Multi-view dimensionality reduction for multi-modal biometrics

Xuran Zhao

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0061

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal and Image »

présentée et soutenue publiquement par

Xuran ZHAO

24.10.2013

Réduction de la Dimensionalité Multivue pour la Biométrie Multimodal

Directeur de thèse : **Nicholas Evans & Jean-Luc Dugelay**

Jury

M. Fabio Roli, Professeur, University of Cagliari	Président
M. Christophe Rosenburger, Professeur, ENSICAEN	Rapporteur
M. John Mason, Professeur, Swansea University	Rapporteur
M. Vincent Despiegel, Chercheur, MORPHO	Examinateur
M. Nicholas Evans, Maitre de conférences, EURECOM	Directeur de thèse
M. Jean-Luc Dugelay, Professeur, EURECOM	Directeur de thèse

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

T H È S E

Acknowledgements

Three years ago when I got my master degree, I said to myself: "I don't want to be a software engineer, so I will do some research instead", and that was how I started my Ph.D. I then found that research is pretty like a game. I was assigned with a topic, namely multimodal biometrics, much like being told: "This is your game for the next three years". So I figure out the rules, look at how others play, and try to play better. After three years, I am still not sure if I can really play better than the others, but am particularly happy with two things: first, I enjoyed the game; and second, I am willing to continue playing as a post-doc rather than to say "I quit".

I had a very good time studying and working at EURECOM, and living under the sunshine of Cote-d'Azur. I would like to thank a lot of people. Of course, I want to first express my greatest gratitude to my thesis supervisors Dr. Nicholas Evans and Dr. Jean-Luc Dugelay, who gave me this opportunity and helped me in all aspects. I remember a Friday late night three years ago, when Nick was at home but still working with me on my first paper by sending his revisions by E-mails. At 10:30 PM, he E-mailed me: "I have to leave for a while to give my baby girl a shower, and will come back in half an hour." It is hard to express my gratitude to him, but I know that I owe him so much. It is also my greatest pleasure to work under the supervision of Prof. Dugelay who shared with me his rich experience in both research and teaching, and gave me a lot of wonderful ideas on my research. My sincere appreciation also goes to my thesis committee members for their supports. I would like to thank all the jury members for their precious time to read my manuscript.

I also want to thank my current and previous colleges, including Antitza Dantcheva, Nesli Erdogmus, Carmelo Velardo, Rui Min, Hajer Fradi, Neslihan Kose, Andrea MELLE, Asmaa Fillatre, Christelle Yemdji Tchassi, Federico Alegre, Giovanni Soldi and Leela Gudupudi who shared knowledge and experience with me had have pleasant discussions. I must also thank the members of "EURECOM Chinese team", including Xiaolan Sha, Heng Cui,

Xueliang Liu, Rui Min, Kaijie Zhou, Shengyun Liu, Xiaohu Wu, Haifan Yin, Xinping Yi, Qianrui Li, Jinyuan Chen from whom I have learnt enormous knowledge in both research and life. I would like to thank all my friends at EURECOM for the colourful life I enjoyed during the past three years.

At last, I would give my heartfelt thanks to my family. My parents always encourage me when I experienced a difficult time. Finally, my love goes to my wife Jiongjiong Mo, who has shared 10-years time with me and always supported me.

Abstract

Biometric data is often represented by high-dimensional feature vectors which contain significant inter-session variation. Efficient dimensionality reduction techniques are thus needed in order to extract class-discriminative, low-dimensional features and to attenuate unwanted variations which is redundant to recognition. Such discriminative dimensionality reduction techniques generally follow a supervised learning scheme, in which a subspace projection is learned with feature-label pairs. However, labelled training data is generally limited in quantity and often does not reliably represent the inter-session variation encountered in test data. The limited size of labelled training set often leads to biased projection matrices and degraded recognition performance.

This thesis proposes to use multi-view dimensionality reduction (MVDR) which aims to extract discriminative features in multi-modal biometric systems, where different modalities are regarded as different views of the same data. Instead of training on feature-label pairs, MVDR projections are trained on feature-feature pairs where label information is not required. Since unlabelled data is easier to acquire in large quantities, and because of the natural co-existence of multiple views in multi-modal biometric problems, discriminant, low-dimensional subspaces can be learnt using the proposed MVDR approaches in a largely unsupervised manner.

According to different functionalities of biometric systems, namely recognition (including identification and verification), clustering, and retrieval, we propose three MVDR frameworks which meet the requirements for each functionality. The proposed approaches, however, share the same spirit: all methods aim to learn a projection for each view such that a certain form of agreement is attained in the subspaces across different views. The proposed MVDR frameworks can thus be unified into one general framework for multi-view dimensionality reduction through subspace agreement. We regard this novel concept of subspace agreement to be the primary contribution of this thesis.

Contents

Chapter 1

Introduction

Biometrics refers to the recognition of humans by their physical or behavioural traits. In every-day life, people deal with a multitude of problems concerning personal identities and the recent innovations in biometric systems provide one solution to simpler, faster, and more secure identification. For example, biometrics systems are widely used in access control, including either physical access to a specific resource, location or territory (access control to a building, immigration border control etc.), or virtual access to a computer network, online bank account, for instance. In these applications, biometric traits such as face, iris, fingerprints and voice could be used to replace (or to complement) passwords or ID cards, which could be either forgotten or stolen [Jain et al., 1999]. Biometrics systems based on face or gait recognition could used to identify individuals (e.g. criminals) in video surveillance systems, since they need minimal user cooperation [Gafurov, 2007, Cucchiara, 2005]. Human face and voice informations can also helps the management of multimedia data, in order to make the retrieval or indexing of multimedia files more accurate and efficient [Sargin et al., 2009, Jain et al., 2012, Lee et al., 2011].

Whatever the applications, from a computer science perspective, biometrics is a pattern recognition problem. Biometric systems commonly contain two operation modules, feature extraction and comparison (or classification). In the feature extraction module, biometric samples are represented by numerical features which can be processed by computer programs; in the comparison or classification module, extracted feature from a test sample is compared to one or several features obtained from the enrollment samples (known as template) to determine if the test sample have the claimed identity (verification mode)

or which enrolled identity the test sample belongs to. In most state-of-the-art biometric systems, data samples are often represented by high-dimensional feature vectors (e.g., local binary patterns (LBP) for face recognition [Ahonen et al., 2006] and Gaussian mixture model (GMM) supervectors for speaker recognition [Campbell et al., 2006]). The high dimensionality of biometric features incur heavy storage and computational burdens, and more importantly, the so-called *curse-of-dimensionality* [Bellman, 1961] can impact on the recognition performance in the following comparison or classification module.

Difficulties associated with the high dimensionality are generally overcome through the application of dimensionality reduction (DR) techniques [Jolliffe, 2005, Bartlett et al., 2002, Niyogi, 2004, Belhumeur et al., 1997], which look for a lower dimensional representation of the high-dimensional data. Depending on whether label information is needed, DR techniques can be broadly categorized into supervised or unsupervised methods. A dilemma arises from the trade-off between the availability of label information and the discriminative power of the extracted low-dimensional features. Supervised methods such as Linear Discriminant Analysis (LDA) [Belhumeur et al., 1997] have high discriminative power, but require large quantity of manually labelled training data. Unsupervised methods such as Principle Component Analysis (PCA) [Jolliffe, 2005] do not require class labels, but generally lack discriminative power. In biometric identification and verification settings, manually labelled data is normally limited in number while large amount of unlabelled data can be easily acquired during the normal system use.

In multi-modal biometrics, different biometric modalities can form different inputs to classification algorithms. Multimodal biometric systems can obtain multiple sets of information from the same modality (i.e., 2D+3D face recognition) [Bowyer et al., 2006] or information from different biometric modalities (i.e. biometric system with face and voice). The fusion of modalities remains a challenging problem and is generally treated in isolation to that of high dimensionality [Ross et al., 2008].

This thesis tackles the high dimensionality problem and the multi-modal fusion problem in a unified framework. Under a multi-modal biometric setting and given abundant unlabelled data, we aim to extract highly-discriminative features from multiple modalities in an unsupervised manner.

1.1 Contributions

In this section, we briefly summarize the content of the thesis and the contributions.

Multi-modal biometric systems utilize two or more individual modalities to improve the

recognition accuracy of conventional uni-modal methods. In a bimodal biometric system which employs two different modalities, data samples can be represented by paired features $< X^{(1)}, X^{(2)} >$ and a subject identity Y as a target variable. State-of-the-art biometric systems often make use of high-dimensional features, thus dimensionality reduction (DR) techniques are often applied to alleviate the so-called curse-of-dimensionality problem in the following classification step. This thesis reports a study of DR approaches for multi-modal biometrics. Commonly referred to as Multi-view Dimensionality Reduction (MVDR), this field has attracted considerable research interest in recent years. Most existing MVDR algorithms are based on Canonical Correlation Analysis (CCA) [Foster et al., 2008] and its variants. These algorithms generally aim to learn projections $P^{(1)}$ and $P^{(2)}$ such that the projected samples $P^{(1)}X^{(1)}$ and $P^{(2)}X^{(2)}$ are maximally correlated. When applied to paired features, the main advantage of MVDR methods over single-view DR is that one view can be regarded as weak labels for the other. Accordingly, class-specific discriminative features can thus be extracted even if labelled training samples are either limited in number or entirely absent. In contrast to such previous approaches, the new work reported in this thesis addresses the MVDR problem from a different angle. Inspired by the pioneering semi-supervised learning method *co-training* [Blum and Mitchell, 1998b], this thesis introduces a new concept of *subspace structure agreement*. The main idea involves learning projections $P^{(1)}$ and $P^{(2)}$, such that the data structure of projected samples $P^{(1)}X^{(1)}$ and $P^{(2)}X^{(2)}$ is as similar as possible. According to different definitions of *agreement* and different applications, namely semi-supervised classification, unsupervised clustering and retrieval, we propose three different MVDR approaches, which are described in the following.

The first approach is a direct extension of incremental co-training to semi-supervised MVDR problems via the co-training of Linear Discriminant Analysis (LDA) [Belhumeur et al., 1997] projections. The algorithm is referred to as Co-LDA and was published in the proceedings of International Conference on Multimedia and Exposition (ICME) in 2012. The input involves a small set of two-view labelled training data $\{X_L^{(1)}, X_L^{(2)}; Y\}$ and a larger pool of unlabelled data $\{X_U^{(1)}, X_U^{(2)}\}$. While the larger pool is unlabelled, its size is more representative of the underlying data distribution. LDA projections $P^{(1)}$ and $P^{(2)}$ are initially trained on each view of the labelled training set. The unlabelled set is then projected into the same subspaces, and samples are assigned labels according to a nearest-centroid classifier. For each view, the subset of unlabelled samples which are most confidently labelled are removed from the unlabelled set and added to the labelled set. The LDA projections and classifiers are then retrained. The process iterates until the unlabelled set is empty. The new algorithm is successful in utilizing unlabelled data to avoid

over-fitting to the smaller hand-labelled dataset. When tested on the bimodal, face-voice MOBIO database [McCool et al., 2012], the proposed Co-LDA algorithm raises a baseline identification rate from 71% to 99% while in a verification task the Equal Error Rate (EER) is reduced from 16% to less than 1%. In an extension of this work which was published in the proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2013, we show that Sparse Representation Classifier (SRC) could be employed to reject out-of-class samples which belong to none of the registered classes. In related work published in proceedings of International Conference on Image Processing (ICIP) 2011 and European Signal Processing Conference (EUSIPCO) 2011, we also proposed a self-training version of the algorithm which could be applied to single-modal systems.

The standard co-training algorithm is semi-supervised and needs at some labelled data for initialization. For purely unsupervised clustering problems, we proposed a multi-view subspace clustering algorithm which is based on a multi-view cluster agreement assumption. We consider the problem of clustering a group of unlabelled two-view, high-dimensional data $X^{(1)}, X^{(2)}$ into k clusters, and samples in the same class are expected to be assigned to the same cluster. Since $X^{(1)}$ and $X^{(2)}$ are different representations of the same underlying class Y, in ideal conditions, identical clustering results are expected irrespective of the view used for clustering. However, this is unlikely if clustering is performed in the original feature spaces $X^{(1)}$ and $X^{(2)}$ since they are corrupted by different intra-class variation. This thesis introduces a new approach to multi-view subspace clustering which seeks projections $P^{(1)}$ and $P^{(2)}$ such that the clustering results are in maximal agreement in the subspaces across each view. We show that this objective can be obtained by combining the simplicity of k-means clustering and Linear Discriminant Analysis (LDA) within a co-training scheme. The new algorithm exploits cluster indicators obtained from k-means clustering in one view to learn discriminative subspaces in another. The algorithm is referred to as CoKmLDA. We show mathematically how LDA projections learned from samples with random label noise are probabilistically equivalent to those learned with clean labels and that cross-view labelling, or co-training, is efficient in correcting erroneous sample labels. Of particular merit, the algorithm does not require the optimization of any hyper-parameters. The effectiveness of the proposed algorithm is demonstrated not only in speaker clustering experiments on the same bimodal, face-voice MOBIO database, but also in more general clustering tasks such as handwritten digit clustering and text document clustering. Significant improvement over alternative multi-view clustering approaches such as CCA and co-spectral clustering are reported. This work has been submitted to a Special Issue on Unsupervised and Supervised Learning in Pattern Recognition Letters.

The proposed CoKmLDA algorithm is suitable for clustering problems but is not well adapted to other unsupervised learning problem such as retrieval, since it needs the number of clusters as an input parameter. In this thesis, we further proposed multi-view dimensionality reduction algorithm for retrieval problems, based on similarity graphs. Graph-based dimensionality reduction methods have recently emerged as a powerful tool for analysing high-dimensional data. Example algorithms include non-linear methods such as ISOMAP [Tenenbaum et al., 2000], Local Linear Embedding (LLE)[Roweis and Saul, 2000], Laplacian eigenmaps [Belkin and Niyogi, 2001] and linear methods such as Locality Preserving Projection (LPP) [Niyogi, 2004]. These methods begin with the construct of a similarity graph S in which the nodes represent data samples whereas the edges s_{ij} represent the similarity measure between the i^{th} and j^{th} sample. Graph-based dimensionality reduction methods are largely aiming to reveal the low-dimensional manifold structure of the original data, but are not capable of extracting class-specific discriminative features due to their unsupervised nature. Due to significant intra-class variations, s_{ij} could be very low measured in original spaces even if sample i and j belong to the same class. Unreliable estimation of similarity will influence projections and thus lead to sub-optimal subspaces. We consider graph-based dimensionality reduction in a two-view setting, where data samples can be again represented in the form of $< X^{(1)}, X^{(2)} >$ and the two views exhibit a certain level of conditional independence, as is often the case with biometrics. If similarity matrix $S^{(1)}$ and $S^{(2)}$ are constructed with $X^{(1)}$ and $X^{(2)}$ respectively, then they are expected to be different since $X^{(1)}$ and $X^{(2)}$ contains different intra-class variation. Assume that there exist optimal projections $P^{(1)}_{opt}$ and $P^{(2)}_{opt}$ such that in the two projected subspaces, same-class samples are located closed to each other where different-class samples are located apart and additionally $S^{(1)}$ and $S^{(2)}$ are constructed with the projected samples $P^{(1)}_{opt}\ X^{(1)}$ and $P^{(2)}_{opt}X^{(2)}$, $S^{(1)}$ and $S^{(2)}$ are expected to be similar. Based on this logic, we propose to approximate $P^{(1)}_{opt}$ and $P^{(2)}_{opt}$ by finding $P^{(1)}$ and $P^{(2)}$ which minimize the difference between $S^{(1)}$ and $S^{(2)}$. This objective could be achieved through the graph-based co-training of LPP, this thesis reports such an approach referred to as Co-LPP. Co-LPP is ideal for metric learning in retrieval problems, and its effectiveness is demonstrated with experiments on audio-visual person retrieval from videos and human face retrieval with multiple facial features. This work has been submitted to IEEE International Workshop on Information Forensics and Security (WIFS), 2013. The proposed subspace graph agreement is highly flexible and can be used to extend other single-view graph-based dimensionality reduction to a multi-view setting.

In summary, the contributions of this thesis are:

- A review of state-of-the-art Multi-View Dimensionality Reduction (MVDR) algorithms;
- A new approach to MVDR focusing on the novel concept of subspace structure agreement;
- Three new MVDR algorithms based on different definitions of subspace structure agreement;
- Applications of proposed algorithms to semi-supervised classification, unsupervised clustering, and retrieval problems in biometrics, especially in the context of audio-visual person recognition;
- Applications of proposed algorithms to more general pattern recognition problems for non-biometric data, such as image and text clustering and retrieval.

The work presented in this thesis has been published by the candidate in the following conferences and journals:

[ICIP2011] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Semi-supervised face recognition using LDA self-training", in Proceedings of IEEE International Conference on Image Processing (ICIP), September, 2011.

[EUSIPCO2011] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "A co-training approach to semi-supervised automatic face recognition", in Proceedings of European Signal Processing Conference (EUSIPCO), September, 2011.

[ICME2012] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Co-LDA: a semi-supervised approach to audio-visual person recognition", in International Conference of Multimedia and Exposition (ICME), July, 2012.

[EUSIPCO2012] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Multi-view Semi-supervised Dimensionality Reduction", in 2012 European Conference on Signal Processing (EUSIPCO), August, 2012.

[ICASSP2013] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Open-set semi-supervised audio-visual person identification using co-training LDA and sparse representation classifiers", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[PRL2013] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "A subspace co-training framework for multi-view clustering", submitted to Pattern Recognition Letters, under review.

[WIFS2013] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Unsupervised Multi-view Dimensionality Reduction with Application to Audio-Visual Speaker Retrieval", accepted in IEEE International Workshop on Information Forensics and Security (WIFS),2013

1.2 Outline

The thesis is organized as follows:

Chapter 2 discusses the background of biometric systems, including the biometric functionality, feature extraction, and the dimensionality reduction techniques. Under a multi-biometrics scenario, we further discuss the need for multi-view dimensionality reduction (MVDR).

Chapter 3 reviews State-of-the-art MVDR algorithms. Existing MVDR methods are categorized and discussed separately. In the end of the chapter, we propose our own MVDR scheme: MVDR based on data structure agreement in subspace.

Chapter 4 introduces our MVDR solutions for semi-supervised audio-visual speaker identification and verification problems. Co-LDA algorithm which is a direct extension of incremental co-training [Blum and Mitchell, 1998a] into the MVDR problem. We further presents the Co-LDA-SRC algorithm, which is an open-set extension of Co-LDA algorithm which is able to deal with out-of-sample data in the unlabelled dataset.

Chapter 5 presents our CoKmLDA algorithm for the clustering of multi-modal biometric data, especially audio-visual speaker clustering in videos. This algorithm can also be applied to more general problems of clustering multi-view high-dimensional data, such as text clustering and image clustering.

Chapter 6 presents the third MVDR approach Co-LPP based on similarity graph agreement and its application to metric learning for multi-view biometric data retrieval.

Conclusions, suggestions and final remarks are made in Chapter 7.

Chapter 2

Biometric Systems and Dimensionality Reduction

In this chapter, we first introduce some background information about biometric systems, including the biometric functionality, feature extraction and dimensionality reduction. Traditional dimensionality reduction techniques, including unsupervised and supervised approaches, are then presented and their relative advantages and disadvantages are analysed. Under a multi-modal biometrics scenario, we further discuss the need for multi-view dimensionality reduction (MVDR), which aims to learn discriminative subspaces in an unsupervised way.

2.1 Biometrics Systems

Just as there are no two identical leaves on a tree, there are no two identical persons in the world. As illustrated in Figure 2.1, different people are differentiated in their physiological traits such as face, fingerprints, iris, DNA, etc., and behaviour traits such as gait , speech, signature, and typing rhythm. *Biometric is the science to understand how to measure these person-specific characteristics and how to use them to distinguish individuals.* Humans have developed such skills during the evolution. For example, the brain has specialized areas to recognize faces [Nelson, 2001] and link identities with specific patterns. Researchers in biometrics try to automatize such processes and make them suitable to be run on a computer or a device by a biometric system.

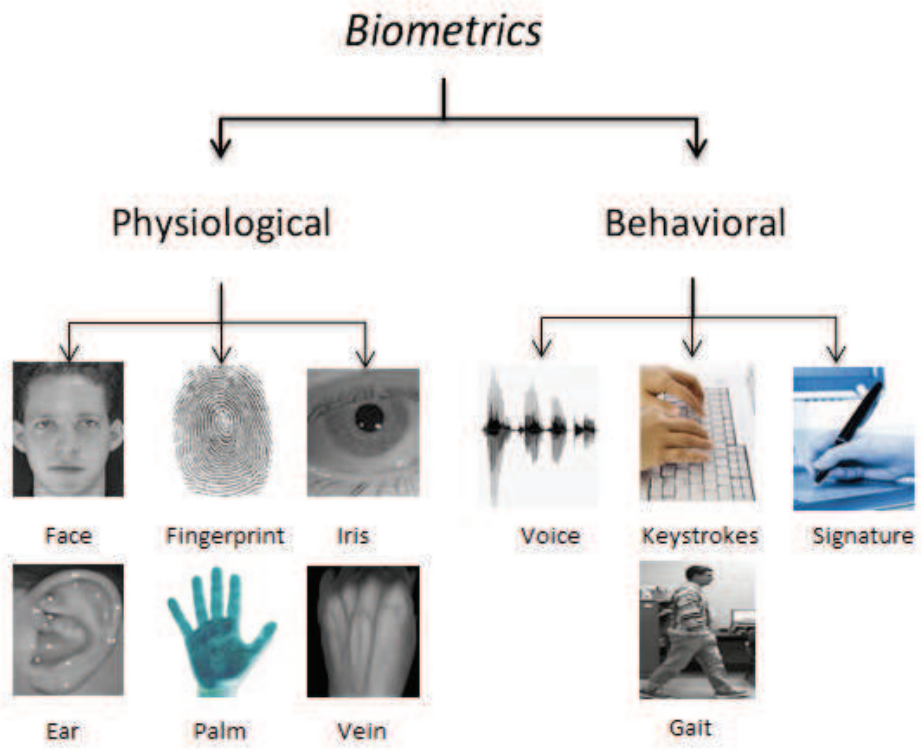


Figure 2.1: Examples of physiological and behavioral biometric traits which can be used to recognize people.

As shown in Figure 2.2, a typical biometric system contains four basic modules, namely sensor module, feature extraction module, comparison module, and decision module.

- **Sensor module**: Row biometric data is acquired form the users by a suitable biometric sensor, e.g. camera, microphone, fingerprint scanner, etc..

- **Feature extraction module**: a feature extractor is then used to extract salient information from the row biometric data to get a new representation of the data which is more favourable for the automated process of distinguishing individuals. For example, the extracted feature set should retain the difference between individuals while suppressing the difference between two samples of the same user. Moreover, the feature set should be in a compact form in order to reduce the computational and storage burdens. These feature sets of each client are stored in a database and referred to as templates or client models.

- **Comparison module**: the feature sets extracted from the test data are compared to one (verification) client model or all the client models (identification) to generate a similarity score or a set of scores according to some specific matching algorithm.
- **Decision module**: According the similarity score(s), the decision module decides if the test data has its claimed identity (verification) or determines its identity (identification).

In summary, biometrics is essentially a pattern recognition problem in which a to predict the subject identity according to some predefined features.

2.2 Functionalities of Biometrics

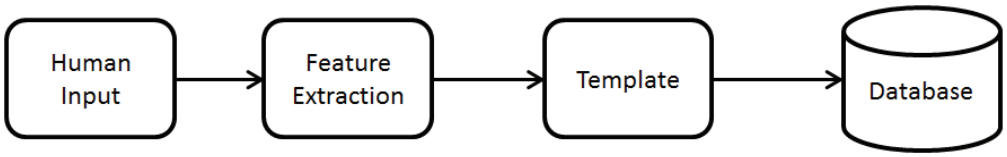
In order to discuss the functionalities of biometric systems, we would like to first define it in a *strict* or *broad* sense. In a strict sense, the final objective of a biometric system is to recognize people according to their biometric traits. In this case they and can be categorised into either *verification* or *identification* modes. In this sense, a biometrics system often has special data acquisition hardware and is closely related to security applications. In a broad sense, a biometric system may refer to any applications which makes use of biometrics features to facilitate other tasks such as data management or mining. This thesis discusses the functionality of biometrics in the two senses separately.

2.2.1 Verification and identification

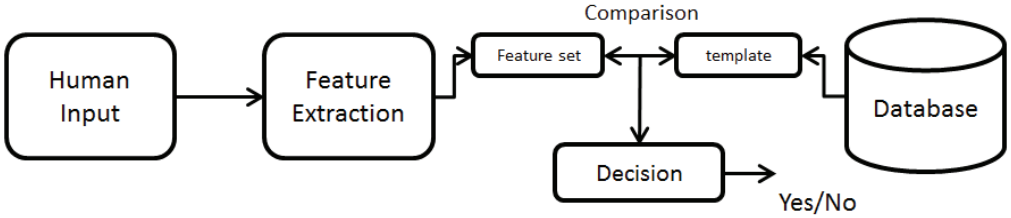
In the strict sense, a biometric system manage the identity of its users by their biometric traits, in order to allow or deny access to restricted areas or some devices, e.g. computers or mobile phones. Biometric enabled devices can also determine the identity of a user, in order to provide user-adapted solutions. In this sense, a biometric system works either in verification or identification modes.

Verification mode: In the verification mode, the system validates a person's identity by comparing the captured biometric data with her own biometric template(s) stored in the system database. In such a system, an individual who desires to be recognized claims an identity, and the system conducts a one-to-one comparison to determine whether the claim is true or not. Verification is typically used for positive recognition, where the aim is to prevent multiple people from using the same identity.

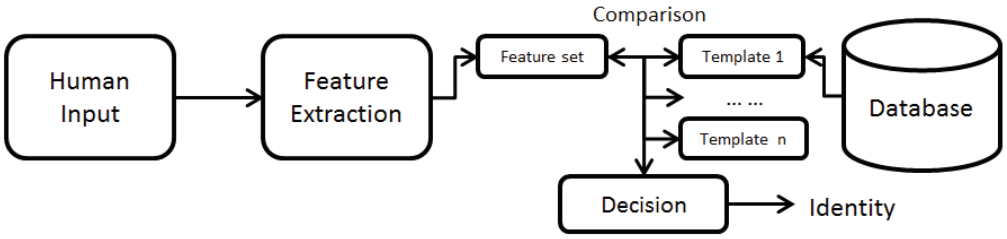
Identification mode: In the identification mode, the system recognizes an individual by searching the templates of all the users in the database for a match. Therefore, the system



(a) Flowchart of biometric enrolment process.



(b) Flowchart of biometric verification



(c) Flowchart of biometric identification

Figure 2.2: Scheme of a general biometric system and its modules: enrolment (a), verification (b), and identification (c).

conducts a one-to-many comparison to establish an individual's identity without a claim of the identity. Identification is a critical component in negative recognition applications where the system establishes whether the person is who he or she denies to be. The purpose of negative recognition is to prevent a single person from using multiple identities. Identification may also be used in positive recognition for convenience.

From a machine learning point of view, biometric verification and identification systems are actually a *supervised* two-class or multi-class classification problem. Client models or classifiers are trained with labelled training data acquired during the enrolment session, and used to classify test samples. Recent research has shown that unlabelled test samples accumulated during a period of normal system use can improve system performance [Rattani et al., 2009, Poh et al., 2009, Bhatt et al., 2011].

