05
	10NN-LDA	MSM	CMSM	OSM	DCC
1st half	0.98±0.01	0.94±0.03	0.98±0.01	0.98±0.01	0.98±0.01
2nd half	0.87±0.07	0.91±0.02	0.93±0.06	0.94±0.06	0.95±0.04

Table 6.2: Evaluation results. The mean and standard deviation of recognition rates of different methods. The results are shown separately for the first (easier) and the second parts (more difficult) of the experiments.

Number of partitions	2	3	4
exp. 1	93.19±0.46	92.86±0.78	92.75±0.77
exp. 2	95.93±0.53	95.60±0.00	95.71±0.35

Table 6.3: Example results for random partitioning. The mean and standard deviation (%)of recognition rates of 10 random trials for two example experiments.

case in the following experiments.

In Table 6.2, most of the methods generally had lower recognition rates for the experiments with larger KL-Divergence between the training and test data. The KLD method achieved by far the worst recognition rate. Considering that the illumination conditions varied across data and that the face motion was largely unconstrained, the distribution of within-class face patterns was very broad, making this result unsurprising. In the methods of non-parametric sample-based matching, the Hausdorff-Distance (HD) measure provided far poorer results than the k-Nearest Neighbors (kNN) methods defined in the PCA subspace. 10NN-PCA yielded the best accuracy of the sample-based methods defined in the PCA subspace, which is on average worse than MSM by 8.6%. Its performance greatly varied across the experiments. Note that MSM showed robust performance with a large margin over the kNN-PCA method under the different experimental conditions. The improvement of MSM over both KLD and HD/kNN-PCA methods was very impressive. The

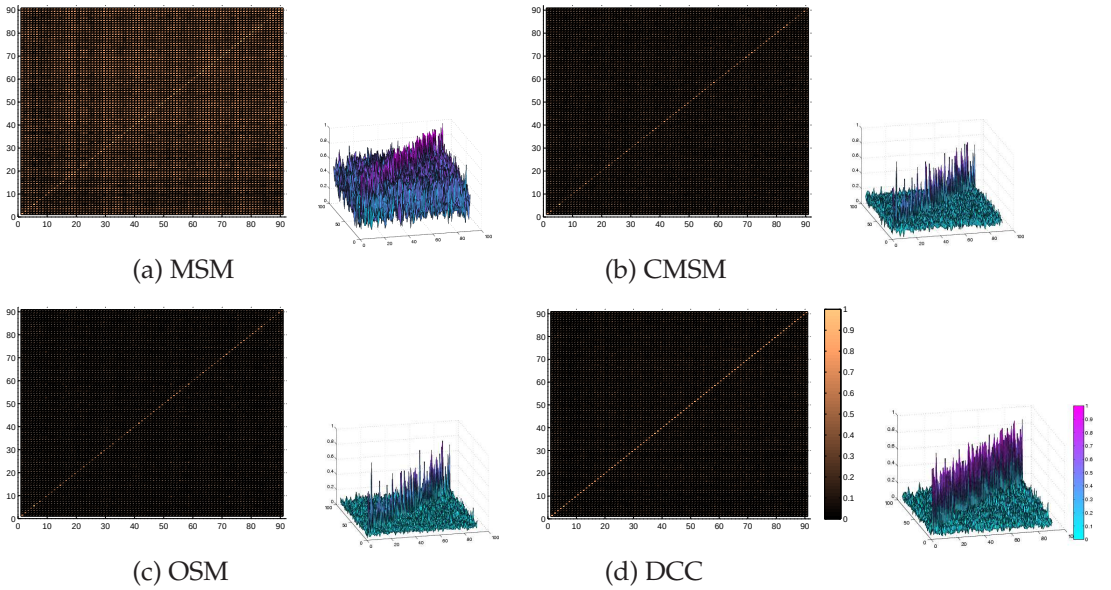
benefits of using canonical correlations over both classical approaches for set classification, which have been explained throughout the previous sections, were confirmed.

The commercial face recognition software FaceIt (v.5.0) yielded a performance between those of kNN-PCA and kNN-LDA methods on average. Although the NN method using FaceIt is based on individual sample matching, it delivered more robust performance for the data changes (the difference in accuracy between the first and the second halves is not as large as those of kNN-PCA/LDA methods). This is reasonable, considering that FaceIt was trained independently of the training images used for other methods.

Table 6.2 also gives a comparison of the methods combined with discriminative learning. kNN-LDA yielded a big improvement over kNN-PCA but the accuracy of the method again greatly varied across the experiments. Note that 10NN-LDA outperformed MSM for similar conditions between the training and test sets, but it became noticeably inferior as the conditions changed. It delivered similar accuracy to MSM on average, which is also shown in Figure 6.9. The proposed methods DCC and OSM and CMSM constantly provided a significant improvement over both MSM and kNN-LDA methods as shown in Table 6.2 as well as in Figure 6.9.

Note that CMSM/OSM may be regarded as simply measuring correlation between subspaces defined by the basis matrix $\mathbf{P}$, which is different from the canonical correlations defined by $\mathbf{PQ}$. In spite of this difference, the accuracy of CMSM/OSM was impressive in this experiment. As explained above, when an ideal solution of CMSM/OSM is available and $\mathbf{Q}$ only provides a rotation within the subspace, the solution of CMSM/OSM can be close to that of the proposed optimal solution DCC. However, if class subspaces cannot be made orthogonal to each other, then the direct optimization of canonical correlations offered by DCC is preferred. Note that the DCC method was better than those of CMSM/OSM for the second half of the experiments in Table 6.2. The differences in the three methods are apparent from the associated similarity matrices of the training data. We trained the three methods using both training and test sets of the worst experimental case for the methods (See the last two of Figure 6.7 (b)), and compared their similarity matrices of the total class data with that of MSM, as shown in Figure 6.10. Both OSM and CMSM considerably improved the ability of class discrimination over MSM, but they were inferior to the optimal achieved by DCC for the given data. As discussed above, both of the proposed methods, DCC, and OSM are preferable to CMSM as they do not involve the selection of dimensionality of the discriminative subspaces. While the best dimension for CMSM had to be identified with reference to the test results, the full dimension of the discriminative space can simply be adopted for any new test data in the DCC and OSM methods.

We designed another face experiment with more face image sets in the Cambridge-Toshiba face video database. The database involves two sets of videos acquired at different times, each of which consists of seven different illumination sequences for each person. We used one-time set for training and the other set for testing, thus producing more variations between the training and testing (See Figure 6.11 for an example of the two sets acquired in the same illumination at different times). Note the training and testing sets in the previous experimental setting were drawn from the same time set. In this experiment using a single illumination set for training, the full 49 combinations of the lighting settings were exploited. We also increased the number of image sets per class for training. We randomly drew a combination of illumination sequences for training and used all 7 illumination sequences for testing. 10-fold cross validation was performed for these ex-



(a) MSM (b) CMSM

(c) OSM (d) DCC

Figure 6.10: Similarity matrices for MSM/CMSM/OSM/DCC methods. Two graphs (for top-view and 3D-diagonal-view) are shown for each method. The diagonal and off-diagonal values in the DCC matrix can be much better distinguished.

Figure 6.11: Example of the two time sets (top and bottom) of a person acquired in a single lighting setting. They contain significant variations in pose and expression.

periments. Figure 6.12 shows the mean and standard deviations of recognition rates of all experiments. The proposed method DCC significantly outperformed OSM/CMSM methods when the test sets differed greatly from the training sets. These results are consistent with those of the methods in the 2nd part of the experiment in Table 6.2 (but the difference is much clearer here). Overall, all three methods improved their accuracy by using more image sets in training.

Matching complexity. The complexity of the methods based on canonical correlations (MSM/ CMSM/OSM/DCC), $O(d^3)$, is much lower than that of the sample-based matching methods (kNN-PCA/LDA), $O(m^2n)$, where d is the subspace dimension of each set, m is the number of samples of each set and n is the dimensionality of feature vectors, since $d \ll m, n$. In the face experiments, the unit matching time of comparing the two image sets which contain about 100 images is 0.004 for the canonical correlations based method and 1.1 seconds for the kNN method.

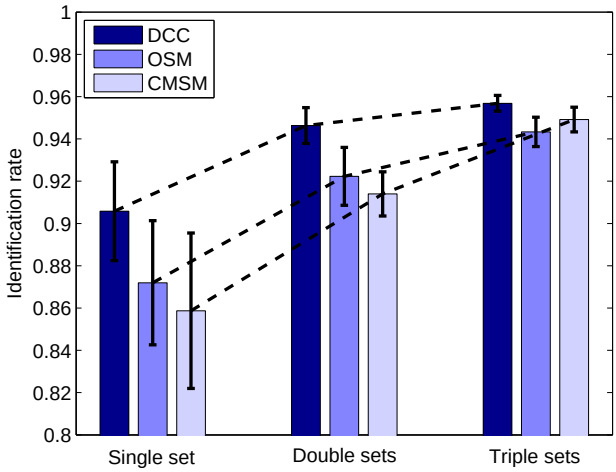
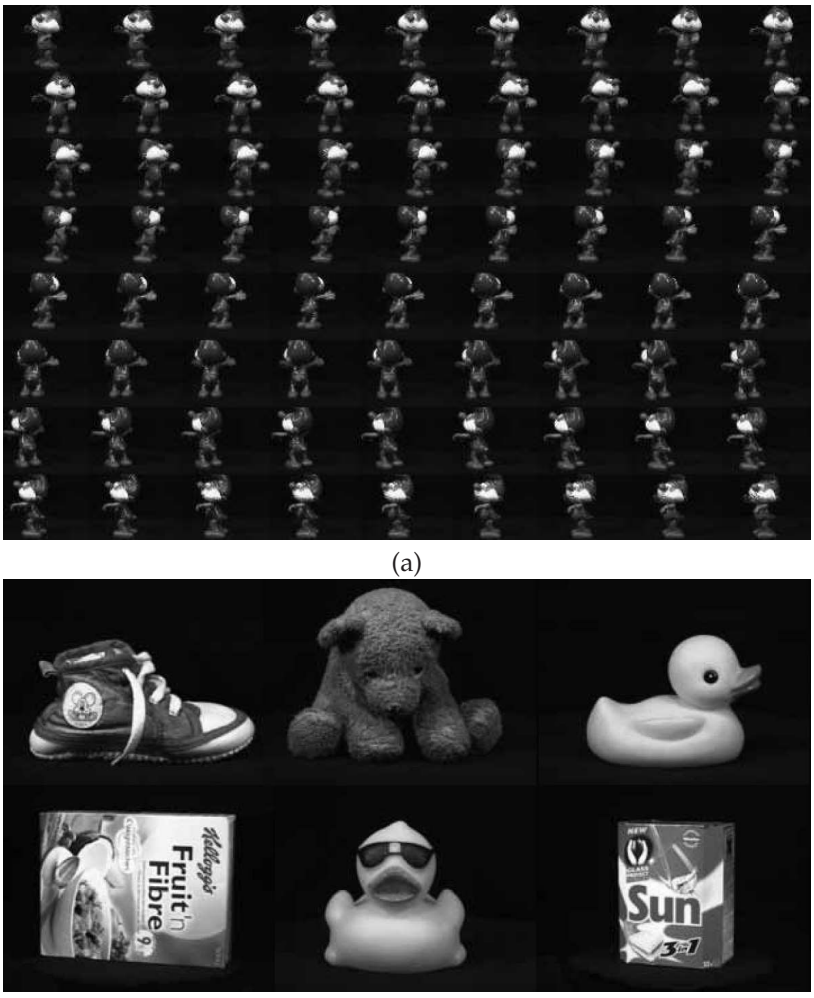


Figure 6.12: Recognition rates of the CMSM/OSM/DCC methods when using a single, double and triple image sets in training.

6.4.4 Experiment on Large Scale General Object Database

The ALOI database [50] with 500 general object categories taken from different viewing angles provides another experimental data set for the proposed method (See Figure 6.13 for some examples). Object images were segmented from the simple background and scaled to 20×20 pixel size. A training set and five test sets were set up with different viewing angles of the objects as shown in Figure 6.14 (a) and (b). Note that the pose of all the images in the test sets differed by at least 5 degree from every sample of the training set. The methods of MSM, kNN-LDA and CMSM were compared with the proposed methods DCC and OSM in terms of identification rate. The parameters were selected in the same way as in the face recognition experiment. The dimension of the linear subspace of each image set was fixed at 5, representing more than 98% data energy in MSM/CMSM/OSM/DCC methods. The best number of nearest neighbors in the kNN-LDA method was found to be five.

Judging from Figure 6.15 and Figure 6.16, kNN-LDA yielded better accuracy than MSM in all the cases. This contrasted with the findings in the face recognition experiment. This may have been caused by the somewhat artificial experimental setting. The nearest neighbours of the training and test set differed only slightly due to the five degree pose difference. Please note that the two sets had no changes in lighting and had accurate localization of the objects as well. Further note that the accuracy of MSM could be improved by using only the first canonical correlation, as with the results shown in Figure 6.8 (b). Here again, CMSM, OSM and the proposed method DCC were substantially superior to MSM. Overall, the accuracy of CMSM/OSM was similar to that of kNN-LDA method, as shown in Figure 6.16. The proposed iterative method, DCC, constantly outperformed all the others including OSM/CMSM as well as kNN-LDA. Please note this experiment involved a larger number of classes, as compared with the face experiments. Furthermore, the set of images of the training class had quite different pose distributions from those of the test set.



(a)

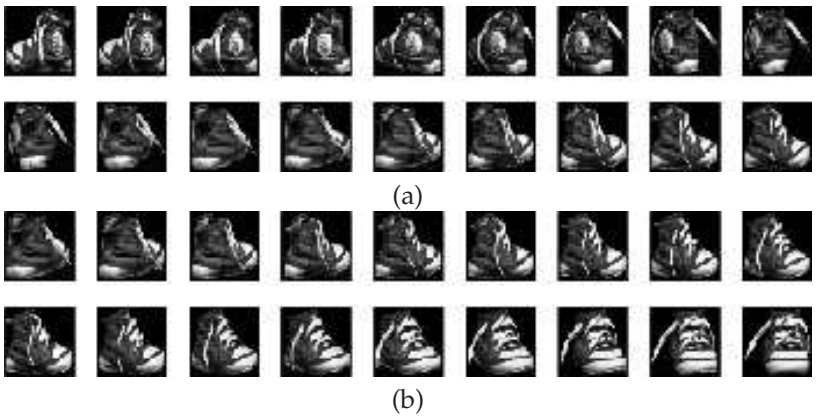
(b)

Figure 6.13: Example images of the ALOI Database. (a) Each object has 72 images taken at every five degree views in the round. (b) Examples of six different objects.

6.4.5 Object Category Recognition using ETH80 database

An interesting experiment of object category recognition was performed using the public ETH80 data base. As shown in Figure 6.17, there are 8 categories which contain 10 objects each, with 41 images of different views. More details of the data base can be found in [117]. We randomly partitioned 10 objects into two sets of five objects for training and testing. In Experiment 1, we used all 41 view images of objects. In Experiment 2, we used all 41 views for training but a random subset of 15 view images for testing. 10-fold cross-validation was carried out for both experiments. Parameters such as the dimension of the linear subspaces, the number of principal angles and nearest neighbors were selected as in the previous experiment. The dimension of the constrained subspace of CMSM was also best optimised.

From Table 6.4, it is worth noting that the accuracy of kNN-PCA method is similar (but slightly inferior) to that of the PCA method reported in [117]. Note that we used only 5 objects per category, in contrast to [117] where 9 objects were used for training. The recognition rates for individual object categories also showed similar behavior to those



(a)

(b)

Figure 6.14: ALOI experiment. (a) The training set consists of 18 images taken at 10 degree intervals. (b) Two test sets are shown. Each test set contains 9 images at 10 degree intervals, different from the training set.

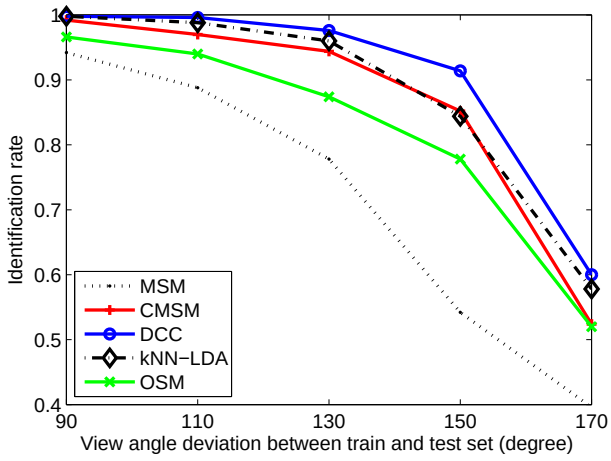


Figure 6.15: Identification rates for the 5 different test sets. The object viewing angles of the test sets differ from those of the training set to a varying extent.

of [117].

As shown in Table 6.4, the kNN methods were much inferior to the those based on canonical correlation. The sample-based matching method was highly sensitive to the variations in different objects of the same categories, failing in object categorisation. The methods using canonical correlations provided much more accurate results. The proposed method (DCC) delivered the best accuracy of all tested methods. The improvement of DCC over CMSM/OSM was greater in the second experiment where only a subset of images of objects was involved in the testing. Note that this makes the testing set very different from the training set. The major principal components of the image sets are highly sensitive to the variations in pose. The accuracy of CMSM/OSM methods was considerably decreased in the presence of this variation, while the DCC method maintained almost the same accuracy.

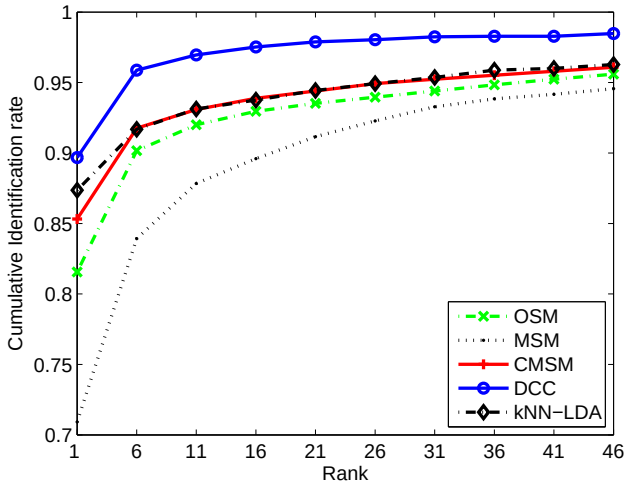
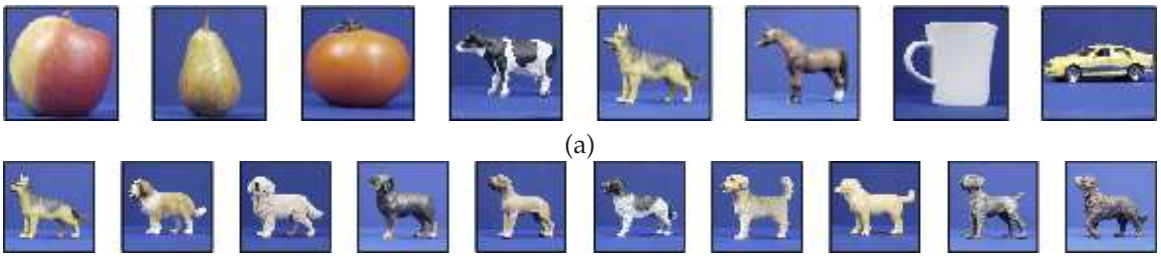


Figure 6.16: Cumulative recognition rates of the MSM/kNN-LDA/CMSM/OSM/DCC methods for the ALOI experiment.



(a)

(b)

Figure 6.17: Object category database (ETH80) contains (a) 8 object categories and (b) 10 objects for each category.

	kNN-PCA	kNN-LDA	MSM	CMSM	OSM	DCC
exp.1	0.762±0.21	0.752±0.17	0.865±0.13	0.897±0.10	0.905±0.09	0.917±0.09
exp.2	-	-	-	0.852±0.21	0.865±0.18	0.912±0.13

Table 6.4: Evaluation results of object categorisation. The mean recognition rate and its standard deviation for all experiments.

6.5 Summary

This study has addressed the question of how to exploit set-information for robust object recognition. Any assembly algorithms which combine individual sample matches in ad-hoc manner were shown to be very poor.

We demonstrated that the canonical correlation is a robust similarity measure of image sets, yielding much higher recognition rates than the traditional probability density-based set-similarity e.g. Kullback Leibler-Divergence (KLD), which is sensitive to simple transformations of input data. On the other hand, object images are well-constrained to low-dimensional subspaces and Canonical Correlation Analysis is affine invariant, effectively placing a uniform prior over the subspaces. CCA provides robust matching of the image sets about the pattern variations on the subspaces. The canonical-correlation based meth-

ods were shown to be highly time-efficient in matching, thus offering an attractive tool for a large-scale recognition task.

A novel discriminative learning framework has been proposed for set classification based on canonical correlations. The novel iterative learning yields the optimal discriminant transformation. The Orthogonal Subspace Method (OSM) has been also explored as an alternative method of improving set-classification by canonical correlations. The proposed methods have been evaluated for various object and object category recognition problems. The new techniques enable discriminative learning over sets, and exhibit an impressive set classification accuracy. They significantly outperformed the probability density based methods, the classical assembly methods based on PCA/LDA or a commercial face recognition software (which was the top in the Face Recognition Vendor Test 2000 and 2002 [154, 13]), and the simple aggregation method of canonical correlations. While OSM/CMSM delivered comparable accuracy to the optimal method DCC in particular cases, they generally lagged behind the method DCC. Compared with the prior-art CMSM method, both of the proposed methods, OSM and DCC, had a benefit in feature selection.

Further necessary studies could be conducted in theory and mathematics. First of all, the proposed image-set similarity is not a metric. By making it a metric or distance function, the method could more conveniently be plugged into other studies without losing generality. Secondly, the proposed iterative learning for DCC should receive closer examination on convergence either by proofs or through more extensive experiment with real-data sets which include outliers.

Interesting research directions include non-linear extension, which would allow us to capture discriminatory information of image sets contained in higher-order statistics. It may also prove beneficial to make the proposed learning more time-efficient so as to be incrementally updated for new training sets. See Chapter 9 for these extensions. Despite its success in image-set based object recognition, CCA is still insufficient for Video-to-Video matching for action/gesture classification. CCA simply takes a video as a set of images not encoding temporal (ordering) information. This will be dealt with in the next chapters.

CHAPTER 7

Tensor Canonical Correlation Analysis for Action Classification

In this chapter, we introduce a new method, namely Tensor Canonical Correlation Analysis (TCCA) which extends classical Canonical Correlation Analysis (CCA) to multidimensional data arrays (also called tensors) for ***Video-to-Video Matching***. The proposed extension is seen as the aggregation of many different sub-CCAs, including the conventional CCA of two image sets as a part. The proposed method inherits the benefits of CCA, robustness about intra-class variation by affine-invariance and considers full space-time information for action classification. Compared with state-of-the-art methods for action classification, this method avoids the difficult problem of explicit motion estimation or significant meta-parameter setting (for e.g. about space-time interest points) and delivers maximum discrimination information in all useful aspects of video data.

The derived tensor canonical correlations are combined with a weak discriminative feature selection and a Nearest Neighbor classifier for action recognition. In addition, we propose a time-efficient action detection and alignment method based on incremental learning of subspaces for Tensor CCA when actions are not aligned in the input space-time domain. In the experiments on the public action data set (KTH) as well as the self-recorded hand gesture data set, the proposed method showed significantly better accuracy at comparable detection speed than state-of-the-art action recognition methods.

The discriminant analysis method developed in the previous chapter for ***Set-to-Set Matching*** is readily plugged into the Tensor CCA method for further improving action classification accuracy, which will be explained in the following chapter.

7.1 Overview of Tensor Canonical Correlation Analysis

In this study, a novel statistical method of pairwise feature extraction from video data is proposed for human action/gesture categorisation. We extend the classical canonical correlation analysis [5, 60] (see Section 6.1) -a standard tool for inspecting linear relations between two sets of vectors- to that of multi-dimensional data arrays (also called high-order tensors), which is for analysis of the similarity of video data/space-time volumes. Figure 7.1 illustrates the action classification problem using CCA [5].