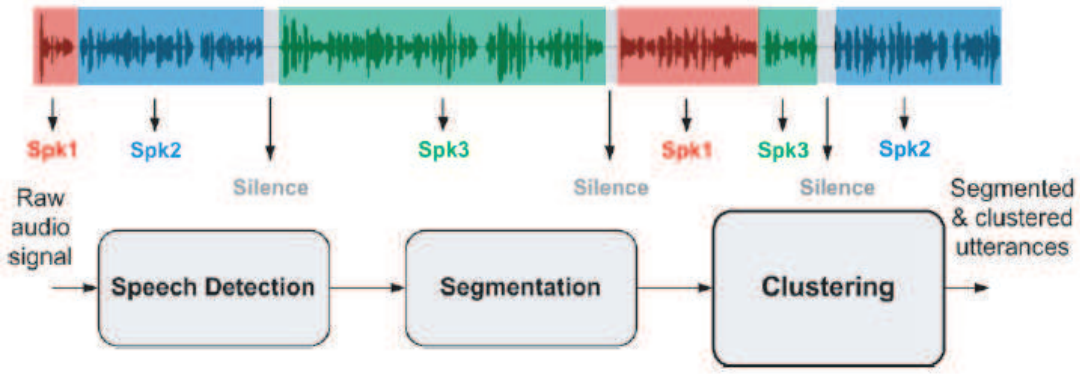


Figure 2.3: The process of speaker diarization (image excerpted from [Tang et al., 2012]).A typical speaker diarization system consists of a speech detection stage, a segmentation stage, and a clustering stage.

2.2.2 Clustering and retrieval

Nowadays the ever-growing mass of information and availability of digital image and video databases demand efficient data management and data mining tools. Biometrics traits, especially face and voice, typically exist in images, sound files and videos and thus speaker and face recognition has been exploited in applications such as multimedia database management and searching, as well as data mining. For example, Facebook[1] uses face recognition to automate user tagging in photographs. Each time an individual is tagged in a photograph, the software application stores information for that person's facial characteristics. When sufficient data has been collected about a person to identify them, the system uses that information to identify the same face in different photographs, and will subsequently suggest the tagging of those pictures with that person's name. Facial and vocal features has been used in tagging celebrities in Youtube videos [Sargin et al., 2009] in order to increase search accuracy. Different from the standard biometric verification and identification applications where client models or classifiers are trained with labelled data in a supervised manner, the billions of image or video samples on the internet are largely unlabelled. As a result, unsupervised learning methods such as clustering and retrieval are attracting more and more attention.

Clustering: Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other

[1] www.facebook.com

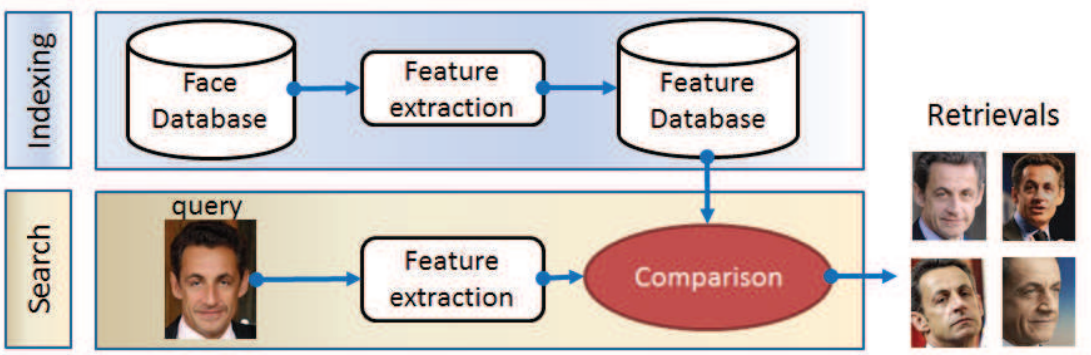


Figure 2.4: Face image retrieval flowchart.

than to those in other clusters. Face clustering is used in many digital album management applications. For example, Google Picasa[2] detects human faces in the user's digital album, clusters them into groups of different subjects and asking the user to label each group. In this way, the number of faces the users have to label is minimized. Speaker clustering is an essential element of speaker diarization [Anguera Miro et al., 2012, Tranter and Reynolds, 2006, Barras et al., 2006]. Here the task is to determine "Who speaks when?" and is accomplished by by partitioning an input audio stream into temporal regions contributed from the same speakers. The process of a typical spearker diarization system is shown in Figure 2.3. From the flowchart, we can see that speaker diarization is actually a speaker segmentation plus a speaker clustering problem. In such clustering problems, features are extracted from different image samples or acoustic frames, and similarity scores between those samples have to be calculated, samples having high similarity should be clustered together and ideally, high similarity should infer same subject identity.

Retrieval: Data retrieval is an important task in database management which involves extracting the wanted data from a database. In terms of biometrics, it can involve the retrieval of samples containing the same subject as that in a query sample according to certain biometric features, e.g. face and voice features in particular. Face image retrieval [Jain et al., 2012, Lee et al., 2011] constitute a significant research issue in many practical applications such as mug shot searching and surveillance systems, while speaker retrieval [Yang et al., 2005, Huijbregts and van Leeuwen, 2010] makes it possible to search for recordings of specific speakers in audio archives. A typical retrieval scheme involves the calculation of similarity between the query sample and each of the target samples in the database, and the a pre-specified number of most similar samples are retrieved.

[2] picasa.google.com

To sum up for Section 2.2, despite the different functionalities of biometrics systems, namely verification, identification, clustering and retrieval, they share at least two basic modules: *feature extraction* and *similarity comparison*. The feature extraction module involves the representation of biometric data samples by numerical features which could serve as inputs for machine learning algorithms, while the similarity comparison module involves comparing data samples measuring similarities.

Biometric data, however, often contains significant amount of intra-class variation while the inter-class variation can be small. Take human faces for example, as shown in Figure 2.5, significant intra-class variation may come from differences in illumination or pose, the presence of facial accessories (glasses or piercings), and ageing over an extended time period, while inter-class variation can be small due to reasons such as biological ties. In this case, similarity in extracted feature space may not reflect same identity, which can cause deteriorated recognition/clustering/retrieval performance.

Accordingly, *How to extract good features* becomes a critical issue in biometric research.

2.3 Biometric Feature Representation

Biometric feature extraction is a process applied to a biometric sample to determine repeatable, distinctive and efficient representations suited for subsequent comparison step. The extraction of biometric features from a biometric sample aims to suppress superfluous information which does not contribute to biometric recognition while simutaneously retaining that pertinent to recognition. How to extract *good* features from biometric samples constitutes a major research issue in the biometric community. A *good* biometric feature should have at least two characteristics: first, a good biometric feature should be *discriminative* between samples from different subjects; second, it should be *persistent* amount samples from the same subjects. In other words, inter-subject variation should be maximized while intra-subject variation should be minimized. Moreover, a good biometric feature should be ideally represented in a compact form, otherwise heavy computational burden will induce a prolonged time in the comparison module, and reduce user satisfaction.

A lot of research has been devoted to the extraction of inter-class-variation-discriminative and intra-class-variation-robust features. For example, in automatic face recognition, one of the most popular features involve Local Binary Patterns [Ahonen et al., 2006] which discriminate different subjects according to differences of texture between different facial images and is insensitive to inter-session variations caused by illumination changes. In automatic speaker recognition, a speech utterance is often represented by a Gaussian Mixture

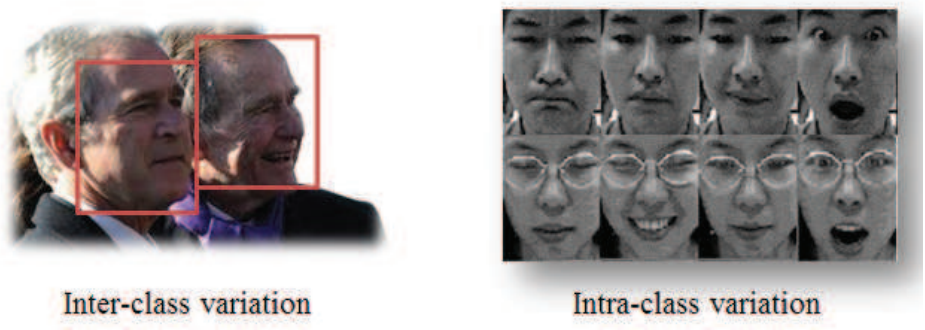


Figure 2.5: Inter and intra class variation in face recognition problems. Left: inter-class variation could be small between persons with biological ties. Right: Intra-class variation could be significant due to different poses, illumination, and expression.

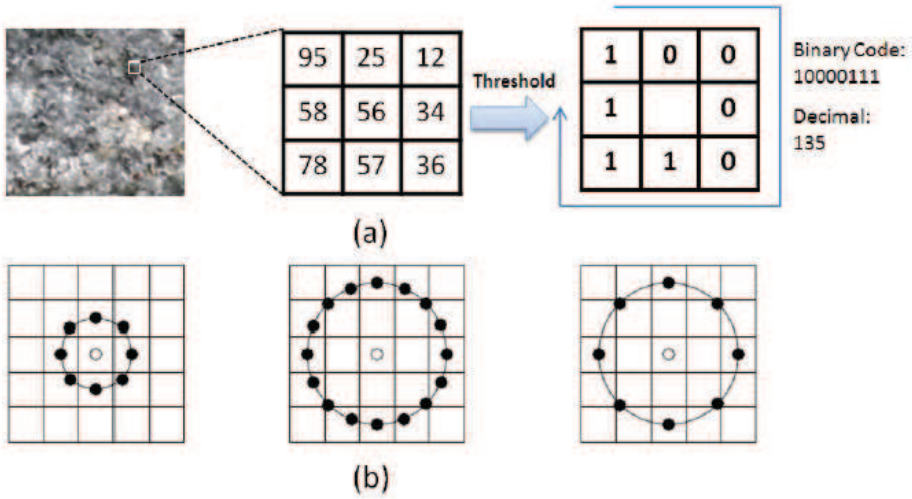


Figure 2.6: (a) basic LBP operator, (b) the circular (8,1),(16,2) and (8,2) neighbourhood

Model (GMM) super-vector [Reynolds et al., 2000a] which describes the distribution of Mel-Frequency Cepstrum Coefficients (MFCC) from each acoustic frame. This feature discards intra-class variations such as pitch and volume while keeping the frequency information which is intrinsic to the vocal tract of a subject. This thesis reports experiments with both speaker and face recognition using GMM super-vector modelling and LBP respectively.

2.3.1 Local binary patterns (LBP) for face representation

The Local Binary Pattern (LBP) operator was introduced by Ojiala et al [Ojala et al., 1996] as a method of texture analysis. The LBP feature extraction process is illustrated in

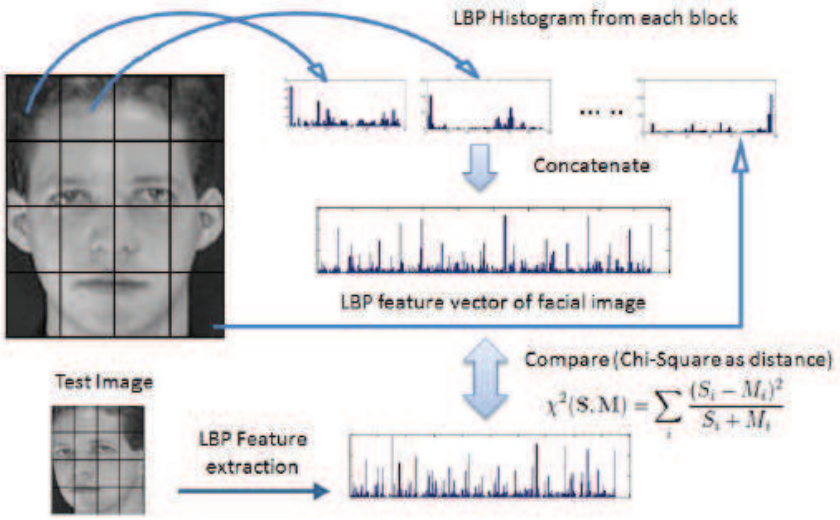


Figure 2.7: LBP face recognition

Figure2.6(a). For each pixel of a given image, the operator considers a 3×3 neighbourhood of pixels and compares their intensity to that in the center, before the difference is thresholded by 0 two form a binary code. Formally, the LBP operator takes the form:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{2.1}$$

where c is an index to the center pixel and n is an index to one of the 8 neighbouring pixels, and $i_{c/n}$ is their corresponding intensity. $s(u)$ is 1 if $u \geq 0$ and 0 otherwise. The result $LBP(x_c, y_c)$ is considered as an 8-bit binary number and is assigned to the center pixel. As a result, each pixel of the image has an LBP value between 0 and 255. Subsequently, a 256-bin histogram of LBP values is calculated and used as a feature vector representing the entire image. The LBP concept was later extended in two ways [Ojala et al., 2002]. First, in order to deal with texture at different scales, the LBP operator was extended to use neighbourhoods of different sizes. The local neighbourhood is defined as a set of sampling points evenly spaced on a circle, and binary interpolation is applied when the sample point does not fall in the center of a pixel. The notation (P, R) implies P sampling points on a circle of radius R. See Figure 2.6(b) for an example. The second extension defined the so-called *uniform patterns*: an LBP is "uniform" if it contains at most one 0-to-1 and one 1-to-0 transition when viewed as a circular bit string. For example, the LBP code in

Figure 2.6(a) is uniform. It is noticed that only 57 of the 256 8-bit patterns are uniform, but they typically account for 90% of all patterns [Ojala et al., 2002]. In the computation of LBP histograms, uniform patterns are used so that the histogram has a separate bin for every uniform pattern and all non-uniform patterns are assigned to a single bin. In this way, the number of bins are significantly reduced without losing too much information.

The application of LBP in face recognition was first introduced by Ahonen et al. in [Ahonen et al., 2006]. The LBP face recognition process is illustrated in Figure 2.7. The facial image is divided into local regions and texture descriptors are extracted from each region independently. The descriptors are then concatenated into a single long vector to form a global description of the face. In the recognition step, the LBP feature vector of the test image is extracted and compared to the LBP features of training images, and a Chi-square distance metric is often applied. Distance between two vectors x and σ is defined as:

$$\chi^2(x, \sigma) = \sum_i \frac{(x_i - \sigma_i)^2}{x_i + \sigma_i} \tag{2.2}$$

.

Suppose the face image is divided into M blocks and the LBP pattern has N bins for each block. The LBP face descriptor is an $M \times N$ dimensional feature vector. In case of using a 8×8 face division and regular 256-bin LBP code, the dimensionality of the LBP face descriptor mounts to 16384 dimensions. To enhance the discriminative capability, some variants of LBP add more patterns or information into the basic LBP face descriptor, which results in even higher dimensional feature vectors [Jin et al., 2004, Chan et al., 2007, Zhang et al., 2005]. For example, Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [Zhang et al., 2005] extracts LBP descriptors from 40 Gabor Magnitude Pictures of the original face image and concatenates them, resulting a feature dimensionality 40 times larger than the original LBP descriptor. This high-dimensionality can cause significant computational burden in the test phase.

2.3.2 Gaussian mixture models (GMM) for voice representation

Gaussian mixture model with universal backgrounds (GMM-UBM) approach is standard in speaker recognition [Reynolds et al., 2000a]. It was first used as a generative probabilistic model, and was then re-interpreted using a single vector, a so-called super-vector, as a more efficient way to represent speech utterances [Campbell et al., 2006]. These super-vectors also form a basis for more recent progresses in feature extraction for speaker recognition

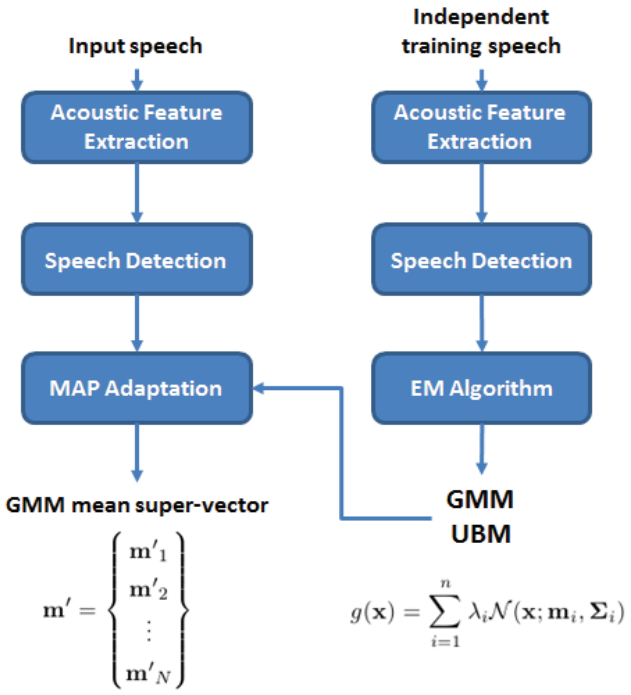


Figure 2.8: The generation of a GMM mean super-vector. A GMM for a speech utterance is obtained by MAP adapting only the component means of a UBM. The means of the N Gaussian mixtures are then concatenated into a long GMM mean super-vecotr.

such as joint factor analysis (JFA) [Kenny et al., 2007] and i-vectors [Dehak et al., 2011]. We review the GMM super-vector feature extraction process as follows.

As shown in Figure 2.8, the generation of a GMM super-vector from a given speech utterance can be divided into two phases, a Universal Background Model (UBM) training phase and a speaker model adaptation phase. The UBM training phase involves the training of a Universal Background Model (UBM) from a large collection of speech data independent from those used in the recognition step. Several processing steps occur in this UBM training phase. First, the speech is segmented into small acoustic frames (by a 20-ms window, for example), and acoustic features such as Mel-Frequency Cepstrum Coefficients (MFCC) are extracted from each frame. Then a voice activity detector (VAD) is applied to discard non-speech frames. Many existing algorithms for VAD use features that depend on energy [Evangelopoulos and Maragos, 2005], zero-crossing rate [Kotnik et al., 2001] or more sophisticated statistical machine learning approaches [Shin et al., 2010]. Remaining speech frames are used to train an N-component GMM with an Expectation-Maximization (EM)

algorithm to represent the speaker-independent distribution of features. The GMM-UBM model is represented mathematically as:

$$g(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \mathcal{N}(\mathbf{x}; \mathbf{m}_i, \mathbf{\Sigma}_i) \tag{2.3}$$

where λ_i is the mixture weights for the i-th component, $\mathcal{N}()$ is a Gaussian distribution, and $\mathbf{m}_i$ and $\mathbf{\Sigma}_i$ are the mean and covariance of each of the i-th Gaussian, respectively. Diagonal covariance matrices are normally used.

In the speaker modal adaptation phase, the same acoustic feature extraction and VAD processes are performed on the training data. The GMM model for the target speaker is obtained by adapting the UBM to the acoustic features of the utterance by maximum a posteriori (MAP) adaptation [Reynolds et al., 2000a]. Adapting the means only has been found to work well in practice [Reynolds et al., 2000b]. Assuming that $\{\mathbf{m'}_1, \ldots, \mathbf{m'}_N\}$ are the means of the new GMM model after adaptation, a so-called "super-vector" is formed by concatenating them into a high dimensional vector,

$$\mathbf{m}' = \begin{Bmatrix} \mathbf{m'}_1 \\ \mathbf{m'}_2 \\ \vdots \\ \mathbf{m'}_N \end{Bmatrix} \tag{2.4}$$

to represent the speech utterance. These super-vectors can be used as inputs to classifiers such as support vector machine (SVM) [Campbell et al., 2006].

GMM super-vectors are generally high-dimensional. Suppose that the acoustic feature extracted from each speech frame are M-dimensional (normally ranges from 30 to 60 component) and the GMM UBM model has N Gaussian components (normally ranges from 64 to 2048), the length of the resulting GMM super-vector is $M \times N$, which could easily reach several thousand dimensions.

2.4 Curse of Dimensionality and Dimensionality Reduction

As discussed in Section 2.3, biometric samples are often represented by high-dimensional feature vectors. high-dimensionality of can incur many problems in the context of statistical pattern recognition as discussed in Section 2.2, where it is commonly refereed to as *curse of dimensionality*. In this section, we will discuss this problem and its potential solution–

dimensionality reduction.

2.4.1 Curse of dimensionality

The term "curse of dimensionality" was coined by Richard Bellman [Bellman, 1961] to describe the problem caused by the exponential in volume associated with adding extra dimensions to a mathematical space. When referring to the biometric functionalities discussed in Section 2.2, namely verification, identification, clustering and retrieval, the curse of dimensionality causes problems in at least three aspects: computational complexity, lost of discriminative power in distance function and over-fitting.

Computational complexity: The computational complexity (running time) of an algorithm typically grows as some function of data dimensionality d. For example, the complexity of typical classifiers such as nearest neighbour and SVM grows linearly with d, whereas for some algorithms involving the computation of covariance matrices, the complexity grows linearly with d^2. Increase in computation time reduces the system usability in a biometric recognition system and makes clustering and retrieval systems particularly less efficient.

Lost of discriminative power in distance function: According to the analysis in [Beyer et al., 1999], when a measure such as an Euclidean distance is defined in a high-dimensional space, there is little difference in the distances between different pairs of samples. Given a single distribution, the minimum and the maximum occurring distances become indiscernible as the difference between the minimum and maximum value compared to the minimum value converges to 0:

$$\lim_{d\to\infty} \frac{\text{dist}_{max} - \text{dist}_{min}}{\text{dist}_{min}} \to 0 \tag{2.5}$$

where dist_{min} and dist_{max} are the minimum and maximum value of a pair of data samples given a certain distribution. In a biometric verification system, client scores are often associated with the distances between pairs of same-subject samples while imposter scores are associated with the distances between pairs of different-subject samples. In a high-dimensional space, those scores tend to overlap. This lose of discriminative power of distance measures in high-dimensional spaces also makes distance-based method such as nearest-neighbour unreliable.

Over-fitting : In machine learning, overfitting can occur when estimating a model, a complex model is learned with insufficient data. A model which has been overfitted will generally have poor predictive performance, as it can exaggerate minor fluctuations in the

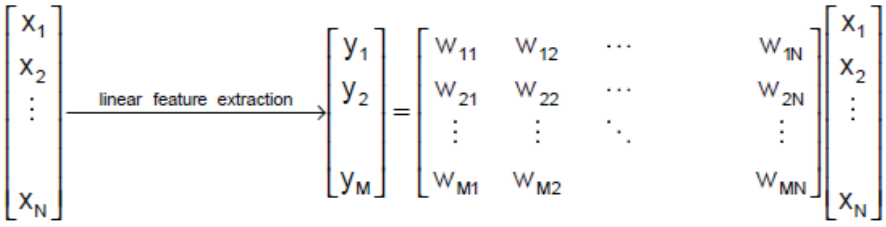


Figure 2.9: An illustration of linear dimensionality reduction.

data. As the dimensionality of the data increases, more training data is required to avoid overfitting. The collection of sufficiently large quantities of data can be difficult in many biometric settings.

2.4.2 Dimensionality Reduction

Dimensionality reduction is typically used to overcome this the curse-of-dimensionality by reducing the number of features in the data set. Dimensionality reduction can transform high dimensional data set into a lower dimensional space, while retaining most of the useful information in the original data. It is commonly assumed that underlying that the high-dimensional data points do not lie randomly in the original feature space, but there is a certain structure in the locations of the data points that can be exploited, and the useful information in high dimensional data can be summarized by a small number of features.

The problem of feature extraction can be stated as: given a feature space $\mathbf{x}_i \in \mathcal{R}^N$ find a mapping $\mathbf{y} = f(\mathbf{x}) : \mathcal{R}^N \rightarrow \mathcal{R}^M$ with $M < N$ such that the transformed feature vector $\mathbf{y}_i \in \mathcal{R}^M$ preserves (most of) the information or structure in $\mathcal{R}^N$. In general, the optimal mapping $\mathbf{y} = f(\mathbf{x})$ will be a non-linear function. However, there is no systematic way to generate non-linear transforms. The selection of a particular subset of transforms is problem dependent. For this reason, feature extraction is commonly limited to linear transforms: $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{y}$ is a linear projection of $\mathbf{x}$, which is illustrated in Figure 2.9.

The selection of the feature transformation matrix $\mathbf{W}$ is guided by an objective function which we seek to maximize (or minimize). Depending on the criteria measured by the objective function, dimensionality reduction techniques are grouped into two categories:

- **Signal representation:** In this case, the goal of the feature extraction mapping is simply to represent the samples in a lower-dimensional space. Since there is no concept of "classes" in the objective function, no class labels are required during

the training process. This category of techniques is referred to as **unsupervised dimensionality reduction**.

- **Classification:** Here the aim is to enhance the class-discriminatory information in the lower-dimensional space, class labels are thus required along with training samples. This category of techniques is also referred to as **supervised dimensionality reduction**.

Within the realm of linear feature extraction, two techniques are commonly used: unsupervised dimensionality reduction via Principle Component Analysis (PCA) and supervised dimensionality reduction via Linear Discriminant Analysis (LDA). We briefly review the principles of PCA and LDA in the following.

PCA: Consider a set of N centred samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ located in an n-dimensional feature space, and also consider a linear transformation which maps the original n-dimensional image space into an m-dimensional subspace, where $m < n$. The new feature vectors $\mathbf{y}_k \in \Re^m$ are defined by the following linear transformation:

$$\mathbf{y}_k = W^T(\mathbf{x}_k - \mu) \; k = 1, \ldots, N \tag{2.6}$$

where $\mu \in \Re^n$ is the mean of all samples and, where $W \in \Re^{n \times m}$ is a matrix with orthonormal columns. If the total scatter matrix S_t is defined as:

$$S_T = \sum_{k=1}^{N} (\mathbf{x}_k - \mu)(\mathbf{x}_k - \mu)^T, \tag{2.7}$$

after applying the linear transformation W^T, the scatter of the transformed feature vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$ is $W^T S_T W$. In PCA, the projection W_{opt} is chosen to maximize the the total scatter matrix of the projected samples:

$$W_{opt} = \arg\max W^T S_T W = [w_1 \, w_2 \, \ldots \, w_m] \tag{2.8}$$

where $\{w_i | i = 1, \ldots, w_m\}$ is the set of n-dimensional eigenvecotrs of S_T corresponding to its m largest eigenvalues.

LDA: Linear Discriminant Analysis is a well-known simple and efficient approach to dimensionality reduction, and is widely used in various classification problems. It aims to find an optimised projection $\boldsymbol{W_{opt}}$ which projects t dimensional data vectors $\boldsymbol{x}$ into a g dimensional space by $\boldsymbol{y} = \boldsymbol{W_{opt}} \boldsymbol{x}$, in which intra-class scatter (S_W) is minimized while the

intra-class scatter (S_B) is maximized. S_W and S_B are determined according to:

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{l_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T, \tag{2.9}$$

and

$$S_B = \sum_{j=1}^{c} l_j (\mu_j - \mu)(\mu_j - \mu)^T, \tag{2.10}$$

where x_i^j is the i_{th} sample of of class j, μ_j is the mean of class j, c is the number of classes, and l_j is the number of samples in class j. $\boldsymbol{W_{opt}}$ is obtained according to the objective function:

$$\boldsymbol{W_{opt}} = \arg\max_W \frac{W^T S_B W}{W^T S_W W} = [w_1, \ldots, w_g] \tag{2.11}$$

where $\{w_i | i = 1, \ldots, g\}$ are the eigenvectors of S_B and S_W which correspond to the g largest generalized eigenvalues according to:

$$S_B w_i = \lambda_i S_W w_i, i = 1, \ldots, g \tag{2.12}$$

Note that there are at most $c-1$ non-zero generalized eigenvalues, so g is upper-bounded by $c-1$. Since $S_W$is often singular, it is common to first apply Principal Component Analysis (PCA) to reduce the dimension of the original vector. LDA has been applied to AFR and ASR and is often referred to as *Fisherface* [Belhumeur et al., 1997] and *Fishervoice* [Li et al., 2010].

While LDA can extract discriminant information from high dimensional feature vectors when labelled training data is abundant, but when training data is scarce, the projections can be significantly biased, which generally leads to reduced performance.

Figure 2.10 illustrates a comparison of PCA and LDA in a two-class problem in which the samples from each class are sampled from a two-dimensional multi-variate Gaussian distribution. Both PCA and LDA have been used to project the points in 2D down to 1D. Comparing the two projections in the figure, PCA actually smears the classes together so that they are no longer linearly separable in the projected space. It is clear that, although PCA achieves larger total scatter, LDA achieves greater between-class scatter, and, consequently better classification potential.

To conclude, both unsupervised dimensionality reduction and supervised dimensionality

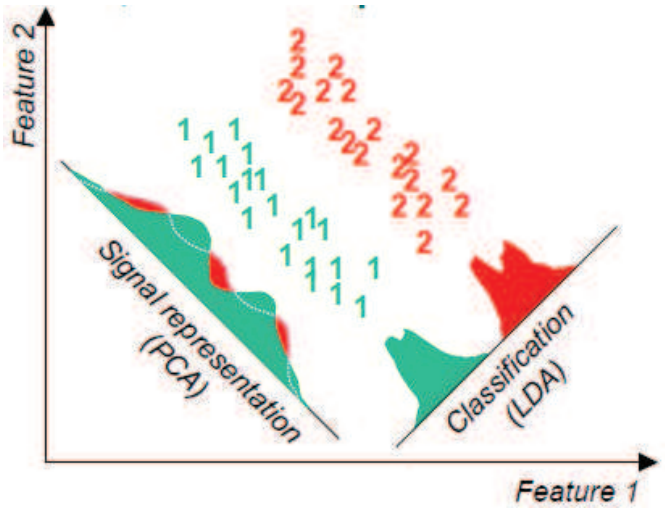


Figure 2.10: An illustration of linear dimensionality reduction.

reduction aim to reduce the volume of the data while keeping the most *useful* information. There difference, however, lies in the definition of *useful* information. With an unsupervised approach, useful information refers to the necessary information to reconstruct data; while for supervised approach, useful information refers to the information needed to discriminate different classes.

As discussed in Section 2.2, biometrics is itself a classification problem. In this sense, supervised dimensionality reduction methods are intrinsically preferred over unsupervised methods since they extract class-specific information necessary for classification. However, supervised feature extraction methods require large quantity of manually labelled data. In practical biometric applications, however, labelled data are often scarce and unlabelled data can often be easily acquired . For example, in verification and identification problems, labelled data typically acquired during an enrolment session, is limited in quantity and may not be representative of the general distribution of the data. Unlabelled data, on the other hand, could be easily obtained during a period of normal system use and be more representative of the intra-session variation. When trained with insufficient labelled data, supervised feature extraction techniques such as LDA tend to overfit and perform even worse than unsupervised methods such as PCA. This phenomenon has been reported in literature [Martinez and Kak, 2001, Delac et al., 2005], and will be further analysed in Chapter 4. In clustering and retrieval problems, the class labels are completely absent, supervised feature extraction methods cannot be applied. Discriminant feature extraction is however still important for clustering and retrieval performance, where same-class sample

pairs are considered more "similar" than different-class samples in the reduced feature space.

The analysis above illustrates the first problem tackled in this thesis: *How to extract discriminative features using unlabelled data?*

2.5 Multiple Representation of Biometric Data

As discussed in Section 2.3, biometric data is often represented by a single high-dimensional feature vector. In this section, we discuss multi-modal biometrics, where a biometric sample can be represented by multiple feature vectors.

2.5.1 Multi-modal biometrics

Most biometric systems typically use only one single biometric trait to establish identity. Such systems are referred to as uni-modal biometric system. However, even the most researched biometric modalities to date are facing numerous performance issues, some of them inherent to the technology itself. In some circumstances a single biometric is not sufficient to meet the variety of requirements including matching performance, robustness to spoofing attacks, etc.. For a given application, multi-modal biometric systems [Jain and Ross, 2004] address some of the drawbacks of the uni-modal biometric systems by using *multiple sources of information*. These systems can improve the recognition performance of a biometric system by increasing robustness to spoof attacks, and reducing the failure-to-enroll rate due to the presence of multiple sources of information.

In a strict sense, a multi-modal biometric system establish identity based on the acquisition of multiple biometric traits. For example, some of the multi-modal biometric systems utilized face and voice to recognize identity of an individual, which is commonly referred to as audio-visual speaker recognition [Chibelushi et al., 1997, Nefian et al., 2003]. Physically uncorrelated traits (e.g., fingerprint and iris) are expected to result in better performance than correlated traits (e.g., voice and lip movement).

In a broader sense, multi-modal biometric system can refer to any biometric system which utilize different sources of information [Ross et al., 2008], which are not necessarily different biometric traits. The multiple information sources can also refers to multiple sensors and multiple features.

Multi-sensor systems employ multiple sensors to capture a single biometric trait of an individual. For example, a face recognition system may deploy multiple 2D cameras to acquire the face image of a subject [Lee et al., 2004]; an infrared sensor may be used in

conjunction with a visible-light sensor to acquire the subsurface information of a person's face [Chen et al., 2005]; or an optical as well as a capacitive sensor may be used to image the fingerprint of a subject [Marcialis and Roli, 2004].

Even only using a single biometric sensor, in the feature extraction module, different features can be extracted from the same biometric raw data. These features often have different properties and are robust to different type of intra-class variations. A Combination of these features may result in improved matching performance. For example Lu et al. [Lu et al., 2003] discuss a face recognition system that combines three different feature extraction schemes (Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA)).

2.5.2 Fusion

In multi-modal biometrics, fusion techniques are intensely studied to combine different sources of information to improve the performance of uni-modal systems. Fusion can be integrated in several different levels in a multi-modal biometric system:

1. **Sensor level**: The raw data acquired from multiple sensors can be processed and integrated to generate new data from which features can be extracted. For example, in the case of fingerprint biometrics, the fingerprint image acquired from both optical and solid state sensors may be fused to generate a single image which could then be subjected to feature extraction and matching.

2. **Feature level**: Information extracted from the different sources is concatenated into a joint feature vector, which is then compared to an enrolment template (which itself is a joint feature vector stored in a database) and assigned a matching score as in a single biometric system.

3. **Score level**: Feature vectors are created independently for each modality and are then compared to the enrolment templates which are stored separately for each biometric trait. Based on the proximity of feature vector and template, each subsystem computes its own matching score. These individual scores are finally combined into a total score, which is passed to the decision module.

4. **Decision level**: A separate authentication decision is made for each biometric trait. These decisions are then combined into a final vote. Fusion at the decision level is considered to be rigid due to the availability of limited information.

No matter in which level the fusion is performed, the objective of fusion is to add information from multiple sources together. As far as our knowledge, no fusion process can remove harmful information (intra-class variance) from the features.

2.6 Problem Statement: Discriminative Feature Extraction from Unlabelled, Multi-view Data

As shown in Section 2.4.2, supervised dimensionality reduction methods learn a projection $\mathbf{P}$ with a feature-label pairs $< \mathbf{X}, \mathbf{Y} >$ where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a set of feature vectors and $\mathbf{Y} = \{y_1, \ldots, y_n\}$ is a set of corresponding class labels for each vector. They have high discriminative power but Y is obtained through an expensive manual labelling. Unsupervised dimensionality reduction methods, on they other hand, only need the feature set $\mathbf{X}$ and can thus exploit the large amount of unlabelled data which could be easily acquired, but have relatively low discriminative power. In a multi-modal biometric system which make use of two biometric traits, data samples are represented by feature-feature pairs $< \mathbf{X}^{(1)}, \mathbf{X}^{(2)} >$. Compared to supervised methods, the class labels $\mathbf{Y}$ acquired by human manual labelling are replaced by another naturally co-existed feature.

This thesis studies the problem of extracting lower-dimensional class-discriminative features from paired high-dimensional biometric features, while minimizing the effort of expensive manual labelling. We refer to the process of extraction discriminative low-dimensional features from unlabelled feature pairs as ***Multi-view Dimensionality Reduction (MVDR)****.*

Depending on different biometric applications, training data can be *partially labelled* or *completely unlabelled*. In identification and verification applications, training data involves a small set of labelled training data $\mathbf{L} =< \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y} >$ acquired during a limited number of enrolment sessions, and a larger set of unlabelled training data $\mathbf{U} =< \mathbf{X}^{(1)}, \mathbf{X}^{(2)} >$ obtained during a period of normal system use. We refer to the MVDR in this scenario as semi-supervised multi-view dimensionality reduction (SSMVDR). In retrieval and clustering applications, the training set only involves unlabelled data $\mathbf{X} =< \mathbf{X}^{(1)}, \mathbf{X}^{(2)} >$. We refer to the MVDR in this setting as unsupervised multi-view dimensionality reduction. Later in the thesis, we propose adapted solutions.

2.7 Summary

This chapter provideds an introduction to biometrics together with some challenges and limitations. We have the following conclusions:

1. Biometrics refers to the recognition of individuals according to their physical or behavioural traits. Most biometric systems share two common operation modules, feature extraction and comparison.

2. Biometric data is often represented by high-dimensional feature vectors, and contains significant amount of intra-class variations. This leads to the curse of dimensionality in the following comparison module.

3. Dimensionality reduction techniques deal with this problem. Supervised dimensionality reduction methods has high discriminative power, but need expensive manual labelling. Unsupervised methods can make use of unlabelled data which is easy to acquire, but commonly lack discriminative power.

4. In multi-modal biometrics, a data sample can be represented by different features. Typical multi-modal fusion process add information from multiple sources together, but are not capable of removing irrelevant components within features themselves.

5. In multi-modal biometrics, given feature-feature pairs without class label, we aim to learn discriminative low-dimensional representations for each feature, as if we have feature-label pairs in a supervised learning setting.

Chapter 3

State-of-the-art in MVDR

As discussed in Chapter 2, multi-modal biometric system represents a single biometric sample by multiple feature vectors, where each has the potential to contain different intra-class variation. The extraction of discriminative, lower-dimensional features in such a multi-view scenario is referred to as multi-view dimensionality reduction (MVDR). In this chapter, we first review the existing state-of-the-art in MVDR methods, analyse their relative advantages and disadvantages, and then propose our own MVDR framework.

3.1 Multi-View Dimensionality Reduction

In real-world practical problems, a single object may be readily represented by two or more types of distinct features, e.g.: gene can be represented by the genetic activity feature and text information feature; people have both facial and vocal features; webpages can be represented by the text in the page and hyper-links to the page. This kind of data is usually called multi-modal or multi-view data. Analysing such multi-view data to acquire useful information and knowledge has attracted more and more research attention over recent years. Related research includes dimensionality reduction, regression and clustering. In this chapter, we focus on MVDR problems with the aim to avoid the curse of dimensionality [Bellman, 1961] and overfitting brought by high dimensionality.

Depending on whether label information is needed, existing MVDR approaches can be divided into supervised methods [Diethe et al., 2008], unsupervised methods [Foster et al., 2008, Lai and Fyfe, 2000, Long et al., 2008, Han et al., 2012], and semi-supervised

methods [Hou et al., 2010, Blaschko et al., 2011]. In supervised MVDR methods, training samples are represented by multiple features and class labels are also available. Unsupervised MVDR methods work with only features in the absence of labels. Semi-supervised methods assume that the training samples are largely unlabelled, while label information is available for a small portion of the training data, or some link information is available (e.g. some pairs of data samples are known to belong to the same class, while some other pairs are known to be in different classes). Most research efforts in MVDR are devoted to unsupervised and semi-supervised scenario, since supervised methods which learn from feature-label pairs $< X, Y >$ is already a type of multi-view learning, in the sense that the label Y can be regarded as another view. This label information is so strong that the improvement by adding another view can be insignificant. This thesis deals with the problem of extracting discriminative information for biometric samples when labelled data is scarce; we concentrate only on unsupervised and semi-supervised MVDR problems.

According to our knowledge, most existing MVDR approaches can be loosely divided into two categories. The first category is based on Canonical Correlation Analysis (CCA) and its variants. The second category is based on multi-view graph embedding. Due to their distinct nature and relative advantages, we review each of them separately.

3.2 MVDR Based on Canonical Correlation Analysis

Proposed by H. Hotelling in 1936 [Hotelling, 1936], CCA is a two-view dimesnionality reduction method which is able to find basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximised. It is commonly used as a multi-view dimensionality reduction approach for multi-view learning tasks such as classification, regression, clustering and retrieval.

3.2.1 Principles of CCA

Formally, consider a set of n samples represented in two views: $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$, where $\boldsymbol{X}^{(v)} = \{\boldsymbol{x}_1^{(v)}, \ldots, \boldsymbol{x}_n^{(v)}\}$, $v = 1, 2$. $\boldsymbol{X}^{(v)}$ is first centred so that $\bar{\boldsymbol{X}}^{(v)} = \sum_i \boldsymbol{x}_i^{(v)}/n = 0$. CCA computes two projection matrix $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ such that the correlation coefficient in the projected subspaces is maximized. The objective function is formulated as:

$$\underset{\mathbf{P}^{(1)},\mathbf{P}^{(2)}}{\arg\max} \frac{\mathbf{P}^{(1)T}\mathbf{X}^{(1)}\mathbf{X}^{(2)T}\mathbf{P}^{(2)}}{\sqrt{\mathbf{P}^{(1)T}\mathbf{X}^{(1)}\mathbf{X}^{(1)T}\mathbf{P}^{(1)}}\sqrt{\mathbf{P}^{(2)T}\mathbf{X}^{(2)}\mathbf{X}^{(2)T}\mathbf{P}^{(2)}}} \tag{3.1}$$

If we denote $\mathbf{C}_{11} = \mathbf{X}^{(1)}\mathbf{X}^{(1)T}$ and $\mathbf{C}_{22} = \mathbf{X}^{(2)}\mathbf{X}^{(2)T}$ as the auto-covariance matrix of the two views and $\mathbf{C}_{12} = \mathbf{C}_{21}^T = \mathbf{X}^{(1)}\mathbf{X}^{(2)T}$ as the correlation matrix between the two views, then Equation 3.1 can be rewritten as:

$$\underset{\mathbf{P}^{(1)},\mathbf{P}^{(2)}}{\arg\max} \frac{\mathbf{P}^{(1)T}\mathbf{C}_{12}\mathbf{P}^{(2)}}{\sqrt{\mathbf{P}^{(1)T}\mathbf{C}_{11}\mathbf{P}^{(1)}}\sqrt{\mathbf{P}^{(2)T}\mathbf{C}_{22}\mathbf{P}^{(2)}}} \tag{3.2}$$

Since the correlation coefficient is invariant to the scale of $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$, CCA can be formulated equivalently as:

$$\begin{aligned} \arg\max_{\mathbf{P}^{(1)},\mathbf{P}^{(2)}} \quad & \mathbf{P}^{(1)T}\mathbf{C}_{12}\mathbf{P}^{(2)} \\ \text{subject to} \quad & \mathbf{P}^{(1)T}\mathbf{C}_{11}\mathbf{P}^{(1)} = 1 \\ & \mathbf{P}^{(2)T}\mathbf{C}_{22}\mathbf{P}^{(2)} = 1 \end{aligned} \tag{3.3}$$

It can be shown that $\mathbf{P}^{(1)} = \{\mathbf{p}_1^{(1)}, \ldots, \mathbf{p}_m^{(1)}\}$, where $\mathbf{p}_i^{(1)}$ is the i-th generalized eigenvector of the generalized eigenvalue problem:

$$C_{12}C_{22}^{-1}C_{21}\mathbf{p} = \lambda^2 C_{11}\mathbf{p} \tag{3.4}$$

Correspondingly, $\mathbf{P}^{(2)} = \{\mathbf{p}_1^{(2)}, \ldots, \mathbf{p}_m^{(2)}\}$ where

$$\mathbf{p}_i^{(2)} = \frac{C_{22}^{-1}C_{21}\mathbf{p}_i^{(1)}}{\lambda_i} \tag{3.5}$$

where λ_i is the i-th generalized eigenvalue in Equation 3.4.

3.2.2 Application of CCA

CCA has long been used as a tool to discover shared information between multiple information sources, but its application in feature extraction only became popular after 2000, when the demand to process multi-view, high-dimensional data begin to increase dramatically on account of the explosion of internet information. CCA and its kernelized version KCCA [Lai and Fyfe, 2000] has been used in many multi-view learning problems such as image retrieval [Hardoon et al., 2004] or clustering [Blaschko and Lampert, 2008] from image and associated text, speaker identification using audio and visual information [Sargm et al., 2006], and document retrieval cross different languages [Li and Shawe-Taylor, 2006]. People generally found that given a two-view representation of the same data, performing pattern recognition tasks such as classification, retrieval or clustering in a CCA or KCCA

space generally lead to improved performance over that obtained in PCA or kernal PCA (KPCA) subspaces of a single view. Foster et al. [Foster et al., 2008] gave a theoretical treatment of the application of CCA in multi-view pattern recognition problems, showing that, given a conditional independence assumption and a redundancy assumption, CCA can significantly reduce the number of labelled samples needed in a regression problem. Chaudhuri et al. [Chaudhuri et al., 2009] showed that, given the independence and redundancy assumption, the most class-specific discriminative information resided in the first $c-1$ directions of the CCA subspaces, where c is the number of underlying classes.

3.3 MVDR Based on Similarity Graphs

Just as CCA can be regarded as a multi-view extension of the linear dimensionality reduction method PCA, another important class of MVDR approaches involves MVDR based on similarity graphs which are multi-view extensions of non-linear dimensionality reduction techniques.

3.3.1 Multi-view spectral embedding

We first review some principles of non-linear dimensionality reduction. Linear dimensionality reduction assumes that the intrinsic data representation lies in a linear, lower-dimensional subspace of the original feature space $\mathbf{X}$ and thus this lower-dimensional representation can be recovered by a linear projection $\mathbf{P}^T\mathbf{X}$. This assumption of linearity, however, does not always hold especially, when the numbers samples is much higher than the feature dimensionality and the classes are not linearly separable. For example, in the left picture of Figure 3.1, n data samples are located in two half-moon shapes in a 2-D feature space. In this case, linear dimensionality reduction techniques will not be effective since any linear projection of it into a lower dimensional space will lead to the loss of class-discriminative information. In contrast, non-linear dimensionality reduction techniques, can readily cope with this problem. Typical non-linear dimensionality reduction methods include Isomap [Tenenbaum et al., 2000], Locally Linear Embedding (LLE) [Roweis and Saul, 2000], Laplacian Eigenmaps [Belkin and Niyogi, 2001] and Kernel PCA (KPCA) [Scholkopf et al., 1999]. All these methods begin with the construction of a $n \times n$ similarity graph $\mathbf{S}$, where each component s_{ij} is a similarity measure between the i-th and the j-th sample. Subsequently, a certain transformation is found to embed the $n \times n$ similarity graph into a $n \times t$ $(t < n)$ matrix $\mathbf{U}$. Each row of $\mathbf{U}$ is regarded as the final embedding for a data

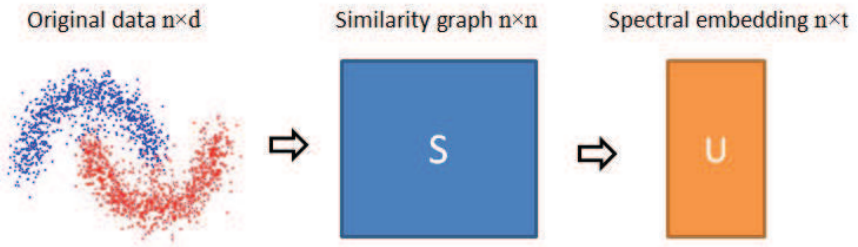


Figure 3.1: Illustration of non-linear dimensionality reduction

sample.

Different non-linear dimensionality reduction techniques have different objective functions and use different ways to find the transformation from **S** to **U**. For example, kernel principle component analysis (KPCA) embeds data samples into the first k eigenvectors of similarity graph S (it is referred to as a kernel in this case), whereas spectral clustering first gets a normalized similarity graph according to $L = D^{-1/2}SD^{-1/2}$, where D is a diagonal matrix with diagonal elements $D_{ii} = \sum_j S_{ij}$, before embed data samples into the first k eigenvectors of L.

Recently, efforts have been made to extend non-linear dimensionality reduction methods into a multi-view setting. One of the first works in this domain involves Distributed Spectral Embedding (DSE) [Long et al., 2008]. Given a multi-view dataset with n objects having m views, i.e., a set of matrices $X = \{X^{(i)} \in \Re^{m_i \times n}\}_{i=1}^m$, each representation $X^{(i)}$ is a feature matrix for view i. DSE assumes that the low-dimensional embedding of each view $X^{(i)}$ is already known, $U = \{A^{(i)} \in \Re^{n \times k_i}\}_{i=1}^m$. DSE focuses on how to learn a consensus, low-dimensional embedding $V \in \Re^{n \times k}$ based on U. The objective function of DSE is defined as:

$$\arg\min_{V,P} \sum_{i=1}^{m} ||U^{(i)} - VP^{(i)}||^2 \text{ s.t.} V^T V = I \tag{3.6}$$

where $P = \{P^{(i)} \in \Re^{k \times k_i}\}_{i=1}^m$ is a set of mapping matrices. This method aims to find a consensus pattern V which can optimally reconstruct spectral embeddings $U^{(i)}$ from each view. Some later works such as Multi-view Spectral Embedding (MVSE) [Xia et al., 2010] and Sparse Spectral Multi-view Embedding (SSMVE) [Han et al., 2012] follow similar principle but add a smoothness and sparsity constraints to the objective function.

Algorithm 1 Spectral Clustering algorithm according to [Ng et al., 2002]

Input: a set of points $\mathbf{X} = \{x_1, \ldots, x_n\} \in \Re^{n \times d}$ that we want to cluster in k groups.

- Construct a $n \times n$ positive similarity matrix (kernel) $\mathbf{S}$, where $\mathbf{S}_{ij}$ quantifies the similarity between sample i and sample j;
- Compute the normalized graph Laplacian $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$.
- Let $\mathbf{U}$ denote a $n \times k$ matrix with columns as the top k eigenvectors of $\mathbf{L}$;
- Normalize each row of $\mathbf{U}$ to obtain $\mathbf{V}$;
- Run the k-means algorithm to cluster the row vectors of $\mathbf{V}$;

Output: Assign example i to cluster c if the i-th row of $\mathbf{V}$ is assigned to cluster c by the k-means algorithm.

3.3.2 Multi-view Spectral clustering

Another separate line of work involves multi-view spectral clustering. Spectral clustering is a technique that exploits the properties of the Laplacian of the graph, whose edges denote the similarities between data points. The top k eigenvectors of the normalized graph Laplacian are relaxations of the indicator vectors that assign each node in the graph to one of the k clusters. Apart from being theoretically well-motivated, spectral clustering has the advantage of performing well on arbitrary shaped clusters, which is otherwise a shortcoming with several other clustering algorithms such as the k-means algorithm. The spectral clustering algorithm is briefly outlined in Algorithm 1. For a detailed introduction to both theoretical and practical aspects of spectral clustering, the reader is referred to [Von Luxburg, 2007]. Spectral clustering can be regarded as a dimensionality reduction technique which embeds data samples into the first k eigenvectors of the Laplacian matrix. In the embedded space, the data structure become clearer and the data non-linearity is attenuated.

Efforts in extending spectral clustering to a multi-view setting has been made in [Kumar and Daumé III, 2011] and [Kumar et al., 2011]. Both algorithms work with the assumption that the true underlying clustering would assign corresponding points in each view to the same cluster. Given this assumption, multi-view clustering problem is approached by limiting our search to clusterings that are compatible across the graphs defined over each of the views.

The co-training spectral clustering approach in [Kumar and Daumé III, 2011] used an iterative approach to use spectral embedding of one view to refine the similarity graph of the other. The proposed algorithm is summarized in Algorithm 2. By projecting the similarity graph $\mathbf{S}^{(1)}$ to directions indicated by the first k eigenvectors of the graph Laplacian $\mathbf{L}^{(2)}$

Algorithm 2 Co-training spectral clustering algorithm [Ng et al., 2002]

Input: Similarity matrix $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$.
Output: Assignments to k clusters.
Initialize: $\mathbf{L}^{(v)} = {\mathbf{D}^{(v)}}^{-\frac{1}{2}}\mathbf{S}^{(v)}{\mathbf{D}^{(v)}}^{-\frac{1}{2}}$ for $v = 1, 2$ solve $\mathbf{U}^{(v)}$ as the first k eigenvectors of $\mathbf{L}^{(v)}$
for $i = 1$ to *iter* **do**

- $\mathbf{K}^{(1)} = sym(\mathbf{U}^{(2)}{\mathbf{U}^{(2)}}^T\mathbf{S}^{(1)})$; $\mathbf{K}^{(2)} = sym(\mathbf{U}^{(1)}{\mathbf{U}^{(1)}}^T\mathbf{S}^{(2)})$;
- Use $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$ as new similarity graphs to compute $\mathbf{L}^{(1)}$ and $\mathbf{L}^{(2)}$, update $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ as the first k eigenvectors of the new $\mathbf{L}^{(1)}$ and $\mathbf{L}^{(2)}$ respectively.

end for

- Row-normalize $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ to obtain $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$.
- Run k-means clustering on the concatenation of $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$.

and projecting back, the modified similarity graph $\mathbf{K}^{(1)}$ tends to have a similar structure as $\mathbf{S}^{(1)}$. The result from the iterative training process will result in final similarity graphs $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ similar to each other.

The co-training spectral clustering method is heuristic and no theoretical proof of convergence is provided. The authors further proposed a co-regulation solution in [Kumar et al., 2011], which have a closed-form solution. Consider a set of multi-view data, and denote $\mathbf{S}^{(v)}$ as the similarity graph in the v-th view. The single view spectral clustering algorithm solves the following optimization problem for the graph Laplacian $\mathbf{L}^{(v)}$:

$$\underset{\mathbf{U}^{(v)} \in \Re^{n \times k}}{\arg\max}\ tr({\mathbf{U}^{(v)}}^T\mathbf{L}^{(v)}\mathbf{U}^{(v)}),\ \text{s.t. } {\mathbf{U}^{(v)}}^T\mathbf{U}^{(v)} = I; \tag{3.7}$$

where tr denotes the matrix trace. The rows of matrix $\mathbf{U}^{(v)}$ are the embeddings of the data points that can be used by the k-means algorithm to obtain cluster memberships. In the multi-view scenario, the pairwise similarities of examples under the new representation (in terms of rows of $\mathbf{U}^{(v)}$'s) is encouraged to be similar across all views. This amounts to enforcing the $\mathbf{U}^{(v)}$ to be the same across all views. A new objective function of multi-view co-regularized spectral clustering is defined as:

$$\begin{aligned}\arg\max_{U^{(1)},U^{(2)}}\ & tr({\mathbf{U}^{(1)}}^T\mathbf{L}^{(1)}\mathbf{U}^{(1)}) + tr({\mathbf{U}^{(2)}}^T\mathbf{L}^{(2)}\mathbf{U}^{(2)}) + \lambda tr(\mathbf{U}^{(1)}{\mathbf{U}^{(1)}}^T\mathbf{U}^{(2)}{\mathbf{U}^{(2)}}^T) \\ & \text{s.t. } {\mathbf{U}^{(1)}}^T\mathbf{U}^{(1)} = I,\ {\mathbf{U}^{(2)}}^T\mathbf{U}^{(2)} = I,\end{aligned} \tag{3.8}$$

In this objective function, the first and second terms are the objective function of the

standard spectral clustering in the first and second single view, and the third term is a regularization term which requires the embedded structure $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ to be similar. The hyperparameter λ trades-off the spectral clustering objectives and the spectral embedding agreement term. For a given $\mathbf{U}^{(1)}$, optimization problem in $\mathbf{U}^{(2)}$ is:

$$\underset{\mathbf{U}^{(2)} \in \Re^{n \times k}}{\arg\max} \; tr\{{\mathbf{U}^{(2)}}^T(\mathbf{L}^{(2)} - \lambda {\mathbf{U}^{(1)}}^T \mathbf{U}^{(1)})\mathbf{U}^{(2)}\} \tag{3.9}$$

This is a standard spectral clustering objective on view 2 with graph Laplacian $\mathbf{L}^{(2)} - \lambda {\mathbf{U}^{(1)}}^T \mathbf{U}^{(1)}$. The joint optimization problem given by Equation 3.8 can be solved using alternating maximization with respect to $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$. The main steps involve:

- Fixing $\mathbf{U}^{(1)}$ and solve for $\mathbf{U}^{(2)}$ as the first k eigenvectors of $\mathbf{L}^{(2)} - \lambda {\mathbf{U}^{(1)}}^T \mathbf{U}^{(1)}$.
- Fixing $\mathbf{U}^{(2)}$ and solve for $\mathbf{U}^{(1)}$ as the first k eigenvectors of $\mathbf{L}^{(2)} - \lambda {\mathbf{U}^{(1)}}^T \mathbf{U}^{(1)}$.