The proposed method called Tensor Canonical Correlation Analysis (TCCA) is general

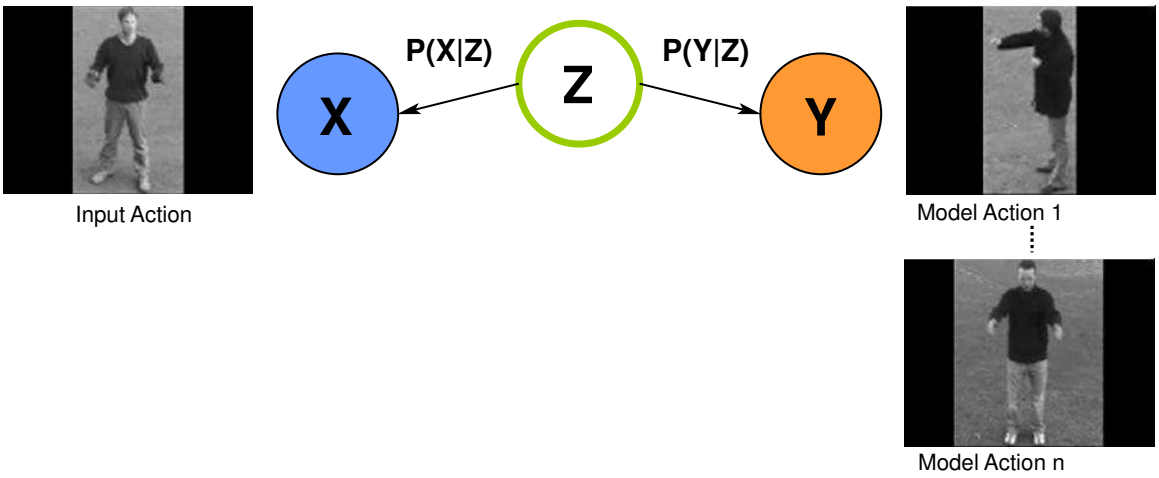


Figure 7.1: Probabilistic Canonical Correlation Analysis tells how well two random variables $\mathbf{x}, \mathbf{y}$ are represented by a common source variable $\mathbf{z}$ [5].

and may be applied to many other tasks requiring tensor matching (e.g. a single color image [7] or filter banks applied to a single gray image [210] can also yield a high-order tensor). Harshman has also presented a concept of Canonical Correlation Analysis of multi-dimensional arrays in his recent work [61](not published as a full paper). Although it was carried out independently of our work, it describes some common concepts which support our new ideas. This chapter comprises not only our new TCCA framework but also new applications of CCA to action classification and efficient action detection algorithms.

This work was encouraged by our previous success [104](See Chapter 6), where Canonical Correlation Analysis (CCA) is adopted to measure the similarity of two image sets for robust object recognition. Image sets are collected either from a video or multiple still shots of objects, containing changes in object appearance due to different lighting and pose. Each image in the two sets is vectorized and classical CCA applied to the two sets of vectors. Object recognition is performed from canonical correlations, also called principal angles, where higher canonical correlations indicate higher similarity of two given image sets. The canonical correlation-based method yielded much higher object recognition rates than the traditional set-similarity measures based on pdfs e.g. Kullback Leibler-Divergence (KLD) in [104]. The KLD-based matching is highly subjective to simple transformations of data (e.g. global intensity changes and variances), which are clearly irrelevant to classification, resulting in poor generalisation to novel data. Compared with traditional methods, a key function of CCA is its affine invariance in matching, which allows great flexibility and yet keeps sufficient discriminative information. The affine-invariance can be explained by the geometrical interpretation of CCA as the angles between two hyper-planes (or linear subspaces). Canonical correlations are the cosine of principal angles and smaller angular planes are thought to be more alike. It is well known that object images are class-wise well-constrained to lie on hyper-planes or low-dimensional subspaces. This subspace-based matching effectively gives affine-invariance, i.e. invariant matching of the image sets to the pattern variations subject to the subspaces. For details, refer to Chapter 6. There are also ample previous studies demonstrating the advantages of the classical subspace concept in various visual recognition tasks.

Despite its success, CCA is still insufficient in action/gesture classification tasks as it

simply represents a video as a set of images not encoding temporal (ordering) information. The proposed Tensor Canonical Correlation Analysis (TCCA) has many favorable characteristics, chief of which are that:

- TCCA of videos yields novel pairwise features reflecting similarity in the joint spatial and temporal domains of videos.
- The new features are flexible up to affine transformation for possible data variations in each domain.
- Action is analysed as global space-time volumes, avoiding the challenging problems of explicit motion estimation.
- The proposed learning-based method does not require any significant tuning parameters.
- The tensor CCA method can be partitioned into sub-TCCAs as each canonical correlation explains different aspects of the multi-dimensional data. For example, previous work on object recognition [104, 200, 207] by CCA tackles a sub-problem of this study.

The quality of TCCA features is demonstrated in terms of action classification accuracy's being combined with a simple feature selection scheme and Nearest Neighbor (NN) classification. Additionally, time-efficient detection of a target video is proposed by incrementally learning the space-time subspaces for TCCA. The proposed method significantly outperformed the state-of-the-art action classification methods in accuracy on the KTH data set [165] as well as our own hand-gesture data set. The proposed detection method could also yield economical computations of TCCA for the case where action is not aligned in the space-time domain, delivering reasonable detection speed compared to the state-of-the-art method in [168].

The rest of the chapter is organised as follows: Notations are given in Section 7.2 and the framework and solution for Tensor CCA are given in Section 7.3. Sections 7.4 and 7.5 are devoted to the proposed discriminative feature selection and action detection method respectively. The results are shown in Section 7.6. We summarise in Section 7.7.

7.2 Notations

7.2.1 Matrix Representation of Canonical Correlation Analysis

The matrix notation of CCA helps explanation of the proposed Tensor CCA. Given two data sets as matrices $\mathbf{X} \in \mathbb{R}^{N \times m_1}$, $\mathbf{Y} \in \mathbb{R}^{N \times m_2}$, canonical correlations are found by the pairs of transformations $\mathbf{u}, \mathbf{v}$. The random vectors $\mathbf{x}, \mathbf{y}$ in (3.26) correspond to the rows of the matrices $\mathbf{X}, \mathbf{Y}$ assuming $N \gg m_1, m_2$. The standard CCA may be written as

$$\rho = \max_{\mathbf{u},\mathbf{v}} \mathbf{X}'^T \mathbf{Y}', \quad \text{where } \mathbf{X}' = \mathbf{X}\mathbf{u}, \mathbf{Y}' = \mathbf{Y}\mathbf{v}. \tag{7.1}$$

Note that the canonical transformations $\mathbf{u}, \mathbf{v}$ hereinafter are defined to be s.t. $\mathbf{XU} = \mathbf{P}^1\mathbf{Q}_1, \mathbf{YV} = \mathbf{P}^2\mathbf{Q}_2$, where $\mathbf{U}, \mathbf{V}$ have $\mathbf{u}, \mathbf{v}$ in their columns respectively and $\mathbf{P}, \mathbf{Q}$ are eigenvector and rotating matrices defined in (6.2) respectively.

7.2.2 Multilinear Algebra

This section briefly introduces basic notation and concepts of multilinear algebra [189, 135], or higher-order tensors for Tensor CCA. For example, a third-order tensor which has the three modes of dimensions I, J, K is denoted by $\mathcal{A} = (\mathcal{A})_{ijk} \in \mathbb{R}^{I \times J \times K}$. The inner product of any two tensors is defined as $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k} (\mathcal{A})_{ijk} (\mathcal{B})_{ijk}$. The *mode*-j vectors are the column vectors of matrix $\mathbf{A}_{(j)} \in \mathbb{R}^{J \times (IK)}$ and the j-mode product of a tensor $\mathcal{A}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J \times N}$ is

$$(\mathcal{B})_{ink} \in \mathbb{R}^{I \times N \times K} = (\mathcal{A} \times_j \mathbf{U})_{ink} = \Sigma_j (\mathcal{A})_{ijk} \mathbf{u}_{jn} \tag{7.2}$$

The j-mode product in terms of j-mode vector matrices is $\mathbf{B}_{(j)} = \mathbf{U}\mathbf{A}_{(j)}$.

7.3 Tensor Canonical Correlation Analysis

7.3.1 Joint and Single-shared-mode TCCA

Many previous studies e.g. [189, 7, 210] have dealt with tensor data in its original form to consider multi-dimensional relations of the data and to avoid the *curse of dimensionality* when the multi-dimensional data array is simply vectorised. We generalise the canonical correlation analysis of two sets of vectors into that of two higher-order tensors having multiple shared modes (or *axes*).

A single channel video volume is represented as a third-order tensor denoted by $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$, which has the three modes, i.e. axes of space (X and Y) and time (T). We assume that every video volume is of uniform size $I \times J \times K$. Thus third-order tensors can *share* any single mode or multiple modes. Note that the canonical transformations are applied to the modes which are not shared. For e.g. in (7.1), classical CCA applies the canonical transformations $\mathbf{u}, \mathbf{v}$ to the modes in $\mathbb{R}^{m_1}, \mathbb{R}^{m_2}$ respectively, having a shared mode in $\mathbb{R}^N$. The proposed Tensor CCA (TCCA) consists of the different architectures according to the number of shared modes. Joint-shared-mode TCCA allows any two modes (i.e. a section of video) to be shared and applies the canonical transformation to the remaining single mode, while the single-shared-mode TCCA shares any single mode (i.e. a scan line of video) and applies the canonical transformations to the two remaining modes. See Figure 7.2 for the concept of the proposed two types of TCCA.

The proposed TCCA for two videos is seen as the aggregation of many different canonical correlation analyses, which are for two sets of XY sections (i.e. images), two sets of XT or YT sections (in the joint-shared-mode), or sets of X,Y or T scan lines (in the single-shared-mode) of the videos.

Joint-shared-mode TCCA. Given two tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K}$, the joint-shared-mode TCCA consists of three sub-analyses. In each, one pair of canonical directions is found to maximise the inner product of the output tensors (called **canonical objects**) by the mode product of the two data tensors by the pair of the canonical transformations. That is, the

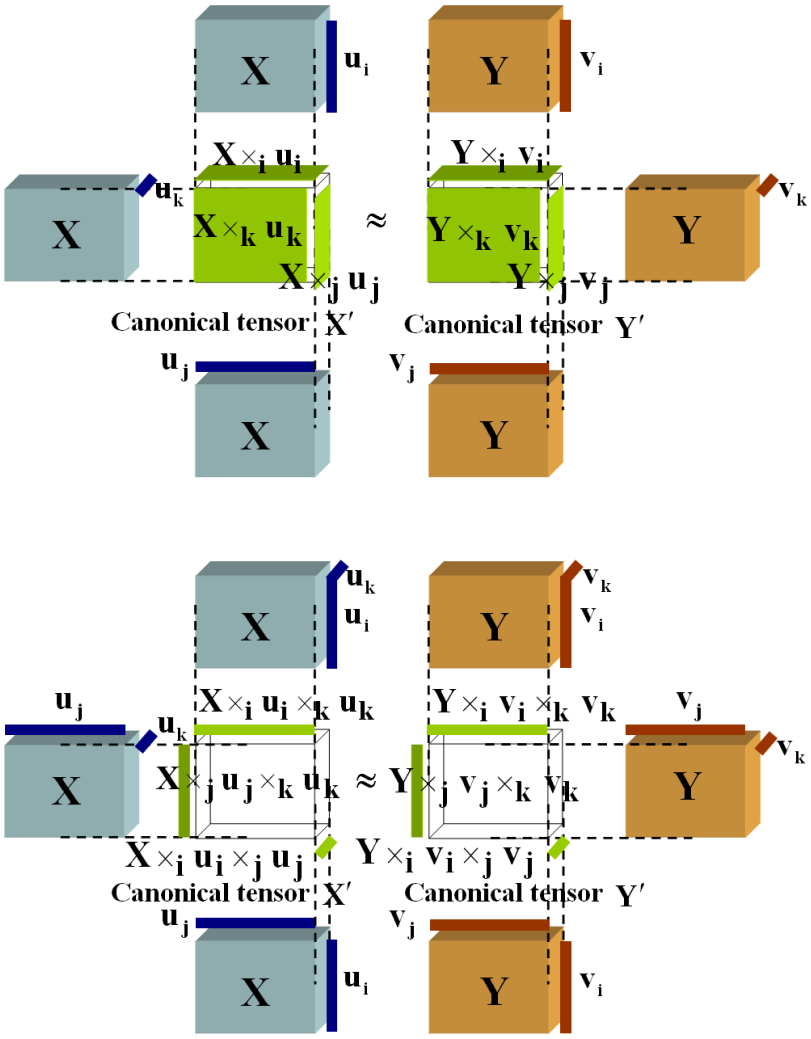


Figure 7.2: Representation of Tensor CCA. Joint-shared-mode TCCA (top) and single-shared-mode TCCA (bottom) of two video volumes (**X,Y**) are defined as the inner product of the canonical tensors (two middle transparent cuboids in each figure), which are obtained by finding the respective pairs of canonical transformations (**u,v**) and canonical objects (green planes in top or lines in bottom figure).

single pair (for e.g. $(\mathbf{u}_k, \mathbf{v}_k)$) in $\Phi = \{(\mathbf{u}_k, \mathbf{v}_k), (\mathbf{u}_j, \mathbf{v}_j), (\mathbf{u}_i, \mathbf{v}_i)\}$ is found to maximise the inner product of the respective canonical objects (e.g. $\mathcal{X} \times_k \mathbf{u}_k$, $\mathcal{Y} \times_k \mathbf{v}_k$) for the IJ, IK and JK joint-shared-modes respectively. The overall process of TCCA can then be given as the optimization problem of the canonical transformations Φ to maximise the inner product of the canonical tensors $\mathcal{X}', \mathcal{Y}'$ which are obtained from the three pairs of canonical objects by

$$\rho = \max_{\Phi} \langle \mathcal{X}', \mathcal{Y}' \rangle, \qquad \text{where} \tag{7.3}$$

$$(\mathcal{X}')_{ijk} = (\mathcal{X} \times_k \mathbf{u}_k)_{ij} \cdot (\mathcal{X} \times_j \mathbf{u}_j)_{ik} \cdot (\mathcal{X} \times_i \mathbf{u}_i)_{jk}$$

$$(\mathcal{Y}')_{ijk} = (\mathcal{Y} \times_k \mathbf{v}_k)_{ij} \cdot (\mathcal{Y} \times_j \mathbf{v}_j)_{ik} \cdot (\mathcal{Y} \times_i \mathbf{v}_i)_{jk}$$

and $\langle , \rangle$ denotes the inner product of tensors defined in Section 7.2.2. Note the mode product of the tensor by the single canonical transformation yields a matrix, a plane as the canonical object. Similar to classical CCA, multiple tensor canonical correlations $\rho_1, ..., \rho_d$ are defined by the orthogonal sets of the canonical directions.

Single-shared-mode TCCA. Similarly, the single-shared-mode tensor CCA is defined as the inner product of the canonical tensors comprising the three canonical objects. The two pairs of the transformations in

$$\Psi = [\{(\mathbf{u}_j^1, \mathbf{v}_j^1), (\mathbf{u}_k^1, \mathbf{v}_k^1)\}, \{(\mathbf{u}_i^2, \mathbf{v}_i^2), (\mathbf{u}_k^2, \mathbf{v}_k^2)\}, \{(\mathbf{u}_i^3, \mathbf{v}_i^3), (\mathbf{u}_j^3, \mathbf{v}_j^3)\}]$$

are found to maximise the inner product of the resulting canonical objects, by the mode product of the data tensors by the two pairs of the canonical transformations, for the I, J, K single-shared-modes. The tensor canonical correlations are

$$\rho = \max_{\Psi} \langle \mathcal{X}', \mathcal{Y}' \rangle, \qquad \text{where} \tag{7.4}$$

$$(\mathcal{X}')_{ijk} = (\mathcal{X} \times_j \mathbf{u}_j^1 \times_k \mathbf{u}_k^1)_i \cdot (\mathcal{X} \times_i \mathbf{u}_i^2 \times_k \mathbf{u}_k^2)_j \cdot (\mathcal{X} \times_i \mathbf{u}_i^3 \times_j \mathbf{u}_j^3)_k$$

$$(\mathcal{Y}')_{ijk} = (\mathcal{Y} \times_j \mathbf{v}_j^1 \times_k \mathbf{v}_k^1)_i \cdot (\mathcal{Y} \times_i \mathbf{v}_i^2 \times_k \mathbf{v}_k^2)_j \cdot (\mathcal{Y} \times_i \mathbf{v}_i^3 \times_j \mathbf{v}_j^3)_k$$

The canonical objects here are the vectors and the canonical tensors are given by the outer product of the three vectors as above.

Note that both joint-shared-mode and single-shared-mode TCCA are natural generalisations of the standard CCA to high-order tensors. Compared with a recent study [61], Harshman only considered a single-shared-mode, while we have proposed joint- as well as single-shared-modes. Our study further gives a general concept of multiple-shared-modes. Moreover, his work was limited to the third-order tensors having two modes of the same size and it can easily be extended to our single-shared-mode which considers all three modes of the same size. A novel alternating solution for the proposed tensor CCA is given in the next section.

7.3.2 Alternating Solution

A solution for both types of TCCA is proposed in a so-called *divide-and-conquer* manner. Each independent process is associated with the respective canonical objects and canonical transformations and also yields the canonical correlation features as the inner products of the canonical objects. This is done by performing SVD for CCA [11] a single time (for the joint-shared-mode TCCA) or several times alternatively (for the single-shared-mode TCCA). This section is devoted to explaining the solution for the I single-shared-mode for example. This involves the orthogonal sets of canonical directions $\{(\mathbf{U}_j, \mathbf{V}_j), (\mathbf{U}_k, \mathbf{V}_k)\}$ which contain $\{(\mathbf{u}_j, \mathbf{v}_j \in \mathbb{R}^J), (\mathbf{u}_k, \mathbf{v}_k \in \mathbb{R}^K)\}$ in their columns, yielding the d canonical correlations $(\rho_1, ...\rho_d)$ where $d < min(K, J)$ for given two data tensors, $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K}$. The solution is obtained by alternating the SVD method to maximise

$$\max_{\mathbf{U}_j, \mathbf{V}_j, \mathbf{U}_k, \mathbf{V}_k} \langle \mathcal{X} \times_j \mathbf{U}_j \times_k \mathbf{U}_k, \ \mathcal{Y} \times_j \mathbf{V}_j \times_k \mathbf{V}_k \rangle. \tag{7.5}$$

Given a random guess for $\mathbf{U}_j, \mathbf{V}_j$, the input tensors $\mathcal{X}, \mathcal{Y}$ are projected as $\widetilde{\mathcal{X}} = \mathcal{X} \times_j \mathbf{U}_j$, $\widetilde{\mathcal{Y}} = \mathcal{Y} \times_j \mathbf{V}_j$. The best pair of $\mathbf{U}_k^*, \mathbf{V}_k^*$ which maximises $\langle \widetilde{\mathcal{X}} \times_k \mathbf{U}_k, \widetilde{\mathcal{Y}} \times_k \mathbf{V}_k \rangle$ are then found. Letting

$$\widetilde{\mathcal{X}} \leftarrow \widetilde{\mathcal{X}} \times_k \mathbf{U}_k^*, \quad \widetilde{\mathcal{Y}} \leftarrow \widetilde{\mathcal{Y}} \times_k \mathbf{V}_k^*, \tag{7.6}$$

then the pair of $\mathbf{U}_j^*, \mathbf{V}_j^*$ are found to maximise $\langle \widetilde{\mathcal{X}} \times_j \mathbf{U}_j, \widetilde{\mathcal{Y}} \times_j \mathbf{V}_j \rangle$. Let

$$\widetilde{\mathcal{X}} \leftarrow \widetilde{\mathcal{X}} \times_j \mathbf{U}_j^*, \quad \widetilde{\mathcal{Y}} \leftarrow \widetilde{\mathcal{Y}} \times_j \mathbf{V}_j^*. \tag{7.7}$$

and repeat the procedures (7.6) and (7.7) until convergence. The solutions for steps (7.6), (7.7) are obtained as follows:

SVD method for CCA [11] is embedded in the proposed alternating solution. First, the tensor-to-matrix and the matrix-to-tensor conversion is defined as

$$\mathcal{A} \in \mathbb{R}^{I \times J \times K} \iff \mathbf{A}_{(ij)} \in \mathbb{R}^{(IJ) \times K} \tag{7.8}$$

where $\mathbf{A}_{(ij)}$ is a matrix which has K column vectors in $\mathbb{R}^{I \times J}$ which are obtained by concatenating all elements of the IJ planes of the tensor $\mathcal{A}$ into vectors. Let $\widetilde{\mathcal{X}}$, $\widetilde{\mathcal{Y}}$ be $\widetilde{\mathbf{X}}_{(ij)}$ and $\widetilde{\mathbf{Y}}_{(ij)}$ respectively. If $\mathbf{P}_{(ij)}^1, \mathbf{P}_{(ij)}^2$ denote two orthogonal bases matrices of $\widetilde{\mathbf{X}}_{(ij)}, \widetilde{\mathbf{Y}}_{(ij)}$ respectively, canonical correlations are obtained as singular values of $(\mathbf{P}^1)^T \mathbf{P}^2$ by

$$(\mathbf{P}^1)^T \mathbf{P}^2 = \mathbf{Q}_1 \mathbf{\Lambda} \mathbf{Q}_2^T, \quad \mathbf{\Lambda} = diag(\rho_1, ...\rho_K). \tag{7.9}$$

The mode products in (7.6) are accordingly given by $\widetilde{\mathcal{X}} \times_k \mathbf{U}_k^* \Longleftarrow \mathbf{G}_{(ij)}^1$, $\widetilde{\mathcal{Y}} \times_k \mathbf{V}_k^* \Longleftarrow \mathbf{G}_{(ij)}^2$ where $\mathbf{G}_{(ij)}^1 = \mathbf{P}^1 \mathbf{Q}_1$, $\mathbf{G}_{(ij)}^2 = \mathbf{P}^2 \mathbf{Q}_2$. The mode products in (7.7) are similarly found by the conversion of $\widetilde{\mathcal{X}} \Longrightarrow \widetilde{\mathbf{X}}_{(ik)}$, $\widetilde{\mathcal{Y}} \Longrightarrow \widetilde{\mathbf{Y}}_{(ik)}$. When it converges, d canonical correlations are obtained from the first d correlations of either $(\rho_1, ...\rho_K)$ or $(\rho_1, ...\rho_J)$, where $d < min(K, J)$. The canonical transformations, for e.g. in (7.6), are also obtained by

$$\mathbf{U}_k^* = (\widetilde{\mathbf{X}}_{(ij)}^T \widetilde{\mathbf{X}}_{(ij)})^{-1} \widetilde{\mathbf{X}}_{(ij)}^T \mathbf{P}^1 \mathbf{Q}_1$$

$$\mathbf{V}_k^* = (\widetilde{\mathbf{Y}}_{(ij)}^T \widetilde{\mathbf{Y}}_{(ij)})^{-1} \widetilde{\mathbf{Y}}_{(ij)}^T \mathbf{P}^2 \mathbf{Q}_2$$

All other component processes of TCCA can be similarly carried out, delivering the $6 \times d$ canonical correlation features in total. The J and K single-shared-mode TCCAs are performed in the same alternating fashion, while the IJ, IK, JK joint-shared-mode TCCA by performing the SVD method a single time without iterations.

7.4 Feature Selection for TCCA

TCCA features. By the proposed tensor CCA, we have obtained $6 \times d$ canonical correlation features in total. (Each joint-shared-mode or single-shared-mode has 3 distinct CCA

processes and each CCA process yields d features). Intuitively, each feature delivers different data semantics in explaining the data similarity. For example, the first few canonical tensors computed by the joint-shared-mode TCCA from the two hand-waving sequences are visualised in Figure 7.3 (a). Canonical objects of IJ, IK, JK joint-shared-mode are the XY, XT, YT planes of the cubes respectively and the canonical tensors are represented as cubes. The canonical objects (XY planes) of the IJ joint-shared-mode show the common spatial components of the two hand-waving videos. Note this mode is independent of temporal information, e.g. temporal ordering of the video frames by applying the canonical transformations to the K axis (time axis). Similarly, the canonical objects of the IK, JK joint-shared-mode reveal the common components of the two videos in the joint space-time domain, which are independent of J, I axis respectively.

Note from Figure 7.3 (a) that the canonical tensors in each pair are very much alike. The two input sequences belong to the same action class, i.e. hand waving, and have different backgrounds and lighting conditions. They are of quite distinct poses of individuals wearing different clothes. However, despite all the differences, the canonical tensors well capture mutual information of the two inputs yielding high correlations. In the canonical objects corresponding to XY planes, the arm movement, hand waving, is emphasised as common information. All other canonical objects (XT, YT planes) are also pairwise similar. On the other hand, the canonical tensors differ significantly from the paired in Figure 7.3 (b) where the two input sequences are of two action classes (one is hand waving and the other walking). Although these sequences were captured in the same environment and posed by the same person, TCCA returns low correlations. The two examples above suggest that the tensor canonical correlations can be good features for class discrimination.

Comparison of joint-shared and single-shared-modes. At this point, it may be worth comparing the proposed two types of TCCA in terms of flexibility and descriptive powers of the original data structures. Generally the single-shared-mode is more flexible and thus preserves less original data structure in matching than the joint-shared-mode. The former involves the two pairs of free transformations, whereas the latter the single pair. In terms of classification, the superiority of one type to the other may depend on applications. Good discriminative features should be well balanced between flexibility and data-descriptive powers. Importantly, in our experiments for action/gesture classification, we have observed that joint-shared-mode TCCA generally delivers better discriminative features than the single-shared-mode TCCA. The plane-like canonical objects in the joint-shared-mode seem to maintain sufficient descriptive information of action classes while giving robustness to data variations within classes (i.e. the flexibility), as shown in Figure 7.3. Our recent success [104] also supports this observation: the CCA of image sets [104] is identical with the IJ joint-shared-mode of the Tensor CCA framework in this paper. Nonetheless, the proposed single-shared-mode TCCA is also important: it helps a more general and unified TCCA framework. The accuracy of the joint-shared-mode TCCA could be improved by the single-shared-mode in our experiments. Furthermore, as discussed above, superiority may depend on applications.

Feature selection. In general, each canonical correlation feature carries a different amount of discrimination information for video classification, depending on applications. In this section, the discriminative boosting method is proposed to select useful tensor canonical correlation features. First, the intra-class and inter-class feature sets (i.e. canonical correlations ρ_i, $i = 1, ..., 6 \times d$ computed from any pair of videos) are generated from the

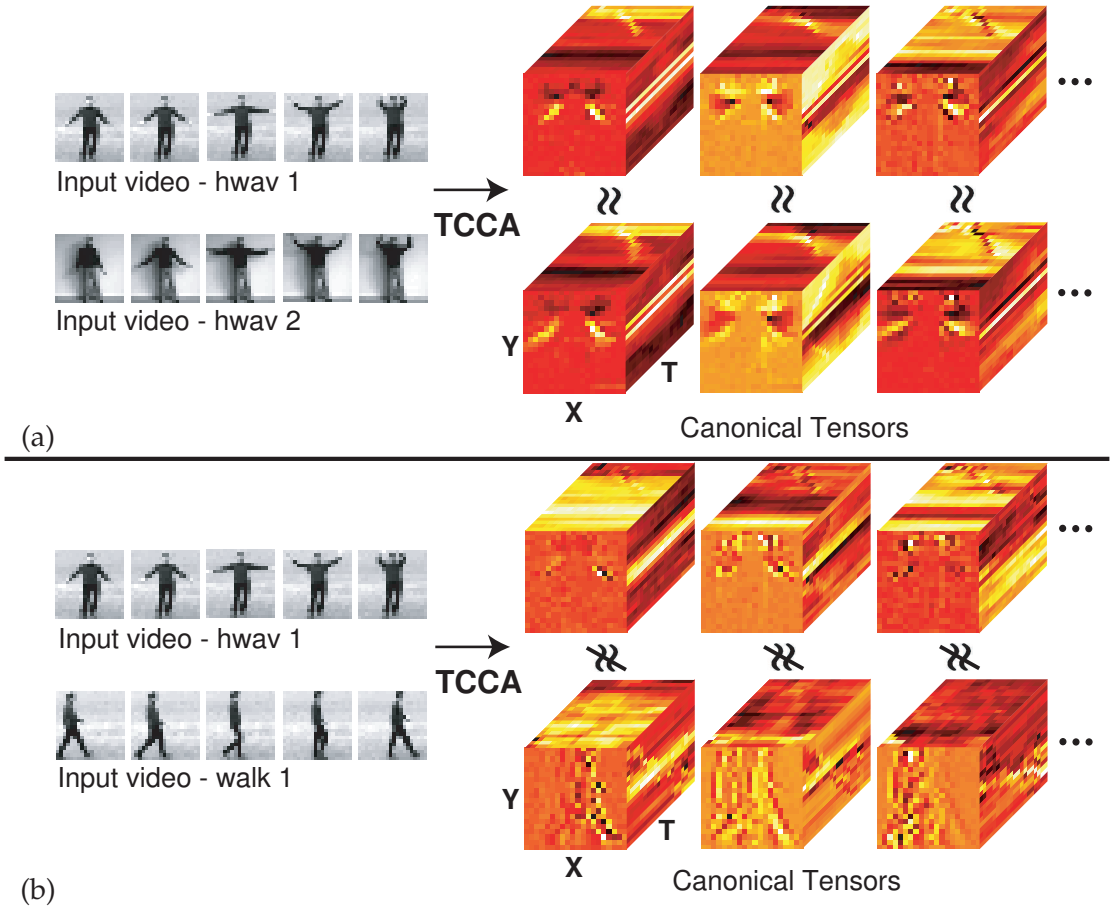


Figure 7.3: Examples of pairwise canonical tensors. This visualises the first few canonical tensors computed for the pair of input sequences of (a) the same action class and (b) the two action classes. Canonical objects of IJ, IK, JK joint-shared-mode are the XY, XT, YT planes of the cubes respectively. Note the canonical tensors in each pair are very much alike in (a) although the two hand-waving sequences are of different environments and poses of individuals wearing different clothes. On the other hand, the canonical tensors in (b) are greatly dissimilar despite the sequences, being of the same person in the same environment.

training data comprising of several class examples. We use each tensor CCA feature to build simple weak classifiers $\mathcal{M}(\rho_i) = sign\,[\rho_i - C]$ and aggregate the weak learners using the AdaBoost algorithm [42]. In an iterative update scheme classifier performance is optimized on the training data to yield the final strong classifier by

$$\mathcal{M}(\rho) = sign\left[\sum_{i=1}^{M} w_{L(i)}\mathcal{M}(\rho_{L(i)}) - \frac{1}{2}\sum_{i=1}^{M} w_{L(i)}\right] \tag{7.10}$$

where w contains the weights and L the list of the selected features. The feature list learnt by Adaboost is finally exploited to select the features for classification. Nearest Neighbor (NN) classification in terms of sum of canonical correlations chosen is performed to categorise a new test video.

Note the feature selection is performed as weak discriminative learning, as the main purpose of this chapter is to see the quality of tensor canonical correlation features themselves. In the next chapter, the combination with the discriminative transformation will be presented.

7.5 Action Detection by Tensor CCA

The proposed TCCA is time-efficient provided that actions or gestures are aligned in the space-time domain. Searching non-aligned actions by TCCA in the three-dimensional (X,Y, and T) input space is, however, still computationally demanding because every possible position and scale of the input space needs to be scanned. By observing that the joint-shared-mode TCCA does not require iterations for the solutions and delivers sufficient discriminative power (See Table 7.1), time-efficient action detection can be done by sequentially applying joint-shared-mode TCCA followed by single-shared-mode TCCA. The joint-shared-mode TCCA can effectively filter out the majority of samples which are far from a query sample then the single-shared-mode TCCA is applied to only few candidates. In this section, we mainly explain the method further to speed up the joint-shared-mode TCCA for action detection by incrementally learning the required subspaces. The following section gives a brief introduction of prior art on incremental Principal Component Analysis [59, 169].

7.5.1 Review on Incremental Principal Component Analysis

An efficient update scheme of eigen-subspaces has been developed in which a new set of vectors is added to an existing data set. This is useful for applications such as object tracking and surveillance where training images are accumulated over time. Given two sets of data represented by eigenspace models $\{\boldsymbol{\mu}_i, M_i, \mathbf{P}_i, \boldsymbol{\Lambda}_i\}_{i=1,2}$, where $\boldsymbol{\mu}_i$ is the mean, M_i the number of samples, $\mathbf{P}_i$ the matrix of eigenvectors and $\boldsymbol{\Lambda}_i$ the eigenvalue matrix of the i-th data set, the combined eigenspace model $\{\boldsymbol{\mu}_3, M_3, \mathbf{P}_3, \boldsymbol{\Lambda}_3\}$ is efficiently computed. The eigenvector matrix $\mathbf{P}_3$ is represented by

$$\mathbf{P}_3 = \boldsymbol{\Phi}\mathbf{R} = h([\mathbf{P}_1, \mathbf{P}_2, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2])\mathbf{R}, \tag{7.11}$$

where $\boldsymbol{\Phi}$ is the orthonormal column matrix spanning the entire combined data space, $\mathbf{R}$ is a rotation matrix, and h is a vector orthonormalization function. Using this representation, an original eigenproblem for $\mathbf{P}_3, \boldsymbol{\Lambda}_3$ is converted into a smaller eigenproblem as

$$\mathbf{S}_{T,3} = \mathbf{P}_3\boldsymbol{\Lambda}_3\mathbf{P}_3^T \;\Rightarrow\; \boldsymbol{\Phi}^T\mathbf{S}_{T,3}\boldsymbol{\Phi} = \mathbf{R}\boldsymbol{\Lambda}_3\mathbf{R}. \tag{7.12}$$

Note the matrix $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{S}_{T,3}\boldsymbol{\Phi}$ has the reduced size $d_{T,1} + d_{T,2} + 1$, where $d_{T,1}, d_{T,2}$ are the number of the eigenvectors in $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively. Thus the eigenanalysis here only takes $O((d_{T,1} + d_{T,2} + 1)^3)$ computations, whereas the eigenanalysis in the l.h.s. (7.12) requires $O(\min(N, M_3)^3)$, where N is the dimension of the input data and M_3 is the total number of data points. Usually, $N, M_3 \gg d_{T,1} + d_{T,2} + 1$.

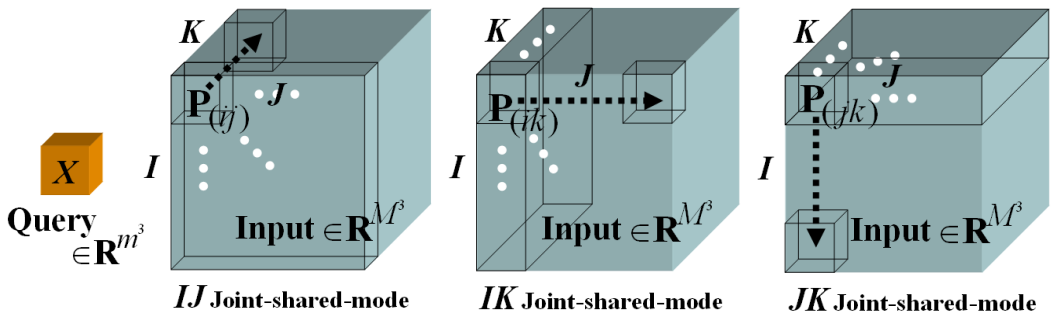


Figure 7.4: Detection scheme. A query video is searched in a large volume input video. TCCA between the query and every possible volume of the input video can be speeded-up by dynamically learning the three subspaces of all the volumes (*cuboids*) for the IJ, IK, JK joint-shared-mode TCCA. While moving the initial slices along one axis, subspaces of every small volume are dynamically computed from those of the initial slices.

7.5.2 Dynamic Subspace Learning for TCCA

The computational complexity of the joint-shared-mode TCCA in (7.9) depends on the computation of orthogonal basis matrices $\mathbf{P}^1, \mathbf{P}^2$ and the Singular Value Decomposition (SVD) of $(\mathbf{P}^1)^T\mathbf{P}^2$. The total complexity trebles this computation for the IJ, IK, JK joint-shared-mode. From the theory of [11], the first few eigenvectors corresponding to most of the data energy, which are obtained by Principal Component Analysis, can be the orthogonal basis matrices. If $\mathbf{P}^1 \in \mathbb{R}^{N\times d}, \mathbf{P}^2 \in \mathbb{R}^{N\times d}$ where d is a usually small number, the complexity of the SVD of $(\mathbf{P}^1)^T\mathbf{P}^2$ taking $O(d^3)$ is relatively negligible. Given the respective three sets of eigenvectors of a query video, time-efficient detection can be performed by incrementally learning the three sets of eigenvectors, the space-time subspaces $\mathbf{P}_{(ij)}, \mathbf{P}_{(ik)}, \mathbf{P}_{(jk)}$ of every possible volume (*cuboid*) of an input video for the IJ, IK, JK joint-shared-mode TCCA respectively. See Figure 7.4 for the concept. There are three separate steps which are carried out in same fashion, each of which is taken to compute one of $\mathbf{P}_{(ij)}, \mathbf{P}_{(ik)}, \mathbf{P}_{(jk)}$ of every possible volume of the input video. First, the subspaces of every cuboid of the initial slices of the input video are learnt, then the subspaces of all remaining cuboids are incrementally computed while moving the slices along one of the axes. For example, for the IJ joint-shared-mode TCCA, the subspaces $\mathbf{P}_{(ij)}$ of all cuboids in the initial IJ-slice of the input video are computed. The subspaces of all next cuboids are then dynamically computed, while pushing the initial cuboids along the K axis to the end as follows (for simplicity, let the size of the query video and input video be $\mathbb{R}^{m^3}, \mathbb{R}^{M^3}$ where $M \gg m$) :

The cuboid at k on the K axis, $\mathcal{X}^k$ is represented by the matrix $\mathbf{X}^k_{(ij)} = \{\mathbf{x}^k_{(ij)}, ..., \mathbf{x}^{k+m-1}_{(ij)}\}$, where $\mathbf{X}^k_{(ij)}$ is obtained by the tensor-to-matrix conversion defined in (7.8). The scatter matrix $\mathbf{S}^k = (\mathbf{X}^k_{(ij)})(\mathbf{X}^k_{(ij)})^T$ is written w.r.t. the scatter matrix of the previous cuboid at $k-1$ as

$$\mathbf{S}^k = \mathbf{S}^{k-1} + (\mathbf{x}^{k+m-1}_{(ij)})(\mathbf{x}^{k+m-1}_{(ij)})^T - (\mathbf{x}^{k-1}_{(ij)})(\mathbf{x}^{k-1}_{(ij)})^T. \tag{7.13}$$

This involves both incremental and decremental learning. A new vector $\mathbf{x}^{k+m-1}_{(ij)}$ is added and an existing vector $\mathbf{x}^{k-1}_{(ij)}$ is removed from the $(k-1)$-th cuboid. The sufficient spanning

set [1] of the current scatter matrix can be $\Upsilon = h([\mathbf{P}_{(ij)}^{k-1}, \mathbf{x}_{(ij)}^{k+m-1}])$ where h is a vector orthogonalization function and $\mathbf{P}_{(ij)}^{k-1}$ is the IJ subspace of the previous cuboid. The eigenvectors of the current scatter matrix can be seen as the product of the sufficient spanning set by an arbitrary rotation matrix $\mathbf{R}$ as $\mathbf{P}_{(ij)}^{k} = \Upsilon\mathbf{R}$. Therefore the original eigen-problem is reduced to the much smaller eigenproblem by

$$\mathbf{S}^k = \mathbf{P}_{(ij)}^{k}\mathbf{\Lambda}^k(\mathbf{P}_{(ij)}^{k})^T \quad \Rightarrow \quad \Upsilon^T\mathbf{S}^k\Upsilon = \mathbf{R}\mathbf{\Lambda}^k\mathbf{R}. \tag{7.14}$$

The matrices $\mathbf{\Lambda}^k, \mathbf{R}$ are computed as the eigenvalue and eigenvector matrix of $\Upsilon^T\mathbf{S}^k\Upsilon$. The final eigenvectors are obtained as $\mathbf{P}_{(ij)}^{k} = \Upsilon\mathbf{R}$ after removing the components in $\mathbf{R}$ corresponding to the least eigenvalues in $\mathbf{\Lambda}^k$, keeping the dimension of $\mathbf{P}_{(ij)}^{k}$ be $R^{m^2\times d}$.

Computational Cost. Similarly, the subspaces $\mathbf{P}_{(ik)}, \mathbf{P}_{(jk)}$ for the IK, JK joint-shared-mode TCCA are computed by moving the all cuboids of the slices along the I, J axes respectively. In this way, the total complexity of learning of the three kinds of the subspaces of every cuboid is significantly reduced s.t.

$$O(M^3 \times m^3) \longrightarrow O(M^2 \times m^3 + M^3 \times d^3) \tag{7.15}$$

as $M \gg m \gg d$. $O(m^3), O(d^3)$ are the complexity for solving eigen-problems in a batch (i.e. the l.h.s. of (7.14)) and the proposed way (the r.h.s. of (7.14)). Efficient multi-scale search may be also plausible, merging two or more subspaces of smaller cuboids in a similar way. This issue is retained as future work.

7.6 Experimental Results

Hand-Gesture Recognition. We acquired *Cambridge-Gesture data base* [2] consisting of 900 image sequences of 9 hand gesture classes, which are defined by 3 primitive hand shapes and 3 primitive motions (see Figure B.1). Each class contains 100 image sequences (5 different illuminations×10 arbitrary motions of 2 subjects). Each sequence was recorded in front of a fixed camera having roughly isolated gestures in space and time. All training was performed on the data acquired in the single plain illumination setting (the leftmost in Figure B.1 (b)) while testing was performed on the data acquired in the remaining settings. See Appendix B for more details on the data set.

All video sequences were uniformly resized into $20 \times 20 \times 20$ in our method. The proposed alternating solution in Section 7.3.2 was applied to obtain the TCCA features of every pairwise training sequences. The alternating solution stably converged, as shown in the left of Figure 7.6. Feature selection was performed for the TCCA features based on the weights and the list of the features learnt from the AdaBoost method in Section 7.4. In the middle of Figure 7.6, it is shown that about the first 60 features contained most of the discriminatory information. Of the first 60 features, the number of the selected features is

[1] The sufficient spanning set is an economical set of bases which can span most data energy, which helps to obtain a small eigen-problem to solve [59, 169].

[2] The database is publicly available at ftp://mi.eng.cam.ac.uk/pub/CamGesData. Contact e-mails: tkk22@cam.ac.uk

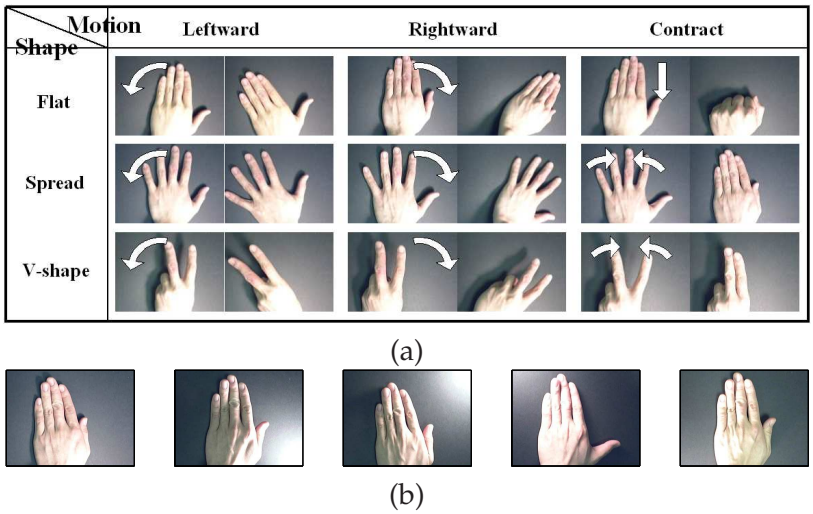


Figure 7.5: Hand-Gesture database. (a) 9 gestures generated by 3 primitive shapes and motions. (b) 5 illumination conditions in the database.

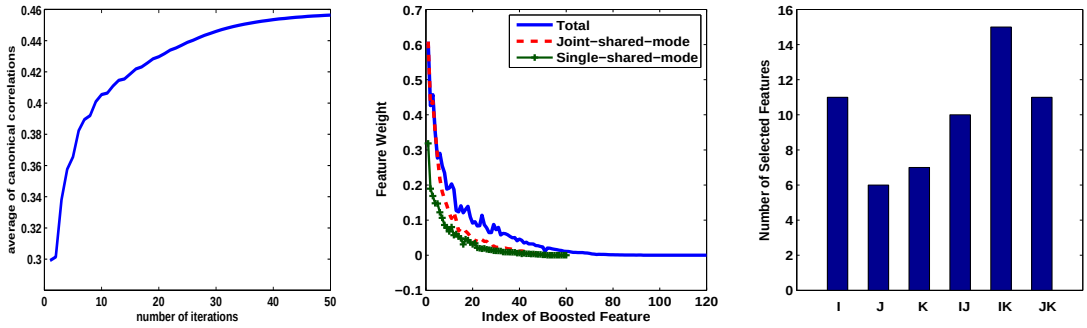


Figure 7.6: Feature selection. (left) Convergence graph of the alternating solution for TCCA. (mid) The weights of TCCA features learnt by boosting. (right) The number of TCCA features chosen for the different shared-modes.

	Joint-shared-mode	Single-shared-mode	Dual-mode
Number of features	01 05 20 - 60	60	60
Accuracy (%)	52 72 76 - 76	52	81

Table 7.1: Accuracy comparison of the joint-shared-mode TCCA and dual-mode TCCA (using both joint and single-shared mode).

shown for the different shared-mode TCCA in the right of Figure 7.6. The joint-shared-mode (IJ, IK, JK) contributed more than the single-shared-mode (I, J, K) but both still kept many features in the selected feature set. From Table 7.1, the best accuracy of the joint-shared-mode was obtained by 20 - 60 features. This is easily reasoned when looking at the weight curve of the joint-shared-mode in Figure 7.6 where the weights of more than 20 features are non-significant. Note that the accuracy monotonically increased delivering the best accuracy at 60 even without feature selection. The single-shared-mode alone gave relatively poor accuracy, which is yet significant compared with those of other methods in Table 7.2. The dual-mode TCCA (using both joint and single-shared mode)

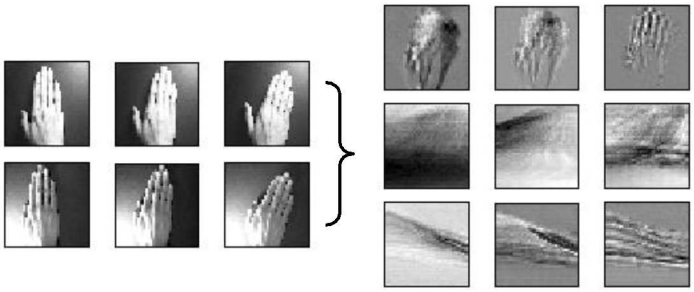


Figure 7.7: Example of canonical objects. Given two different lighting sequences of the same hand gesture class (the left two rows), the first three canonical objects of the IJ,IK,JK joint-shared-mode are shown on the top, middle, bottom rows respectively.

Methods	set1	set2	set3	set4	total
TCCA	81	81	78	86	82±3.5
CCA [104]	63	61	65	69	65±3.2
pLSA [143]	70	57	68	71	66±6.1
MGO/RVM [202]	-	-	-	-	44
MGO/SVM [202]	-	-	-	-	30

Table 7.2: Hand-gesture recognition accuracy (%) of the four illumination sets.

with the same number of features improved the accuracy of the joint-shared mode by 5%. NN classification was performed for a new test sequence based on the selected TCCA features. Figure 7.7 shows the example of canonical objects computed from the two lighting sequences of the same hand gesture class. One of each pair of canonical objects only is shown here, as the other is almost similar.