In each iteration, the objective function in Equation 3.8 monotonously decreases, which guarantees the convergence of the algorithm. Finally, $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ are concatenated and k-means clustering is applied on the resulting spectral embeddings. Compared to co-training spectral clustering algorithm, the co-regularization spectral clustering algorithm is theoretically justified and the convergence is guaranteed. The clustering performance of the two algorithm is comparables, but the co-regularization method requires careful tuning of the hyperparameters λ while the co-training method has no hyperparameter.

3.4 Proposed MVDR Approach

In previous sections we have surveyed some existing MVDR approaches in two major classes: CCA-based approaches and similarity graph based method. The CCA-based methods aimed to find subspace projections so that the projected samples in each view are maximumly correlated. The similarity graph based approaches first represent data samples in each view with a similarity graph, then extract shared patterns from multiple similarity graphs.

CCA-based approach has two major limitations. First, according to the analysis of [Chaudhuri et al., 2009], CCA learns a low dimensional subspace spanned by the means of different classes (equivalent to the maximization of between-class scattering). However, same cluster samples are not necessarily projected near to each other (minimization of with-in class scattering). Second, CCA-based methods rely strongly on the conditional independence assumption, which may not hold in practical problems. According the experimental work

of [Kumar and Daumé III, 2011] and [Kumar et al., 2011], CCA-based method performs poorly when there is some dependence between views.

The disadvantage associated with similarity graph based MVDR approaches is that, features $X^{(v)}$ are not used again after the $S^{(v)}$ is built. In the case that original features $X^{(v)}$ contain substantial number of noisy dimensions which are irrelevant to underlying classes, the estimation of $S^{(v)}$ is intrinsically inaccurate, thus improvements from graph fusion can be sub-optimal.

In contrast to such previous approaches, the new work reported in this thesis addresses the MVDR problem from a different angle. Inspired by the pioneering semi-supervised learning method *co-training* [Blum and Mitchell, 1998b], this thesis introduces a new concept of *subspace structure agreement*. The main idea involves learning projections $P^{(1)}$ and $P^{(2)}$, such that the data structure of projected samples $P^{(1)}X^{(1)}$ and $P^{(2)}X^{(2)}$ is as similar as possible. According to different definitions of *agreement* and different applications, namely semi-supervised classification, unsupervised clustering and retrieval, we propose three different MVDR framework, namely MVDR by incremental co-training, MVDR by subspace clustering agreement, and MVDR by subspace graph agreement.

MVDR by incremental co-training: This framework is designed for biometric recognition problems. We assume that a small quantity of labelled data is available during an enrolment session while a larger pool of unlabelled data can be acquired during a period of normal system use. Following a typical co-training procedure, Linear Discriminant Analysis (LDA) projections initially weakly learnt with the small set of labelled data are incrementally re-learnt with automatically labelled data from the unlabelled dataset. This algorithm is referred to as Co-LDA and is applied to the audio-visual person recognition problem.

MVDR by subspace clustering agreement: This framework is designed for clustering high-dimensional, multi-view data, e.g. facial-vocal biometric data in videos. The framework combines the simplicity of k-means clustering and LDA within a co-training scheme which exploits labels learned automatically in one view to learn discriminative subspaces in another, and this new algorithm is referred to as CoKmLDA. In essence, CoKmLDA algorithm learns a subspace for each view such that the clustering structure is in maximum agreement across each view. We also provide an extension of the two-view CoKmLDA to more than two views. The effectiveness of the proposed algorithm is demonstrated empirically with an audio-visual speaker clustering experiment. Significant improvements over alternative multi-view clustering approaches are reported. The CoKmLDA algorithm is also tested on other multi-view clustering problems such as text clustering and image clustering.

MVDR by subspace graph agreement: This framework is designed for biometric data retrieval problems. The similarity relationship between samples in a dataset can be represented by a similarity graph, thus this framework aims to learn a subspace projection for each view such that the difference between the similarity graphs built on the projected samples in each view is minimized. We have shown that this objective can be achieved by a graph-based co-training process of Locality Preserving Projections (LPP), and the new algorithm is referred to as Co-LPP. The effectiveness of the proposed algorithm is validated by audio-visual speaker retrieval experiment and a face retrieval experiment with two different facial features.

The three MVDR frameworks will be presented in chapter 4, chapter 5 and Chapter 6, respectively.

3.5 Summary

In this chapter, we reviewed literatures which is considered as state-of-the art in MVDR. These existing methods are divided into two categories, CCA-based methods and similarity-graph-based methods. We then proposed our MVDR based on subspace structure agreement.

Chapter 4

MVDR by Incremental Co-training

In this chapter, we present our first MVDR framework, which is adapted to semi-supervised biometric identification and verification systems, and audio-visual person recognition in particular. In this framework, we assume that a small quantity of labelled data is available during the enrolment session while a larger pool of unlabelled data can be acquired during a period of normal system use. Following a typical co-training [Blum and Mitchell, 1998a] procedure, LDA projections initially weakly learnt with the small set of labelled data are incrementally re-learnt with automatically labelled data from the unlabelled dataset. This co-training style, semi-supervised MVDR framework is referred to as MVDR by incremental co-training.

4.1 Motivations

Biometric systems exploit physiological and/or behavioural traits to recognize individuals. Popular traits or modalities include fingerprints, hand-geometry, face, voice, iris, retina, gait, signature, palm-print, ear, etc. Among them, face and voice features have the advantages of non-intrusiveness, easy acquisition and also the possibility of non-cooperative acquisition. Automatic Speaker Recognition (ASR) and Automatic Face Recognition (AFR) have thus attracted a high degree of research interest in the last decade.

ASR and AFR systems generally share the same operational paradigm. During enrolment, training data is collected and client models are learnt or adapted, while under normal use or testing new samples are compared to a single model (verification) or to a group of

models (identification). Under well controlled conditions performance is typically acceptable. In real operational scenarios, however, test data can exhibit substantial differences to that collected during enrolment. In the case of face recognition, so-called inter-session variability may come from differences in illumination or pose, the presence of facial accessories (glasses or piercings), and ageing over an extended time period. Voice features may vary as a consequence of environmental noise or changes to the vocal tract as a consequence of illness or ageing. Unless such variations are captured and represented in the client models, or unless suitably robust features or normalization approaches are applied, recognition performance can deteriorate drastically.

The use of more robust features can ameliorate this problem to some extent. In AFR, for example, Local Binary Pattern (LBP) features [Ahonen et al., 2006] are among the most robust to illumination changes, while SIFT-like features [Bicego et al., 2006] are robust to geometrical transformations. To date, however, there are no "perfect features" universally robust to every foreseeable variation. Another approach involves the decomposition of observed features into session-dependent and session-independent components and the only the later are used for recognition. Decomposition and transformation typically require large quantities of data to learn and some important information is often lost. One such example is Joint Factor Analysis (JFA) [Kenny et al., 2007], which is popular in ASR.

Semi-supervised learning (SSL) is another popular approach to the data insufficiency problem and has experienced a surge in research interest in the machine learning community during the last decade [Zhu, 2005]. Compared to supervised learning (learning from labelled data) and unsupervised learning (clustering unlabelled data), SSL uses a small amount of labelled data and a larger pool of unlabelled data to learn models, thereby avoiding costly manual labelling. SSL can be used to solve the problem of scarce labelled data in AFR and ASR: models weakly trained during enrolment can be enhanced by learning from abundant unlabelled data obtained during normal use or testing, which is inherently rich in variation. Several semi-supervised AFR and ASR systems have been proposed and show the capacity for increasing the performance of supervised systems [Wang et al., 2006][Yamada et al., 2010].

Co-training is one of the most successful examples of SSL and was proposed by Blum and Mitchell [Blum and Mitchell, 1998b] in 1998. The basic assumption is that each data sample can be represented by two independent features, each of which is generally sufficient for correct classification. First, two classifiers are weakly trained using a small number of labelled examples on two different feature sets respectively. Each classifier is then used to classify a larger pool of unlabelled auxiliary data. The most positive examples are then

used to train the other classifier. The process is iterative and is repeated several times. Consequently, both classifiers become more robust with the accumulation of new training data. Blum and Mitchell demonstrated that if the two following assumptions are verified, co-training guarantees improved performance over supervised learning [Blum and Mitchell, 1998b]: (i) sufficiency, which requires each classifier feeds to the other more correctly labelled samples than incorrectly labelled samples, (ii) independency, which requires that samples confidently classified by one classifier are fully informative to train the other.

One of the first applications of co-training to AFR is proposed in [Zhao et al., 2011], but based on two different facial features. The two features are extracted from the same image and thus the assumption of independency is not satisfied; unlabelled samples confidently classified by one system may not help to improve the other, and thus improvements in performance are modest. A template co-update biometric system based on two independent biometric features, face and fingerprints, is proposed in [Roli et al., 2007]. This combination of modalities requires special equipment and thus application is limited.

In this chapter, we propose a co-training type algorithm which exploit the natural co-occurrence of audio-visual data, namely co-Linear Discriminant Analysis (co-LDA), which uses both labelled and unlabelled data to learn discriminative subspaces in which test examples can be better classified. We report its application to audio-visual person recognition in videos. The scenario involves a very limited number of labelled videos and a larger auxiliary pool of unlabelled videos. Each video contains images and audio from a single person, and is parametrized by face and voice feature vectors of high dimension. For each feature, a LDA-based classifier is learnt with the small number of labelled samples and is used to classify the unlabelled samples. The most confident classification results (samples) identified by one classifier are added to the labelled data set, and the corresponding features are then used to train the other LDA subspace and classifier, and vice versa. After several iterations and the accumulation of automatically labelled data, we obtain more reliable subspaces for both face and voice classification.

The remainder of this chapter is organized as follows. In Section 4.2, the principles of LDA and co-training are described and the co-LDA framework are presented and analyzed. The application of the proposed algorithm in audio-visual person recognition is described in Section 4.3. Experiments and results are detailed in Section before our conclusions are presented in Section 4.4. In order to deal with out-of-class samples existed in the unlabelled set, we propose an open-set extension of Co-LDA algorithm by incorporating the sparse representation classifier (SRC) in Section 4.5, and the corresponding experimental work is presented in Section 4.6.

4.2 LDA ,Co-training and Co-LDA

In this section, we first briefly introduce the principles of LDA and co-training in Section 4.2.1 and Section 4.2.2 respectively, and then present the semi-supervised discriminant subspace learning problem, propose and analyse the co-LDA algorithm in Section 4.2.3.

4.2.1 LDA and small sample size problem

Linear Discriminant Analysis is a well-known simple and efficient approach to dimensionality reduction, and is widely used in various classification problems. It aims to find an optimized projection $\boldsymbol{W_{opt}}$ which projects t dimensional data vectors $\boldsymbol{x}$ into a g dimensional space by $\boldsymbol{y} = \boldsymbol{W_{opt}}\boldsymbol{x}$, in which intra-class scatter (S_W) is minimized while the inter-class scatter (S_B) is maximized. S_W and S_B are determined according to:

$$S_W = \sum_{j=1}^{c}\sum_{i=1}^{l_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^T, \tag{4.1}$$

and

$$S_B = \sum_{j=1}^{c} l_j(\mu_j - \mu)(\mu_j - \mu)^T, \tag{4.2}$$

where x_i^j is the i_{th} sample of of class j, μ_j is the mean of class j, c is the number of classes, and l_j is the number of samples in class j. $\boldsymbol{W_{opt}}$ is obtained according to the objective function:

$$\boldsymbol{W_{opt}} = \arg\max_W \frac{W^T S_B W}{W^T S_W W} = [w_1, \ldots, w_g] \tag{4.3}$$

where $\{w_i | i = 1, \ldots, g\}$ are the eigenvectors of S_B and S_W which correspond to the g largest generalized eigenvalues according to:

$$S_B w_i = \lambda_i S_W w_i, i = 1, \ldots, g \tag{4.4}$$

Note that there are at most $c-1$ non-zero generalized eigenvalues, so g is upper-bounded by $c-1$. Since $S_W$is often singular, it is common to first apply Principal Component Analysis (PCA) to reduce the dimension of the original vector. LDA has been applied to AFR and ASR and is often referred to as *Fisherface* [Belhumeur et al., 1997] and *Fishervoice* Li et al. [2010].

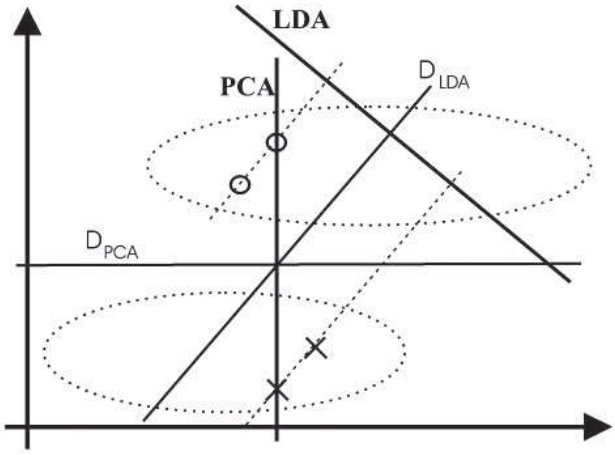


Figure 4.1: Two different classes in Gaussian distributions are represented by doted elliptical curves. However, for each class, only two samples are available for the learning of PCA and LDA. In this case, the classification boundary obtained by a nearest-neighbor classifier in the PCA subspace D_{pca} are more desirable than the classification boundary in LDA subspace D_{LDA} (image excerpted from [Martinez and Kak, 2001]).

As analyzed in [Martinez and Kak, 2001], while LDA can extract discriminant information from high dimensional feature vectors when labelled training data is abundant, but when training data is scarce and not representative of the underlying data structure, the projections can be significantly biased on the small labelled training set, which generally leads to reduced performance, even worse then the unsupervised method PCA. Consider the data distribution illustrated in Figure 4.1. The data points belong to two classes and each class is a underlying Gaussian distribution, as indicated by a elliptical dotted curve. For each class, only two observations are available for the training of PCA and LDA. PCA computes a projection axis which maximizes the variance, which is shown as a vertical line noted as **PCA**. LDA computes a projection direction which maximize the separation of the two classes, the projection axis is shown as the horizontal line noted as **LDA**. If a nearest neighbor classifier is applied, the decision boundary in PCA and LDA subspaces are noted as $\mathbf{D}_{PCA}$ and $\mathbf{D}_{LDA}$. The LDA decision boundary intersect the ellipses of the class distributions, thus PCA will give better classification performance.

4.2.2 Co-training

Co-training belongs to a class of algorithms which combine semi-supervised learning and multi-view learning into one unified framework. The basic assumption of co-training is that

the data samples can be presented with two disjoint views $\mathbf{x_1}$ and $\mathbf{x_2}$. Two classifiers $C_1(\mathbf{x_1})$ and $C_2(\mathbf{x_2})$ are initially learnt with a small set of labelled data $\mathbf{L}$: $\{x_{i1}; x_{i2}, l_i | i = 1, 2, ..., N\}$ where l is the class label, and a large amount of unlabelled data $\mathbf{U}$: $\{x'_{i1}; x'_{i2} | i = 1, 2, ..., M\}$, where N and M denote the size of labelled and unlabelled dataset respectively. At each iteration, the algorithm incorporates samples from the unlabelled set $\mathbf{U}$ into the pool of labelled data $\mathbf{L}$. Typically the selected data are those with the highest prediction confidence for each view. Each classifier is then updated using the augmented labelled data set. The process can be repeated iteratively until all unlabelled auxiliary data is incorporated into labelled dataset. Finally, the outputs of the two classifiers C_1 and C_2 can be weighted and give a single-view classifier C. The intuition of co-training is that each classifier can provide the other with additional, automatically labelled data which might be as informative as some random noisy labelled examples. Based on the analysis of Nigam et al [Blum and Mitchell, 1998b], co-training requires the two views to be conditionally independent in order that each classifier provides informative data to the other.

4.2.3 Co-LDA

In many practical AFR and ASR applications, but unlabelled test data is often abundant, ie. obtained during testing. It typically contains a high degree of intersession variations, from which much more reliable LDA projections can be learnt. We propose a novel co-training framework which is applied to in the discriminant dimensionality reduction problem in two distinct feature spaces, where each classifier iteratively and automatically labels and provide new training data to another.

As illustrated in Figure 4.2, the input of the co-LDA algorithm is a small amount of labelled data and a large pool of unlabelled data, while each sample can be represented with two features, $\boldsymbol{x_1}$ (left in Fig.1)and $\boldsymbol{x_2}$ (right), which are assumed to be independent and sufficient for classification. An LDA projection is learnt on each view respectively. As shown in Fig.1 (a), the labelled dataset is small and is not representative of the general class distribution, so S_B and S_W in Equation (1) and (2) are not well estimated. The LDA projection ($\boldsymbol{W_{opt}}$) learned from this data is illustrated by a solid line. It is biased and leads to an ineffective classification boundary (dashed line). The LDA space of view 1, a classifier is then applied to classify all the unlabelled data, one (or a few) sample that is farthest from the classification boundary is added to the labelled set, and the LDA projection for view 2 is relearned, as shown in Fig.1 (b). Note that, since the two views are assumed to be independent from each other, one point confidently classified in view 1 is highly informative

in view 2 (otherwise if the two views are correlated, that point will be also far from the classification boundary in view 2), and is able to correct to improve the corresponding LDA. In the same way, unlabelled data in view 2 is also classified, and the most confident samples are added to the labelled dataset before the LDA projection for view 1 is also relearned. The process is iterative and as more labelled data is accumulated, the LDA projections are improved and give better results. Of course, one view may feed misclassified samples to the other but according to the sufficiency assumption, classifiers will feed more correctly labelled data than mislabelled data to the other classifier, and thus performance ultimately improves.

4.3 Application to Semi-supervised Audio-visual Speaker Recognition

It is well known that better recognition performance can be achieved through the combination of multiple biometric modalities, through so-called multi-modal systems [Kittler et al., 1998]. With both traits available with standard commercial video capturing devices and on account of their non-intrusive nature, audio-visual person recognition is of natural appeal to both commercial clients and end-users and thus attracted considerable research interest in recent years. Such systems generally involves the score level fusion of AFR and ASR systems. Both are vulnerable to inter-session variations discussed in Section 1, and the proposed co-LDA approach has natural application in audio-visual person recognition scenario: (1) Labelled data is limited while abundant unlabelled data is available during the normal system operation; (2) Peoples's face and voice are naturally available in videos and are independent from each other; (3) Many state-of-the-art ASR & AFR implementations use high-dimensional feature vectors so dimensionality reduction is needed.

The proposed co-LDA audio-visual person recognition system is composed of three steps. First, a facial feature vector and a vocal feature vector are extracted from each video; second, two discriminant subspaces are learned with both labelled and unlabelled face and voice data respectively; third, verification is achieved with accepting or rejecting the claim, while in the identification task, there is no identity claim, and the system is required to establish their identity.

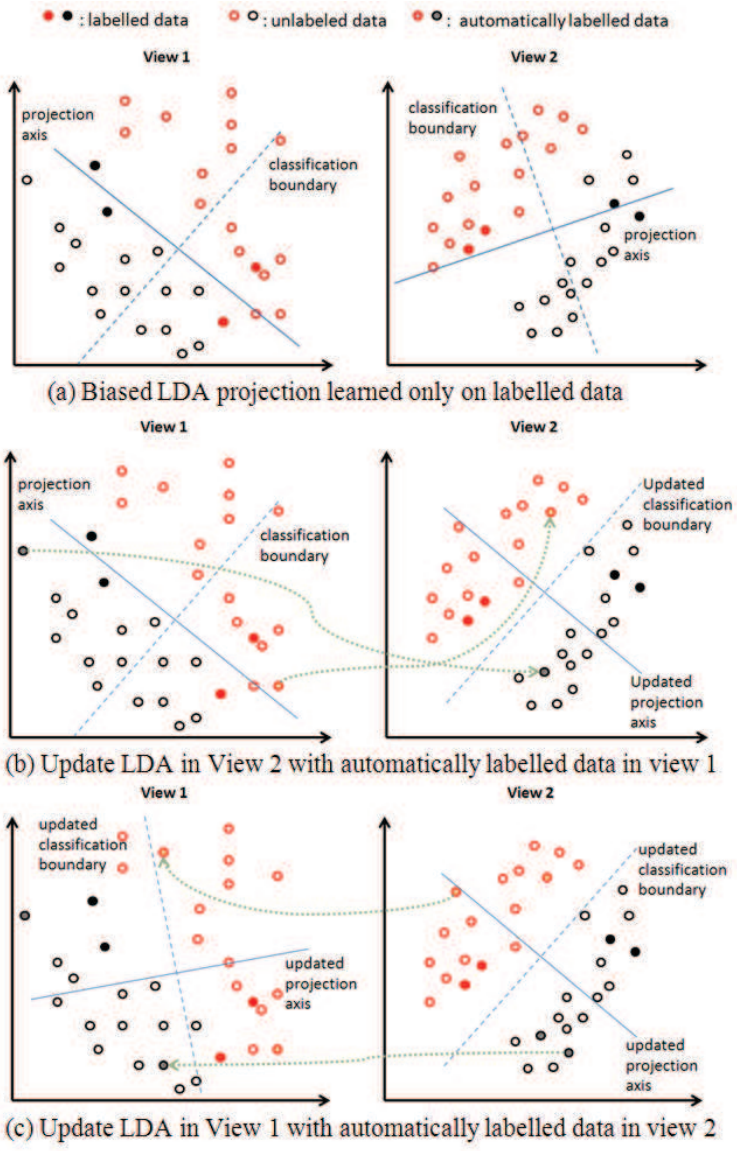


Figure 4.2: Illustration of Co-LDA algorithm

4.3.1 Feature extraction

The process of feature extraction is illustrated in Figure 4.3. For the face modality, face detection is first applied and detected faces are aligned according to detected facial landmark positions. For each video, Local Binary Pattern (LBP) feature vector [Ahonen et al., 2006] is extracted from the most confident detected face. LBP feature extraction divides faces into sub-regions and LBP histograms, which reflect the local texture are extracted from

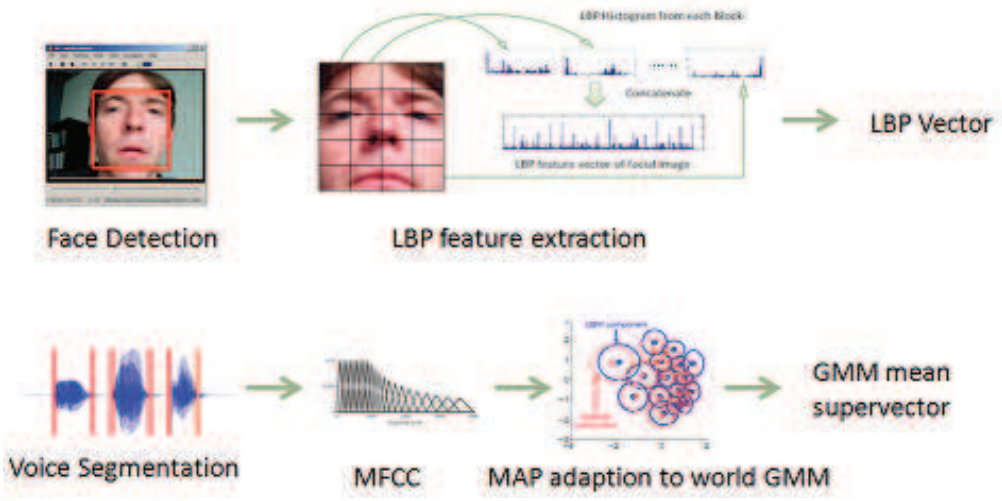


Figure 4.3: Feature vector extraction for face and voice

each region and concatenated to form a high dimensional vector. For the voice modality, voice detection is first applied to eliminate non-speech frames. MFCC coefficients are then extracted from each audio frame and used to determine a Gaussian Mixture Model (GMM) through the Maximum A Posteriori (MAP) adaptation of a speaker independent world model. The means of the GMM model are concatenated into a high-dimensional supervector [Reynolds et al., 2000b]. Accordingly each video is represented by a facial feature vector f_{face} and a voice feature vector f_{voice}.

4.3.2 Subspace learning

The co-LDA system is supplied with a small set of labelled training data acquired during the enrolment session, and a large set of unlabelled data acquired during a period of normal system operation. The dimensionality of the original face and voice feature vector f_{face} and f_{voice} is too great to perform LDA, so a PCA step is first applied to reduce the dimension to n, (x_{face}, x_{voice}) represents the two features in the PCA space. As illustrated in Figure 4.4, the labelled training samples are first used to learn LDA projections with face and voice feature vectors respectively, and then to learn two classifiers C_{face} and C_{voice}. Here we simply use a nearest-template classifier, where a template for each class is calculated as the within-class mean, and the test samples are assigned the label of the closest template according to the label of a test data is determined according to the normalized correlation metrics, which has been demonstrated to be an appropriate similarity measure for LDA space [Kittler et al., 2000]. The similarity between a test point $\boldsymbol{x}$ and a template $\boldsymbol{\mu}$ is

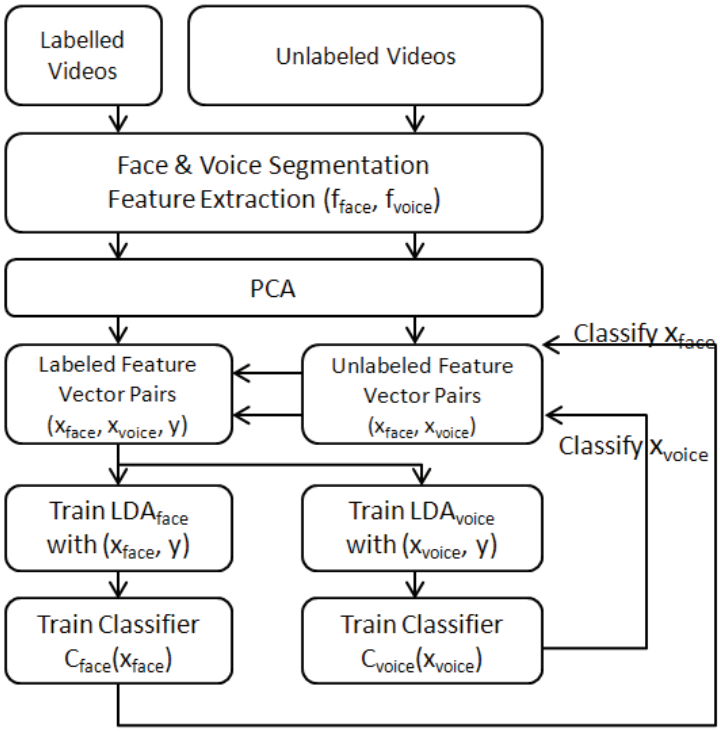


Figure 4.4: Illustration of co-LDA subspace learning

defined as:

$$S_N = \frac{\|\boldsymbol{x}^T\boldsymbol{\mu}\|}{\sqrt{\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{\mu}^T\boldsymbol{\mu}}} \tag{4.5}$$

All unlabelled face and voice samples are projected into their LDA spaces respectively, and classified by C_{face} and C_{voice}. For each classifier and each class, the unlabelled samples closest to the the template are moved from the unlabelled dataset to the labelled dataset with the automatically determined label. We refer to this auxiliary training data as pseudo-labelled data. With the increased pool of labelled data, the two LDA subspaces are relearned, and the templates are recalculated. This process is iterative and is repeated until the unlabelled dataset is empty.

4.3.3 Identification and verification

Both identification and verification tasks can be accomplished using the LDA projections and client templates learned according to the above procedure.