Table 7.2 shows the recognition rates of the proposed TCCA method, the simple CCA method [104], Niebles et al.'s method [143] (the probabilistic Latent Semantic Analysis (pLSA) with the space-time descriptors, which exhibited the best action recognition accuracy among the state-of-the-arts in [143]), Wong et al.'s method (Support Vector Machine/or Relevance Vector Machine (RVM) with the Motion Gradient Orientation image (MGO) [202]).The original codes and the best settings of the parameters (e.g. the size parameters of the space-time descriptors and the size of the code book) were used in the evaluation for the previous works. The two methods of SVM/or RVM on the MGO images turned out far worse. As observed in [202], using RVM improved the accuracy of SVM by about 10%. However, both methods often failed to discriminate the gestures, which have the same motion of the different shapes, as the methods are mainly based on motion information of gestures. The failure of the SVM/RVM methods in this scenario might be partly due to *small sample size* problem. Large difference in illumination conditions of the train and the test sets might have led the performance degradation of the classifiers. Also, the holistic representation of videos seemed too rigid to cope with the intra-class variations in spatial temporal alignment of the gesture sequences. The unsupervised learning method pLSA with the space-time interest points and the simple CCA method achieved the second-rank accuracy by more flexible representation or matching: the pLSA method is based on part-based representation, i.e. distribution of local patterns and CCA provides affine-invariance in matching. However, note that accuracy of the pLSA method is highly dependent on good parameter setting (of the space-time descriptors), which is hard in

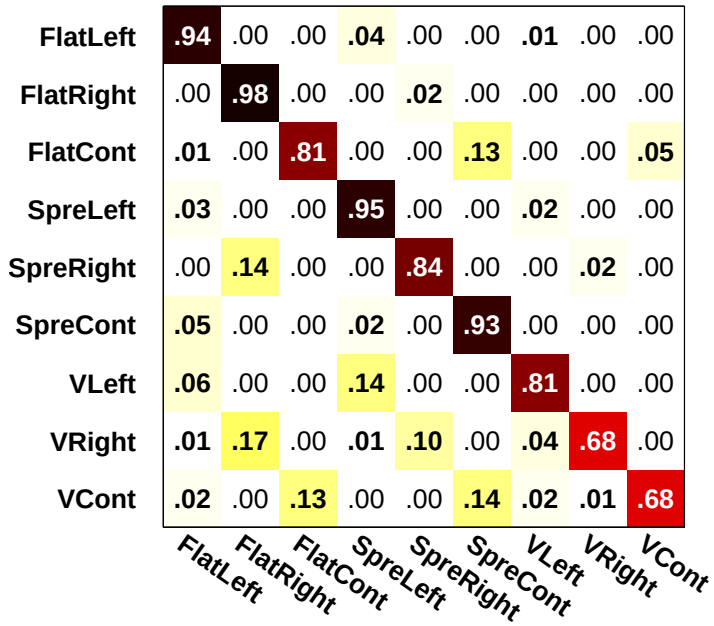


Figure 7.8: Confusion matrix of hand gesture recognition.

practice. In the meantime, neither method takes full video information into account: pLSA does not encode global shape information while CCA temporal information. The proposed method, TCCA, significantly outperformed all comparative methods. The proposed tensor extension of CCA improved around 17% over the simple CCA method. By matching both spatial and temporal information with the affine-invariance, the proposed method is far better at correct identification of the sequences of distinct shapes subject to similar motion as well as the similar shape sequences having different motions. See Figure 7.8 for the confusion matrix of our method.

Action Categorisation on KTH Data Set. We followed the experimental protocol of Niebles et al.'s work [143] on the KTH action data set, which is the largest public action data base [165]. The data set contains six types (boxing, hand clapping, hand waving, jogging, running and walking) of human actions performed by 25 subjects in 4 scenarios. The original input videos contain actions which are not strictly space-time aligned and are repeated several times. Leave-one-out cross-validation was performed to test the methods, i.e. for each run the videos of 24 subjects are exploited for training and the videos of the remaining subject is for testing. Some sample videos are shown in Figure 7.9 with the indication of the action alignment (or cropping). This space-time alignment of actions was carried out manually for accuracy comparison but can also be automatically achieved by the proposed detection scheme, as will be shown below. The defined aligned actions contain unit motions without any repetition. Most competing methods are based on histogram representations of the local Space-Time interest points with Support Vector Machine (ST/SVM) (Dollar et al [32], Schuldt et al. [165]) or pLSA (Niebles et al. [143]). Ke et al. applied the spatio-temporal volumetric features [85]. As these methods do not exploit any global space-time shape information, they do not require alignment of actions in nature. These methods were, therefore, applied to the original unsegmented input videos. For compari-

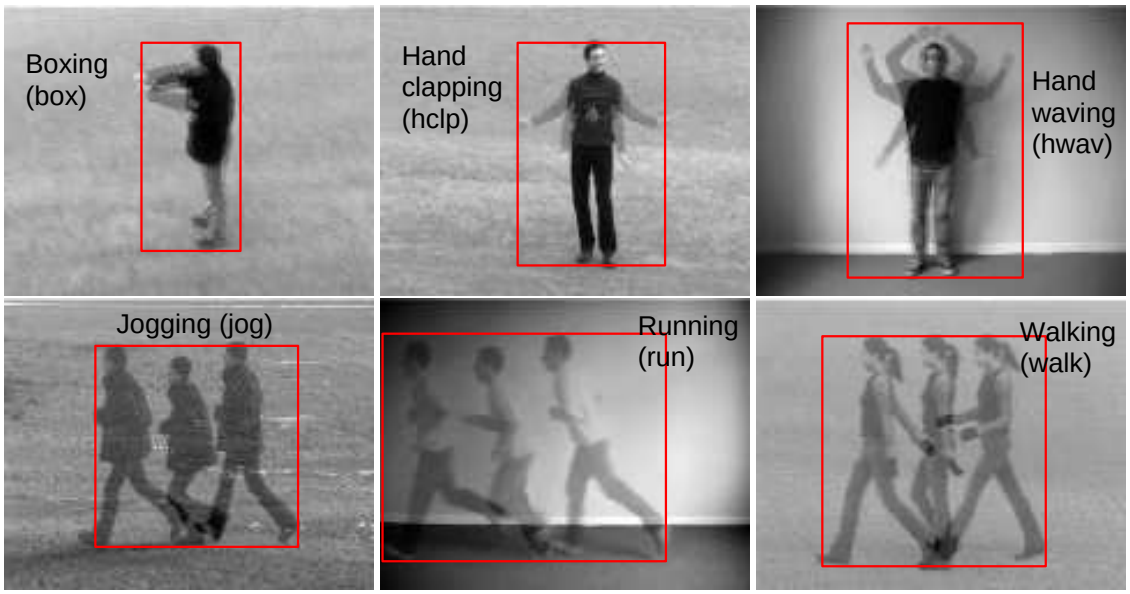


Figure 7.9: Example action videos in KTH data set. The bounding boxes indicate the spatial alignment and the superimposed images of the initial, intermediate and the last frames of each action show the temporal segmentation of action classes.

son, we quoted the accuracy of the methods reported in [143] and further performed testing of the simple CCA method, the pLSA method [143] and the proposed TCCA method on the segmented videos. In TCCA method, the aligned video sequences were uniformly resized to $20 \times 20 \times 20$. See Table 7.3 for accuracy comparison of several methods and Figure 7.10 for the confusion matrices of our TCCA method and the CCA method. The pLSA method on the segmented videos reduced the accuracy of the pLSA on the original input videos by about 10%, maybe due to insufficient number of interest points detected in the segmented videos. Note the original sequences contain several repetitions of the actions giving much more fluent local representation. The SVM applied to the same histogram representation as that of the pLSA [32] delivered similar accuracy to that of pLSA. While most of the histogram-based methods delivered accuracy around 60-80%, the proposed TCCA method and the CCA method achieved impressive accuracy at 95% and 89% respectively. From the good accuracy of the CCA method, one may legitimately presume that the six different action classes of the KTH data set are quite well discriminative in spatial domain. Most of the histogram-based methods lost the important information in the global space-time shapes of actions, resulting in ambiguity for spatial variations of the different action classes. As expected, the TCCA method improved the CCA method by using joint spatial-temporal information, being better particularly in discrimination between the jogging and running actions, which is clear in the confusion matrices in Figure 7.10.

There have been recent attempts to combine the structural information with the local information based on the local Space-Time interest points [203, 163]. As shown in the last row of Table 7.3, they achieved reasonable improvements over the methods based purely on local information, but were still inferior to the method proposed.

Action Detection on KTH Data Set. Action detection was performed by a training set consisting of sequences of five persons, not including any tested persons. Every possi-

Methods	(%)	Methods	(%)
TCCA	95.33	ST/SVM [32]	81.17
CCA [104]	89.50	ST/SVM [165]	71.72
pLSA [143]	81.50	Ke et al. [85]	62.96
pLSA* [143]	68.53		
pLSA-ISM [203]	83.92	Savarese et al. [163]	86.83

Table 7.3: Recognition accuracy (%) on the KTH action data set. pLSA* denotes the pLSA method applied to the segmented videos.

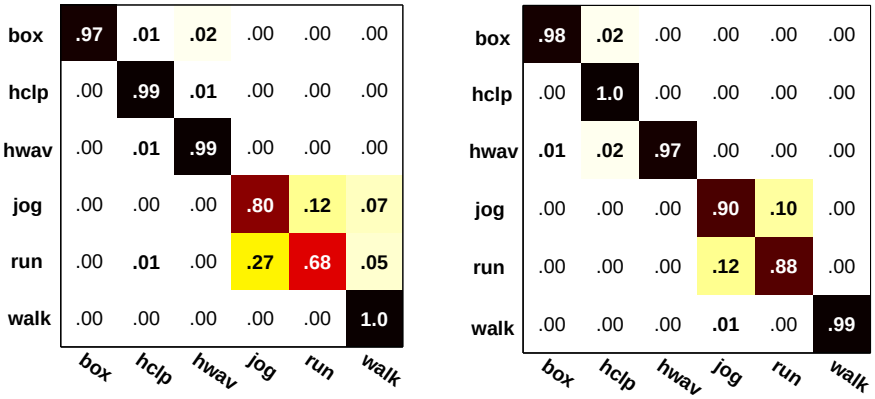


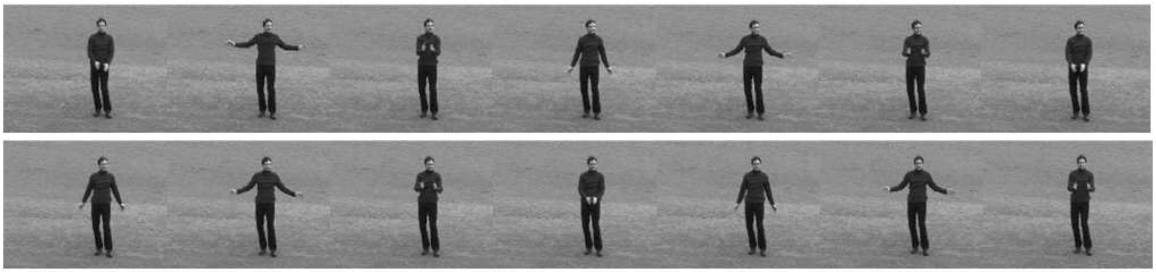
Figure 7.10: Confusion matrix of CCA (left) and TCCA (right) method for the KTH data set. The six action classes of the KTH data set are quite well discriminative in spatial domain by CCA. TCCA improved CCA especially by better discriminating between the jogging and running actions.

action class	box	hclp	hwav	jog	run	walk
dynamic subspace learning	43.01	35.42	19.27	12.60	5.16	10.70
or batch subspace learning	240.26	245.45	47.62	64.10	19.72	40.34
+ TCCA	9.96	8.43	2.26	3.09	1.14	2.21

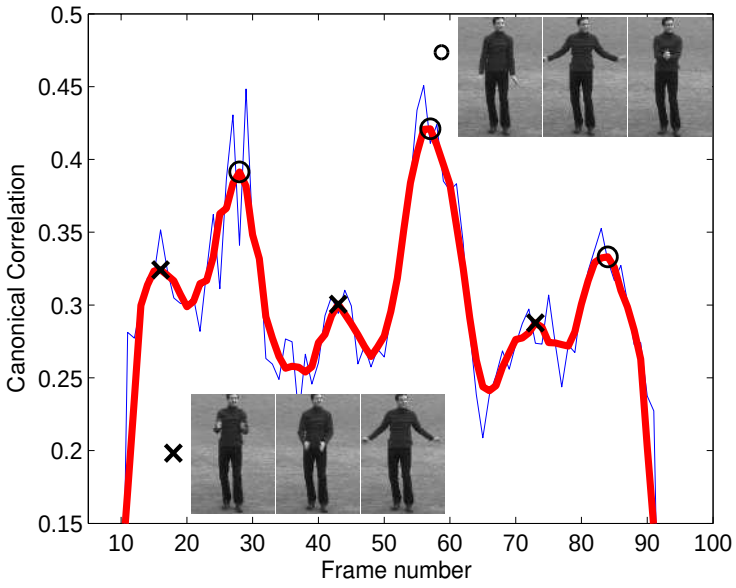
Table 7.4: Action detection time (seconds) for the fixed scale and a single query. The detection speed differs for the size of input volume with respect to the size of query volume.

ble volume in an input video is scanned and is matched with the training sequences by TCCA. Figure 7.11 shows the detection results for the continuous hand clapping video, which comprises the three correct unit clapping actions defined. The maximum canonical correlation value is shown for every frame of the input video. All three correct hand clapping actions are detected at the three highest peaks, with the three intermediate actions at the three lower peaks. The intermediate actions which exhibited local maxima between any two correct hand-clapping actions had different initial and end postures from those of the correct actions.

For the fixed scale search, the detection time of the proposed method is reported in Table 7.4 on a Pentium 4 3GHz using non-optimized Matlab code. The incremental subspace learning much reduced the detection time of the batch computation. The detection speed differs for the size of input volume with respect to the size of query volume. For example, the volume sizes of the input video and the query video for the hand clapping actions



(a)



(b)

Figure 7.11: Action detection result. (a) An example input video sequence of continuous hand-clapping actions. (b) The detection result: all three correct hand-clapping actions are detected at the highest three peaks, with the three intermediate actions at the three lower peaks.

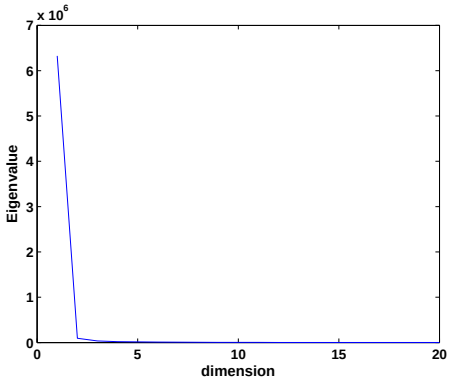


Figure 7.12: Eigenvalue plot. Averaged eigenvalue plot of the three kinds of subspaces of different actions.

(pixels)	box	hclp	hwav	jog	run	walk
X	48±8	60±11	68±10	80±20	101±26	71±18
Y	91±10	87±10	92±12	86±12	82±13	84±13
T	32±8	22±6	19±4	11±2	9±1	15±1

Table 7.5: Average volume size of action classes. The mean and the standard deviations along each axis are shown.

are $120 \times 160 \times 102$ and $92 \times 64 \times 19$ respectively. The dimension of the input video and query video was reduced by the factors $4.6, 3.2, 1$ (for the respective three dimensions). In the reduced dimension, the size of the query video, m in (7.15) was 20. The dimension of the subspaces, d in (7.15) were chosen as 5 to reflect most data energy from the eigenvalue plot in Figure 7.12. If the search area M and the size of the query video m were larger, the computational saving by the proposed method could be even greater. The speed obtained seems to be comparable to that of the state-of-the-art [168] and fast enough to be integrated into a real-time system if provided with a smaller search area either by system setting or by simple video processing techniques for finding the *focus of attention*, e.g. by moving area segmentation.

Figure 7.13 shows some action detection results with the scale variations, which are obtained by the three steps in each axis. The three steps are the mean, mean plus/minus the standard deviation of the scales of all actions as given in Table 7.5. The detection results show the alignment of the best response region in the original input sequences. Despite the small training samples (of the five persons as mentioned) and the coarse scale search, the automatic alignment results were very close to the manual settings shown in Figure 7.9. Note the temporal alignment was as reliable as the spatial localization. The estimated initial, intermediate and last poses of actions, as shown in the superimposed images, look similar to those of the manually defined action classes in Figure 7.9. Some less accurate placement might be caused mostly by the rough three steps in multi-scale search or by insufficient variation contained in the training sequences. Further studies on efficient multi-scale search may help more accurate and yet time-efficient action detection.

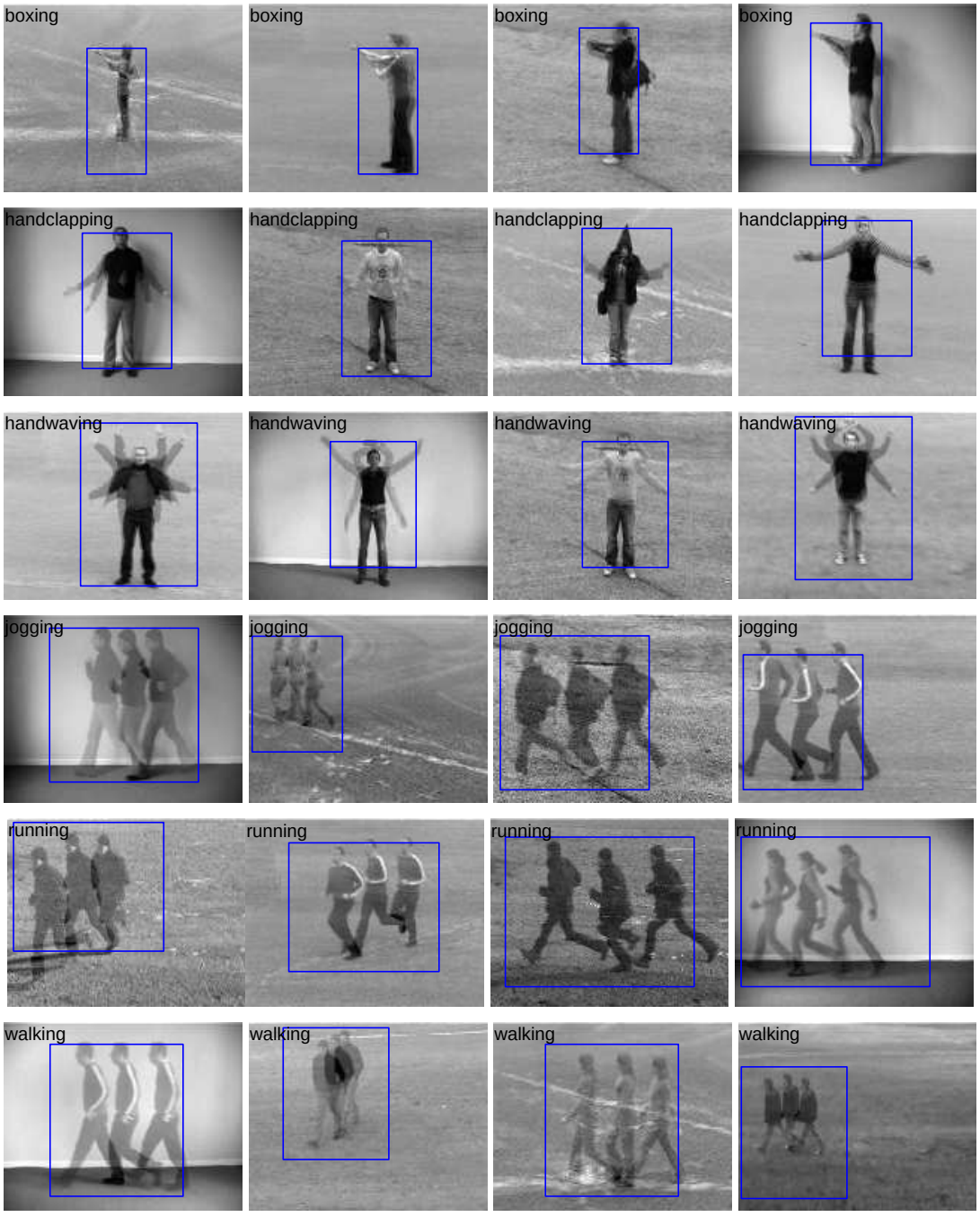


Figure 7.13: Automatic action detection result. The bounding boxes indicate the spatial alignment and the superimposed images of the initial, intermediate and the last frames of each action show the temporal segmentation of action classes.

7.7 Summary

We have proposed a novel method, Tensor Canonical Correlation Analysis (TCCA) which extracts pairwise flexible and yet descriptive correlation features of videos in the joint space-time domain. The proposed features combined with NN classifier significantly improved the accuracy over state-of-the-art action recognition methods. The proposed method is also practically appealing as it does not require any significant tuning parameters. Additionally, the proposed detection method for TCCA could yield time-efficient action detection or alignment in large volume video inputs.

In spite of the efficient detection method, further speeding up the method is needed. The computational complexity of the current detection method can still be demanding in the scenarios which have a much larger search space and/or require multi-scale search in real time. One may try a hierarchical approach which applies one or several simpler but less accurate methods to filter out majority of candidates and then our method, with the benefit of high accuracy. Efficient multi-scale search by merging the space-time subspaces of TCCA would constitute useful future work.

For further enhancement in accuracy, more class priors may be exploited for the TCCA method. The proposed method as a general meta-algorithm may be combined with other methods (e.g. a task-specific representation or segmentation methods) for further improvement. In the next chapter, the raw pixel representation in the TCCA method is replaced with the Scale-Invariant-Feature-Transform (SIFT) based representation for more robust gesture classification. In addition, the discriminative transformation will be combined with the TCCA method.

Part III

Integration

CHAPTER 8

Integrating Discriminant Function and SIFT to Tensor CCA for Gesture Recognition

In this chapter, we propose a discriminant function for Tensor Canonical Correlation Analysis so that a resulting method is more discriminative for better gesture recognition. In the previous chapter, the quality of the derived features by TCCA was demonstrated in terms of action recognition accuracy by being combined with a simple Nearest Neighbor classifier. Since TCCA comprises many sub-CCAs, the discriminant analysis method developed for CCA in Chapter 6 can be conveniently integrated to the TCCA method. Moreover, the proposed discriminant analysis is a general learning method so it can be be further improved by a task-specific representation. This is shown by combining the *Scale-Invariant-Feature-Transform (SIFT)* into the method for more robust gesture representation. The combined method delivered better accuracy in the experiments using 900 videos of 9 hand gesture classes.

8.1 TCCA with Discriminant Function

Canonical Correlation Analysis (CCA) was extended to two multidimensional data arrays in Chapter 7. The method called Tensor Canonical Correlation Analysis has two architectures: the joint-shared and single-shared-modes (See Chapter 7). The method described here is related to the joint-shared-mode which exploits *planes* rather than *scan vectors* of

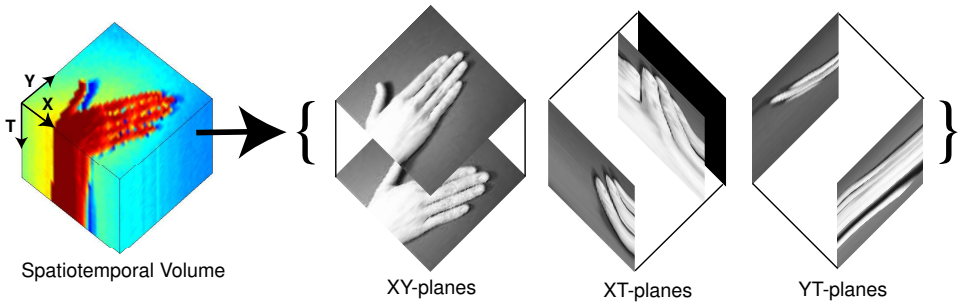


Figure 8.1: Spatiotemporal Data Representation.

two videos. The TCCA method can be re-interpreted as follows: A gesture video is represented by firstly decomposing an input video clip (i.e. a spatiotemporal volume) into three sets of orthogonal planes, namely XY-, YT- and XT-planes, as shown in Figure 8.1. This allows posture information in XY-planes and joint posture/dynamic information in YT and XT-planes. Three kinds of subspaces are learnt from the three sets of planes (which are converted into vectors by raster-scanning). Gesture recognition is then done by comparing these subspaces with the corresponding subspaces from the models by classical Canonical Correlation Analysis, which measures *principal angles* between subspaces (See Chapter for the solution of CCA). By comparing subspaces of an input and a model, robust gesture recognition can be achieved up to pattern variations on the subspaces. The similarity of any model $\mathcal{D}_m$ and query spatiotemporal data $\mathcal{D}_q$ is defined as the weighted sum of the normalized canonical correlations of the three subspaces by

$$\mathcal{F}(\mathcal{D}_m, \mathcal{D}_q) = \Sigma_{k=1}^{3} w^k \mathcal{N}^k(\mathbf{P}_m^k, \mathbf{P}_q^k) \tag{8.1}$$

where,

$$\mathcal{N}^k(\mathbf{P}_m^k, \mathbf{P}_q^k) = (\mathcal{G}(\mathbf{P}_m^k, \mathbf{P}_q^k) - m^k)/\sigma^k, \tag{8.2}$$

$\mathbf{P}^1, \mathbf{P}^2, \mathbf{P}^3$ denotes a matrix containing the first few eigenvectors in its columns of XY-planes, XT-planes, YT-planes respectively and $\mathcal{G}(\mathbf{P}_m, \mathbf{P}_q)$ sum of the canonical correlations computed from $\mathbf{P}_m, \mathbf{P}_q$. The normalization parameters with index k are mean and standard deviation of matching scores, i.e. $\mathcal{G}$ of all pairwise videos in a validation set for the corresponding planes.