In the identification scenario, facial and vocal feature vectors are extracted from each test video in the manner as described in Section 3.1, and each of them is first projected into their PCA subspaces, and then into their LDA subspaces respectively. In each space, the projected point is compared to each of the c templates according to the normalized correlation similarity measure as described above, thus resulting in two sets of c similarity scores $(S^1_{face}, S^2_{face}, \ldots, S^c_{face})$ and $(S^1_{voice}, S^2_{voice}, \ldots, S^c_{voice})$ Corresponding face and voice similarity scores are then averaged to obtain a fused score:

$$S^i_{fused} = \frac{S^i_{face} + S^i_{voice}}{2}, \tag{4.6}$$

and the test sample is assigned the label of the template whose similarity score is highest. The recognition performance is evaluated with the top 1 identification rate.

In the verification scenario, the face and voice feature vectors of a test data sample are extracted and projected into the same LDA space as before, but are compared only to the template corresponding to the claimed identity. Face and voice similarity scores are fused in the same way. The verification performance is evaluated with the Detection Error Trade-offs (DET) plot acquired with client and impostor scores.

4.4 Experimental work

The experiments reported here aim to evaluate the capability of the co-LDA audio-visual person recognition algorithm to use inter-session variations contained in unlabelled data to enhance models which are weakly learned with limited labelled data. In this section, we first discribe the MOBIO database on which the experimental work is carried, and then presented the experiment results in identification and verification mode.

4.4.1 Mobio audio-visual database

The MOBIO database [McCool et al., 2012] consists of bi-modal (audio and video) data taken from 152 people. The database has a female-male ratio or nearly 1:2 (100 males and 52 females) and was collected from August 2008 until July 2010 in six different sites from five different countries. This led to a diverse bi-modal database with both native and non-native English speakers.

In total 12 sessions were captured for each client: 6 sessions for Phase I and 6 sessions for Phase II. The Phase I data consists of 21 questions with the question types ranging from: Short Response Questions, Short Response Free Speech, Set Speech, and Free Speech. The

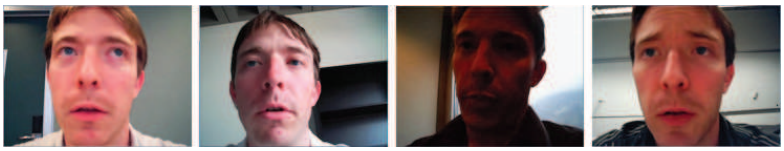


Figure 4.5: Image examples of MOBIO database

Phase II data consists of 11 questions with the question types ranging from: Short Response Questions, Set Speech, and Free Speech. A more detailed description of the questions asked of the clients is provided below.

The database was recorded using two mobile devices: a mobile phone and a laptop computer. The mobile phone used to capture the database was a NOKIA N93i mobile while the laptop computer was a standard 2008 MacBook. The laptop was only used to capture part of the first session, this first session consists of data captured on both the laptop and the mobile phone. Figure 4.5 shows example images which demonstrate typical pose and illumination variability. Similar variability is also presented in the audio streams which contain different environmental noise.

4.4.2 Experimental results

We selected 30 subjects with which to train a GMM world model for speaker recognition, another 30 subjects to conduct co-training experiments, and 15 subjects are selected as imposters in the verification experiment. For subspace learning, one session is randomly selected and used as labelled training data for enrolment, another session is randomly chosen as test data, and the other 10 sessions are used as unlabelled data.

In each video, face images are detected automatically with an OpenCV based face detector. It incorporates eye and nose detection which help to crop detected faces according to facial landmark coordinates. Cropped face images are then resized to 144×128 pixels. For each video, the single most confidently detected face is selected. This face image is divided into 9×8 blocks and $LBP^{u2}_{(8,2)}$features are extracted from each block and concatenated into a 4248-dimensional vector. MFCC acoustic features are extracted over 20ms Hamming windowed frames at a 10ms frame rate. Features are composed of 26 MFCC coefficients augmented with their 26 delta coefficients and the delta energy, resulting in acoustic vectors of 53 coefficients. Informative speech frames are extracted with an acoustic energy based speech detector described in [Besacier et al., 2000] and non-speech frames are discarded. A 64-component speaker model is then adapted from the world model trained with an EM algorithm of the world model subset. MAP adaptation is performed with a relevance factor

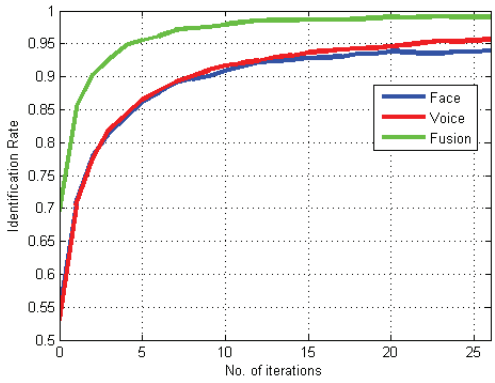
of 14 and only means were adapted. The GMM means are concatenated to form a 3392-dimensional supervector. Each video is thus represented by an LBP feature vector and a GMM voice supervector.

Following co-training as described in Section 3, initial LDA projections and classifiers are learned on the labelled dataset, and iteratively updated with automatically labelled data. After the learning process, data is projected into the learnt LDA spaces and both identification and verification experiments were conducted. The identification rate reported is the average of 50-fold cross-validation. In verification experiment, following the protocol for LDA face verification described in [Kittler et al., 2000], we used an imposter set containing 15 subjects which is independent from the training set used to learn the projections and models. Thus client scores are calculated by comparing the test data of the 30 clients to their true identity models, and imposter scores are calculated by comparing 15 imposters to 30 client models in an exhaustive way. we and the verification performance is reported in terms of Detection Error Trade-off (DET) curves which correspond to these client and imposter scores.

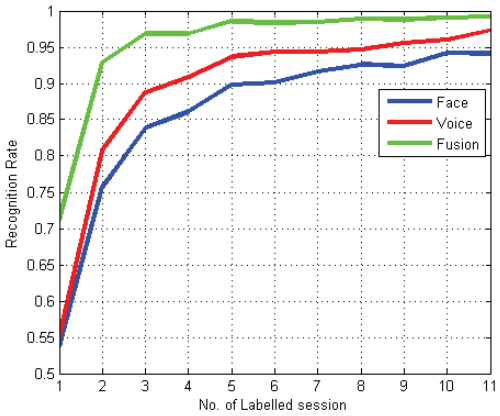
We first report results for the identification task. The identification rate attained by independent face and voice classifiers and their fusion is shown in Figure 4.6 (a). In all cases, performance is shown as function of number of iterations of co-training. Profiles show that the identification rate for both face and voice classifiers increases when a greater number of unlabelled samples is incorporated into the training set through co-training: face identification rate increases from 53% to 96% while the voice identification rate increases from 55% to 94%; and the identification rate for the fused system increases from 70% to 99%. Among the automatically labelled data samples, 98.5% of them are correctly labelled.

We may wonder with purely supervised learning method, how many sessions of labelled data we need in order to achieve the same performance. So we randomly select 1-11 sessions as labelled training data to train the LDA spaces and models, and another session as test data, each experiment is repeated 50 times and the average identification rate with respect to the different number of labelled training sessions is shown in Figure 4.6 (b). The result shows that, with supervised method, at least 10 labelled training sessions are needed to reach the performance of the proposed co-training method, which uses only 1 labelled session accompanied with 10 unlabelled sessions.

In a verification scheme, test data vectors are projected into the LDA subspaces learnt through co-training and are compared to all the client models. The DET curves for Face/Voice/Fusion verification systems before and after co-training are shown in Figure 4.7. The performance for these systems without co-training is generally low due to the large



(a) Identification rate as a function of co-training iterations



(b) Identification rate of as a function of labelled training sessions for baseline system

Figure 4.6: Results for identification task

inter-session variations which are not represented in the low quantity of training data (AFR and ASR verification rates are around 20%). Similar results were reported in [Marcel et al., 2010]. However, after co-training, both single systems achieve below 5% EER while the fusion system achieves an EER of 1.4%. These results demonstrate the effectiveness of the proposed method.

4.5 Coping with Out-of-class Data

The Co-LDA algorithm presented in previous sections introduced a new approach to combine the learning of discriminant features with more robust modelling and classification in a unified co-training framework. However, it assumes a closed-set scenario.

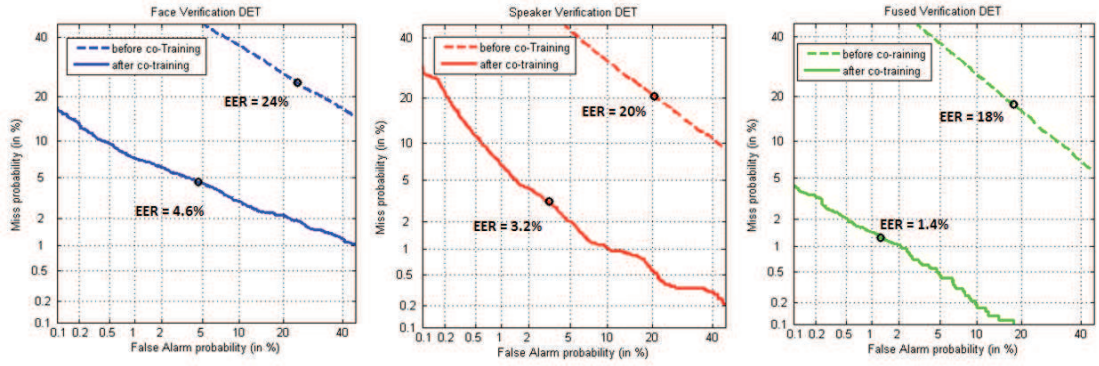


Figure 4.7: DET curves for face (left) voice (middle) and fused (right) verification system

This section presents an extension of our co-training algorithm to open-set scenarios. The new algorithm combines linear discriminant analysis (LDA) with a sparse representation classifier (SRC) [Wright et al., 2009]. While SRC has shown to give state-of-the-art performance in face recognition Wright et al. [2009] and speaker recognition Naseem et al. [2010], it depends upon the availability of large quantities of data, hence its combination with co-training. A sparsity concentration index (SCI) is also effective in rejecting out-of-class data, hence its suitability to open-set problems.

4.5.1 Sparse representation classifier and out-of-class sample detection

Suppose we have c classes, and let $\mathbf{A} = [\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_c}]$ be a set of training samples, where $\mathbf{A_i} = \{\mathbf{v_{i,1}}, \ldots, \mathbf{v_{i,n_i}}\}$ indicates the subset of training samples for class i. A single testing sample $\mathbf{y}$ could be well approximated by the linear combination of training samples from $\mathbf{A_i}$, which could be written as

$$\mathbf{y} = \sum_{j=1}^{n_i} \alpha_{i,j} \mathbf{v_{i,j}}. \tag{4.7}$$

Since $\mathbf{A}$ is the dictionary which includes all the training samples, Equation 4.7 can be rewritten in the form $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha}_0$ where $\boldsymbol{\alpha}_0 = \{0, \ldots, 0, \alpha_{i,1}, \ldots, \alpha_{i,n_i}, 0, \ldots, 0\}^T$ is the coefficient vector in which most coefficients are zero except the ones associated with class i. Due to the fact that a valid test sample $\mathbf{y}$ can be sufficiently represented only using the training samples from the same class, and this representation is the sparsest among all others, to find the identity of $\mathbf{y}$ then equals to find the sparsest solution of $\boldsymbol{\alpha}$. So The four main steps involved in the application of SRC are outlined in the following.

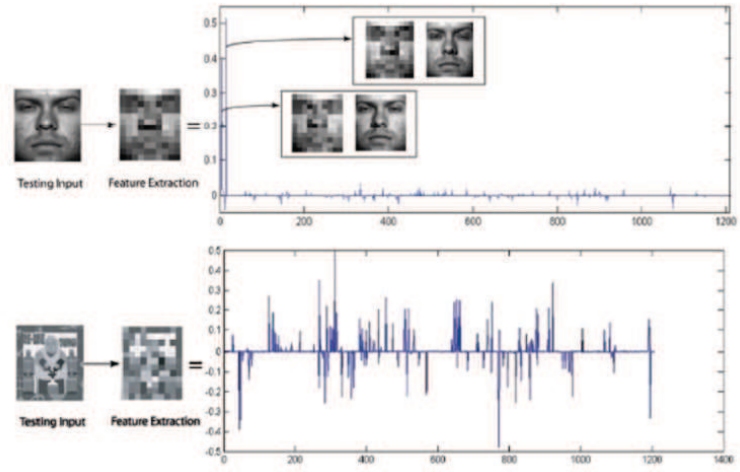


Figure 4.8: Coded coefficients of a in-class test sample (top) and out-of-class test sample (bottom). The coding coefficients of in-class samples are concentrated in a single class, and for out-of-class samples, the coding coefficients are dispersed.(image excerpted from [Wright et al., 2009])

1. Normalize each entry in $\mathbf{A}$ to have unit $l2$-norm;
2. Sparsely code $\mathbf{y}$ on $\mathbf{A}$ via $l1$-norm minimization:

$$\hat{\boldsymbol{\alpha}} = \arg\min \| \boldsymbol{\alpha} \|_1 \text{ subject to } \| \mathbf{y} - \mathbf{A}\boldsymbol{\alpha} \|_2 < \epsilon \tag{4.8}$$

3. Compute the residuals of each class by:

$$r_i(\mathbf{y}) = \| \mathbf{y} - \mathbf{A}\hat{\boldsymbol{\alpha}}_i \| \ (\mathbf{i} = \mathbf{1}, \ldots, \mathbf{c}) \tag{4.9}$$

   where $\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector associated with class i, and $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_c]$,
4. Classification according to:

$$\text{Identify}(\mathbf{y}) = \arg\min_{\mathbf{i}}(\mathbf{r_i}(\mathbf{y})) \tag{4.10}$$

SRC was originally developed for face identification [Wright et al., 2009] and has since been applied in speaker identification [Naseem et al., 2010]. Comparative experiments show that SRC outperforms Nearest Neighbor (NN) and Support Vector Machine (SVM) classifiers.

The original work [Wright et al., 2009] proposed a sparsity concentration index (SCI) which aims to reject invalid test samples. We propose its use to reject out-of-class data.

Since the aim in SRC is to represent each test sample according to a sparse, weighted set of training samples, the representation of within-class samples should be concentrated on a single class. The representation of out-of-class samples, however, is more dispersed. Figure. 4.8 shows the coded coefficients of a in-class face test sample and an out-of-class sample. The SCI score of a coefficient vector $\hat{\boldsymbol{\alpha}}$ is defined as:

$$SCI(\hat{\boldsymbol{\alpha}}) = \frac{c * \max_i \| \hat{\boldsymbol{\alpha}}_{\boldsymbol{i}} \|_1 / \| \hat{\boldsymbol{\alpha}} \|_1 - 1}{c - 1} \tag{4.11}$$

and is bounded between 0 and 1. Out-of-class samples can thus be rejected according to a threshold $\tau \in (0, 1)$ where $SCI(\hat{\boldsymbol{\alpha}}) < \tau$.

4.5.2 Co-LDA-SRC algorithm

As shown in [Martinez and Kak, 2001], LDA projections can be unrepresentative of intersession variations when learned on smaller datasets and thus give unsatisfactory performance. SRC also requires abundant labelled training data so that test samples can be reliably reconstructed from a linear combination of same-class training samples [Wright et al., 2009]. In most biometric applications, however, labelled data acquired during enrollment is generally limited in quantity and the acquisition of more, manually labelled data is usually costly or impractical. In the following we show how both LDA and SRC can be integrated within a unified co-training framework thereby exploiting abundant, unlabelled data to improve performance.

Consider a multi-modal biometric system where different biometric modalities can be considered as independent views of the same data. Also assume that abundant auxiliary data can be acquired over an extended period so that it is representative of intersession variations. According to a general co-training scheme, a classifier in one view can be used to provide automatically labelled, new training data to another, and vis-versa.

The standard co-training algorithm assumes a closed-set scenario, where all unlabelled data belong to one of the registered classes. In practical scenarios, however, and particularly for biometric systems, data acquired automatically during regular use may often contain out-of-class samples (persons not pre-enrolled). Out-of-class samples should not be incorporated into the labelled training set.It is thus necessary to adapt the standard co-training algorithm to reject out-of-class samples. This facility is provided readily through a threshold SCI as discussed in Section 4.5.1.

We assume each data sample is represented by two feature vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ extracted from two independent biometric traits. A small labelled training set of n samples $\mathbf{L}$:

Algorithm 3 Co-LDA-SRC

Input:
- Labelled dataset $\mathbf{L}$ from c classes and unlabelled dataset $\mathbf{U}$;
- SCI Threshold τ and number of samples N to be incorporated into the set of labelled samples.

Output:
- Projection matrix $\mathbf{P_1}$ and $\mathbf{P_2}$;
- Increased labelled training set $\mathbf{L}$.

Initialization: Center $\mathbf{L}$ and $\mathbf{U}$ in both view, apply PCA if the dimensionality is too high;

repeat
  for $v = 1, 2$ **do**
  - Train LDA projections $\mathbf{P_v}$ with samples in the v_{th} view of $\mathbf{L}$ and project samples according to $\mathbf{P_v}$ to form $\mathbf{A_v}$;
  - Project the v-th view of $\mathbf{U}$ into $\mathbf{P_v}$, noted as $\mathbf{Y_v}$;
  - Run SRC on each entry of $\mathbf{Y_v}$ with training set $\mathbf{A_v}$, discard entries with SCI lower than τ.
  - $\mathbf{L_v} \leftarrow \emptyset$

    for $i = 1$ to c **do**
      for each class i, add to $\mathbf{L_v}$ the single sample in $\mathbf{U}$ most confidently labelled (lowest $r_i(\mathbf{y})$).
    end for
  end for
  $\mathbf{L} \leftarrow \mathbf{L} \cup \mathbf{L_1} \cup \mathbf{L_2}$; $\mathbf{U} \leftarrow \mathbf{U} - \mathbf{L_1} - \mathbf{L_2}$
until N pseudo-labelled samples are incorporated into the training set

$\{\mathbf{x_{i1}}, \mathbf{x_{i2}}; l_i | i = 1, 2, ..., n\}$ is acquired during an enrollment session, while a larger unlabelled dataset of m samples $\mathbf{U}$: $\{\mathbf{x'_{i1}}, \mathbf{x'_{i2}} | i = 1, ..., m\}$ is obtained over an extended period of normal use. The entire training set is noted by $\mathbf{X} = \mathbf{L} \cup \mathbf{U}$.

$\mathbf{X}$ is first centred so that $\overline{\mathbf{x}}^{(\mathbf{v})} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_{iv}} = 0, (v = 1, 2)$, and optionally treated conventionally with principal component analysis (PCA) to reduce its dimensionality if to high to be treated directly with LDA. Then, LDA projections $\mathbf{P_1}$ and $\mathbf{P_2}$ are determined for each view using only the set of labelled samples $\mathbf{L}$. The same set is then projected into the new subspaces according to $\mathbf{A_v} = \mathbf{P_v}^T \mathbf{x_v}$. The result forms training examples for SRC in the v-th view.

Both views $\mathbf{x'_1}$ and $\mathbf{x'_2}$ of the set of unlabelled samples U are then projected onto their respective subspaces according to $\mathbf{Y_v} = \mathbf{P_v}^T \mathbf{x'_v}$. Each entry $\mathbf{y}$ of $\mathbf{Y_v}$ is sparsely coded on $\mathbf{A_v}$ according to Equation 4.8, and the reconstruction residues $r_i(\mathbf{y})$ and SCI score are determined according to Equations 4.9 and 4.11 respectively. Those entries whose SCI score is less than a threshold τ are labelled as out-of-class samples, whereas the remaining in-class

samples are assigned to one of the known classes according to Equation 4.10. For each view and each class, the single in-class sample most confidently labelled (with the lowest $r_i(\mathbf{y})$) is removed from $\mathbf{U}$ and incorporated into $\mathbf{L}$. Projections $\mathbf{P_1}$ and $\mathbf{P_2}$ are then re-trained with the now-larger labelled dataset. This process is repeated until a pre-specified number of labelled samples are gathered. The algorithm is summarized in Algorithm 3. In the test phase, the v-th view of a test sample is projected onto $\mathbf{P_v}$ and classified by SRC with the increased training set $\mathbf{A_v}$.

4.6 Experimental Results

In this section, we report an evaluation of the proposed Co-LDA-SRC algorithm through experiments in audio-visual persons identification where the task is to identify the speaker in a video sequence according to acoustic and facial observations. A small sum of labelled training data collected during a single enrollment session is used as labelled data for initial modelling. Comparisons against a baseline system using supervised LDA feature extraction and SRC classification show how learning from a larger pool of unlabelled data acquired during normal system use is effective in capturing intersession variation. We stress, however, that the framework is general and can be applied to any multi-view problems.

4.6.1 Database and protocols

Experiments were conducted with the standard MOBIO database [McCool et al., 2012] as discussed in Section 4.4.1.

We use cropped face images provided with the MOBIO database, one image per video sample. All images are resized to 50×43 pixels and then histogram equalized. Rows of pixel intensities are concatenated to form feature vectors of 2150 dimensions. The speech signal is split into frames of 20ms duration before the extraction of features composed of 26 Mel-scaled frequency cepstral coefficients (MFCCs), their 26 derivatives and the delta energy. Energy-based voice activity detection is then applied to disguard non-speech frames. A 64-component Gaussian mixture model (GMM) is then fitted to remaining speech data through the maximum a posteriori (MAP) adaptation of a speaker-independent world model. The means of the GMM model are then concatenated to form a 3392-dimensional GMM super-vector. Both face and speech feature vectors are first reduced to 100 dimensions through the application of PCA.

To create a pool of in-class samples, we selected only 20 subjects as registered clients.

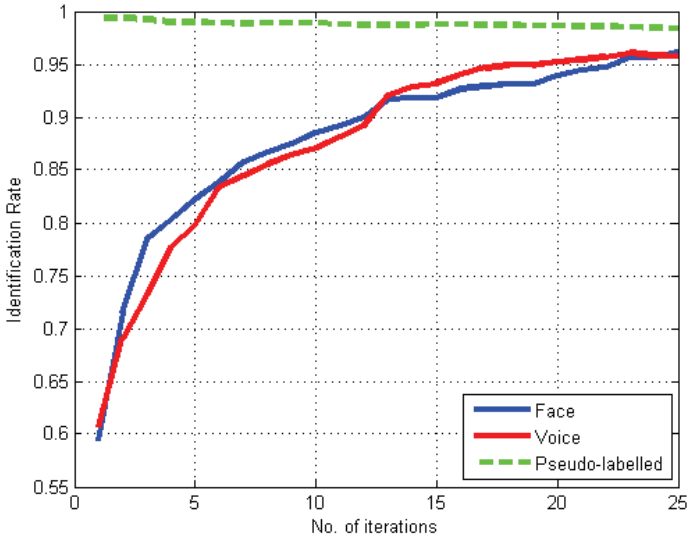


Figure 4.9: Identification rate and pseudo-labelled data accuracy as a function of the iteration number.

For each subject, 5 videos are selected from each of the 12 sessions, which results in 60 videos per subject. For each registered subject, 5 videos in one randomly selected session are used as labelled training data for enrolment, 5 videos from another randomly selected session are used as test data, and the 50 videos in the other 10 sessions are used as unlabelled data. A random selection of 300 videos of the remaining, different subjects were added to the unlabelled dataset as out-of-class samples. Thus, from an unlabelled pool of 1300 samples, just under 25% are out-of-class. The SCI threshold τ is empirically set to 0.4, and the number of pseudo-labelled samples N to be incorporated into the labelled set is set to 90% of the expected number of in-class samples in the unlabelled dataset.

The evaluation is two-fold: first, we report a top-1, closed-set identification experiment performed on the independent test dataset; second, we report the labelling accuracy of automatically labelled data. All results are averaged through 20-fold cross-validation.

4.6.2 Results

Figure 4.9 shows the identification rate of an SRC classifier applied to face and voice observations independently, in addition to the accuracy of the increasing number of pseudo-labelled samples added to the labelled dataset. Between each iteration the size of the labelled dataset increases by about of $20 \times 2 = 40$ samples. While the labelling accuracy of

Id. Rate(std.)	Face	Speech
PCA + SRC	0,670(0,035)	0,652(0,030)
LDA + SRC	0,590(0,046)	0,611(0,048)
SDA[Cai et al., 2007] + SRC	0,725(0,029)	0,759(0,032)
VLR[Nie et al., 2011]	0.772(0,036)	0,863(0,033)
Co-LDA	0.902(0.032)	0.891(0.034)
Co-LDA-SRC	**0,961(0,051)**	**0,962(0,054)**

Table 4.1: Comparison of identification rate and standard deviation of different algorithms on MOBIO database

pseudo-labelled samples is shown to decrease to $98,5\%$, the effect of labelling errors does not outweigh the benefit of modelling intersession variations through the use of additional, automatically labelled data. Profiles show that the identification rate for both face and voice classifiers increases when a greater number of unlabelled samples is incorporated into the training set through co-training.

Table 4.1 shows the mean value and standard deviation of identification rate over 20 runs of different algorithms. The baseline approaches are the SRC classifiers applied to features in PCA and LDA-derived subspaces, where the training samples only include the original, manually labelled dataset. The performance of LDA is even worse than that of unsupervised PCA, most probably due to the effect of over-fitting. We also report results for Semi-supervised Discriminant Analysis (SDA) [Cai et al., 2007] and Virtual Label Regression (VLR) [Nie et al., 2011], two semi-supervised feature extraction methods trained on both labelled and unlabelled data. Due to the use of single views in each case, however, both approaches yield only modest improvements over the PCA and LDA systems. VRL outperforms SDA since it is one of the very few semi-supervised learning approaches where out-of-class samples are modelled independently and excluded from the in-class data to train the projection. Our own previous approach, Co-LDA, out-performs all single view methods on account of the the co-training framework. Finally, the proposed multi-view, co-training algorithm out-performs co-LDA by a large margin. The significant improvement in performance is attributed to the use of an SRC classifier and its capacity to reject out-of-class samples. Compared to the co-LDA algorithm, the error rate is reduced by over 60% relative. The experiment demonstrate the effectiveness of the proposed algorithm to use unlabelled data to enhance the recognition performance of traditional supervised multi-modal biometric systems.

4.7 Summary

In this chapter, we presented our first MVDR framework based on incremental co-training. The proposed Co-LDA and algorithm allows two independent biometric systems to train each other using a large pool of automatically labelled auxiliary training data while equally applicable to any combination of biometric modalities. In order to deal with out-of-class samples existed in unlabelled dataset, we also provided an extension of the Co-LDA algorithm by incorporating a Sparse Representation Classifier (SRC). We demonstrate the utility of proposed algorithms in the scenario of audio-visual person recognition in videos. Automatic speaker and face recognition systems are shown to make efficient use of both labelled and unlabelled data, where unlabelled data are added iteratively to the labelled dataset and are used to improve the discriminative power of LDA. Experimental results on both identification and verification tasks show significant improvements in performance and demonstrate the effectiveness of our algorithm.

Chapter 5

MVDR by Subspace Clustering Agreement

In chapter 4, we presented a semi-supervised MVDR framework based on incremental co-training. However, several questions also arise:

1. Co-training is a semi-supervised algorithm which requires at least some labelled data to initialize. Is it possible to extended to purely unsupervised problems such as clustering?

2. In the co-training framework, the small labelled dataset is iteratively enlarged by incorporating automatically labelled samples, which inevitably contains mis-classified samples which may cause label errors. Can LDA projection be learnt with data with label noise?