The discriminative spatiotemporal canonical correlation is defined by applying the discriminative transformation in Chapter learnt from each of the three data domains as

$$\mathcal{H}(\mathcal{D}_m, \mathcal{D}_q) = \Sigma_{k=1}^{3} w^k \mathcal{N}^k(h(\mathbf{Q}^{kT}\mathbf{P}_m^k), h(\mathbf{Q}^{kT}\mathbf{P}_q^k)), \tag{8.3}$$

where h is a vector orthonormalization function and $\mathbf{Q}^k$ are the discriminative transformation matrix learnt over the corresponding sets of planes. The discriminative matrix is found to maximise the canonical correlations of within-class sets and minimises the canonical correlations of between-class sets by analogy to the optimization concept of Linear Discriminant Analysis (LDA) (See Chapter 6 for details). On the transformed space, gesture video classes are more discriminative in terms of canonical correlations. In this work, this concept has been validated not only for the spatial domain (XY-subspaces) but also for the spatiotemporal domains (XT-, YT-subspaces).

Discussion.

The proposed method is a so-called *divide-and-conquer* approach which partitions original input space into the three different data domains, learning the canonical correlations on each domain, and then aggregating them with proper weights. In this way, the original data dimension N^3, where N is the size of each axis, is reduced into $3\times N^2$ so that the data is conveniently modelled. As shown in Figure 8.2a-c, each data domain is well-characterised by the corresponding low-dimensional subspace (e.g. hand shapes in XY-planes, joint spatial and temporal information in YT-, and XT- planes).

The method is, moreover, robust in using mutual (or canonically correlated) components of the pairwise subspaces. By finding the mutual components of maximum correla-

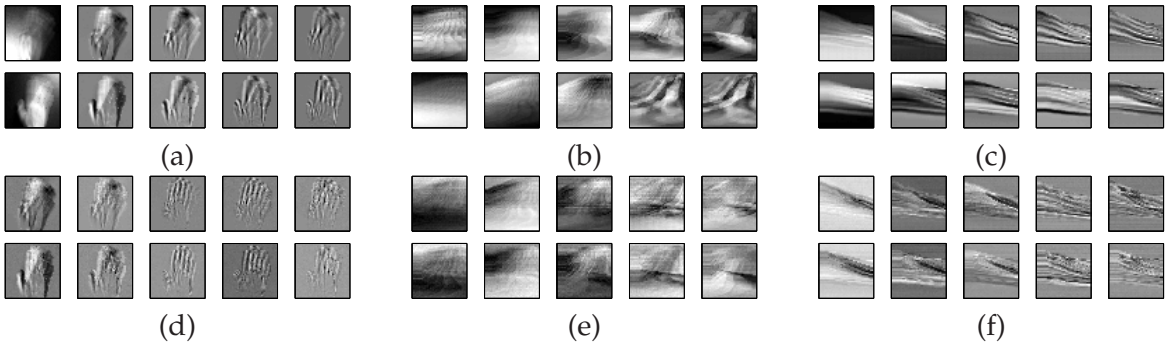


Figure 8.2: Principal Components and Canonical Vectors. The first few principal components of the (a) XY (b) XT (c) YT subspaces of two different illumination sequences of a gesture class are shown at the top and bottom row respectively. The corresponding pairwise canonical vectors are visualised in (d) - (f). Despite the different lighting conditions of the two input sequences, the canonical vectors in the pair (top and bottom) are very much alike, capturing common modes.

tions, which are canonical correlations, some undesirable information for classification can be filtered out. See Figure 8.2 for the principal components and canonical vectors for the given two sequences of the same gesture class which were captured under the differing lighting conditions. Whereas the first few principal components mainly corresponded to the different lighting conditions (in Figure 8.2a-c), the canonical vectors (in Figure 8.2d-f) well captured the common modes of the two sequences, being visually the same in each pair. In other words, the lighting variations across the two sets were removed in the process of CCA, as it is invariant to any variations on the subspaces. Many previous studies have shown that lighting variations are often confined to a low-dimensional subspace.

8.2 SIFT Descriptor for Spatiotemporal Volume Data

Edge-based description of each plane of videos can help the method achieve more robust gesture recognition. In this section we propose a simple and effective SIFT (Scale-Invariant Feature Transform) [130] representation for a spatiotemporal data by a fixed grid. As explained, the spatiotemporal volume is broken down into three sets of orthogonal planes (XY-, YT- and XT-planes) in the method. Along each data domain, there is a finite number of planes which can be regarded as images. Each of these images is further partitioned into $M \times N$ patches in a predefined fixed grid and the SIFT descriptor is obtained from each patch (see Figure 8.3a). For each image, the feature descriptor is obtained by concatenating the SIFT descriptors of several patches in a predefined order. The SIFT representation of the three sets of planes is directly integrated into the proposed method in Section 8.1 by replacing the sets of image vectors with the sets of the SIFT descriptors prior to canonical correlation analysis. The experimental results show that the edge-based representation generally improves the intensity-based representation in both of the joint space-time domain (YT-, XT-planes) and the spatial domain (XY-planes).

SIFT obtained from 3D blocks. This section presents a general 3D extension of SIFT features. Traditional classifiers such as Support Vector Machine (SVM)/ Relevance Vector Machine (RVM) are applied to the video data represented by the 3D SIFT so that they can

be compared with the proposed method (with SIFT) in the same input space. Given a spatiotemporal volume representing a gesture sequence, the volume is firstly partitioned into $M \times N \times T$ tiny blocks. Within each tiny block, further analysis is conducted along XY-planes and YT-planes (see Figure 8.3b). For analysis on a certain plane, say XY-planes, derivatives along X- and Y- dimensions are obtained and accumulated to form several regional orientation histograms (under a 3D Gaussian weighting scheme). For each tiny block, the resultant orientation histograms of both planes are then concatenated to form the final SIFT descriptor of dimension 256. The descriptor for the whole spatiotemporal volume can then be formed by concatenating the SIFT descriptors of all tiny blocks in a predefined order. The spatiotemporal volume is eventually represented as a single long concatenated vector.

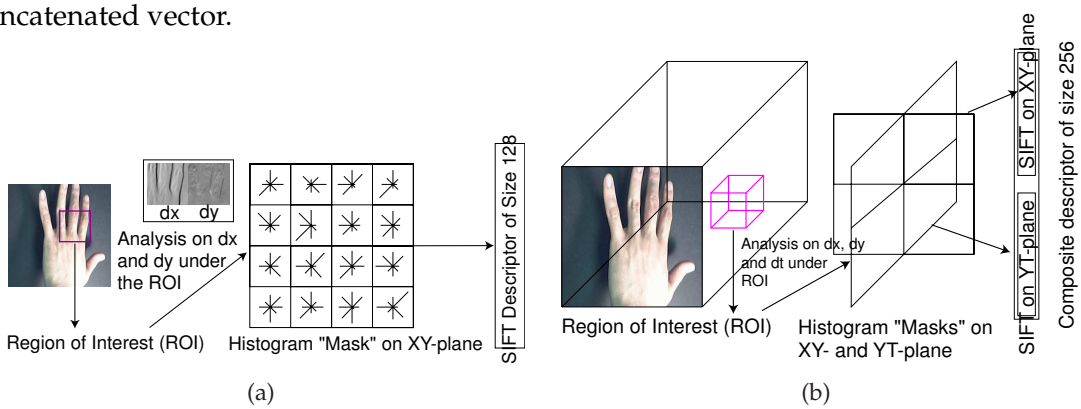


Figure 8.3: SIFT Representation. (a) SIFT used in [130]. (b) SIFT from 3D blocks (refer to text).

8.3 Empirical Evaluation

We have performed the new experiments on the Cambridge Hand Gesture Data Set which was exploited in the previous chapter (See Appendix B for full details). The data set consists of 900 image sequences of 9 gesture classes. Each class has 100 image sequences (5 illuminations $\times$ 10 arbitrary motions of 2 subjects). All training was performed on the data acquired in a single illumination setting while testing was done on the data acquired in the remaining settings. The 20 sequences in the training set were randomly partitioned into the 10 sequences for training and the other 10 for the validation.

We compared the accuracy of 9 methods:

- Applying Support Vector Machine (SVM) or Relevance Vector Machine (RVM) on Motion Gradient Orientation Images [202] (MGO SVM or MGO RVM),
- Applying RVM on the 3D SIFT vectors described in Section 8.2 (3DSIFT RVM),
- Using the canonical correlations (CC) (i.e. the method using $\mathcal{G}(\mathbf{P}_m^1, \mathbf{P}_q^1)$ in (8.1), spatiotemporal canonical correlations (ST-CC), discriminative ST-CC (ST-DCC),
- Using the canonical correlations of the SIFT descriptors (SIFT CC), spatiotemporal canonical correlations of the SIFT vectors (SIFT ST-CC), and SIFT ST-CC with the discriminative transformations (SIFT ST-DCC).

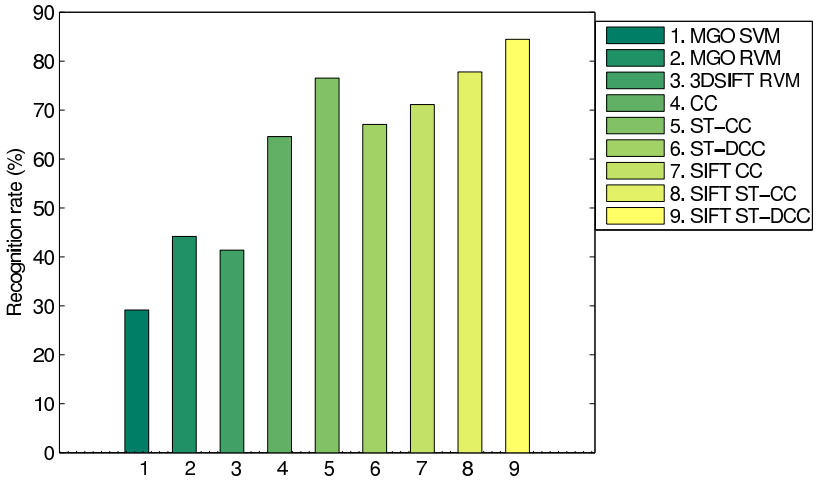


Figure 8.4: Recognition Accuracy. The identification rates (in percent) of all comparative methods are shown for the plain lighting set used for training and all the others for testing.

In the proposed method, the weights w^k were set up proportionally to the accuracy of the three subspaces for the validation set and Nearest Neighbor classification (NN) was conducted with the defined similarity functions.

Figure 8.4 shows the recognition rates of the 9 methods, when the plain lighting set (the leftmost in Figure B.3) was exploited for training and all the others for testing. The approaches of using SVM/RVM on the motion gradient orientation images are the worst. As observed in [202], using RVM improved the accuracy of SVM by about 10% for MGO images. However, we got much poorer accuracy than in those in the previous study [202]. Both methods often failed to discriminate the gestures which exhibit the same motion of the different shapes, as the methods are mainly based on motion information of gestures. A much smaller number of sequences of a single lighting condition (10 sequences per a single class) used in training is another reason for performance degradation. The accuracy of the RVM on the 3D-SIFT vectors was also poor. The high dimension of the 3D-SIFT vectors and small sample size might prevent the classifier from learning properly. We measured the accuracy of the RVM classifier for the different numbers of blocks in the 3D-SIFT representations (2-2-1,3-3-1,4-4-1,4-4-2 for X-Y-T) and obtained the best accuracy for the 2-2-1 case, which yields the lowest dimension of the 3D-SIFT vectors (but it is still huge at 256x4). Canonical correlation-based methods significantly outperformed the previous approaches. The proposed spatiotemporal canonical correlation method (ST-CC) improved the simple canonical correlation method by about 15%. The proposed discriminative method (ST-DCC) unexpectedly decreased the accuracy of ST-CC, possibly due to overfitting of discriminative methods. The train set did not reflect the lighting conditions in the test set. Note, however, that the discriminative method improved accuracy when it was applied to the SIFT representations rather than using intensity images (See SIFT ST-CC and SIFT ST-DCC in Figure 8.4). The proposed three methods using the SIFT representations are better than the respective three methods of the intensity images. The best accuracy was achieved by the SIFT ST-DCC at 85%.

Tables 8.1 and 8.2 show more results on the proposed method, where all 5 experimental results (corresponding to each illumination set used for training) are averaged. As shown in Table 8.1, canonical correlations of the XY subspace obtained better accuracy

	CC				SIFT CC			
(%)	XY	XT	YT	ST	XY	XT	YT	ST
mean	64.5	40.2	56.2	78.9	70.3	61.8	58.3	80.4
std	1.3	5.9	5.3	2.4	2.1	3.3	4.0	3.2

Table 8.1: Evaluation of the individual subspace.

	2-2-1		3-3-1		4-4-1		4-4-2	
(%)	ST-CC	ST-DCC	ST-CC	ST-DCC	ST-CC	ST-DCC	ST-CC	ST-DCC
mean	80.3	80.0	78.9	83.8	80.4	85.1	75.9	83.4
std	1.9	2.5	3.6	2.7	3.2	2.8	2.4	0.7

Table 8.2: Evaluation for different numbers of blocks in the SIFT representation. E.g. 2-2-1 indicates the SIFT representation where X,Y,and T axes are divided into 2,2,1 segments respectively.

with smaller standard deviations than the other two subspaces, but all three are relatively good compared with the traditional methods, MGO SVM/RVM and 3DSIFT RVM. Using the SIFT representation considerably improved the accuracy of the intensity images for each subspace, whereas the improvement for the joint representation was relatively small. Table 8.2 shows the accuracy of ST-CC and ST-DCC for the numbers of blocks of the SIFT representation. The best accuracy was obtained in the case of 4-4-1 for XYT (each number indicates the number of divisions along one axis). Generally, using the discriminative transformation improved the accuracy of ST-CC for SIFT representation. Note that accuracy of the method is not sensitive to settings in number of the blocks, which is practically important.

The proposed approach, based on canonical correlations, is also computationally cheap taking computations $O(3 \times d^3)$, where d is the dimension of each subspace (which was 10), and thus facilitates efficient gesture recognition in a large data set.

8.3.1 Comparison with SVM with Varying Training Data

Given the importance of Support Vector Machine (SVM) in pattern classification studies, more comparisons with SVM are shown in this section. The gesture videos are vectorised by concatenating all pixels in 3D volumes and are fed into SVM. The plain lighting set (the leftmost in Figure B.3) was exploited for training and all the others for testing. The evaluation set was exploited to set up the kernel parameters (RBF kernel) and the thresholds of SVMs. Table 8.3 shows the accuracy of Nearest Neighboring classifier in the sense of Euclidean Distance (ED) or Normalized Correlation (NC) and SVM with/without Histogram Equalization (HE) of the input vectors. Figure 8.5 shows confusion matrices of several methods. The ED/NC, as a similarity between two gesture videos, is far poorer than the proposed Video-to-Video matching method. Given the limited training data, the classical vector similarity is inappropriate for classification of the gesture videos due to its sensitivity to the intra-class variation. The SVM trained on the same input could enhance the accuracy of the NN method, but is still far worse than the proposed method. As shown in Figure 8.6, the SVM sharply dropped its accuracy with less training data while the proposed TCCA method kept high accuracy. The experiments strongly support that

(%)	ED	NC	SVM
With HE	0.2278	0.2278	0.2639
Without HE	0.2944	0.2903	0.4125

Table 8.3: Recognition accuracy (%) of NN classifiers and SVM trained on raw-pixel data. Nearest Neighboring classifier in the sense of Euclidean Distance (ED) and Normalized Correlations (NC) and SVM with/without Histogram Equalization (HE) are evaluated on the raw-pixel data.

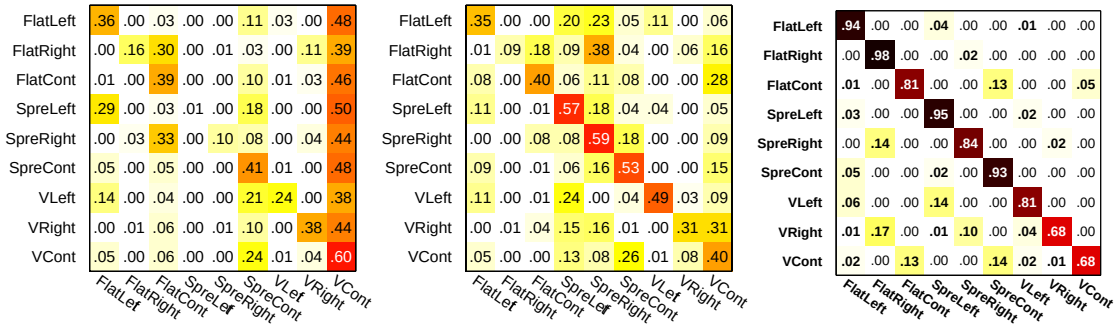


Figure 8.5: Comparison of confusion matrices. (left) NN classifier, (middle) SVM, (right) TCCA.

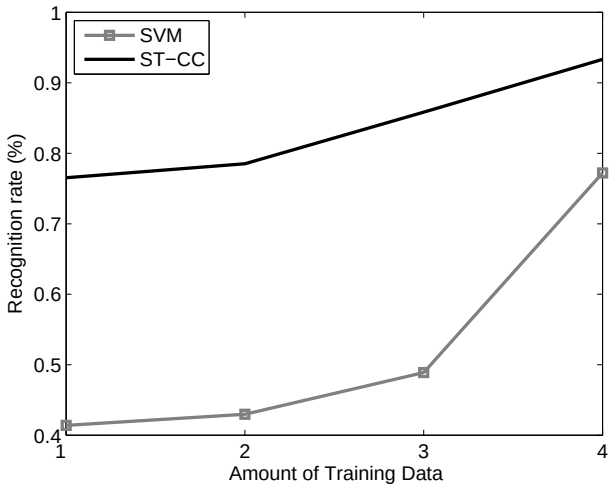


Figure 8.6: Recognition accuracy (%) of SVM and TCCA for different amount of training data. SVM sharply dropped its accuracy with less training data while the proposed TCCA method kept high accuracy.

the proposed Video-to-Video matching is flexible enough to absorb large intra-class variation of gesture videos so that robust classification performance is achieved even under small sample size.

8.4 Summary

The TCCA method reduces input dimension using the three sets of orthogonal planes and provides robust spatiotemporal volume matching by analysing mutual information (or canonical correlations) between any two gesture sequences. The discriminant analysis method has been integrated into the Tensor Canonical Correlation Analysis (TCCA) framework for robust ***Video-to-Video Matching***. The SIFT (Scale-Invariant Feature Transform) [130] based representation combined with the method also improved accuracy. Experiments for the 900 gesture sequences showed that the proposed method much enhanced the accuracy of the TCCA method, which delivered the best accuracy of many state-of-the-art methods in Chapter 7, and significantly outperformed the traditional classifiers such as Support Vector Machine and Relevance Vector Machine learnt in the same input domain. The method is also practically attractive as it does not involve significant tuning parameters and is computationally efficient for given aligned gesture videos.

CHAPTER 9

On-line Learning for Locally Orthogonal Subspace Method for Object Recognition with Image Sets

This chapter presents an integration of several methods and ideas which we have proposed through the dissertation: Orthogonal Subspace Method (OSM) as the Discriminant Analysis method of images sets in Chapter 6 makes different class subspaces orthogonal to each other. We have shown that the Canonical Correlation Analysis (CCA) after the OSM gave a good object recognition solution with image sets. Typically, in the recognition task involving image sets, efficient learning over a volume of image sets is important. In this work, we propose incremental learning for the Orthogonal Subspace Method. Owing to a close relation between the OSM and Linear Discriminant Analysis (LDA), the incremental version of OSM is established similarly to the incremental LDA in Chapter 5. A non-nonlinear OSM is further developed by a set of local linear models inspired by the Locally Linear Discriminant Analysis (LLDA) work in Chapter 4. In the experiments using 700 face image sets, the so called Locally Orthogonal Subspace Method outperformed the OSM in accuracy. The Locally Orthogonal Subspace Method is also amenable to incremental updating due to its linear base structure.

9.1 Orthogonal Subspace Method

The popularity of the methods of object recognition based on image sets has been increasing because of their greater accuracy and robustness as compared with the conventional approaches exploiting a single image as input [167, 4, 162, 207, 199, 101]. Of those methods that compare a set to a set, canonical correlation [1] of linear subspaces has attracted much attention with its benefits of robust and computationally efficient matching when dealing with changing conditions of data acquisition and large volumes of data as input for decision making [207, 199, 101, 144]. In Chapter 6, the optimal linear discrimi-

[1] It is also called canonical angle or principal angle.

nant function is proposed to find the components to maximise the canonical correlations of the within-class subspaces and minimise the canonical correlations of the between-class subspaces. However, the iterative optimization in the method is computationally costly, making incremental update rather difficult. On the other hand, the Orthogonal Subspace Method(OSM), which was proposed as an alternative discriminative method of image sets in terms of canonical correlations, does not require iterations, being simpler in learning. In many cases in the experiment of Chapter 6, this method delivered comparable accuracy to the optimal iterative solution while greatly outperforming the simple canonical correlation method. We also showed that the OSM is closely related to the existing method, Constrained Mutual Subspace Method (CMSM) [144, 46] adopted in a state-of-the-art commercial system called FacePass [186]. Compared with the CMSM, the proposed OSM provides a more solid theoretical framework with a smaller number of parameters to set empirically, as discussed in Chapter 6.

The formulation of OSM is briefly given as follows (See Chapter 6 for details): Denote the correlation matrices of the C classes by $R_i,\ i = 1, ..., C$, where $R_i = 1/M_i \sum xx^T$ and M_i is the number of data vectors in the i-th class. Let w_i denote the respective prior probabilities. Then, matrix $R_T = \sum_{i=1}^{C} w_i R_i$ is the correlation matrix of the mixture of all the classes. The total correlation matrix is decomposed s.t. $P_T^T R_T P_T = \Lambda_T$. Letting $Z = P_T \Lambda_T^{-1/2}$, we have $Z^T R_T Z = I$. Let the matrix, U_i, be constructed from eigenvectors of the i-th class having the eigenvalues equal to unity s.t.

$$w_i U_i^T Z^T R_i Z U_i = I_i, \tag{9.1}$$

then

$$\sum_{j \neq i} w_j U_i^T Z^T R_j Z U_i = O \quad \text{and} \quad w_j U_i^T Z^T R_j Z U_i = O, \text{ for all } j \neq i, \tag{9.2}$$

where O is a zero matrix since every matrix $w_j U_i^T Z^T R_j Z U_i$ is positive semi-definite. Assume that the j-th class is also represented by the eigenvectors of $w_j Z^T R_j Z$ having the eigenvalues equal to one s.t. $w_j Z^T R_j Z \simeq U_j U_j^T$. From (9.2), we have $w_j U_i^T U_j U_j^T U_i = O$, i.e. $U_i^T U_j = O$. This is the definition of the mutually orthogonal subspaces where all the vectors of each subspace are orthogonal to those of the other subspace [145]. NN classification is performed in the sense of the canonical correlations of the orthogonalized subspaces U_i, U_j.

See Table 9.1 for the important notations used throughout the chapter.

9.2 Incremental Orthogonal Subspace Method

In this section, an incremental method of learning orthogonal subspaces is presented. In practice, the eigenvectors having eigenvalues which are exactly equal to one in (9.1), do not often exist. Instead, the eigenvectors corresponding to the largest few eigenvalues can be exploited. Note that in the space projected by matrix Z in Section 9.1, the most important basis vectors for each class which are the eigenvectors corresponding to the largest eigenvalues, are at the same time the least significant basis vectors for the ensemble of the rest of the classes. Thus the classical orthogonal subspaces (9.1) can be generalised

Notations	Descriptions
C, N	number of classes, dimension of input data
M_i, M_T	number of data points of the i-th class and total data
R_i, R_T	correlation matrix of the i-th class and total data
U_i	orthogonal component matrix of i-th class
P_i, Λ_i	eigenvector and eigenvalue matrices of R_i
P_T, Λ_T	eigenvector and eigenvalue matrices of R_T
d_i, d_T	number of sufficient components of the i-th class and total data
U_j^i	locally orthogonal component matrix of j-th class to i-th class

Table 9.1: Notations.

into the subspaces spanned by the components U_i s.t.

$$w_i U_i^T Z^T R_i Z U_i = \Delta_i, \quad \sum_{j \neq i} w_j U_i^T Z^T R_j Z U_i = I - \Delta_i, \tag{9.3}$$

where Δ_i is the diagonal matrix corresponding to the largest few eigenvalues. That is, the generalised orthogonal subspace method seeks the class-specific components which maximise the ratio of the variances of the i-th class correlation matrix over the total correlation matrix.

Similarly to that of the incremental LDA method we proposed in Chapter 5, an incremental OSM solution is proposed by the three steps: update of the principal components of each class correlation matrix, update of those of the total correlation matrix and the computation of the orthogonal components only using both updated principal component sets. The concept of the sufficient spanning set [59] is conveniently exploited in each step to reduce the dimension of the eigenvalue problems. The proposed method provides the same solution as the batch-mode OSM with far lower computational cost. When new data are added to the existing data set, all existing orthogonal subspace models $U_i,\ i = 1, ..., C$ (C is the number of classes) are incrementally updated to get new orthogonal subspaces described by U_i' as follows. Here we assume the equal prior probabilities for all classes for simplicity.

1) Update of principal components of class correlation matrix. Let the number of samples, eigenvector and eigenvalue matrices corresponding to the first few eigenvalues of the i-th class correlation matrix R_i in the existing data be (M_i, P_i, Λ_i) respectively. The set $(M_i^n, P_i^n, \Lambda_i^n)$ similarly denotes those of the new data. The update is defined as the functional form by

$$\mathcal{F}_1(M_i, P_i, \Lambda_i, M_i^n, P_i^n, \Lambda_i^n) = (M_i', P_i', \Lambda_i'). \tag{9.4}$$

Note this is applied only to the classes having the new data. As the updated class correlation matrix is $R_i' \simeq \frac{M_i}{M_i'} P_i \Lambda_i P_i^T + \frac{M_i^n}{M_i'} P_i^n \Lambda_i^n P_i^{nT}$ where $M_i' = M_i + M_i^n$, the sufficient spanning set of R_i' can be given as $\Upsilon_i = \mathcal{H}([P_i,\ P_i^n])$, where $\mathcal{H}$ is an orthonormalisation function of column vectors (e.g. QR decomposition). The function $\mathcal{H}$ also eliminates any zero vectors after the orthonormalisation to further reduce the number of the sufficient components. The updated principal components can then be written as $P_i' = \Upsilon_i Q_i$, where Q_i is a rotation matrix. By this representation, the eigenproblem of the updated class cor-

relation matrix is changed into a new low dimensional eigenproblem

$$R_i' \simeq P_i' \Lambda_i' P_i'^T = \Upsilon_i Q_i \Lambda_i' Q_i^T \Upsilon_i^T \quad \rightarrow \quad \Upsilon_i^T (\frac{M_i}{M_i'} P_i \Lambda_i P_i^T + \frac{M_i^n}{M_i'} P_i^n \Lambda_i^n P_i^{nT}) \Upsilon_i \simeq Q_i \Lambda_i' Q_i^T. \tag{9.5}$$

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 given. The latter term is for computing $(M_i^n, P_i^n, \Lambda_i^n)$ from the new data.

2) Update of principal components of total correlation matrix. The subsequent update is described as

$$\mathcal{F}_2(M_T, P_T, \Lambda_T,\ M_i^n, P_i^n, \Lambda_i^n) = (M_T', P_T', \Lambda_T') \qquad i = 1, ..., C^n, \tag{9.6}$$

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

$$R_T' \simeq \frac{M_T}{M_T'} P_T \Lambda_T P_T^T + \frac{M_T^n}{M_T'} \sum_{i=1}^{C^n} P_i^n \Lambda_i^n P_i^{nT} \tag{9.7}$$

where $M_T' = M_T + M_T^n$, $M_T^n = \sum M_i^n$. The sufficient spanning set of R_T' can be given as

$$\Upsilon_T = \mathcal{H}([P_T, P_1^n, ..., P_{C^n}^n]) \tag{9.8}$$

and $P_T' = \Upsilon_T Q_T$,where Q_T is a rotation matrix. Accordingly, the new small dimensional eigenproblem is obtained by

$$R_T' \simeq P_T' \Lambda_T' P_T'^T \quad \rightarrow \quad \Upsilon_T^T (\frac{M_T}{M_T'} P_T \Lambda_T P_T^T + \frac{M_T^n}{M_T'} \sum_{i=1}^{C^n} P_i^n \Lambda_i^n P_i^{nT}) \Upsilon_T \simeq Q_T \Lambda_T' Q_T^T \tag{9.9}$$

The computation requires $O(d_T^3)$, where d_T^3 is the sufficient number of components of Υ_T. Note that all P_i^n have already been produced at the previous step.

3) Update of orthogonal components of all classes. The final step only exploits the updated principal components of the previous steps, which is defined as

$$\mathcal{F}_3(P_i', \Lambda_i',\ P_T', \Lambda_T') = U_i', \quad i = 1, ..., C. \tag{9.10}$$

where U_i' denotes the updated orthogonal components of the i-th class data. Let $Z = P_T' {\Lambda_T'}^{-1/2}$, then, $Z^T R_T' Z = I$. The problem remaining is to find the components which maximise the variance of the projected data $Z^T R_i' Z$. The sufficient spanning set of the

projection data can be given by $\Phi_i = \mathcal{H}({P'_T}^T P'_i)$. As a result the eigenproblem to solve is

$$Z^T R'_i Z = \Phi_i Q_i \Delta_i Q_i^T \Phi_i^T \quad \rightarrow \quad \Phi_i^T Z^T P'_i \Lambda'_i {P'_i}^T Z \Phi_i = Q_i \Delta_i Q_i^T, \tag{9.11}$$

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

Batch OSM vs. incremental OSM for time and space complexity. 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 9.1 for the batch-mode computation). The batch computation also requires all data vectors or $N \times N$ correlation matrices to be kept track of. 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 space costs, which corresponds to the data $(P_i, \Lambda_i, P_T, \Lambda_T),\ i = 1, ..., C$.

9.3 Locally Orthogonal Subspace Method (LOSM)

In the generalised orthogonal subspaces (9.3), the prior probabilities of classes w_j can better be set up to improve the discriminatory powers of the classes with their rival classes. Rather than equal priors for all classes, higher priors are given to the neighboring classes of the i-th class by

$$w_j \rightarrow w_j^i \quad \begin{cases} \propto \mathcal{S}(U_i, U_j) & \text{for} \quad j = 1, ..., C,\ j \neq i, \\ = 0 & \text{for} \quad j = i \end{cases}$$

where $\mathcal{S}$ is the canonical correlation function defined in Section 6.1. The i-th class locally orthogonal subspace U_i^i is then similarly computed as U_i in Section 9.1 by replacing the total correlation matrix R_T with the class-specific total correlation matrix by $R_T^i = \sum_{j=1}^C w_j^i R_j$ and diagonalizing $Z^T R_i Z$. The weights w_j^i can also be binary-valued in the same concept s.t. $w_j^i = 1$, if $\mathcal{S}(U_j, U_i) > thres$, $w_j^i = 0$ otherwise. In this way, the local orthogonality of classes is more emphasised.

Normalization. When a new test set is given, the locally orthogonal components of the new test set are class-wise extracted with R_T^i for $i = 1, ..., C$. If we let U_q^i as the locally orthogonal components of the new test set for the i-th model class, NN recognition is performed with the normalized scores

$$(\mathcal{S}(U_i^i, U_q^i) - m_i)/\sigma_i, \tag{9.12}$$

where m_i, σ_i are the mean and standard deviation of matching scores of a validation set with the i-th model class. As each class model exploits a different total correlation matrix,

the score normalization process is important.

Time-efficient matching. Batch computation of the C locally orthogonal subspaces of a given new test set is time-consuming, i.e. taking $O(C \times min(N, M_q)^3)$, where M_q is the number of vectors in the new test set. This computational cost can be significantly reduced using the update function $\mathcal{F}_3(P_q, \Lambda_q, P_T^i, \Lambda_T^i)$ in Section 9.2, where P_q, Λ_q are the eigenvector and eigenvalue matrices of the correlation matrix of the new test set and P_T^i, Λ_T^i for the class specific total correlation matrix respectively. Note that this only requires $O(C \times d_q^3)$, d_q is the number of columns of P_q. The subsequent canonical correlation matching with C orthogonal subspace models is not computationally expensive as it costs $O(C \times d^3)$ (Refer to Section 6.1), where d is the dimension of the orthogonal subspaces.

Incremental update of LOSM. The computational cost of the incremental locally OSM is increased by that of the update of the components of the C class-specific total correlation matrices, but it is still much cheaper than the batch OSM. Firstly, the principal components of class correlation matrices are updated by $\mathcal{F}_1$ in the previous section. The update of the principal components of the weighted total correlation matrices defined s.t.

$$\mathcal{F}_2'(M_T, P_T^i, \Lambda_T^i, w_j^i, M_j^n, P_j^n, \Lambda_j^n) = (M_T', P_T^{i\prime}, \Lambda_T^{i\prime}) \qquad i = 1, ..., C, \;\; j = 1, ..., C^n. \tag{9.13}$$

is achieved as follows. The updated weighted total correlation matrix is given as

$$R_T^{i\prime} = \frac{M_T}{M_T'} P_T^i \Lambda_T^i P_T^{i\,T} + \frac{M_T^n}{M_T'} \sum_{j=1}^{C^n} w_j^i P_j^n \Lambda_j^n P_j^{nT}. \tag{9.14}$$

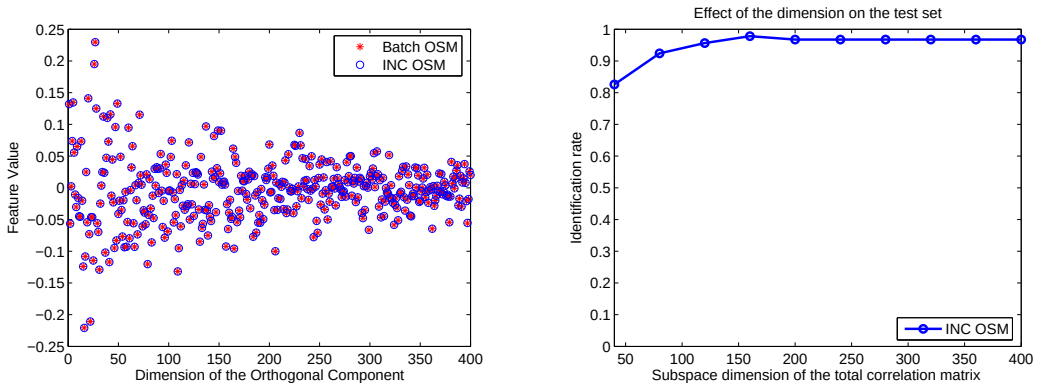
Regardless of the extra weight terms, the sufficient spanning set of $R_T^{i\prime}$ is similarly given by $\Upsilon_T^i = \mathcal{H}([P_T^i, P_1^n, ..., P_{C^n}^n])$. Thus the new eigen-problems and the updated components are similarly given as the second step in Section 9.2. If we assume that the NN recognition has already been performed for the given new test sets by the scores of $\mathcal{S}(U_i^i, U_j^i), \;\; i = 1, ..., C, \;\; j = 1, ...C_n$, the weights w_j^i can be set up proportionally to these scores. The final locally orthogonal components are also similarly updated by $\mathcal{F}_3$, replacing P_T', Λ_T' with $P_T^{i\prime}, \Lambda_T^{i\prime}$.

9.4 Evaluation

We used the Cambridge face video database consisting of 100 subjects. For each person, 7 video sequences of the individual in arbitrary motion were collected. Following automatic localization using a cascaded face detector [193] 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 (See Appendix A).

9.4.1 Accuracy and time complexity of the incremental OSM

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



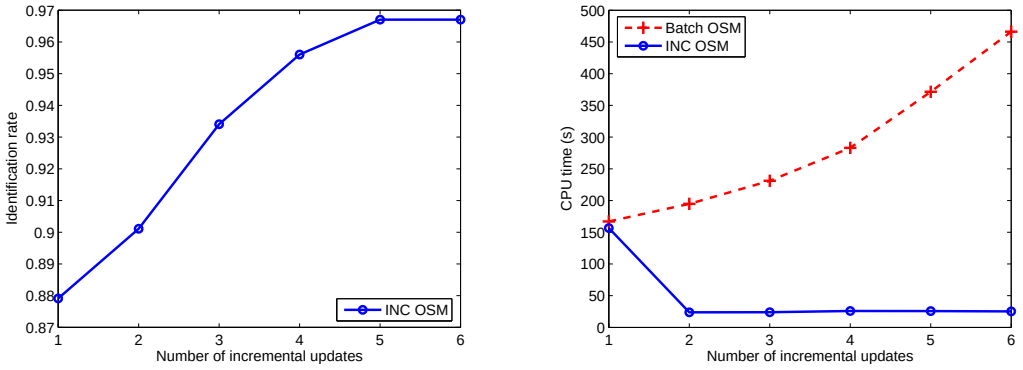
(a) (b)

Figure 9.1: Batch vs. Incremental OSM-1. (a) Example orthogonal components, which are computed by the incremental and the batch-mode, are very alike. (b) Insensitiveness 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, just provided the enough dimensionality of the subspaces.

were initially used for learning the orthogonal subspaces. The sets of the 100 face classes of other illumination settings were then additionally given for the update. We set the total number of updates including the initial batch computation at 6 and the number of images to add at each iteration around 10,000. The dimensionality of the uniformly scaled images was 2,500 and the number of orthogonal components was around 10, which varies for more than 99% of the energy from the eigenvalue plot. See Figure 9.1 (a) for the example orthogonal component computed by the proposed incremental algorithm and the batch-mode. Figure 9.1 (b) shows the insensitivity of the incremental OSM to the dimensionality 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 9.2 (a) shows the accuracy improvement of the incremental OSM according to the number of updates. It efficiently updates the existing orthogonal subspace models over new evidences contained in the additional data sets, giving gradual accuracy improvements. The computational costs of the batch OSM and the incremental OSM are compared in Figure 9.2 (b). Whereas the computational cost of the batch-mode is largely increased as the data is repeatedly added, the incremental OSM keeps the cost of the update low.

9.4.2 Accuracy of Locally OSM

Another experiment was designed to compare accuracy of several methods with the locally orthogonal subspaces. 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 with both training and test sets was exploited for the validation. We compared the performance of Mutual Subspace Method (MSM) [207] as a gauging method, where the dimensionality of each subspace is 10 representing more than 99% energy of the data, CMSM [144] used in a state-of-the-art commercial system FacePass [186], where the dimension of the constrained subspace was determined to be 360, which yielded the



(a) (b)

Figure 9.2: 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.

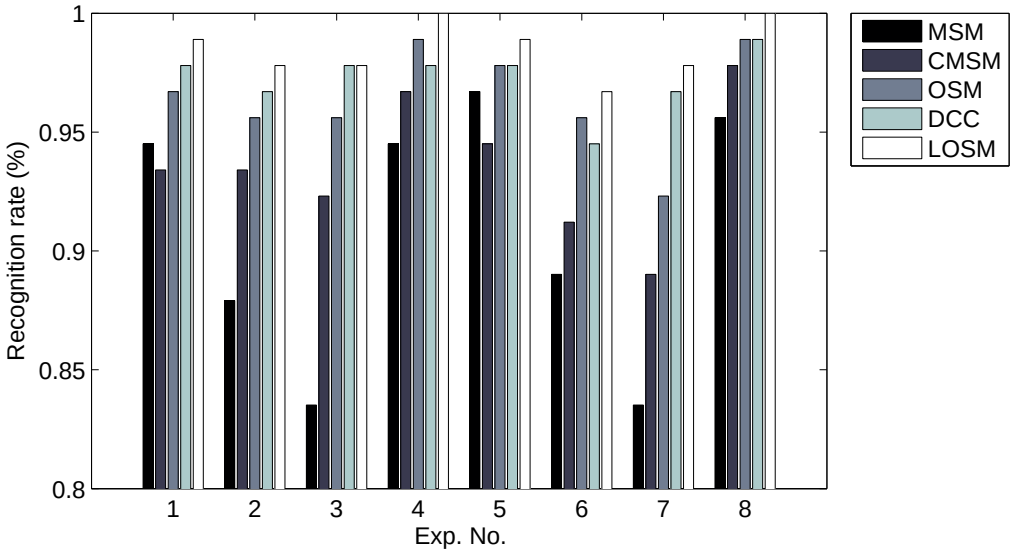


Figure 9.3: Accuracy comparison.

best accuracy for the validation set, canonical correlations of Orthogonal Subspace Method (OSM), and canonical correlations of the Locally Orthogonal Subspace Method (LOSM), where the class prior probabilities were set to be binary-valued by a certain threshold. The threshold typically returned 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 9.3 compares the recognition accuracy of all methods, where the experiment numbers correspond to the combinations of the training/test lighting sets, which were chosen as the most difficult scenarios for MSM. In Figure 9.3, the OSM was superior to CMSM and the proposed locally orthogonal subspace method (LOSM) outperformed all the other methods. Theoretically, the proposed incremental solution of LOSM provides the same solution of the batch computation of LOSM with slightly greater computational cost than that of the incremental OSM.

9.5 Summary

We have shown that the methods developed in Chapter 4 and Chapter 5 for the non-linear and incremental discriminant analysis respectively, could help the recognition task with image sets or videos. The non-linear and incremental version of the discriminant analysis of image-sets have been similarly developed. In the object-recognition task involving image sets, developing an efficient learning method for handling increasing volumes of image sets is important. Image data emanating from environments dramatically changing from time to time should be continuously accumulated. The proposed incremental solution of the Orthogonal Subspace Method and the Locally Orthogonal Subspace Method (as a non-linear model of the OSM) facilitate 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 Locally OSM delivered the best accuracy among competing methods.

Although we have conducted experiments only for the recognition task with image sets, the proposed methods could be readily extended to the recognition task with videos for action and gesture classification.

CHAPTER 10
Conclusion

10.1 Concluding Remarks

This study has proposed methods of Discriminant Analysis (DA) for three visual classification tasks; face recognition with a single-per-class image, object recognition by image sets (or ensembles) and action categorisation in videos. Contributions of this work include not only obtaining algorithms which outperform those of state-of-the-art methods in each of the three tasks, but also developing methods for the tasks in a single Discriminant Analysis framework.

We have followed the MPEG-7 protocol [1, 134] for face image retrieval, i.e. the recognition task with a single-per-class image. Two updates on our existing method which won the MPEG-7 standard competition [99, 98, 102], have been proposed for non-linear classification and on-line learning problems. A novel method of non-linear DA has been proposed for tackling a challenging problem, novel-view face recognition with a single model image, by *aligned local discriminative models*. It outperforms conventional Linear Discriminant Analysis (LDA), the LDA mixture model, Kernel Discriminant Analysis (KDA) and a commercial face recognition system. The method is also computationally efficient as compared with the KDA. The proposed method for the on-line learning showed close agreement with the batch LDA with far lower complexity in time and space for experiments using thousands of face classes. Our on-line method guarantees an accurate LDA solution provided sufficient components spanning most energy of data space.

We have demonstrated that Canonical Correlation Analysis (CCA) yields an image-set based object recognition solution which has good generalisation over novel data. CCA yielded much higher recognition rates than traditional probability density based set-similarity measures, which are highly sensitive to simple transformations of input data [101, 104]. It is well known that images are well confined to lie on low-dimensional subspaces. The CCA, a subspace-based set-similarity, effectively places a uniform prior over the subspaces and provides invariant matching up to the pattern variations on the subspaces. We have proposed a novel method of discriminant analysis of image sets for optimal classification by CCA. It has been evaluated for various object recognition problems, using face image sets with arbitrary motion captured under different lighting conditions, image sets of five hundred general objects taken at different views and object category recognition using the ETH-80 database [117]. We have extended the CCA into high-order tensor data for

analysing human actions/gestures in videos. Our Tensor CCA method absorbs large intra-class variation of actions and facilitates robust action recognition under small sample size. The TCCA method notably outperformed various state-of-the-art action recognition methods on the largest public action data base (KTH) as well as a self-recorded hand-gesture data set (See Appendix).

10.2 Observations

Here, we suggest some important observations and insights obtained from the work.

Unsupervised vs. supervised learning for recognition task with a single model image: The recognition task with a single-per-class image is categorised into an unsupervised learning problem (in the sense that class labels are not given), as there is no intra-class information available from given classes, each of which has just a single image. Naturally, many previous studies have tackled the problem with Principal Component Analysis (PCA), which is a representative unsupervised learning method [204, 224, 215, 195, 81, 82, 87]. This work has, on the other hand, proposed a supervised learning method (in the sense that class labels are exploited) using an independent training set which contains prototype classes of multiple-per-class samples. We have shown that the Discriminant Analysis learnt from the prototype classes works very well for novel classes having single-per-class images, significantly surpassing the PCA-based methods. Achieving good generalisation of discriminative information across different classes would be an interesting topic for object recognition with limited training samples.

Robust gesture recognition under small sample size: The Support Vector Machine (SVM) [188] has been a state-of-the-art classification method of various applications. In this dissertation, we have compared our method with various methods based on SVM for action and gesture recognition tasks. In particular, we have obtained a notable improvement over SVM in the gesture recognition experiment where the gesture videos were vectorized by concatenating all pixels in 3D volumes and were fed into SVM (See Section 8.3.1). The accuracy of SVM was far poorer than that of the proposed Tensor CCA method. The SVM dropped sharply in accuracy with fewer training data, while the proposed TCCA method retained high accuracy. The comparison supports our argument that the proposed Video-to-Video matching absorbs large intra-class variation of gesture so that the classification method yields good generalisation for novel data for robust classification under small sample size.

Holistic vs. part-based methods for action classification: There has been a popular line of action-recognition methods which is based on space-time interest points and visual code words [143, 165, 32, 113]. Although these part-based approaches have yielded good accuracy mainly due to the high discrimination power of individual parts, they ignore global structural information. Recently, a few methods have attempted to combine the structural information with the local information [203, 163]. However, their performance depends highly on proper setting of the parameters of the space-time interest points and the code book. A holistic method has, on the other hand, been proposed in this thesis. The Tensor CCA method directly analyses global space-time volumes without significant

tuning parameters. Whereas conventional holistic representations are sensitive to simple transformations of input data, the method is invariant up to affine transformation of the input data. We have obtained a large accuracy gain over various part-based methods as well as other holistic methods on the largest public action data set (KTH).

10.3 Limitations

A major limitation of this work is in the requirement of pattern registration in images. Objects are isolated in images and normalized in position and scale for input of the proposed methods. Thanks to previous efforts on face detection [193], registration is quite simple for face images, but not for general object categories. This work has assumed that general objects are conveniently segmented from simple backgrounds, which is often too strong in real-world applications.

The proposed action classification method also requires spatiotemporal registration of actions/gestures. Despite the automatic action detection method for the registration, it is computationally demanding in the scenarios that require multi-scale search and a large search space. The method is limited to cases of stationary cameras and recognition of unit actions to meet the difficulty of the spatiotemporal registration.

Our methods have been mostly applied to holistic raw-pixel representation of images and videos, which can be sensitive to e.g. cluttered backgrounds, occlusion and geometric variation of patterns. The methods proposed as a general meta-algorithm may be combined with better representation for further improvement, depending on application. As an example, we have combined the Scale Invariant Feature Transform (SIFT)-based representation with the methods for gesture recognition.

See the summary of each chapter for further limitations and concerns.

10.4 Future Work

Interesting work directions are summarized in the following.

- **Automatic registration of objects and actions.** As mentioned above, the proposed methods depend on segmented objects or actions. An efficient registration method is required. To speed up detection, one may try a hierarchical approach which applies one or several simpler but less accurate methods to filter out the majority of candidates and then apply our method with the benefit of high accuracy. For action detection in videos, efficient multi-scale search by merging the space-time subspaces of TCCA should be investigated in the future.