3. Co-training assumes a conditional independence between different views. What if this assumption is violated?

4. Co-training deals with two input views. Can it be extended to more than two views?

5. Is the proposed MVDR framework suitable for non-biometric data?

In this chapter, we try to address all these open questions. We propose a new unsupervised MVDR algorithm for clustering multi-view, high-dimensional data, which learns

discriminative subspaces in an unsupervised fashion based upon the assumption that a reliable clustering should assign same-class samples to the same cluster in each view. The framework combines the simplicity of k-means clustering and Linear Discriminant Analysis (LDA) within a co-training scheme which exploits labels learned automatically in one view to learn discriminative subspaces in another. The proposed method can be extended to multi-view settings where more than two input views are available. The effectiveness of the proposed algorithm is demonstrated empirically with multi-modal biometric data for which the conditional independence assumption is fully satisfied and with non-biometric data for which the independence assumption is only partially satisfied. Significant improvements over alternative multi-view clustering approaches are reported in both cases. In essence, the proposed MVDR method proposed in this chapter learns a subspace for each view such that the clustering structure is in maximum agreement across each view, it is referred to as *MVDR by subspace clustering agreement.*

5.1 Motivation

The recent explosion of multimedia information on the Internet demands effective clustering techniques capable of handling huge quantities of potentially complex data. First, multimedia data are generally represented in high-dimensional spaces in which the so-called *curse-of-dimensionality* makes the application of many clustering techniques somewhat troublesome. Second, by its very nature, multimedia data is multi-modal, for example audio and video information can form two independent clustering inputs. The fusion of modalities remains a challenging problem and is generally treated in isolation to that of high dimensionality.

Difficulties associated with the high dimensionality are generally overcome through the application of dimensionality reduction (DR) techniques, such as Principle Component Analysis (PCA) [Jolliffe, 2005] and related approaches. Dimensionality reduction can either be applied in a pre-processing step prior to clustering, or be integrated into the clustering framework itself. The latter is referred to as subspace clustering (see a survey [Kriegel et al., 2009]). Whatever the technique, however, the goal is always to identify a subspace in which clusters are maximally separated.

Research in multi-modal fusion, which aims to optimally combine information in different views of the same data, has led to a number of multi-view clustering algorithms, e.g. [Bickel and Scheffer, 2004, Chaudhuri et al., 2009, Kumar and Daumé III, 2011]. The goal with all such methods is to identify a clustering result which agrees across different views (samples clustered together in one view are also clustered together in other views).

This chapter presents our efforts to address the problems of high-dimensionality and multi-modal fusion in a unified framework. We assume that each data sample is represented by two feature vectors corresponding to two independent views. We further assume significant information in each feature vector to be unrelated to the underlying class label and that there exists a lower dimensional subspace in which classes are maximally separated. Inspired by the concept of *co-training* [Blum and Mitchell, 1998a], we describe a new multi-view subspace clustering algorithm which reflects the intuition that a true underlying clustering should assign samples to the same cluster irrespective of the view. It seeks a discriminant subspace for each view which results in a clustering policy with maximal agreement across views. Discriminant subspaces in one view are learned using cluster labels for the same samples in another view, and vice versa. The process is iterative and is repeated until a maximum agreement is achieved. The proposed algorithm simultaneously outputs cluster indicators, discriminant subspaces for each view, and compact models of different clusters. As a result, the algorithm copes naturally with out-of-sample data and is readily extended to semi-supervised classification.

The remainder of this chapter is organized as follows. Section 2 analyses three essential components of the proposed algorithm: LDA, k-means, and co-training. Section 3 presents the proposed clustering algorithm and extensions to cosine distance, non-linear case and semi-supervised settings. Section 4 describes the proposed algorithm in the context of existing literature. Section 5 presents experimental evaluations in audio-visual speaker clustering. Section 6 presents our conclusions.

5.2 LDA, k-means, and co-training

In this section we provide an analysis of the three essential components of the proposed algorithm: LDA, k-means and co-training. Although the principles of LDA and co-training has already been presented in Section 4.2, here we provide a deeper analysis. We show that the objective functions of LDA and k-means are compatible, and that the essence of co-training is to attain an agreement between two predictors.

5.2.1 LDA and k-means

As discussed in [Ding and Li, 2007], the objective function of LDA and k-means are closely related. Consider a set of centred input data $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ such that $\bar{\mathbf{x}} = \sum_i \mathbf{x_i}/\mathbf{n} = \mathbf{0}$. Let the class labels be given by $H = \{h_1, \ldots, h_n\}$, and define matrices of between-class

scatter S_b, within-class scatter S_w and total scatter S_t as:

$$\begin{aligned} S_b &= \sum_k n_k \mathbf{m_k} \mathbf{m_k^T} \\ S_w &= \sum_k \sum_{i \in C_k} (\mathbf{x_i} - \mathbf{m_k})(\mathbf{x_i} - \mathbf{m_k})^\mathbf{T} \\ S_t &= \sum_{i=1}^{n} \mathbf{x_i} \mathbf{x_i^T} \end{aligned} \tag{5.1}$$

where n_k is the number of samples in class k, $\mathbf{m_k}$ is the mean of class k, and C_k is the set of samples belonging to k-th class ($l_i = k$) and $S_t = S_w + S_b$. LDA seeks a projection P which maximizes the ratio between S_b and S_w. The objective function is thus:

$$\begin{aligned} &\arg\max_P \mathrm{Tr} \frac{P^T S_b P}{P^T S_w P} = \arg\max_P \mathrm{Tr} \frac{P^T S_b P}{P^T S_w P} + 1 \\ &= \arg\max_P \mathrm{Tr} \frac{P^T S_t P}{P^T S_w P} = \arg\min_P \mathrm{Tr} \frac{P^T S_w P}{P^T S_t P} \end{aligned} \tag{5.2}$$

Where $\mathrm{Tr}\{\cdot\}$ is the trace of a matrix.

On the other hand, the k-means objective function is give by:

$$\arg\min_H \sum_k \sum_{i \in C_k} \| \mathbf{x_i} - \mathbf{m_k} \|^\mathbf{2} \tag{5.3}$$

where H represents a cluster indicator and $\mathbf{m_k}$ is the mean of the k-th cluster. In most cases same-class samples should be assigned to the same cluster, i.e. cluster labels should be indicative of the class label L. In this case, the k-means objective function is equivalent to the minimisation of the trace of the within-class scatter matrix so that:

$$\arg\min_H \mathrm{Tr}\ S_w = \arg\min_H \mathrm{Tr}\ (S_t - S_b) \tag{5.4}$$

Equations 5.2 and 5.4 thus reveal that the LDA and k-means objective functions are compatible: k-means aims to minimize within-class scatter while LDA minimizes the within-class scatter and maximize total scatter in the same time.

5.2.2 Co-training

Co-training [Blum and Mitchell, 1998a] is one of the most acclaimed approaches to semi-supervised learning. In co-training, data samples are assumed to be represented by two conditionally independent features X_1 and X_2. Two predictors f_1 and f_2 assign to each

X a class label Y ($f : X \rightarrow Y$) and are trained according to each view using a small pool of labelled data. The two predictors are used to assign labels to a larger pool of unlabelled data. A subset of samples with which the predictors have the most confidence in label assignments is added to the pool of labelled data. The predictors are then iteratively re-learned and applied to the remaining unlabelled data. Co-training essentially learns two different predictors f_1 and f_2 which *agree* on unlabelled data across different views. A theoretical treatment of convergence is given in the original paper Blum and Mitchell [1998a] and shows that, under the assumption of conditional independence, a weak predictor f_1 in view X_1 which can tolerate random label noise can learn from automatically labelled samples provided by f_2 in view X_2.

Here we presents the extension of co-training predictors to co-training subspaces. LDA is a supervised method which requires class labels, while k-means is a unsupervised method which generates cluster indicators. Under the assumption of conditional independence between views, they can be regarded as class labels corrupted with random noise for the other view. The two methods are combined with the idea of co-training.

5.3 Multi-view subspace clustering: a co-training algorithm

In this section, we apply the concept of co-training to the problem of discriminant subspace learning for multi-view clustering. Since we assume unsupervised clustering, the standard semi-supervised co-training algorithm cannot be applied directly. However, the goal remains the same, i.e. to learn a subspace for each view which results in a common clustering policy. For clarity, samples assigned to the same cluster in the subspace of one view should be assigned to the same cluster in the subspace of the other view and, conversely, samples assigned to different clusters in the subspace of one view should be assigned to different clusters in the subspace of the other view.

5.3.1 An algorithm: CoKmLDA

We first define a *Cluster Agreement Index* (CAI). Let $H^{(1)}$ and $H^{(2)}$ represent the assignment of samples in views $v = 1$ and $v = 2$ to one of K clusters. The CAI is defined as:

$$CAI(H^{(1)}, H^{(2)}) = \frac{1}{n} \sum_{i=1}^{n} \delta \left(h_i^{(1)}, \text{map}(h_i^{(2)}) \right) \tag{5.5}$$

where n is the total number of samples and $\delta(a, b)$ is a function equal to unity if $a = b$ and zero otherwise. The $map()$ function returns an optimal mapping between cluster identifiers in view 1 to those in view 2 in order that the CAI is maximized. This is achieved with a classical Hungarian algorithm [Steiglitz and Papadimitriou, 1982].

We then seek two LDA projections $P^{(1)}$ and $P^{(2)}$ such that the CAI resulting from k-means on both subspaces is maximized. The objective function is givne by:

$$\arg\max_{P^{(1)},P^{(2)}} CAI(H^{(1)}, H^{(2)}) \tag{5.6}$$

where $H^{(v)}$s are further dependent on $P^{(v)}$s

$$H^{(v)} = \arg\max_{H^{(v)}} \sum_{k=1}^{K} \sum_{h_i^{(v)}=k} \parallel {P^{(v)}}^T \mathbf{x_i} - {P^{(v)}}^T \mathbf{m_k} \parallel^2 (v = 1, 2). \tag{5.7}$$

In the following we propose an algorithm that alternatively solves Equation 5.6 and Equation 5.7 for $P^{(v)}$ and $H^{(v)}$ according to a modified co-training approach. We use cluster indicators generated by k-means in one view as label information to train LDA projections in the other view, and vis-versa. While the essential elements of the proposed algorithm are relatively straightforward, the algorithm tends to converges given that LDA can learn approximately good projections with some extent of label noise (mathematical proof given in section 5.3.3). The new algorithm is referred to as co-k-means linear discriminant analysis (CoKmLDA). The main steps of the iterative algorithm are as follows:

1. **k-means clustering** Solve Equation 5.7 with fixed $P^{(1)}$ and $P^{(2)}$ by determine cluster indicators $H^{(1)}$ and $H^{(2)}$ with a k-means algorithm operating in each view. In the initialization step, k-means is applied on original features. If the dimensionality of the original feature is too high, PCA is applied as a preprocessing step.
2. **Cross-labelling** Label samples in view 1 according to $H^{(2)}$, and vis-versa.
3. **LDA training** Learn LDA projection $P^{(1)}$ with original or PCA processed features $X^{(1)}$ and labels corresponding to view 2, and vis-versa. This step optimizes Equation 5.6 in the sense that, in view 2, samples belongs to the same cluster indicated by $H^{(1)}$ will be projected near each other while samples belongs to different clusters indicated by $H^{(1)}$ will be projected apart. So the data structure in view 2 will be constrained to be more compatible with view 1, vis-versa.
4. **Iterate** Return to step 1, perform k-means again in projected subspace. We compute the objective function in Equation 5.6 for each iteration. The iteration process can be terminated either after a fixed number of iterations, or when the objective function

Algorithm 4 CoKmLDA

Input: a set of n multi-view samples $X = \{X^{(v)}|v = 1, 2\}$, where $X^{(v)} = \{x_1^{(1)}, \ldots, x_n^{(v)}\}$, and the expected number of clusters K.

Output: view dependent cluster indicators $H^{(v)} = \{h_1^{(v)}, \ldots, h_n^{(v)}\}$, and projection matrices $P^{(1)}, P^{(2)}$

Initialize:

1. Center the feature vectors in each view and apply PCA if the dimensionality of the feature space is too high;
2. Perform k-means clustering in each view to estimate cluster indicators $H^{(v)} = \{h_1^{(v)}, \ldots, h_n^{(v)}\}$;
3. For each view v, identify the single sample closest to each of the K clusters, $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$.

for $t = 1$ **to** *iter* **do**
  for $v = 1$ **to** 2 **do**
    1. Use $X^{(v)}$ and $H^{(3-v)}$ to train LDA projections $P^{(v)}$ and project samples into the LDA subspace;
    2. Using seeds $S^{(v)}$, perform k-means clustering on projected samples to estimate new cluster indicators $H^{(v)}$;
    3. Update seeds $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$.
  end for
end for

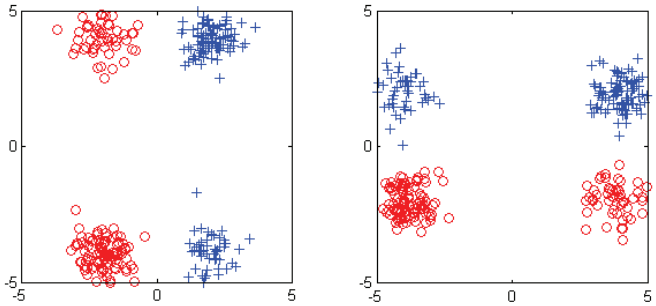
did not reach a new minimum for a fixed number of iterations in each view.

It is well known that the performance of k-means is sensitive to the quality of its initialization (seeds). Accordingly it is common to run k-means several times with random initialization, and to retain the clustering result which minimizes Equation 5.4. This approach is computationally demanding and thus we utilize *seed inheritance* to reduce the computational burden. After each application of k-means, we identify in each view the single sample closest to each of the K cluster centroids, denoted $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$. In subsequent iterations, k-means applied in view v is initialized with the K seeds in $S^{(v)}$. The CoKmLDA algorithm is formally summarized in Algorithm 4.

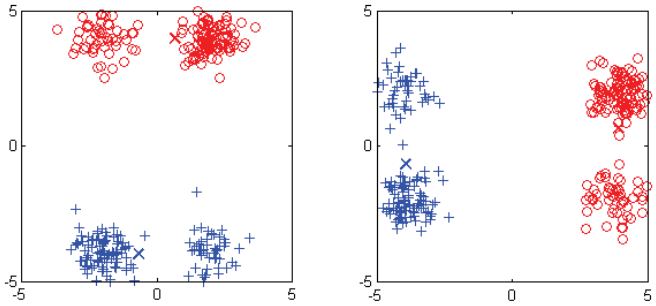
The computational complexity of single iteration is in the order of $O(pn)$ for k-means, and $O(p^2n)$ for LDA, where p is the feature dimensionality and n is the number of samples. For t iterations the complexity of CoKmLDA algorithm is hence $O(pnt + p^2nt)$.

5.3.2 An illustrative example

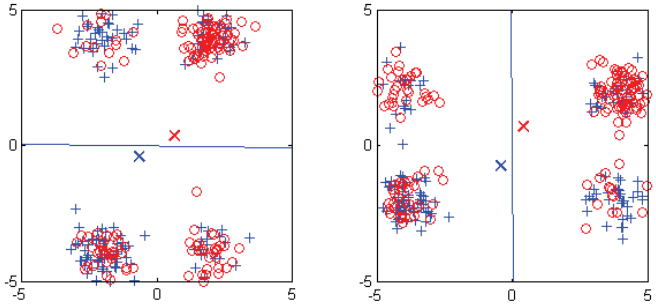
Here we illustrate the behaviour and merits of the proposed CoKmLDA algorithm using synthetic data of 300 samples represented in two views, each of two dimensions. Each sample



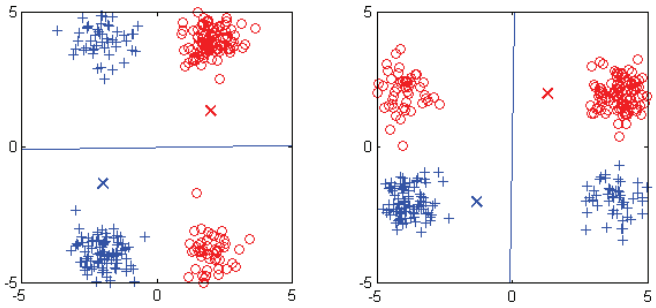
(a) data structure in two views



(b) k-means results on original space



(c) cross-labelling and LDA training



(d) k-means clustering in LDA space

Figure 5.1: illustration of a test run of CoKmLDA on synthetic dataset

Table 5.1: (x,y) locations of Gaussian centroids and number of samples

	class 1 (red)	class 2 (blue)
view 1	(-2,4) 50 smpl. (-4,-4) 100 smpl.	(2,4) 100 smpl. (2,-4) 50 smpl.
view 2	(-4,-2) 100 smpl. (4,-2) 50 smpl.	(-4,2) 50 smpl. (4,2) 100 smpl.

belongs to one of two classes, where each class is a two component Gaussian mixture. All four Gaussian components have a covariance matrix of $\Sigma = diag(0.3, 0.3)$. The means of each Gaussian component are detailed in Table 5.1. The two views are conditionally independent, i.e. two samples generated by the same Gaussian component in one view can belong to different Gaussian components in the other view. Finally, the number of samples corresponding to each Gaussian components, also illustrated in Table 5.1, is intentionally unbalanced in order that, for initialisation, k-means gives better-than-random accuracy relative to real class labels.

Scatter plots of generated data in 2 views are show in Figure 5.1 (a). Figure 5.1(b) illustrates clustering results after the initial application of k-means in the original feature space. We note a high degree of error; two of the four Gaussian components are incorrectly clustered. The result of cross-labelling, where samples in each view are labelled according to the clustering indicators in the other view, is shown in Figure 5.1 (c). The two crosses at centre of each plot represents the two cluster centroids in each view. The resulting LDA projections (1-dimensional for this trivial example) are shown by the solid, straight lines in Figure 5.1 (c). After the samples are projected into the new subspaces and k-means is reapplied, the clustering results are greatly improved as illustrated in Figure 5.1 (d). The new cluster centroids and LDA projections normally used in the second iteration are also illustrated in Figure 5.1 (d). Whereas several iterations are required in practice, the new clustering result is fully representative of the true underlying class structure and the algorithm converges in a single iteration for this illustrative example.

5.3.3 Mathmatical analysis

The above example illustrates the behaviour of the algorithm for a trivial example. Given the assumption of conditional independence between the two views, clustering indicators in one view can be utilised as class labels in another view, but with random label noise. Here we aim to show mathematically that *LDA projection can be learned with labelled samples*

with random label noise.

We first define a hypothetical level of label noise λ. Let there be n centred data samples, $X = [x_1, \ldots, x_n]$ and let X_k and n_k be the subset and number of samples in the k-th class respectively. For each class, $(1-\lambda)n_k$ and λn_k points are randomly sampled from X_k and $X - X_k$ respectively to form a new subset X_k^* for the k-th class with random label noise. In the following we show that the expected LDA projection trained with X_k^* is equivalent to the LDA projection trained with true labels.

Trained on X_k^* with noisy labels, the LDA projection P is determined according to:

$$\max_P \mathrm{Tr} \frac{P^T S_b^* P}{P^T S_t^* P} \tag{5.8}$$

where S_b^* and S_t^* are the between-class and total scatter estimated with noisy data. It is clear that $S_t^* = S_t$ since its calculation do not need label information, whereas S_b is defined as:

$$S_b^* = \sum_k n_k \mathbf{m_k^*} \mathbf{m_k^*}^\mathbf{T}, \tag{5.9}$$

where $\mathbf{m_k^*}$ is the mean of X_k^*. Its value in the sense of statistical expectation is given by:

$$\begin{aligned} \mathbb{E}(\mathbf{m_k^*}) &= \mathbb{E}(\frac{1}{n_k}(\sum_{i=1}^{(1-\lambda)n_k} x_{ki}^+ + \sum_{i=1}^{\lambda n_k} x_{ki}^-)) \\ &= (1-\lambda)\mathbb{E}(x_{ki}^+) + \lambda \mathbb{E}(x_{ki}^-) \end{aligned} \tag{5.10}$$

where x_{ki}^+ is the i-th sample from X_k and x_{ki}^- is the i-th sample from $X - X_k$. It is straightforward that

$$\begin{aligned} \mathbb{E}(x_{ki}^+) &= \mathrm{mean}(X_k) &&= \mathbf{m_k}, \\ \mathbb{E}(x_{ki}^-) &= \mathrm{mean}(X - X_k) &&= -\tfrac{n_k}{n-n_k}\mathbf{m_k} \end{aligned} \tag{5.11}$$

Combining Equation 5.10 and 5.11, we obtain:

$$\mathbb{E}(\mathbf{m_k^*}) = (1 - \frac{\lambda n}{n - n_k})\mathbf{m_k} \tag{5.12}$$

From Equation 5.12 we observe that the expected value of m_k^* estimated with X_k^* containing noisy labels lies in the same direction relative to the origin as in the case where it is estimated with clean labels, but with a shorter vector norm. This can be observed in Figure 5.1 (c) and (d) in which the two class means in each view lie in the same direction, but different distances from the origin.

Upon substitution of Equation 5.11 into Equation 5.9, we obtain the expectation of S_b^*:

$$\begin{aligned}\mathbb{E}(S_b^*) &= \sum_k n_k \mathbb{E}(\mathbf{m_k^*}) \mathbb{E}(\mathbf{m_k^*}^T) \\ &= \sum_k n_k (1 - \frac{\lambda n}{n - n_k})^2 \mathbf{m_k} \mathbf{m_k}^T \end{aligned} \tag{5.13}$$

If we assume an equal number of sample per class, i.e. a constant $n_k = n/K$, then:

$$\mathbb{E}(S_b^*) = (1 - \frac{\lambda K}{K-1})^2 S_b \tag{5.14}$$

and if S_b^* and S_t^* in Equation 5.8 are replaced with their expectated values, we obtain:

$$\max_P \mathrm{Tr} \frac{C^2 P^T S_b P}{P^T S_t P} \tag{5.15}$$

where $C = (1 - \frac{\lambda K}{K-1})$ is a constant. Equations 5.15 shows that LDA objective function in the case of sample with random label noise is equivalent to the objective function in the case of clean labels.

5.3.4 Extensions of CoKmLDA

In this chapter we present the idea of unsupervised subspace clustering using co-training. The framework is entirely flexible and may combine different clustering methods and supervised dimensionality reduction algorithms according to specific application and nature of related data. For example, to cluster text data, cosine distance is a more appropriate distance measure, and for non-linear separable data, kernel methods are often applied. In this section, we first presents three extensions related to clustering, namely cosine k-means, kernel approach and semi-supervised extension. We also provide multi-view extension to adapt to the situation where the data is represented by more than two views.

Cosine distance extension: The standard k-means algorithm uses a Euclidean distance metric. In some experiments in multi-modal face and speaker recognition, however, we observe that the cosine distance normally gives better performance when used in LDA subspace. Tang et al. [Tang et al., 2012] report similar findings in the context of speaker clustering. The use of a cosine distance metric in clustering problems is proposed in [Dhillon and Modha, 2001] which reports a spherical k-means algorithm which maximizes the sum of the cosine similarity between samples and related cluster centroids. Spherical k-means follows a similar iterative process as standard k-means, except that feature vectors are first normalized to have unit length and, in the assignment step, samples are assigned to the cluster centroid which has the highest cosine similarity. The power of spherical k-

means clustering is brought to CoKmLDA simply by replacing the standard k-means step in Algorithm 4.

Kernel extension: LDA learns a subspace in which classes are better separated. In the event that classes are not linearly separable in the original space, then performance is usually poor. Using a kernel trick similar to that employed in Support Vector Machines (SVM), LDA can be implicitly performed in a new feature space, which allows non-linear mappings to promote maximum separability of different classes. This approach is commonly referred to as Generalized Discriminant Analysis (GDA) [Baudat and Anouar, 2000]. By replacing standard LDA by GDA, the proposed algorithm may be applied to clustering problems in which multi-view data is not linearly separable.

Semi-supervised extension: The algorithm is also readily extended to semi-supervised clustering when a subset of manually labelled data in addition to a larger subset of unlabelled data are available. In this case the initial k-means step uses centroid statistics acquired from the manually labelled data as proposed in [Basu et al., 2002]. In our approach the k-means algorithm is seeded in each iteration with labelled data. In the case where the number of classes is high, and where random initialization often generates several seeds corresponding to some classes whereas none for others, this seemingly naive method often brings significant improvements in performance in our framework. The proposed algorithm simultaneously determines discriminant subspaces in addition to compact cluster/class models and is naturally equipped to handle out-of-sample data. Unseen test data can be projected into the relevant subspaces and classified according to the nearest centroid.

Multi-view extension: Finally, it is possible to extend the proposed two-view CoKmLDA algorithm to multi-view clustering. Assuming that each data sample is represented by m-views ($m > 2$), subsequent to the initialization and each iteration in Algorithm 4, m sets of cluster indicators are generated. In the two-view setting, an LDA projection in one view is learnt using cluster indicators in the other view to enforce a similar data structure in each subspace. Extending to an m-view setting, a straight forward solution involves the learning of an LDA projection in one view using cluster indicators of *all other views* as class labels.

Traditional LDA accepts only a single label vectors. In order to deal with multiple label vectors, the traditional LDA algorithm is modified as follows. Assume a set of centred input data $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and m sets of class indicators $\{H^{(1)}, \ldots, H^{(m)}\}$. We first calculate the within-class scatter $S_w^{(v)}$ and the between-class scatter $S_b^{(v)}$ using each class indicator $H^{(v)}$ according to Equation 5.1. The overall between-class scatter S_b and within-class

scatter S_w are then defined as:

$$S_b = \sum_{v=1}^{m} S_b^{(v)}; \quad S_w = \sum_{v=1}^{m} S_w^{(v)}. \tag{5.16}$$

Finally, the optimal projection P is obtained in the same way as for the traditional LDA by optimizing the objective function in Equation 5.2. Despite the different formulation, this method is have similar effect to the Multi-label Linear Discriminant Analysis (MLDA) proposed in [Wang et al., 2010]. The proposed method is still referred to as MLDA for simplicity.

To conclude, multi-view CoKmLDA differs from the two-view CoKmLDA in that in step 1 of the iterative process of Algorithm 4, an MLDA projection $P^{(v)}$ is learnt using $X^{(v)}$ and the cluster indicators $\{H^{(1)}, \ldots, H^{(v-1)}, H^{(v+1)}, \ldots, H^{(m)}\}$ from all other views, while all other operations remain the same.

5.4 Related works and analysis

Several different approaches to subspace and multi-view clustering have been reported in the open literature. Here we discuss their relationship with the new algorithm proposed in this chapter.

Subspace clustering [Kriegel et al., 2009] refers to a general class of clustering methods which aim to discover a subspace more amenable to clustering. These methods are largely uni-modal. Among numerous other examples, the most relevant to the proposed algorithm are the LDA-Km algorithm [Ding and Li, 2007] and DisKmeans [Ye et al., 2007] which use cluster indicators generated by k-means to learn an LDA projection. As a form of self-training, such approaches do not generally lead to significant improvements in clustering performance over a baseline k-means. The proposed CoKmLDA algorithm can be regarded as a co-training extension of [Ding and Li, 2007].