- **An integrated system for a long-term learning.** The human visual system handles various types of visual data and learns over a long period. A machine that works well for whatever type of inputs given, must be valuable. Inputs may vary from a single model image to image sets and videos. Long-term learning with the different types of visual data would be interesting. For this purpose, further speeding up and a scalability check of the algorithms are required.

- **Semi-supervised and active learning** are approaches which can help to minimise human intervention in incremental learning for model reinforcement [237]. Several

important issues arise here: time-efficient Semi-Supervised Learning (SSL), SSL for multiple classes, use of temporal information of video input and temporal/spatial weighting for robust SSL.

- **Co-training with multi-channel data.** Videos are comprised of color and sound channels as well as an intensity channel. Rather than exploiting only intensity information of videos, co-training with other sources, colors and sounds may help to boost accuracy. Note the proposed tensor framework is advantageous in this aim as it can simply add more modes (or dimensions) to the current third-order tensors.

APPENDIX A

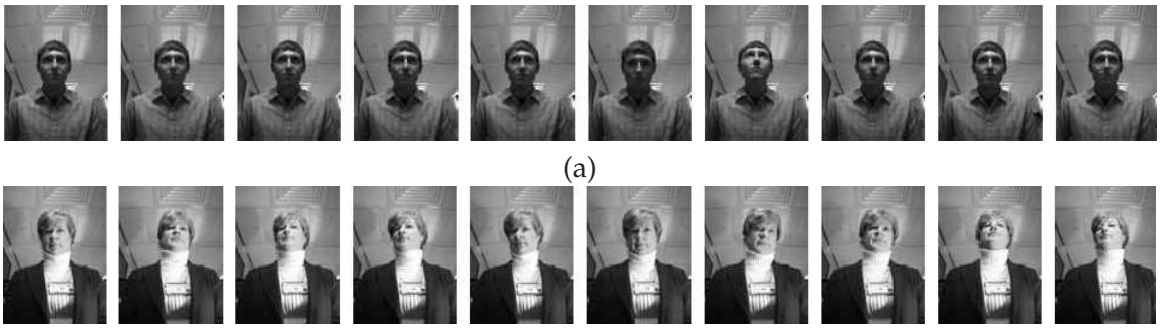
Cambridge-Toshiba Face Video Data Set

For the recognition task with image sets, we have collected a database called the *Cambridge-Toshiba Face Video Database* with 100 individuals of varying age and ethnicity, and equally represented gender (See Table A.1 and Figure A.1). For each person, 14 (7 illuminations $\times$ two recordings) video sequences of the person in arbitrary motion were collected. Each sequence was recorded in a different illumination setting for 10s at 10fps and at 320$\times$240 pixel resolution. See Figure A.2 for some samples of original image sequences. The motion of the user was not controlled, leading to different motion patterns and poses. As shown in Figure A.3, two time sets of a subject in the same lighting conditions exhibit significant variations in pose and expression. See Figure A.4 for seven lighting prototypes. Following automatic localization using a cascaded face detector [193] and cropping to a uniform scale of 20$\times$20 pixels, images of faces were histogram equalized. Note that the face localization was performed automatically on the images of uncontrolled quality. Thus it was not as accurate as any conventional face registration with either manual or automatic eye positions performed on high-quality face images. Our experimental conditions are closer to those given for typical surveillance systems. Figure A.5 (a) shows the preprocessing. Note the typical outliers contained in image sets, which are caused mostly by errors in automatic localization.

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

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

Figure A.1: Examples of Cambridge Face Video Database. The data set contains 100 face classes with varying age, ethnicity and gender. Each class has about 1400 images from the 14 image sequences captured under 7 lighting settings.



(a)

(b)

Figure A.2: Raw data. Frames from two typical video sequences from the database used for evaluation. The motion of the user was not controlled, leading to different motion patterns and poses.

Figure A.3: Example of the two time settings (top and bottom) of a subject acquired in a single lighting setting. They contain significant variations in pose and expression.

Figure A.4: Illumination. 7 illumination settings in the database. Note that in spite of the same spatial arrangement of light sources for a particular illumination setting, its effect on the appearance of faces changes significantly due to variations in subject's height and their *ad lib* position relative to the camera.

(a)

(b)

Figure A.5: Data preprocessing. (a) Left to right – typical input frame from a video sequence of a subject performing unconstrained head motion (320×240 $pixels$), output of the face detector (72×72 pixels) and the final image after resizing to uniform scale and histogram equalization. (b) Typical outliers – face detector false positives – present in our data.

APPENDIX B
Cambridge Hand Gesture Data set

For evaluating gesture classification algorithms, we acquired a set of *Cambridge-Gesture data base* [1], which consists of 900 image sequences of 9 gesture classes which are defined by 3 primitive hand shapes and 3 primitive motions (see Figure B.1). The target task for this data set is accordingly to classify different shapes as well as different motions. Each class contains 100 image sequences (5 different illuminations$\times$10 arbitrary motions of 2 subjects). Each sequence was recorded in front of a fixed camera having roughly isolated gestures in space and time. Thus fairly large intra-class variation in spatial and temporal alignment is reflected to the data set. See Figure B.2 for typical sample sequences of the 9 classes and Figure B.3 for 5 illumination prototypes.

Protocol. All training was performed on the data acquired in the single plain illumination setting (leftmost in Figure B.3) while testing was done on the data acquired in the remaining settings. The 20 sequences in the training set were randomly partitioned into the 10 sequences for training and the other 10 for the validation.

[1]The database is publicly available at ftp://mi.eng.cam.ac.uk/pub/CamGesData. Contact e-mails: tkk22@cam.ac.uk

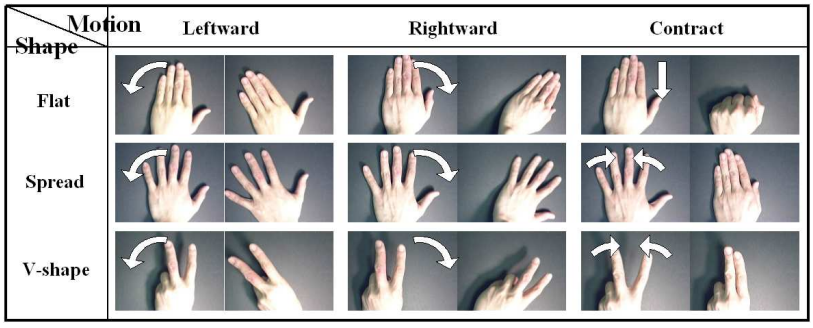


Figure B.1: Hand-Gesture Database. 9 gesture classes are generated by 3 primitive shapes and motions.

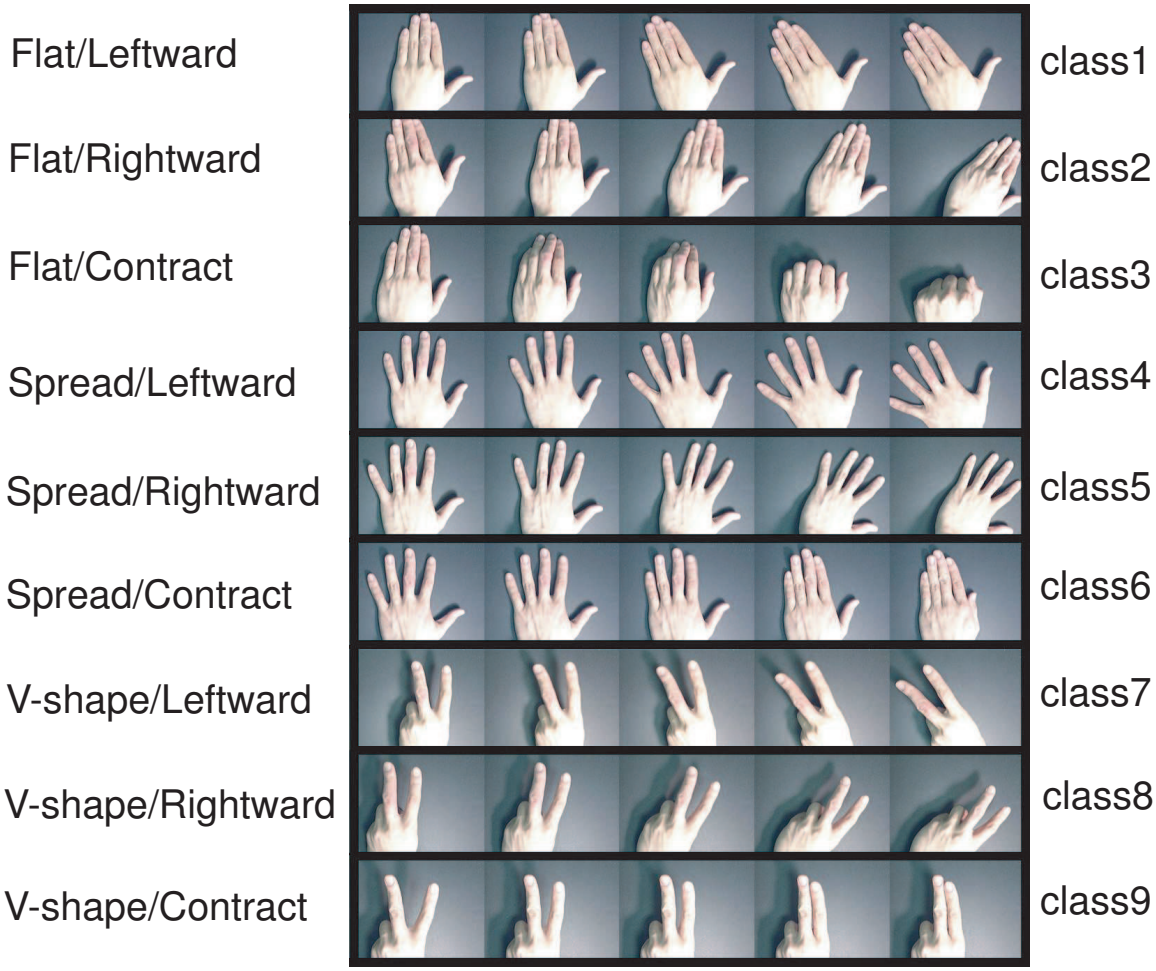


Figure B.2: Sample sequences of the 9 gesture classes.

Figure B.3: 5 lighting conditions in the database.

APPENDIX C

Equivalence of SVD solution to Mutual Subspace Method

There are many solutions for Canonical Correlation Analysis. Here, we briefly show that the SVD solution [11] is equivalent to the Mutual Subspace Method [207]. In Mutual Subspace Method (MSM), canonical correlations are defined as the eigenvalues of the matrix $\mathbf{P}_1\mathbf{P}_1^T\mathbf{P}_2\mathbf{P}_2^T\mathbf{P}_1\mathbf{P}_1^T \in \mathbb{R}^{N\times N}$, where $\mathbf{P}_i \in \mathbb{R}^{N\times d}$ is a basis matrix of a data set i. The SVD solution in (6.2) for computing canonical correlations is symmetric. That is,

$$\mathbf{Q}_{12}^T\mathbf{P}_1^T\mathbf{P}_2\mathbf{Q}_{21} = \mathbf{\Lambda}$$

$$\mathbf{Q}_{21}^T\mathbf{P}_2^T\mathbf{P}_1\mathbf{Q}_{12} = \mathbf{\Lambda}$$

By multiplying the above two equations, we obtain

$$(\mathbf{Q}_{12}^T\mathbf{P}_1^T\mathbf{P}_2\mathbf{Q}_{21})(\mathbf{Q}_{21}^T\mathbf{P}_2^T\mathbf{P}_1\mathbf{Q}_{12}) = \mathbf{\Lambda}^2$$

$$\rightarrow \mathbf{Q}_{12}^T\mathbf{P}_1^T\mathbf{P}_2\mathbf{P}_2^T\mathbf{P}_1\mathbf{Q}_{12} = \mathbf{\Lambda}^2$$

$$\rightarrow \mathbf{P}_1\mathbf{P}_1^T\mathbf{P}_2\mathbf{P}_2^T\mathbf{P}_1\mathbf{P}_1^T = \mathbf{P}_1\mathbf{Q}_{12}\mathbf{\Lambda}^2\mathbf{Q}_{12}^T\mathbf{P}_1^T$$

as $\mathbf{Q}_{12}\mathbf{Q}_{12}^T = \mathbf{Q}_{21}\mathbf{Q}_{21}^T = \mathbf{I}$. $\mathbf{P}_1\mathbf{Q}_{12}$ and $\mathbf{\Lambda}^2$ are the eigenvector matrix and eigenvalue matrix respectively of the matrix $\mathbf{P}_1\mathbf{P}_1^T\mathbf{P}_2\mathbf{P}_2^T\mathbf{P}_1\mathbf{P}_1^T$. That is, the canonical correlations of MSM simply assume the square value of the canonical correlations of the SVD solution. Please note that the dimension of the matrix $\mathbf{P}_1^T\mathbf{P}_2 \in \mathbb{R}^{d\times d}$ is relatively low compared with that of $\mathbf{P}_1\mathbf{P}_1^T\mathbf{P}_2\mathbf{P}_2^T\mathbf{P}_1\mathbf{P}_1^T \in \mathbb{R}^{N\times N}$.

Bibliography

[1] M. Abdel-Mottaleb, J.H. Connell, R.M. Bolle, R. Chellappa, *Face descriptor syntex*, Merging proposals P181, P551, and P650, ISO/MPEG M5207, Melbourne, 1999.

[2] B. Adhikara and D. Joshi, Distance discrimination et resume exhaustif, *Publs. Inst. Statis.*, 5:57-74, 1956.

[3] O. Arandjelovic and R. Cipolla, Face recognition from face motion manifolds using robust kernel resistor-average distance, *IEEE Workshop on Face Processing in Video*, pages 88-97, Washington D.C., USA, June 2004.

[4] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *Proc. Computer Vision and Pattern Recognition*, pp. 581–588, San Diego, USA, June 2005.

[5] 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.

[6] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahalanobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005.

[7] C. Bauckhage, T. Kaster and J.K. Tsotsos. Applying Ensembles of Multilinear Classifiers in the Frequency Domain. In *Proc. Computer Vision and Pattern Recognition*, pages 95–102, 2006.

[8] G. Baudat and F. Anouar, Generalised Discriminant Analysis Using a Kernel Approach, *Neural Computation*, 12:2385-2404, 2000.

[9] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, *IEEE Trans. on PAMI*, 19(7):711–720, July 1997.

[10] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

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

[12] M.J. Black. Explaining optical flow events with parameterized spatio-temporal models. In *Proc. of CVPR*, pages 1326–1332, 1999.

[13] D.M. Blackburn, M. Bone, and P.J. Phillips, *Facial Recognition Vendor Test 2000: Evaluation Report*, 2000. http://www.frvt.org/default.htm

[14] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *Proc. CVPR*, pages 1395–1402, 2005.

[15] V.Blanz, S.Romdhani, and T.Vetter, Face identification across different poses and illuminations with a 3D morphable model, In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 192-197, May 2002.

[16] A. Bobick and J. Davis. The recognition of human movements using temporal templates. *IEEE Trans. on PAMI*, 23(3):257–267, 2001.

[17] M. Borga, *Canonical Correlation a Tutorial*, On-line tutorial, Dept. of Biomedical Engineering, Linkoping University, Sweden. http://people.imt.liu.se/∼magnus/cca/

[18] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *Proc. ECCV*, pages 390–401, 2004.

[19] G.E.P. Box and G.C.Tiao, *Bayesian Inference in Statistical Analysis*, John Wiley and Sons, Inc., New York, 1992.

[20] M. Bressan, J. Vitria Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24(15):2743–2749, 2003.

[21] B. Boser, I. Guyon, and V. Vapnik, A training algorithm for optimal margin classifiers, In *Proc. Fifth Annual Workshop on Computational Learning Theory*, Pages 144 - 152, Pittsburgh, USA, 1992.

[22] C. J. C. Burges, A tutorial on support vector machines for pattern recognition, *Data mining and knowledge discovery*, 2:121-167, 1998.

[23] H-T. Chen, H-W. Chang and T-L. Liu, Local Discriminant Embedding and Its Variants, In *Proc. of CVPR*, volume 2, pages 846-853, 2005.

[24] L. Chen, H. Liao, M. Ko, J. Lin and G. Yu, New LDA-based face recognition system which can solve the small sample size problem, *Pattern Recognition*, 33(10):1713-1726, 2000.

[25] H. Chernoff, A measure of asymptotic efficiency of tests for a hypothesis based on a sum of observations, *Annals of Mathametical Statistics*, 23:493-507, 1952.

[26] T.-J. Chin and D. Suter. Incremental Kernel PCA for Efficient Non-linear Feature Extraction. In *Proc. BMVC*, volume 3, pages 939–948, 2006.

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

[28] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based methods*, Cambridge University Press, 2000.

[29] J. Dai, S. Yan, X. Tang, James. Kwok. Locally Adaptive Classification Piloted by Uncertainty. In *Proc. of International Conference on Machine Learning*, pages 225–232, 2006.

[30] T. Darrell and A. Pentland. Space-time gestures. In *Proc. of CVPR*, pages 335–340, 1993.

[31] G.K. Demir and K. Ozmehmet, Online local learning algorithms for linear discriminant analysis, *Pattern Recognition Letters*, 26(4):421-431, 2005.

[32] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65-72, 2005.

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

[34] C. Eckes, S. Eickeler, M. Larson, J. Loffler, K. Biatov, J. Kohler, Proposal of a face recognition descriptor based on Pseudo2D-HMMs, ISO/IEC JTC1/SC21/WG11 M8394, Fairfax, VA, May 2002.

[35] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. of ICCV*, pages 726–733, 2003.

[36] J. Eichhorn and O. Chapelle, *Object categorisation with SVM: Kernels for Local Features*, Technical Report No. 137, Max Planck Institute for Biological Cybernetics, July 2004

[37] Face Recognition Grand Challenge. http://bbs.bee-biometrics.org.

[38] A. Fitzgibbon and A. Zisserman, Joint manifold distance: a new approach to appearance based clustering, In *Proc. Computer Vision and Pattern Recognition*, pages 26–33, Madison, USA, 2003.

[39] V. Franc and J. Matas, An extension of the component-based LDA descriptor by the Generalised Discriminant Analysis, ISO/IEC/JTC1/SC21/WG11 M8727, Klagenfurt, AT, July 2002.

[40] R. Fransens, J. Prins, and L.V. Gool, SVM-based Nonparametric Discriminant Analysis, An Application to Face Detection, In *Proc. of International Conference on Computer Vision*, volume 2, pages 1289–1296, 2003.

[41] W. Freeman and M. Roth. Orientation histogram for hand gesture recognition, In *Int'l Conf. on Automatic Face and Gesture Recognition*, pages 296–301, 1995.

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

[43] J.H. Friedman, Regularized Discriminant Analysis. *Journal of Amer. Statist. Assoc.*, 84:165-175, 1989.

[44] O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson, Detection of Neural Activity in Functional MRI Using Canonical Correlation Analysis, *Magnetic Resonance in Medicine* 45:323–330, 2001.

[45] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Proc. Int'l Symp. of Robotics Research*, pages 192–201, 2003.

[46] K. Fukui, B. Stenger, O. Yamaguchi, A framework for 3D object recognition using the kernel constrained mutual subspace method In *Proc. Asian Conference on Computer Vision*, pages 315-324, 2006.

[47] K. Fukunaga, *Introduction to statistical pattern recognition*, (2nd ed.), Academic Press, 1990.

[48] C. Fyfe and P.L. Lai, Canonical Correlation Analysis Neural Networks, In *Proc. International Conference on Pattern Recognition*, volume 2, pages 2977–2980, 2000.

[49] A.S.Georghiades, P.N.Belhumeur, and D.J.Kriegman, From few to many: illumination cone models for face recognition under varialbe lighting and pose, *IEEE Trans. on PAMI*, 23(6):643–660, 2001.

[50] J.M. Geusebroek, G.J. Burghouts, and A.W.M. Smeulders. The Amsterdam library of object images. *Int'l J. Computer Vision*, 61(1):103–112, January 2005.

[51] R. Gittins. *Canonical analysis: A review with applications in ecology.* Springer-Verlag, Berlin, Germany, 1985.

[52] S. Gong, S. McKenna, and J. Collins, An investigation into face pose distributions, In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 265–270, Vermont, USA, October 1996.

[53] S. Gong, S.J.McKenna and A.Psarrou, *Dynamic Vision : From Images to Face Recognition*, Imperial College Press, 2000.

[54] D.B. Graham, N.M. Allinson, Automatic Face Representation and Classification, In *Proc. British Machine Vision Conference*, pages 64–73, 1998.

[55] K. Grauman and T. Darrell, Pyramid Match Kernels: Discriminative Classification with Sets of Image Features, Technical report, AIM-2005-007, MIT, 2005.

[56] R. Gross, I. Matthews, S. Baker, Appearance-Based Face Recognition and Light-Fields, *IEEE Trans. on PAMI*, 26(4):449–465, 2004.

[57] S. Gutta and H. Wechsler, Face Recognition Using Asymmetric Faces, In *Proc. of International Concerence on Biometric Auentication*, pages 162–168, 2004.

[58] A. Hadid and M. Pietikainen. From Still Image to Video-Based Face Recognition: An Experimental Analysis. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 813–818, 2004.

[59] P. Hall, D. Marshall, and R. Martin. Merging and splitting eigenspace models. *IEEE Trans. on PAMI*, 22(9):1042–1049, 2000.

[60] D. Hardoon, S. Szedmak and J. Shawe-Taylor Canonical correlation analysis; An overview with application to learning methods *Neural Computation*, 16(12):2639–2664, 2004.

[61] R. Harshman. Generalisation of Canonical Correlation to N-way Arrays. Poster at *Thirty-fourth Annual Meeting of the Statistical Society of Canada*, May 2006.

[62] T. Hastie, R. Tibshirani and A. Buja, Flexible discriminant analysis by optimal scoring, *Journal of the American Statistical Association*, 89:1255–1270, 1994.

[63] T. Hastie, A. Buja and R. Tishibarani, Penalized Discriminant Analysis, *Annals of Statistics*, 23:73-102, 1995.

[64] T. Hastie and R. Tibshirani, Discriminant Analysis by Gaussian Mixtures, *Journal of the Royal Statistical Society series B*, 58:158–176, 1996.

[65] X.He, S.Yan, Y.Hu and H. Zhang, Learning a Locality Preserving Subspace for Visual Recognition, In *Proc. ICCV*, pages 385-392, Nice, France, 2003.

[66] X. He, X. Yan, and Y. Hu, Face Recognition Using Laplacianfaces, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3):328-340, 2005.

[67] B. Heisele, P. Ho and T. Poggio, Face Recognition with Support Vector Machines: Global versus Component-based Approach, In *Proc. International Conference on Computer Vision*, volume 2, pages 688-694, Vancouver, Canada, 2001.

[68] K. Hiraoka, K. Hidai, M. Hamahira, H. Mizoguchi, T. Mishima, and S. Yoshizawa. Successive learning of linear discriminant analysis: Sanger-type algorithm. In *Proc. ICPR*, volume 2, pages 664–667, 2000.

[69] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, 42(1-2):177–196, 2001.

[70] S.C.H. Hoi and M.R. Lyu. A Semi-Supervised Active Learning Framework for Image Retrieval. In *Proc. CVPR*, Pages 302–309, 2005.

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

[72] P. Howland and H. Park, Generalising discriminant analysis using the generalised singular value decomposition, *IEEE Trans. on PAMI*, 26(8):995-1006, 2004.

[73] J. Hua, Z. Xiong and E.R. Dougherty, Determination of the optimal number of features for quadratic discriminant analysis via the normal approximation to the discriminant distribution, *Pattern Recognition*, 38(3):403-421, 2005.

[74] J. Huang, P.C. Yuen, W.S. Chen, J.H. Lai, Component-based LDA Method for Face Recognition with One Training Sample, In *Proc. of AMFG*, pages 120–126, 2003.

[75] Aapo Hyvarinen, Juha Karhunen and Erkki Oja, *Independent Component Analysis*, John Wiley and Sons, Inc. 2001.

[76] S. Ioffe, Probabilistic Linear Discriminant Analysis, In *Proc. of ECCV*, pages 531–542, 2006.

[77] A.K. Jain and B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, In *Handbook of Statistics*, Edited by P.R. Krishnaiah and L.N. Kanal, 2:835–855, 1982.

[78] A. Just, Y. Rodriguez, and S. Marcel. Hand posture classification and recognition using the modified census transform. In *Proc. Int'l Conf. on Automatic Face and Gesture Recognition*, pages 351–356, 2006.

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

[80] T. Kamei, A. Yamada, T-K. Kim, H. Kim, Wonjun Hwang, S. Cheol Kee, *Advanced face descriptor using Fourier and intensity LDA features*, ISO/IEC JTC1/SC29/WG11 M8998, Oct 2002.

[81] T. Kamei and A. Yamada, Report of core experiment on Fourier spectral PCA based face description, ISO/IEC JTC1/SC21/WG11 M8277, Fairfax, VA, May 2002.

[82] T. Kamei, Face retrieval by an adaptive Mahalanobis distance using a confidence factor, In *Proc. of IEEE International Conference on Image Processing*, volume 1, pages 153–156, Rochester, USA, 2002.

[83] T. Kamei, A. Yamada, T-K. Kim, H. Kim, W. Hwang, S.C. Kee, Advanced face descriptor using Fourier and intensity LDA features, ISO/IEC JTC1/SC29/WG11 M8998, Oct 2002.

[84] T. Kanade and A. Yamada, Multi-Subregion Based Probabilistic Approach Toward Pose-Invariant Face Recognition, In *Proc. Of IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 954-959, Kobe, Japan, 2003.

[85] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. ICCV*, pages 166-173, 2005.

[86] J. R. Kettenring, Canonical analysis of several sets of variables, *Biometrika*, 58:433–451, 1971.

[87] H.C. Kim, D. Kim, and S.Y. Bang, Face retrieval using 1st- and 2nd-order PCA mixture model, In *Proc. of International Conference on Image Processing*, volume 2, pages 605-608, Rochester, USA, 2002.

[88] H.C. Kim, D. Kim, S.Y. Bang, Face Recognition Using LDA Mixture Model, In *Proc. International Conference on Pattern Recognition*, volume 2, pages 486–489, Canada, 2002.

[89] M.-S. Kim, D. Kim, S. Lee, and S.J. Kim, Experiment results of face descriptor using the embedded hmm with the 2nd-order block-specific eigenvectors, ISO/IEC JTC1/SC21/WG11 M8328, Fairfax, VA, May 2002.

[90] S.J. Kim, A. Magnani and S.P. Boyd, Robust Fisher Discriminant Analysis, In *Proc. of Neural Information Processing Systems*, pages 659–666, 2005.

[91] T-K. Kim, H. Kim, W. Hwang, S.C. Kee, J.H. Lee, Component-based LDA Face Descriptor for Image Retrieval, In *Proc. British Machine Vision Conference*, pages 507–526, 2002.

[92] T-K. Kim, D. Kong and S-R. Kim, Learning a Decision Boundary for Face Detection, In *Proc. of International Conference on Image Processing*, volume 1, pages 902-923, Rochester, USA, 2002.

[93] T-K. Kim, S-U. Lee, J-H. Lee, S. Kee and S-R. Kim, Integrated Approach of Multiple Face Detection for Video Surveillance, In *Proc. of International Conference on Pattern Recognition*, volume 2, pages 394–397, Quebec, Canada, 2002.

[94] T-K. Kim, H. Kim, W. Hwang, S-C. Kee and J. Kittler, Independent Component Analysis in a Facial Local Residue Space, In *Proc. IEEE Computer Vision and Pattern Recognition*, volume 1, pages 579-586, Madison, USA, 2003.

[95] T-K. Kim, J. Kittler, H-C. Kim and S-C. Kee, Discriminant Analysis by Multiple Locally Linear Transformations, In *Proc. British Machine Vision Conference*, pages 123–132, Norwich, UK, 2003.

[96] T-K. Kim, H. Kim, W. Hwang and S-C. Kee, Face Description based on Decomposition and Combining of a Facial Space with LDA, In *Proc. of International Conference on Image Processing*, volume 2, pages 877–880, 2003.

[97] T-K. Kim, H. Kim, W. Hwang and J. Kittler, Independent Component Analysis in a Local Facial Residue Space for Face Recognition, *Pattern Recognition*, 37(9):1873–1885, 2004.

[98] T-K. Kim and J. Kittler. Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005.

[99] T-K. Kim, H. Kim, W. Hwang, and J. Kittler. Component-based LDA face description for image retrieval and MPEG-7 standardisation. *Image and Vision Computing*, 23:631–642, 2005.

[100] T-K. Kim, O. Arandjelović and R. Cipolla, Learning over Sets using Boosted Manifold Principal Angles (BoMPA). In *Proc. British Machine Vision Conference*, pages 779–788, 2005.

[101] T-K. Kim, J. Kittler and R. Cipolla, Learning Discriminative Canonical Correlations for Object Recognition with Image Sets. In *Proc. European Conf. Computer Vision*, pages 251–262, Graz, Austria, 2006.

[102] T-K. Kim and J. Kittler. Design and Fusion of Pose Invariant Face Identification Experts. *IEEE Trans. on Circuit and System for Video Technology*, 16(9):1096-1106, 2006.

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

[104] T-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. on PAMI*, 29(6):1005–1018, 2007.

[105] T-K. Kim, S. Wong and R. Cipolla. Tensor Canonical Correlation Analysis for Action Classification. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[106] T-K. Kim, S-F. Wong, B. Stenger, J. Kittler and R. Cipolla, An Accurate and Efficient Solution of Incremental Linear Discriminant Analysis and its Applications, In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[107] T-K. Kim and R. Cipolla Gesture Recognition Under Small Sample Size, In *Proc. of Asian Conference on Computer Vision*, pages 335-344, Tokyo, Japan, Nov 2007.

[108] T-K. Kim, O. Arandjelović and R. Cipolla, Boosted Manifold Principal Angles for Image Set-Based Recognition. *Pattern Recognition*, 40(9):2475-2484, 2007.

[109] H. Kong, L. Wang, E.K. Teoh, J-G. Wang and R. Venkateswarlu, A framework of 2D Fisher discriminant analysis: application to face recognition with small number of training samples, In *Proc. CVPR*, volume 2, pages 1083-1088, 2005.

[110] T. Kozakaya, O. Yamaguchi, K. Fukui, Development and Evaluation of Face Recognition System Using Constrained Mutual Subspace Method *IPSJ Journal*, 45(3):951–959, 2004.

[111] M. Kuss and T. Graepel, *The Geometry Of Kernel Canonical Correlation Analysis*, Technical Report No. 108, Max Planck Institute for Biological Cybernetics, May 2003.

[112] P.L. Lai, Probabilistic Derivation and Multiple Canonical Correlation Analysis, In *Proc. European Symposium on Artificial Neural Networks*, pages 445–452, 2002.

[113] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. of ICCV*, pages 432-439, 2003.

[114] D.D. Lee and H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, 401:788–791, 1999.

[115] D.D. Lee and H.S. Seung, Algorithms for non-negative matrix factorization, In *Proc. NIPS*, pages 556-562, 2001.

[116] K. Lee, M. Yang, and D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 313–320, Madison, USA, 2003.

[117] B. Leibe and B. Schiele, Analysing appearance and contour based methods for object categorisation. In *Proc. Computer Vision and Pattern Recognition*, pages 409–415, Madison, USA, 2003.

[118] Y. Li, S. Gong, and H. Liddell. Recognising the dynamics of faces across multiple views. In *Proc. British Machine Vision Conference*, pages 242–251, 2000.

[119] Y. Li, S. Gong, and H. Liddell, Constructing facial identity surfaces in a nonlinear discriminating space, In *Proc. CVPR*, volume 2, pages 258–263, 2001.

[120] R.-S. Lin, D. Ross, J. Lim, and M.-H. Yang. Adaptive discriminative generative model and its applications. In *Proc. NIPS*, pages 801–808, 2005.

[121] Y. Lin and Y. Jeon, Discriminant analysis through a semiparametric model, *Biometrika*, 90(2):379-392, 2003.

[122] T. Lissack and K. Fu, Error estimation in pattern recognition via L-distance between posterior density functions, *IEEE Trans. Information Theory*, 22:34-45, 1976.

[123] C. Liu and H.-Y. Shum, Kullback-Leibler Boosting, In *Proc. of CVPR*, volume 1, pages 587–594, 2003.

[124] Q. Liu, R. Huang, H. Lu and S. Ma, Face recognition using Kernel-based Fisher Discriminant Analysis, In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 205–211, 2002.

[125] X. Liu and T. Chen. Video-Based Face Recognition Using Adaptive Hidden Markov Models. In *Proc. Computer Vision and Pattern Recognition*, pages 340–345, Madison, USA, 2003.

[126] M. Loog, R.P.W. Duin, and R.H. Umbach, Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria, *IEEE Trans. on PAMI*, 23(7):762–766, 2001.

[127] M. Loog and R.P.W. Duin, Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion, *IEEE Trans. PAMI*, 26(6):732–739, 2004.

[128] M. Loog, D. Ridder, Local Discriminant Analysis, In *Proc. of International Conference on Pattern Recognition*, pages 328–331, 2006.

[129] D. Lowe, Object recognition from local scale-invariant features, In *Proc. International Conference on Computer Vision*, pages 1150–1157, 1999.

[130] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[131] D.J.C.Mackay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[132] K. Maeda, O. Yamaguchi, and K. Fukui, A Fundamental Discussion on 3-Dimensional Pattern Matching Using Canonical Angles between Subspaces, *IEICE Trans. on Information and Systems*, J89-D(6):1288–1296, 2006.

[133] B.S. Manjunath, R. Chellappa, and C.V.D. Malsburg, A feature based approach to face recognition, In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 373-378, 1992.

[134] B.S. Manjunath, P. Salembier, T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley, New York, 2002.

[135] C.D.M. Martin. *Tensor Decompositions Workshop Discussion Notes*, American Institute of Mathematics (AIM), Palo Alto, CA, July 2004. http://aimath.org/WWN/tensordecomp/tensordecomp.pdf

[136] A.M. Martinez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(6):748–763, 2002.

[137] A.M. Martinez, Recognizing Expression Variant Faces from a Single Sample Image per Class, In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 353–358, Madison, USA, 2003.

[138] K. Matusita, Decision rules based on the distance for problems of fit, two samples and estimation, *Annals of Mathematical Statistics*, 26:631-640, 1955.

[139] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K-R.Muller, Fisher Discriminant Analysis with Kernels, In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pages 41–48, 1999.

[140] B. Moghaddam, and A. Pentland, Probabilistic visual learning for object representation, *IEEE Trans. on PAMI*, 19(7):696–710, 1997.

[141] A. Nefian and M. Hayes, An embedded hmm-based approach for face detection and recognition, In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3553–3556, 1999.

[142] A. Nefian and B. Davies, Standard support for automatic face recognition, ISO/IEC JTC1/SC21/WG11/M7251, Sydney, July 2001.

[143] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words, In *Proc. BMVC*, volume 3, pages 1249–1258, 2006.

[144] M. Nishiyama, O. Yamaguchi and K. Fukui, Face Recognition with the Multiple Constrained Mutual Subspace Method. In *Proc. of Audio- and Video-based Biometric Person Authentication*, pages 71–80, 2005.

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

[146] K. Okada, C.V.D. Malsburg, Analysis and synthesis of human faces with pose variations by a parametric piecewise linear subspace method, In *Proc. CVPR*, pages 761–768, 2001.

[147] T. Okada and S. Tomita, An Optimal Orthonormal System for Discriminant Analysis, *Journal of Pattern Recognition*, 18:139–144, 1985.

[148] E. Osuna, R. Freund and F. Girosi, Training support vector machines: an application to face detection, In *Proc. CVPR*, pages 130–136, June 1997.

[149] S. Pang, S. Ozawa, and N. Kasabov. Incremental linear discriminant analysis for classification of data streams. *IEEE Trans. on System, Man and Cybernetics*, 35(5):905–914, 2005.

[150] E. Patrick and F. Fisher, Nonparametric feature selection, *IEEE Trans. Information Theory*, 15:577-584, 1969.

[151] A. Pentland, B. Moghaddam, and T. Starner, View-based and Modular Eigenspaces for Face recognition, In *Proc. CVPR*, pages 84–91, 1994.

[152] A. Pezeshiki, M.R.A-Sadjadi and L.L. Scharf, Undersea Target Classification Using Canonical Correlation Analysis, *IEEE Journal of Oceanic Engineering*, to appear 2008.

[153] P.J. Phillips, H. Moon, S. Rizvi, and P.J. Rauss, The FERET evaluation methodology for face-recognition algorithm, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[154] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone, *FRVT 2002: Evaluation Report*, Mar 2003. http://www.frvt.org/FRVT2002/.

[155] A.K. Qin, P.N. Suganthan, M. Loog, Uncorrelated heteroscedastic LDA based on the weighted pairwise Chernoff criterion. *Pattern Recognition*, 38(4):613-616, 2005.

[156] S.J. Raudys and A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(3):252-264, 1991.

[157] M. Reiter, R. Donner, G. Langs, and H. Bischof, Estimation of Face Depth Maps from Color Textures using Canonical Correlation Analysis, In *Proc. Computer Vision Winter Workshop*, pages 17–21, Czech Republic, February 2006.

[158] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge university press, Cambridge, 1997.

[159] S.T. Roweis and L.K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding, *Science*, 290(5500):2323–2326, 2000.

[160] M.T. Sadeghi and J.V. Kittler. Decision Making in the LDA Space: Generalised Gradient Direction Metric. In *Proc. Int'l Conf. on Automatic Face and Gesture Recognition*, pages 248–253, 2004.

[161] P. Saisan, G. Doretto, Y.N. Wu and S. Soatto, Dynamic Texture Recognition, In *Proc. CVPR*, volume 2, pages 58–63, 2001.

[162] S. Satoh. Comparative Evaluation of Face Sequence Matching for Content-based Video Access. In *Proc. Int'l Conf. on Automatic Face and Gesture Recognition*, pages 163–168, 2000.

[163] S. Savarese, 2D and 3D spatial reasoning for object categorisation, Tutorial in *International Computer Vision Summer School*, Sicily, Italy, July 2007.

[164] R.E. Schapire, The boosting approach to machine learning: An overview, In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, *Nonlinear Estimation and Classification*, Springer, 2003.

[165] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc. ICPR*, pages 32–36, 2004.

[166] A. Shabbar, and A.G. Barnston, Skill of Seasonal Climate Forecasts in Canada Using Canonical Correlation Analysis, *Journal of monthly weather review*, 124(10):2370–2385, 1996.

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

[168] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. CVPR*, pages 405–412, 2005.

[169] D. Skocaj and A. Leonardis. Weighted and robust incremental method for subspace learning. In *Proc. ICCV*, pages 1494–1501, 2003.

[170] D. Skocaj, A. Leonardis, Appearnce-based localization using CCA, In *Proc. Computer Vision Workshop*, pages 205–214, Slovenia, 2004.

[171] T. Starner, A. Pentland, and J. Weaver. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1371–1375, 1998.

[172] G. Strang, *Linear algebra and its applications*, Harcourt, Inc., USA, 1988.

[173] M. Sugiyama, Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction, In *Proc. of International Conference on Machine Learning*, pages 905–912, Pittsburgh, USA, 2006

[174] Q-S. Sun, P-. Heng, Z. Jin and D-S. Xia, Face Recognition Based on Generalised Canonical Correlation Analysis, In *Proc. International Conference on Intelligent Computing*, pages 958-967, 2005.

[175] B.G. Tabachnick and L.S. Fidell, *Using Multivariate Statistics*, Harper Collins College Publishers: New York, 1996.

[176] F. Tang and H. Tao, Fast LDA using binary bases, In *Proc. of International Conference on Pattern Recognition*, volume 2, pages 52–55, 2006.

[177] F. Tang and H. Tao, Binary Principal Component Analysis, In *Proc. of British Machine Vision Conference*, volume 1, pages 377–386, 2006.

[178] X. Tao, J. Ye, Q. Li, R. Janardan, and V. Cherkassky. Efficient Kernel Discriminant Analysis via QR Decomposition. In *Proc. NIPS*, pages 1529–1536, 2005.

[179] M.E. Tipping and C.Bishop, Probabilistic Principal Component Analysis, *Journal of the Royal Statistical Society Series B*, 61(3):611–622, 1999.

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

[181] M.E. Tipping, The Relevance Vector Machine, In *Proc. Advanced in Neural Information Processing Systems*, pages 652–658, San Mateo, CA, 2000.

[182] F. D. Torre and M.J. Black, Robust principal component analysis for computer vision, In *Proc. of Int. Conf. on Computer Vision*, pages 362–369, 2001.

[183] F.D. Torre and T. Kanade, Multimodal oriented discriminant analysis, In *Proc. International Conference on Machine Learning*, pages 177–184, 2005.

[184] F.D. Torre, R. Gross, S. Baker, and V. Kumar. Representational oriented component analysis (ROCA) for face recognition with one sample image per training class. In *Proc. CVPR*, volume 2, pages 266–273, 2005.

[185] F.D. Torre, Component Analysis for Computer Vision, Tutorial in *European Conference on Computer Vision*, Graz, Austria, May 2006.

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

[187] M. Turk and A. Pentland, Eigenfaces for Recognition, *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[188] V.Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.

[189] M.A.O. Vasilescu and D. Terzopoulos. Multilinear Analysis of Image Ensembles: TensorFaces. In *Proc. ECCV*, pages 447–460, 2002.

[190] J.J. Verbeek, S.T. Roweis and N. Vlassis, Non-linear CCA and PCA by Alignment of Local Models, In *Proc. NIPS*, pages 297–304, 2003.

[191] T. Vetter and T. Poggio, Linear Object Classes and Image Synthesis From a Single Example Image, *IEEE Trans. PAMI*, 19(7):733–742, 1997.

[192] J. Via, I. Santamaria, J. Perez, A learning algorithm for adaptive canonical correlation analysis of several data sets, *Neural Networks*, 20(1):139–152, 2007.

[193] P. Viola and M. Jones. Robust real-time face detection. *Int'l J. Computer Vision*, 57(2):137–154, 2004.

[194] H. Wang, W. Zheng, Z. Hu, S. Chen, Local and Weighted Maximum Margin Discriminant Analysis. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[195] L. Wang and T.K. Tan, Experimental results of face description based on the 2nd-order eigenface method, ISO/IEC JTC1/SC21/WG11/M6001, Geneva, May 2000.

[196] X. Wang and X. Tang. Random Sampling LDA for Face Recognition. In *Proc. Computer Vision and Pattern Recognition*, pages 259–265, 2004.

[197] M. Welling, Kernel Canonical Correlation Analysis, Technical Note, Department of Computer Science, University of Toronto. http://www.ics.uci.edu/∼welling/classnotes/papersclass/

[198] L. Wiskott, J.-M. Fellous, N. Kruger, C.V.D. Malsburg, Face recognition by elastic bunch graph matching, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775-779, 1997.

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

[200] L. Wolf and A. Shashua. Kernel Principal Angles for Classification Machines with Applications to Image Sequence Interpretation. In *Proc. CVPR*, pages 635-640, 2003.

[201] L. Wolf, H.Jhuang and T.Hazan, Learning Appearances with Low-Rank SVM, In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007

[202] S-F. Wong and R. Cipolla. Real-time interpretation of hand motions using a sparse Bayesian classifier on motion gradient orientation images. In *Proc. BMVC*, pages 379–388, 2005.

[203] S-F. Wong, T-K. Kim and R. Cipolla, Learning Motion Categories Using Both Semantic and Structural Information, In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[204] J. Wu and Z.-H. Zhou, Face Recognition with one training image per person, *Pattern Recognition Letters*, 23(14):1711-1719, 2002.

[205] Y. Wu and S. Huang. View-independent recognition of hand postures. In *Proc. CVPR*, pages 2088–2094, 2000.

[206] T. Xiong, J. Ye, and V. Cherkassky, Kernel Uncorrelated and Orthogonal Discriminant Analysis: A Unified Approach. In *Proc. Computer Vision and Pattern Recognition*, pages 125–131, 2006.

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

[208] J. Yan, B. Zhang, S. Yan, Q. Yang, and H. Li. IMMC: Incremental maximum margin criterion. In *Proc. Int'l Conf. Knowledge Discovery and Data Mining*, pages 725–730, 2004.

[209] S. Yan, H. Zhang, Y. Hu, B. Zhang and Q. Cheng, Discriminant Analysis on Embedded Manifold, In *Proc. of European Conference on Computer Vision*, pages 121–131, Prague, May 2004.

[210] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang and H. Zhang. Discriminant Anlysis with Tensor Representation. In *Proc. CVPR*, pages 526–532, 2005.

[211] M.-H. Yang, Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods, In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 215–220, 2002

[212] Q. Yang, and X. Ding, Symmetrical PAC in Face Recognition, In *Proc. of ICIP* volume 2, pages 97–100, 2002.

[213] M.H. Yang, D. Kriegman, and N. Ahuja, Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[214] J. Yang and J-Y. Yang, Why can LDA be performed in PCA transformed space?, *Pattern Recognition*, 36(2):563–566, 2003.

[215] J. Yang, D. Zhang, A.F. Frangi, Two-dimensional PCA: A new approach to appearance based face representation and recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(1):131-137, 2004.

[216] J. Ye, R. Janardan, and Q. Li, Two-Dimensional Linear Discriminant Analysis. In *Proc. of NIPS*, pages 1569–1576, 2004.

[217] J. Ye, R. Janardan, C. Park, and H. Park, An optimization criterion for generalised discriminant analysis on undersampled problems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(8):982-994, 2004.

[218] J. Ye and Q. Li, A two-stage linear discriminant analysis via QR decomposition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(6):929-941, 2005.

[219] 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.

[220] J. Ye and T. Wang, Regularized discriminant analysis for high dimensional, low sample size data, In *Proc. of 12th ACM SIGKDD*, pages 454–463, Philadelphia, USA, 2006.

[221] J. Ye, Least Squares Linear Discriminant Analysis. In *Proc. of International Conference on Machine Learning*, pages 1087–1093, 2007.

[222] X. Yin, Canonical correlation analysis based on information theory, *Journal of Multivariate Analysis*, 91(2):161–176, 2004.

[223] H. Yu and J. Yang, A Direct LDA Algorithm for High-Dimensional Data - with Application to Face Recognition, *Pattern Recognition* 34(10):2067–2070, 2001.

[224] D.Q. Zhang, S.C. Chen, and Z.-H. Zhou, A new face recognition method based on SVD perturbation for single example image per person, *Applied Mathematics and Computation*, 163(2):895–907, 2005.

[225] W. Zhao, R. Chellappa and N. Nandhakumar, Empirical Performance Analysis of Linear Discriminant Classifiers, In *Proc. CVPR*, pages 164–169, Santa Barbara, CA, June 1998.

[226] W.Y. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant Analysis of Principal Components for Face Recognition. In *Proc. Int'l Conf. on Automatic Face and Gesture Recognition*, pages 336–341, 1998.

[227] W. Zhao, R. Chellappa, and P.J. Phillips, Subspace Linear Discriminant Analysis for Face Recognition, Technical Report CAR-TR-914, Center for Automation Research, University of Maryland, 1999.

[228] W. Zhao, Discriminant Component Analysis For Face Recognition, In *Proc. International Conference on Pattern Recognition*, volume 2, pages 818–821, 2000.

[229] W. Zheng, C. Zou, L. Zhao, Weighted maximum margin discriminant analysis with kernels, *Neurocomputing*, 67:357-362, 2005.

[230] W. Zheng, X. Zhou, C. Zou, L. Zhao, Facial Expression Recognition Using Kernel Canonical Correlation Analysis (KCCA), *IEEE Trans. Neural Networks*, 17(1):233–238, 2006.

[231] S. Zhou and R. Chellappa, Probabilistic human Recognition from video, In *Proc. European Conference on Computer Vision*, volume 3, pages 681–697, Copenhagen, Denmark, 2002.

[232] 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.

[233] S. Zhou and R. Chellappa, Probabilistic distance measures in reproducing kernel Hilbert space, SCR Technical Report, University of Maryland, 2004.

[234] S. Zhou and R. Chellappa, From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(6):917–929, June 2006.

[235] M. Zhu and A.M. Martinez, Selecting Principal Components in a Two-Stage LDA Algorithm, In *Proc. of CVPR*, pages 132–137, 2006.

[236] M. Zhu and A.M. Martinez, Subclass discriminant analysis, *IEEE Trans. on PAMI*, 28(8):1274-1286, 2006.

[237] X. Zhu. *Semi-Supervised learning literature survey*. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2006.