In the multi-view clustering setting, the general objective is to find certain kind of agreement between different views. Recent approach to multi-view clustering can be roughly divided into two major categories. The first category of algorithms is multi-view spectral clustering based on similarity graphs. As shown in Figure 5.2 (a), a similarity graphs (matrix) $S^{(v)}$ is first constructed for each view $X^{(v)}$ where $S_{ij}^{(v)} = exp(< x_i, x_j >^2 / t^2)$, where $< . >$ is a certain distance measure and t is the Gaussian bandwidth, thus $S_{ij}^{(v)}$ represents the similarity between i-th and j-th sample in the v-th view. The original similarity graphs $S^{(v)}$ are then transformed so that the difference between the transformed similarity graphs $S^{*(v)}$s is minimized across each view. Such transformations can be learnt by different ap-

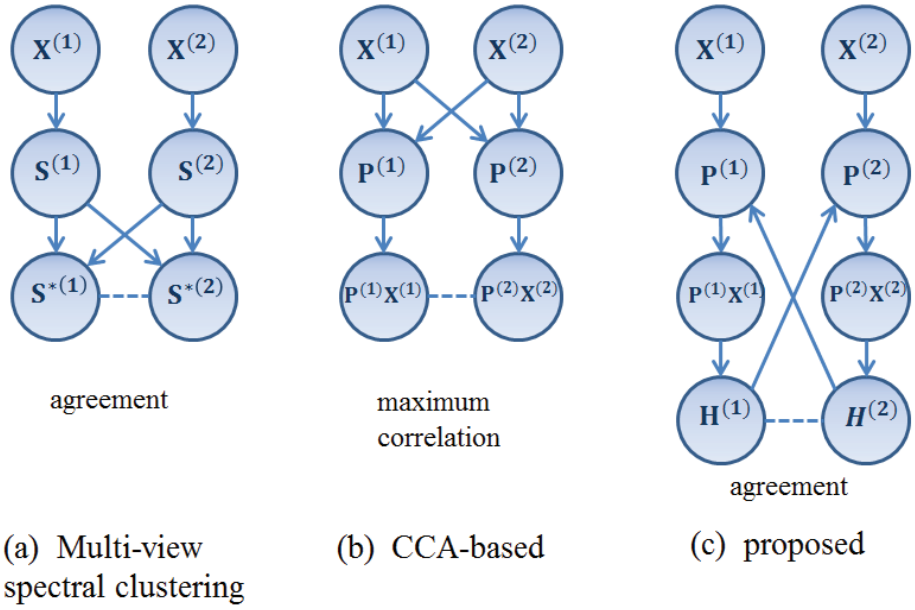


Figure 5.2: Working flowchart of different multi-view clustering algorithms

proaches such as Min-Disagreement [de Sa, 2005], co-training [Kumar and Daumé III, 2011] or co-regularization [Kumar et al., 2011]. Finally, standard spectral clustering [Ng et al., 2002] can be applied to the transformed graph of the most informative view to obtain the final clustering result. This class of algorithms is related to the CoKmLDA in the sense that both approaches aim to identify clusters which are in consensus across each view, such that pairs of samples which are considered similar in one view should be considered similar in other views. However, the disadvantage of this class of algorithms is that, features $X^{(v)}$ are not used again after the $S^{(v)}$ is built. In the case that original features $X^{(v)}$ contain substantial number of noisy dimensions which are irrelevant to underlying classes, the estimation of $S^{(v)}$ is intrinsically inaccurate, thus improvements from graph fusion can be sub-optimal.

The second category of clustering approaches based on Canonical Correlation Analysis (CCA), i.e.[Chaudhuri et al., 2009, Blaschko and Lampert, 2008] aim to cope with multi-view, high-dimensional data. As illustrated in Figure 5.2(b), the general idea involves jointly learning projections $P^{(1)}$ and $P^{(2)}$ with $X^{(1)}$ and $X^{(2)}$ such that the correlation between the projected samples in two views are maximized. Standard clustering algorithms such as k-means can then be applied to projected samples. The objective function is formulated

as:

$$\underset{P^{(1)},P^{(2)}}{\arg\max} \frac{\mathbf{E}(P^{(1)}X^{(1)})\mathbf{E}(P^{(2)}X^{(2)})}{\sqrt{\mathbf{E}(P^{(1)}X^{(1)})^2\mathbf{E}(P^{(2)}X^{(2)})^2}} \tag{5.17}$$

We foresee two disadvantages of CCA based algorithms. First, according to the analysis of [Chaudhuri et al., 2009], CCA learns a low dimensional subspace spanned by the means of different clusters (equivalent to the maximization of S_b). However, same cluster samples are not necessarily projected near to each other (minimization of S_w). Second, CCA-based methods rely strongly on the conditional independence assumption, which may not hold in practical problems. According the experimental work of [Kumar and Daumé III, 2011] and [Kumar et al., 2011], CCA-based method performs poorly when there is some dependence between views; this can be expected from Equation 5.17. In the worst case, if $X^{(1)}$ and $X^{(2)}$ are fully correlated ($X^{(1)} = \alpha X^{(2)}$), any projections $P^{(1)} = P^{(2)}$ will maximize the objective function to its maximum value 1.

The framework of proposed CoKmLDA algorithm is illustrated in Figure 5.2(c). CoKmLDA requires a maximum agreement between clustering results $H^{(1)}$ and $H^{(2)}$ on projected views $P^{(1)}X^{(1)}$ and $P^{(2)}X^{(2)}$. Compared to graph-based multi-view clustering algorithms, CoKmLDA reduces noise existed in features $X^{(1)}$ and $X^{(2)}$ through the iterative learning of projections $P^{(1)}$ and $P^{(2)}$ whereas graph-based methods reduce noises in similarity graphs. Compared to CCA-based multi-view clustering algorithms, CoKmLDA directly requires maximum agreement of clustering results rather than maximum correlation in projected spaces. Moreover, CoKmLDA is less sensitive to the view dependency. After all, CoKmLDA algorithm is equivalent to single-subspace clustering algorithm LDA-Km proposed in [Ding and Li, 2007]. Finally, as we will shown in Section 5.5.3, CoKmLDA can exploit the existed independence between views even if it is weak.

5.5 Experiments and discussions

In this section, we evaluate the effectiveness of the proposed algorithm on 3 independent datasets under 2 scenarios, when the conditional independence assumption is fully satisfied or only partially satisfied. For the former, its performance is assessed with audio-visual person clustering based on facial and speech features on the MOBIO database [1][McCool et al., 2012]. For the later, we report its application to image clustering using the UCI handwrit-

[1] https://www.idiap.ch/dataset/mobio

ten digit dataset [2], and text document clustering using BBC News Synthetic multi-view text dataset [3]. Note that the complexity of the proposed CoKmLDA algorithm grows linearly with the square of feature length (as discussed in Section 5.3.1), so for the efficiency of computation, in all experiments, all features are reduced to 100 dimensions by a PCA preprocessing step. All results are averaged by across independent trials of random initialization.

The performance of CoKmLDA is compared to four baseline systems: conventional k-means in PCA space, the LDA-Km single-view subspace clustering algorithm [Ding and Li, 2007] and two other recently proposed multi-view clustering algorithms, namely Canonical Correlation Analysis (CCA) [Chaudhuri et al., 2009] and co-training spectral clustering (CoSC) [Kumar and Daumé III, 2011]. These latter two algorithms represents the two different approaches to multi-view clustering algorithms discussed in Section 5.4.

5.5.1 Evaluation metrics

The clustering performances of the proposed CoKmLDA algorithm and other compared methods are assessed using two different metrics. The first is referred to as the clustering accuracy and is given by:

$$CA = \frac{1}{n}\sum_{i=1}^{n}\delta(h_i, \mathrm{map}(l_i)) \tag{5.18}$$

where n is the number of samples, h_i is the cluster indicator estimated for the i-th sample, l_i is the corresponding true label, and $\delta(a,b)$ is a function which returns 1 if $a = b$ and 0 otherwise. The *map*() function represents the mapping between cluster indicators and true labels as determined according to a Hungarian algorithm [Steiglitz and Papadimitriou, 1982].

The normalized mutual information (NMI) [Strehl and Ghosh, 2003] is another popular clustering metric derived from information theory and given by:

$$NMI = \frac{I(H,L)}{\sqrt{E(H)E(L)}} \tag{5.19}$$

where $I(H,L)$ is the mutual information between H and L and $E(H)$ and $E(L)$ are the respective entropies. The NMI lies between 0 and 1 and larger values indicate more accurate clustering indicators. Please see [Strehl and Ghosh, 2003] for more details.

[2] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[3] http://mlg.ucd.ie/datasets/segment.html

5.5.2 Audio-visual speaker clustering (conditional independence assumption satisfied)

We first evaluate the effectiveness of the proposed algorithm through experiments in audio-visual speaker clustering. In this case, each view is conditionally independent and represented with features of high dimensionality. Facial features are corrupted by inter-session variations such as illumination, expression and pose whereas vocal features are corrupted by different phonemes pronounced in a short speech episode, which are expected to be independent from each other.

Database and feature extraction

We consider speaker clustering using speech and facial images. Experiments are conducted with the same MOBIO database [McCool et al., 2012] as presented in Section 4.4.1, which contains videos of 150 subjects captured in real-world, challenging conditions. For computational efficiency, we test our algorithm using a subset of data from 40 male subjects and for each of them, 5 videos are selected from each of the 12 sessions. This results in a pool of 2400 video samples.

We use cropped face images provided with the MOBIO database, one image per video sample. All images are resized to 50×43 pixels and then histogram equalized. Rows of pixel intensities are concatenated to form feature vectors of 2150 dimensions. The speech signal is split into frames of 20ms duration before the extraction of features composed of 26 Mel-scaled frequency cepstral coefficients (MFCCs), their 26 derivatives and the delta energy. Energy-based voice activity detection is then applied to discard non-speech frames. A 64-component Gaussian mixture model (GMM) is then fitted to remaining speech data through the maximum a posteriori (MAP) adaptation of a speaker-independent world model. The means of the GMM model are then concatenated to form a 3392-dimensional GMM super-vector [Reynolds et al., 2000a]. Both face and speech feature vectors are first reduced to 100 dimensions through the application of PCA.

Results

The performance of CoKmLDA is assessed in terms of clustering accuracy and NMI. The proposed CoKmLDA algorithm and all compared methods require the expected number of clusters K as an input parameter, which is set to be the number of subjects. In our experiments we observed that, for all linear subspace methods (PCA, LDA-Km, CCA and CoKmLDA), the use of cosine-distance-based spherical k-means [Dhillon and Modha, 2001]

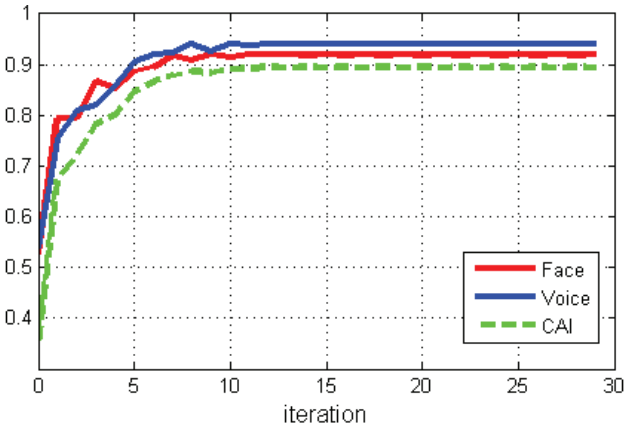


Figure 5.3: Clustering accuracy for face modality (red), voice modality (blue) and CAI score (green) v.s. number of iterations of co-training

consistently out-performs Euclidean-distance-based k-means. As a result, we report the results obtained with spherical k-means in these methods whereas for CoSC, we report the results obtained with conventional k-means which achieves the best performance in this case.

Table 5.2 summarizes the mean and standard deviation of the clustering accuracy and NMI score obtained with 20 different runs of k-means with random initialization. It is observed that, for both metrics, multi-view clustering methods CCA, CoSC and CoKmLDA perform significantly better than the PCA baseline, whereas the single view LDA-Km method only gives modest improvements over the PCA baseline. Finally, the proposed CoKmLDA algorithm outperforms the closest-performing method CCA by a significant margin (over 10% gain in clustering accuracy and approximately 5% in NMI). Figure 5.3 shows the variation in accuracy and CAI scores as a function of the number of iterations. Convergence is seen to occur in fewer than 15 iterations. In practice we have not encountered any cases where convergence does not occur.

Clustering visualisations and discussion

All the approaches compared above embed data samples into lower dimensional spaces in which clustering is then performed with a standard k-means algorithm. PCA, LDA-Km, CCA, and the proposed CoKmLDA algorithm embed original data into linear subspaces, while co-training spectral clustering embeds data samples into the first K eigenvectors of the graph Laplacian [Kumar and Daumé III, 2011]. It is informative to visualize

Table 5.2: Performance comparison

(a) mean and std. of clustering accuracy

Accuracy(std.)	Face	Speech
PCA	0,530(0,029)	0,512(0,021)
LDA-Km [Ding and Li, 2007]	0,712(0,042)	0,688(0,045)
CCA [Chaudhuri et al., 2009]	0,825(0,046)	0,798(0,050)
CoSC [Kumar and Daumé III, 2011]	0,785(0,036)	0,799(0,039)
CoKmLDA	**0,934(0,029)**	**0,910(0,024)**

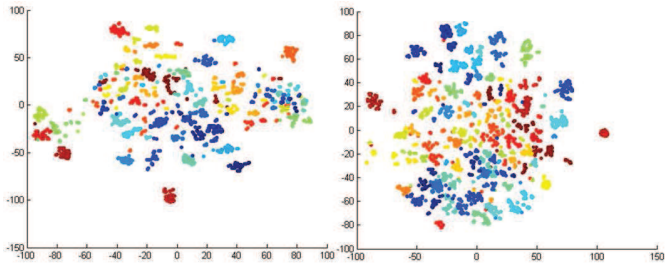
(b) mean and std. of NMI score

NMI(std.)	Face	Speech
PCA	0,665(0,013)	0,667(0,011)
LDA-Km [Ding and Li, 2007]	0,842(0,022)	0,815(0,023)
CCA [Chaudhuri et al., 2009]	0,915(0,018)	0,924(0,018)
CoSC [Kumar and Daumé III, 2011]	0,895(0,011)	0,895(0,009)
CoKmLDA	**0,970(0,008)**	**0,959(0,009)**

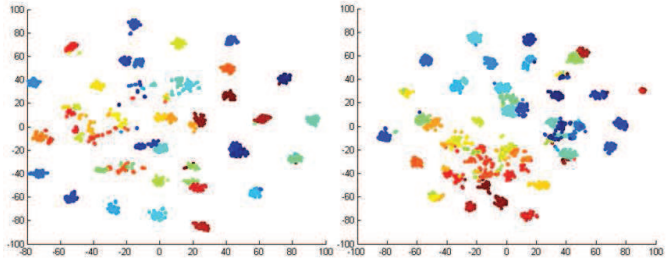
the embedded data structure and thus to observe the relationship between the embedded structure and clustering performance. However, the embedded subspaces are still high-dimensional and cannot be visualized directly. T-distributed Stochastic Neighbour Embedding (t-SNE) [van der Maaten and Hinton, 2008] is a powerful tool used to visualize high-dimensional data via the embedding of data into a 2-D or 3-D space while respecting relative distances between data samples.

Figure 5.4 illustrates 2-D scatter plots of projected data for PCA, LDA-Km, CCA, CoSC and CoKmLDA after the application of t-SNE. In all cases, samples belonging to different classes are represented by different colours. The features processed by PCA is shown in Figure 5.4(a). The sample distribution is especially noisy which explains the poor clustering performance. In Figure 5.4(b), clearer cluster structures are observed in LDA-Km subspaces but the confusion between several classes is still high, due to its single-view nature. In CCA subspaces (Figure 5.4(c)), cluster structure is not visually obvious. Same-class samples are approximately located in one single Gaussian distribution, but the variance is relatively high, since CCA does not minimize with-in class scatter, as discussed in Section 5.4. Both CoSC and the proposed CoKmLDA produce large between-class/within-class scatter ratio, as shown in Figure 5.4 (d) and (e) respectively. However, the clustering purity in CoKmLDA subspaces is significantly better.

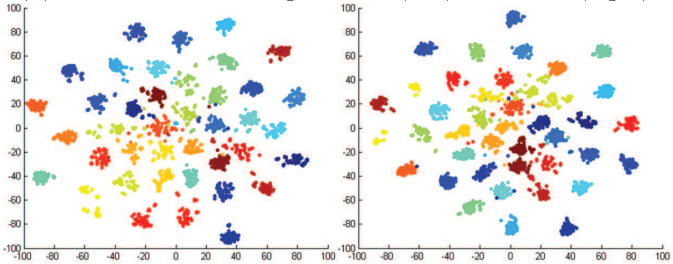
In the following we address some potential anomalies in the reported results. Figure 5.4



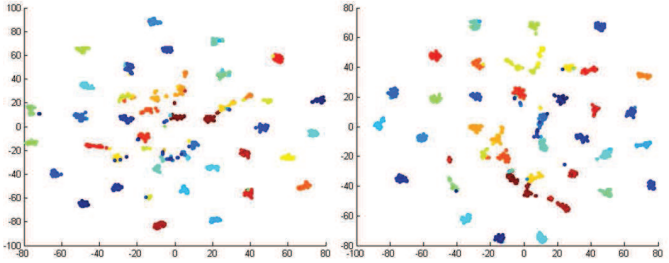
(a) PCA embeddings for face(left) and voice(right)



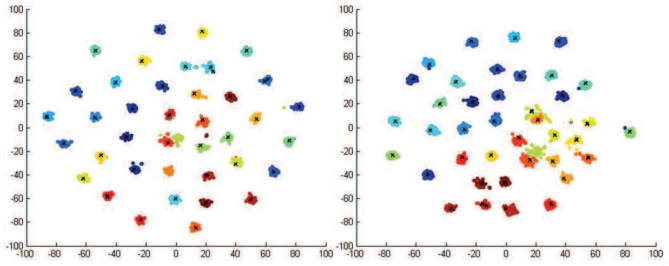
(b) LDA-Km embeddings for face(left) and voice (right)



(c) CCA embeddings for face(left) and voice (right)



(d) CoSC embeddings for face(left) and voice (right)



(e) CoKmLDA embeddings face(left) and voice (right). Black crosses represents k-mean seeds inherited

Figure 5.4: 2-D t-SNE visualizations of data structures for PCA, CCA, CoSC, and CoKmLDA subspaces. Different subjects/classes are represented by different colors.

and Table 4.1 show that, while CoSC gives better cluster separation, performance is worse than that of CCA. Even though CoSC produces a subspace in which different clusters are better separated, the data structure produced with CCA is cleaner with respect to the true labels. However, with a better separated cluster structure, more sophisticated initialization method for the k-means algorithm may deliver improved clustering performance.

It is also of interest to reflect on the reasons why CoKmLDA delivers such significantly better performance than other approaches. We attribute the superior performance of CoKmLDA to two main factors. First, CoKmLDA learns discriminative subspaces in which the cluster structure is in agreement for each view. Insodoing, the influence of feature dimensions which are unrepresentative of the underlying class structure is greatly reduced. Second, as discussed in Section 5.3.1, seeds used for k-means in each iteration are inherited from samples closest to the K centroids identified in the preceding iteration and the algorithm tends to give one seed per compact cluster. This fact is shown in Figure 5.4(e), where the black crosses represents the seeds of k-means automatically learnt by CoKmLDA algorithm.

5.5.3 Handwritten digit clustering and text document clustering (Conditional independence assumption not fully satisfied)

Co-training-style algorithms generally assume the conditional independence between the multiple features in use. However, in many practical problems, this assumption is not fully validated. As opposed to different features from different sources (as with visual and audio sources in the previous example), when both features come from the same source, they are expected to be correlated to some extent. To assess the CoKmLDA algorithm in such settings, we report further experiments with the clustering of image-only and text-only documents.

Databases

The proposed algorithm is assessed using two different databases: the UCI handwritten digits dataset [4] for image clustering with different features, and the BBC News Synthetic multi-view text dataset [5] for text document clustering.

The UCI handwritten digits dataset consists of images of handwritten numerals ('0'–'9') extracted from a collection of Dutch utility maps. Some sample images are shown in

[4] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[5] http://mlg.ucd.ie/datasets/segment.html

(a) Visualization of some sample images.

(b) Feature extraction.

Figure 5.5: Sample images and feature extraction for UCI handwritten digits dataset

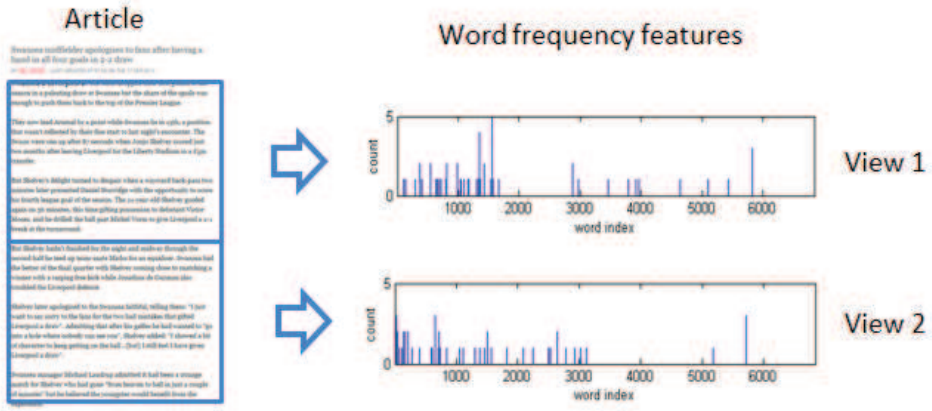


Figure 5.6: Feature extraction and view assignment for BBC dataset. An article is divided into two segments. Word frequency features are calculated for each segment and used as two input views.

Figure 5.5 (a). 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. The database provides multiple type of features extracted from the images. We used the 76-dimensional Fourier coefficients (Fou) as view 1 and the 216-dimensional profile correlations (Fac) as view 2. In order to assess the effectiveness of the extension of CoKmLDA to more than two views, we further choose the pixel intensities (Pix) as the third view, which is a 240-dimensional feature vector. Note that the first view is intrinsically less informative than the two others since the Fourier coefficients are rotation invariant hence cannot distinguish between the digit '6' and the digit '9'. Both profile correlation and pix intensity features are reduced to 100 dimensions by PCA as in audio-visual person clustering experiment, while for Fourier coefficient features are only pre-processed by removing the mean since it has only 76 dimensions.

The BBC News synthetic multi-view text dataset consist of term frequency features from news articles from the BBC [Greene and Cunningham, 2005], as illustrated in Figure 5.6. BBC data contains 2225 complete news articles corresponding to stories in five topical areas (business, entertainment, politics, sport and technology). Each document is segmented into two parts, and word frequency features are computed from each part, which constitute the two views (seg1of2 & seg2of2)[Greene and Cunningham, 2009]. The feature dimension is 6838 and 6790 for the two views, respectively. Both features are reduced to 100 dimensions by PCA.

Results and discussions

Clustering performance is again assessed in terms of clustering accuracy and NMI. Table 5.3 shows results for the UCI handwritten dataset. The number in the parentheses (2 or 3) after CoSC and CoKmLDA indicates the number of views used in the algorithm. Since [Chaudhuri et al., 2009] does not provide an extension to more than two views for CCA-based clustering, in this case results are reported for the first two views only. The proposed CoKmLDA algorithm gives the best performance among all methods compared in terms of both metrics. For the two-view setting, and the most informative view (Fac), the single-view LDA-Km algorithms performs closest to the CoKmLDA algorithm. However, CoKmLDA is still successful in utilizing information in the first view (Fou), even if it is less informative, and performs marginally better than the single-view algorithm. This observation shows that the proposed algorithm can exploit even-marginal independence between views. CCA only provides marginal improvements over the PCA baseline, due to its sensitivity to correlated features. When the additional third view is used, the CoKmLDA algorithm gives a further 12% increase in terms clustering accuracy for the most informative view (Fac), which demonstrates the effectiveness of the multi-view extension proposed in Section 5.3.4.

Table 5.4 summarizes results for BBC News dataset. The proposed algorithm still outperforms all the compared methods and the co-spectral-clustering algorithm is the second-best performing algorithm. Note that the CCA method performs even worse than the PCA baseline. These results confirm the analysis in Section 5.4 that CCA method strongly relies on the assumption of conditional independence between views and is risky to use when this assumption no longer holds. The proposed CoKmLDA algorithm, on the other hand, is more reliable when the conditional independence assumption is weak.

Table 5.3: Performance comparison on UCI Handwritten digits dataset

(a) mean and std. of clustering accuracy.
Number (2) or (3) indicates the number of views used in the approach.

Accuracy(std.)	View 1 (Fou)	View 2 (Fac)	View 3 (Pix)
PCA	0.525(0,029)	0,603(0,032)	0.601 (0.034)
LDA-Km [Ding and Li, 2007]	0,576(0,042)	0,750(0,045)	0.771 (0.043)
CCA [Chaudhuri et al., 2009]	0,542(0,031)	0,644(0,030)	N.A.
CoSC (2) [Kumar and Daumé III, 2011]	0,702(0,036)	0,748(0,034)	N.A.
CoSC (3) [Kumar and Daumé III, 2011]	0,740(0,037)	0,772(0,032)	0,764 (0,035)
CoKmLDA (2)	**0,725(0,029)**	**0,761(0,045)**	N.A.
CoKmLDA (3)	**0,720(0,024)**	**0,892(0,046)**	**0,845 (0,044)**

(b) mean and std. of NMI score.
Number (2) or (3) indicates the number of views used in the approach.

NMI(std.)	View 1 (Fou)	View 2 (Fac)	View 3 (Pix)
PCA	0,603(0,027)	0,651(0,026)	0,642(0,025)
LDA-Km [Ding and Li, 2007]	0,677(0,039)	0,798(0,042)	0,804(0,041)
CCA [Chaudhuri et al., 2009]	0,647(0,031)	0,687(0,027)	N.A.
CoSC (2) [Kumar and Daumé III, 2011]	0,752(0,011)	0,774(0,023)	N.A.
CoSC (3) [Kumar and Daumé III, 2011]	0,773(0,021)	0,793(0,027)	0,782(0,025)
CoKmLDA (2)	**0,769(0,042)**	**0,810(0,033)**	N.A.
CoKmLDA (3)	**0,759(0,041)**	**0,852(0,036)**	**0,844(0,039)**

Table 5.4: Performance comparison on BBC News dataset

(a) mean and std. of clustering accuracy

Accuracy(std.)	View 1 (seg1of2)	View 2 (seg1of2)
PCA	0.852(0,049)	0,863(0,042)
LDA-Km [Ding and Li, 2007]	0,877(0,024)	0,882(0,023)
CCA [Chaudhuri et al., 2009]	0,725(0,031)	0,746(0,028)
CoSC [Kumar and Daumé III, 2011]	0,886(0,021)	0,887(0,035)
CoKmLDA	**0,912(0,029)**	**0,915(0,036)**

(b) mean and std. of NMI score

NMI(std.)	View 1 (seg1of2)	View 2 (seg2of2)
PCA	0,701(0,027)	0,713(0,026)
LDA-Km [Ding and Li, 2007]	0,762(0,019)	0,755(0,021)
CCA [Chaudhuri et al., 2009]	0,688(0,031)	0,692(0,027)
CoSC [Kumar and Daumé III, 2011]	0,762(0,019)	0,775(0,021)
CoKmLDA	**0,788(0,032)**	**0,803(0,035)**

5.6 Summary

In this chapter, we presents our second MVDR framework for clustering multi-view, high-dimensional data. It applies the results of unsupervised clustering in one view to learn discriminant subspaces in another. The general framework assumes conditionally independent views. We show, however, that the new algorithm still performs well when the conditional independence is weak. Furthermore, the framework is straightforward and combines well-known, even trivial algorithms to positive effect. The chapter also presents a theoretical treatment which shows how LDA projections learned from samples with random label noise are equivalent to those learned with entirely clean labels and that the cross-view labelling, or co-training, is efficient in correcting erroneous sample labels. Experiments in audio-visual speaker clustering, multi-view handwritten digit clustering and text document clustering demonstrate the effectiveness of our algorithm and superior performance to existing state-of-the-art approaches.

In the end of this chapter, we would like to highlight the following conclusions:

1. The CoKmLDA algorithm proposed in this chapter extends the semi-supervised Co-LDA algorithm proposed in the previous chapter into a purely unsupervised setting, which is in particular suitable for multi-view clustering applications.
2. Compared to the semi-supervised MVDR framework based on co-training proposed in Chapter 4, the unsupervised MVDR framework proposed in this chapter adopts a modified co-training scheme: instead of increasing the size of labelled dataset at each iteration, it enforces a similar clustering structure across each view.
3. In this chapter, we also present a theoretical treatment which shows how LDA projections learned from samples with random label noise are equivalent to those learned with entirely clean labels and that the cross-view labelling, or co-training, is efficient in correcting erroneous sample labels.
4. In this chapter, we also experimentally show that, although co-training assumes a conditional independence between different views, the proposed algorithm is still reliable when this assumption is partially violated.
5. We also provide an extension of the proposed MVDR algorithm to more than two views.
6. Despite that the CoKmLDA algorithm is initially designed to cluster multi-modal biometric data, it equally works for non-biometric problems, such as text clustering and image clustering.

Chapter 6

MVDR by Subspace Graph Agreement

In Chapter 5, we presented our CoKmLDA algorithm, which is an unsupervised MVDR approach to clustering high-dimensional, multi-view data. Note that this MVDR approach is designed specifically for clustering problems, but is not well adapted to biometric retrieval problems discussed in Section 2.2.2. CoKmLDA algorithm aims to learn a subspace projection for each view, such that the clustering results are in maximal agreement across each view. In retrieval problems, however, there is no explicit concept of clusters. Moreover, the CoKmLDA algorithm is based on LDA, which is only optimal for Gaussian distributions and cannot capture complex, non-Gaussian data structures. On the other hand, similarity graphs are widely used to encode similarity relationships between data samples and are able to deal with non-linear data structures. In this chapter, we propose our third MVDR framework based on similarity graphs, which aims to learn a subspace projection for each view, such that the similarity graphs built in subspaces of different views maximally agree with each other. This framework is referred to as *MVDR by subspace graph agreement*.

6.1 Motivations

Graph-based dimensionality reduction methods have recently emerged as a powerful tool for analysing high-dimensional data. Example algorithms include ISOMAP [Tenenbaum et al., 2000], Local Linear Embedding (LLE) [Roweis and Saul, 2000], Laplacian eigenmaps [Belkin

and Niyogi, 2001], etc., and spectral clustering [Ng et al., 2002] can also be regarded as a type of graph-based dimensionality reduction technique. Given a set of n data points $\{x_1, \ldots, x_n\}$, these methods begin with the construct of a local similarity graph S, with each vertex in this graph represents a data point. Two vertices are connected if the two data points x_i and x_j are considered "close", and the edge is weighted by s_{ij}. A typical way to define such a graph involves a k Nearest Neighbour (kNN) graph with either a binary weigh:

$$S_{ij} = \begin{cases} 1, \text{ if } x_i \in kNN(x_j) \text{ or } x_j \in kNN(x_i) \\ 0, \text{ otherwise} \end{cases} \tag{6.1}$$

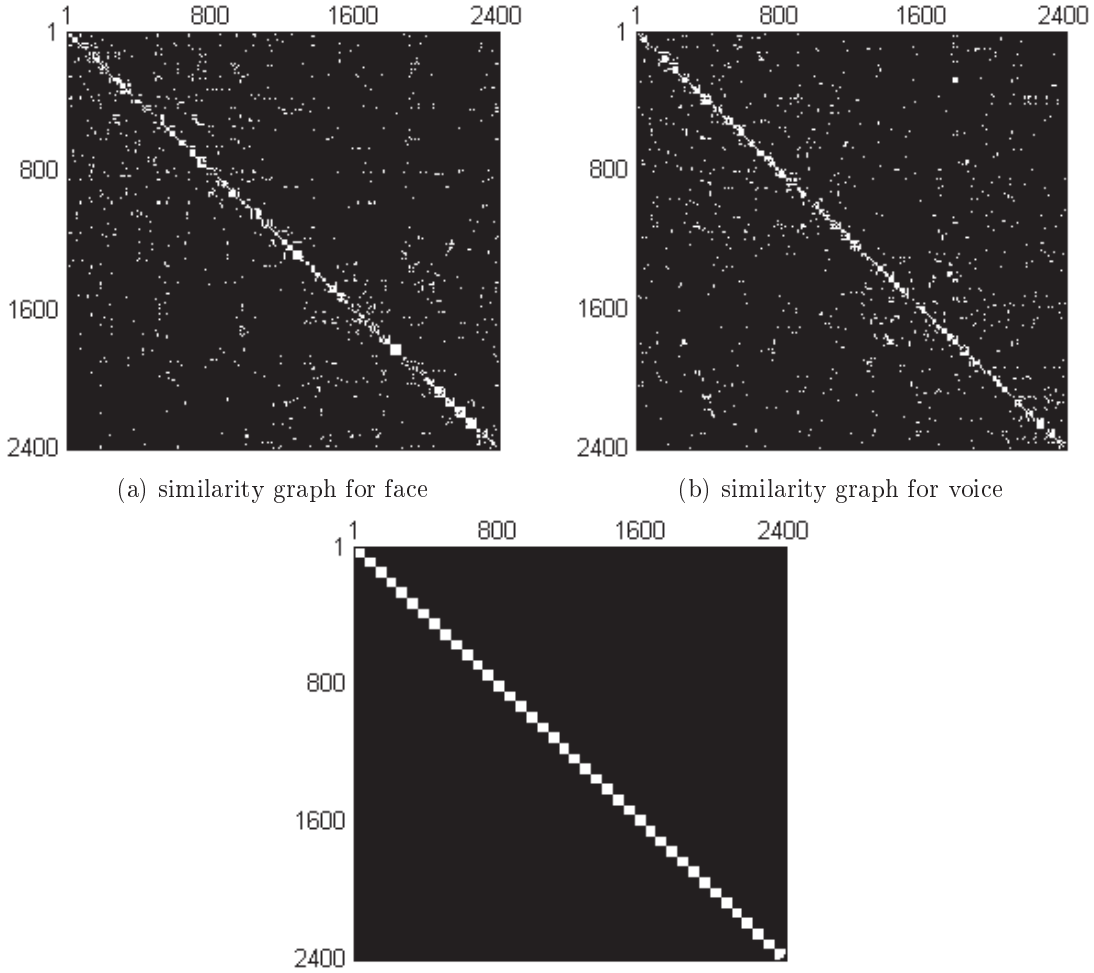
or heat kernel weight:

$$S_{ij} = \begin{cases} exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}), \text{ if } x_i \in kNN(x_j) \text{ or } x_j \in kNN(x_i) \\ 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{6.2}$$

where $kNN(\boldsymbol{x}_i)$ denotes the set of k nearest neighbours of sample $\boldsymbol{x}_i$.

Graph-based dimensionality reduction methods such as ISOMAP [Tenenbaum et al., 2000], Local Linear Embedding (LLE) [Roweis and Saul, 2000], and Laplacian eigenmaps [Belkin and Niyogi, 2001] aim to reveal the low-dimensional manifold structure of the original data, but are not capable of extracting class-specific discriminative features due to their unsupervised nature. In biometric problems, due to significant intra-class variations and possible small intra-class variations (as analysed in Section 2.3), the similarity between two samples could be very low in original spaces even if they belong to the same class. This can lead to missing links, and the similarity between some different-class samples can, thereby leading to wrong links. For illustration, we built a 59-NN binary weight similarity graph on visual and audio features of the the 2400 samples of MOBIO database described in Section 5.5.2. The results are shown in Figure 6.1 (a) and (b) respectively. Illustrated in Figure 6.1 (c) is a reference graph, in which every sample is linked only to all 59 other same-class samples. We observe that the graphs built on original visual and audio features contain significant amount of missing and wrong links.

We consider graph-based dimensionality reduction in a two-view setting. The two views exhibit a certain level of conditional independence, as is often the case with multi-modal biometrics. If similarity graphs are constructed with original features of the two views respectively, then they are expected to be different since the two views contain different intra-class variation, as shown in the case of audio-visual features in Figure 6.1(a) and (b). However, if we assume that there exist optimal discriminative projections $P_{opt}^{(1)}$ and $P_{opt}^{(2)}$ such



(a) similarity graph for face

(b) similarity graph for voice

(c) reference similarity graph

Figure 6.1: Bineary kNN Similarity graphs (k=59) of 2400 samples of 40 subjects from MOBIO database based on original facial (a) and vocal (b) features. (c) indicates a reference graph, where $s_{ij} = 1$ for all same-class samples and $s_{ij} = 0$ for all different-class samples

that in the two projected subspaces, same-class samples are located closed to each other where different-class samples are located apart, the two similarity graphs constructed with the two subspaces are expected to be similar, both close to the reference graph shown in Figure 6.1(c). Based on this logic, we propose to approximate $P_{opt}^{(1)}$ and $P_{opt}^{(2)}$ by finding $P^{(1)}$ and $P^{(2)}$ which minimize the difference between the two similarity graphs when constructed using projected samples in the two views.

In this chapter, we show that the objective of minimizing the difference between similarity graphs built within the subspaces of multiple views can be achieved through the graph-based co-training of Locality Preserving Projections (LPP) [Niyogi, 2004]. This new approach is referred to as Co-LPP. Co-LPP is unsupervised but exhibits discriminative characteristics and is thus well suited to applications such as biometric data retrieval where class labels are generally unavailable. The effectiveness of the proposed algorithm is evaluated under an audio-visual speaker retrieval experiment with the MOBIO database and a single-modal face retrieval experiment with two different facial features with the AR face database. The retrieval performance of the proposed approach out-performs other single-view or multi-view dimensionality reduction methods by a significant margin.

6.2 Locality Preserving Projection

LPP belongs to the family of manifold (or local) dimensionality reduction techniques, and seeks to preserve intrinsic geometric structure by learning a locality preserving sub-manifold [Niyogi, 2004]. In simpler words, LPP seeks to find an optimal projection $\boldsymbol{P}$ such that the neighbouring samples in the original space remain closely located in the projected space. The objective function of LPP is formulated as:

$$\arg\min_{\boldsymbol{P}} \sum_{i,j} (\boldsymbol{P}^T \boldsymbol{x_i} - \boldsymbol{P}^T \boldsymbol{x_j})^2 S_{ij}, \tag{6.3}$$

where S is a local similarity matrix constructed following Equation 6.1 or Equation 6.2, which reflects the similarity of any pair of samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Equation 6.3 shows that, if two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are considered similar in the original space ($S_{ij} > 0$), projecting them far apart will incur a high penalty.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ be a matrix of n samples. Through some straightforward algebraic manipulation (interested readers are referred to [Niyogi, 2004] for details), the objective function in Equation 6.3 can be re-written as:

$$\arg\min_{P} (\boldsymbol{P}^T \boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^T \boldsymbol{P}), \tag{6.4}$$

where $\boldsymbol{L}$ is the graph Laplacian matrix and where:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}, \tag{6.5}$$

in which $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. The projection is obtained by $\boldsymbol{P} = [\boldsymbol{p_1}, \ldots, \boldsymbol{p_k}]$ where $\boldsymbol{p_1}, \ldots, \boldsymbol{p_k}$ are the eigenvectors corresponding to the k smallest eigenvalues of the generalized eigenvalue problem:

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T\boldsymbol{p} = \lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\boldsymbol{p}. \tag{6.6}$$

Although LPP has been successfully applied in automatic face recognition problems [He et al., 2005], it has relatively low discriminative power. Biometric data often contain significant intra-class variation, causing data sample from the same subject located far-apart in the original feature space. This is likely to be the same in the projected space, due to the data structure preserving nature of LPP.

6.3 Co-LPP

In this section, we apply the idea of co-training to multi-view unsupervised subspace learning. Analogous to the *predictor agreement assumption* in co-training, we propose a *subspace data structure agreement assumption* in the UMVDR problem. Given paired features $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ and, assuming that there exist optimal projections $\mathbf{P}^{(1)}_{opt}$ and $\mathbf{P}^{(2)}_{opt}$ which can remove the intra-class variation while retaining inter-class variation, two closely-located data samples in one projected space should be also closely-located in the other projected space. Since the similarity relationships between data samples can be represented by similarity graphs, we proposed to approximate $\mathbf{P}^{(1)}_{opt}$ and $\mathbf{P}^{(2)}_{opt}$ by $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ which minimize the differences between the two similarity graphs constructed in the subspace of each view.

6.3.1 Objective function

Consider a set of n samples represented in two views: $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$, where $\boldsymbol{X}^{(v)} = \{\boldsymbol{x}_1^{(v)}, \ldots, \boldsymbol{x}_n^{(v)}\}$, $v = 1, 2$. $\boldsymbol{X}^{(v)}$ is first centred so that $\bar{\boldsymbol{X}}^{(v)} = \sum_i \boldsymbol{x}_i^{(v)}/n = 0$. Given two projections $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$, we define local similarity matrices $S^{(1)}$ and $S^{(2)}$ in subspaces for each view such that:

$$S_{ij}^{(v)} = \begin{cases} 1, & \text{if } \boldsymbol{P}^{(v)T}\boldsymbol{x}_i^{(v)} \in KNN(\boldsymbol{P}^{(v)T}\boldsymbol{x}_j^{(v)}) \\ & \text{or } \boldsymbol{P}^{(v)T}\boldsymbol{x}_j^{(v)} \in KNN(\boldsymbol{P}^{(v)T}\boldsymbol{x}_i^{(v)}) \\ 0, & \text{otherwise} \end{cases} \tag{6.7}$$

Equation 6.7 is similar to Equation 6.1 excepted that the KNN function is performed in the projected subspace rather than the original feature space. For simplicity, here we employ a binary weight rather than a heat kernel weight to avoid the optimization of parameter σ in Equation 6.2. $S^{(v)}$ encodes local similarity relationships between data samples in the subspace of the v-th view. The i-th and j-th sample are considered similar in the v-th view if $S_{ij}^{(v)} = 1$ and dissimilar otherwise.

We further define a multi-view local structure agreement index as:

$$Agr(S^{(1)}, S^{(2)}) = 1 - \frac{2 \times \sum_{ij} |S_{ij}^{(1)} - S_{ij}^{(2)}|}{\sum_{ij} S_{ij}^{(1)} + \sum_{ij} S_{ij}^{(2)}} \tag{6.8}$$

which is upper-bounded by 1 if $S^{(1)} = S^{(2)}$ and lower-bounded by 0 if $S_{ij}^{(1)} \neq S_{ij}^{(2)}$ for all pairs of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We seek projections $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ such that the agreement in the local data structure is maximized. The objective function thus is given by:

$$\underset{\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}}{\arg\max}\ Agr(S^{(1)}, S^{(2)}) \tag{6.9}$$

Note that $S^{(v)}$ is solely determined by $\boldsymbol{P}^{(v)}$ if the number of neighbours K in the KNN function is fixed.

6.3.2 Algorithm

In the following we propose an algorithm that optimizes $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ by a cross-view training of LPP. The main steps of the iterative algorithm are as follows:

1. Fix $\boldsymbol{P}^{(1)}$ and construct $S^{(1)}$ according to Equation 6.7. Solve for $\boldsymbol{P}^{(2)}$ according to Equation 6.5 and 6.6 by setting $\boldsymbol{X} = \boldsymbol{X}^{(2)}$ and $S = S^{(1)}$;
2. Fix $\boldsymbol{P}^{(2)}$ and construct $S^{(2)}$ according to Equation 6.7. Solve for $\boldsymbol{P}^{(1)}$ according to Equation 6.5 and 6.6 by setting $\boldsymbol{X} = \boldsymbol{X}^{(1)}$ and $S = S^{(2)}$;
3. Go back to step 1 and iterate. At the end of each iteration, calculate the agreement score using Equation 6.8. Stop when the agreement score converges or after a fixed number of iterations;

In other word, we iteratively use the similarity matrix generated in the subspace of one view as a constraint to train LPP projections in the other view. Sometimes the dimensionality of the original features is very high (more than several thousand), and, in this case, PCA can be applied to each view as a preprocessing step as suggested with the Laplacianface method [He et al., 2005]. Since the cross-view training process of LPP projections is similar to the co-training process of predictors, we refer to the new approach as Co-LPP algorithm.

Algorithm 5 Co-LPP

Input: A set of n multi-view samples $\boldsymbol{X} = \{\boldsymbol{X}^{(v)}|v = 1, 2\}$, number of neighbourhood K.
Output: Projection matrices $\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}$
Initialize:

- Center the feature vectors in each view and apply PCA if the dimensionality of the feature space is too high;
- Constructed Similarity graphs $S^{(1)}$ and $S^{(2)}$,

repeat
for $v = 1$ **to** 2 **do**

- Use $\boldsymbol{X}^{(v)}$ and $S^{(3-v)}$ to train LPP projections $\boldsymbol{P}^{(v)}$ and project $\boldsymbol{X}^{(v)}$ into this subspace;
- Update $S^{(v)}$ with projected samples $\boldsymbol{P}^{(v)T}\boldsymbol{X}^{(v)}$

end for
until $Agr(S^{(1)}, S^{(2)})$ does not reach a new maximum within a fixed number of iterations;

The algorithm is formally summarized in Algorithm 5.

Here we justify the proposed co-training approach to optimize the objective function in Equation 6.9. Step 1 of the co-training process optimizes the following objective function:

$$\arg\min_{\boldsymbol{P}^{(2)}} \sum_{i,j} (\boldsymbol{P}^{(2)T}\boldsymbol{x}_i^{(2)} - \boldsymbol{P}^{(2)T}\boldsymbol{x}_j^{(2)})^2 S_{ij}^{(1)} \tag{6.10}$$

Accordingly, if two samples are considered similar in view 1 ($S_{ij}^{(1)} = 1$), they are required to be projected close to each other in view 2, otherwise a penalty is incurred. As a result, the new similarity matrix $S^{(2)}$ determined from $\boldsymbol{P}^{(2)T}\boldsymbol{X}^{(2)}$ is forced to have a similar structure to $S^{(1)}$. The same logic applies in step 2 where $\boldsymbol{P}^{(1)}$ is obtained by training LPP with $\boldsymbol{X}^{(1)}$ and $S^{(2)}$. We acknowledge that the proposed optimization approach is heuristic, as is the case with the original co-training method [Blum and Mitchell, 1998a]. While we do not present a strict proof of convergence, we did not observe divergence in any of our experiments.

6.3.3 Application to multibiometric retrieval

Because of the unsupervised nature of the proposed Co-LPP algorithm, it is particularly suitable for biometric applications where no class labels are available, i.e. retrieval and clustering for instance. Here we discuss its applications to a multibiometric data retrieval problem.

Given a pool of unlabelled biometric data consisting of n samples represented in two

views $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ where $\boldsymbol{x}_i = \{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}\}$ and one query sample $\boldsymbol{q} = \{\boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}\}$, a retrieval algorithm is expected to return t (retrieval window size) samples from $\boldsymbol{X}$ which are considered to contain the same subject as in the query $\boldsymbol{q}$. In our approach, PCA is first performed separately on each view of $\mathbf{X}$ as a preprocessing step. The obtained PCA projection matrix for two views are noted as $\boldsymbol{P}_{pca}^{(1)}$ and $\boldsymbol{P}_{pca}^{(2)}$ respectively. The Co-LPP projection matrix $\boldsymbol{P}_{colpp}^{(1)}$ and $\boldsymbol{P}_{colpp}^{(2)}$ are then jointly learned with PCA embeddings of both views. The final embedding of the i-th sample in the v-th view is:

$$\mathbf{y}_i^{(v)} = \boldsymbol{P}_{colpp}^{(v)T} \boldsymbol{P}_{pca}^{(v)T} (\boldsymbol{x}_i^{(v)} - \boldsymbol{\mu}^{(v)}), \ (v = 1, 2) \tag{6.11}$$

where $\boldsymbol{\mu}^{(v)}$ is the mean of $\boldsymbol{x}^{(v)}$. Similarly, the two views of the query sample are projected into the obtained subspaces by:

$$\mathbf{z}^{(v)} = \boldsymbol{P}_{colpp}^{(v)T} \boldsymbol{P}_{pca}^{(v)T} (\boldsymbol{q}^{(v)} - \boldsymbol{\mu}^{(v)}), \ (v = 1, 2) \tag{6.12}$$

Then the cosine similarity score between $\boldsymbol{z}$ and each target $\boldsymbol{y}_i$ in the v-th view is calculated as:

$$s_i^{(v)} = \frac{\boldsymbol{z}^{(v)} \cdot \boldsymbol{y}_i^{(v)}}{||\boldsymbol{z}^{(v)}||\ ||\boldsymbol{y}_i^{(v)}||}, \ (v = 1, 2) \tag{6.13}$$

This similarity score is bounded between -1 and 1. In each view, t samples in the database with the largest similarity score are returned. Since our approach learns similar local data structures across different views, the retrieval results in each view also tend to be similar. However, they are not necessarily the same. Fusion could be performed to further improve the performance. Since this paper is focused on discriminant feature extraction rather than fusion, we apply a simple weighted sum score level fusion:

$$s_i = \alpha s_i^{(1)} + (1 - \alpha) s_i^{(2)} \tag{6.14}$$

where $0 < \alpha < 1$ is a weighting parameter.

6.4 Experimental results

In this section, the effectiveness of the proposed algorithm is evaluated with two sets of experiments. The first experiment involves an audio-visual speaker retrieval experiment where each sample is represented by a vocal feature vector and a facial feature vector, which is a typical multi-modal biometric setting. The second experiment involves retrieval of human face images. For each face image, two different features are extracted and used as two views for Co-LPP. In the first case, the conditional independence assumption of co-

training is fully satisfied. In the second case, the two input views are correlated since they are extracted from the same face image. However, our experimental result shows that the proposed Co-LPP algorithm is robust to partial violation of this conditional independence assumption.

6.4.1 Databases and Protocol

The audio-visual speaker retrieval experiment is conducted still on with the same MOBIO database [McCool et al., 2012] presented in Section 5.5.2. For computational efficiency, we test our algorithm using a subset of data from 40 male subjects and for each of them, 5 videos are selected from each of the 12 sessions. This results in a pool of 2400 video samples.

We use cropped face images provided with the MOBIO database, one image per video sample. All images are resized to 50×43 pixels and then histogram equalized. Rows of pixel intensities are concatenated to form feature vectors of 2150 dimensions. The speech signal is split into frames of 20ms duration before the extraction of features composed of 26 Mel-scaled frequency cepstral coefficients (MFCCs), their 26 derivatives and the delta energy. Energy-based voice activity detection is then applied to disguard non-speech frames. A 64-component Gaussian mixture model (GMM) is then fitted to remaining speech data through the maximum a posteriori (MAP) adaptation of a speaker-independent world model [Reynolds et al., 2000b]. The means of the GMM model are then concatenated to form a 3392-dimensional GMM supervector [Reynolds et al., 2000a]. As a result, each video is represented by a face feature vector and a vocal feature vector.

The face retrieval experiment is conducted using the AR face database [Martinez, 1998] which contains over 4,000 face images from 126 people recorded during 2 sessions. We use only non-occluded face images and randomly selected 100 subjects (50 male, 50 female). The resulting subset contains 14 images per subject. The 14 face image samples for a subject is shown in Figure 6.2. As we can see, those images contains significant expression and illumination variations. These intra-class variations incur considerable difficulties for retrieval, since these same-class samples is can be dissimilar in the feature space. All images were manually cropped according to eye coordinates and resized to 128×128 pixels. Rows of pixel intensities are concatenated to form feature vectors of 16384 dimensions and are used as the first view in the application of Co-LPP algorithm. For the second view, each face image is divided into 8×8 blocks and $LBP^{u2}_{(8,2)}$ [Ahonen et al., 2006] features are extracted from each block and concatenated into a 3776-dimensional feature vector.

Figure 6.2: 14 non-occluded face images for a subject of AR face database

In our experiment, we adopted a *leave-one-out* strategy for separation the dataset into query and target database. Each time, one video sample is randomly selected as a query while the left 2399 samples are used as the target database. Note that the query sample is not included in the subspace training process. Commonly used evaluation metrics for an information retrieval system involves *Precision* and *Recall*, which are defined as:

$$\text{Precision} = \frac{\text{Number of relevant samples retrieved}}{\text{Retrieval window size}}, \tag{6.15}$$

$$\text{Recall} = \frac{\text{Number of relevant samples retrieved}}{\text{Total number of relevant samples in database}}. \tag{6.16}$$

. The larger the retrieval window, the lower the precision score and the higher the recall score. In our experiment, we chose a window size equals to the total number of relevant samples in database. In this case the precision and the recall are the same, which is similar to the concept of the equal error rate (EER) in biometric verification. As a result, the retrieval window size is set to be 59 in the audio-visual person retrieval experiment and 13 in the face retrieval experiment. We use the corresponding precision/recall score as an evaluation metric. The experiment is repeated 50 times with the random selection of the query sample and the mean precision/recall is reported.

The performance of the proposed Co-LPP algorithm is compared to four alternative dimensionality reduction approach: single-view approach PCA [Jolliffe, 2005], LPP [Niyogi, 2004], as well as multi-view approach CCA [Chaudhuri et al., 2009] and KCCA [Hardoon and Shawe-Taylor, 2003]. Note that different dimensionality reduction algorithms are used to determine subspace projections, while the retrieval processes follow the same protocol as presented in Section 6.3.3.

Here we declare the parameter selections in our experiments:

Dimensionality of subspaces: According to the analysis in [Chaudhuri et al., 2009], in the case of CCA, most discriminative information resides in the first *number of classes*

Precision	Face	Speech	Fusion
PCA	0,478	0,452	0,615
LPP	0,710	0,675	0,847
CCA	0,858	0,784	0,898
KCCA	0,879	0,796	0,910
Co-LPP	**0.984**	**0,952**	**0,994**

(a) Average retrieval precision for audio-visual retrieval experiments on the MOBIO database

Accuracy	LBP	Pixel	Fusion
PCA	0,445	0,612	0,525
LPP	0,489	0,710	0,632
CCA	0,752	0,715	0,787
KCCA	0,772	0,725	0,801
Co-LPP	**0.902**	**0,881**	**0,944**

(b) Average retrieval precision for face retrieval experiments on the AR database

Table 6.1: retrieval performance comparison

- *1* eigenvectors corresponding to the largest eigenvalues. For simplicity, in all compared methods, the dimensionality of subspaces is set to be *number of classes - 1*.

Number of neighbours in KNN graphs: Both LPP and Co-LPP need to specify the number of neighbours in the kNN graphs. Here we adopt a rule of thumb $K = log(n)$ where n is the total number of samples, as suggested in [Von Luxburg, 2007]. In our experiments, this choice leads to reasonable performance for both LPP and Co-LPP.

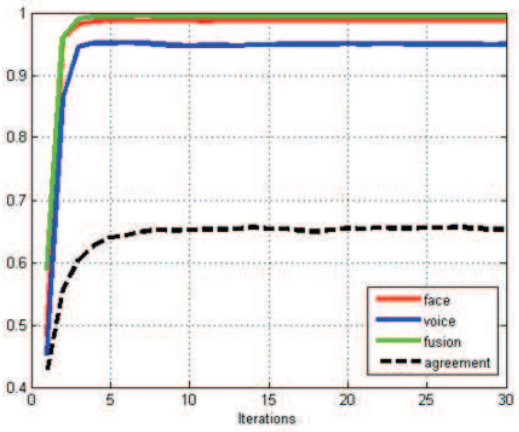
PCA pre-processing: As discussed in Section 6.3.2, in all our experiment, the original features are pre-processed by PCA while keeping 90% of information in the sense of reconstruction error.

Fusion: In the score level fusion process, for each method, the weighting parameter α in Equation 6.14 is set to 100 values equally distributed in $[0, 1]$ region and the best accuracy is reported.

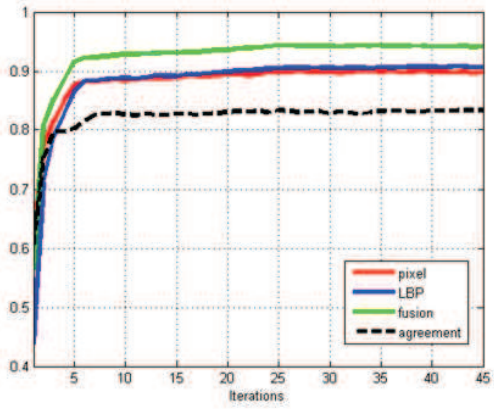
6.4.2 Results and analysis

The retrieval precision for audio-visual retrieval and face retrieval experiment for each single modality and score-level fusion scheme is reported in Table 6.1. We observed that the performance of all multi-view approaches (CCA, KCCA, and Co-LPP) out-performs single-

view approaches (PCA and LPP) in each single modality and the fusion scheme. If for each approach, we compare the score-level fusion accuracy (the fourth column) with the best single-view accuracy, we notice that the fusion process provides less relative gain in multi-view approaches than in single-view approaches. This could be expected since the information fusion process has already been integrated in the multi-view dimensionality reduction process, and the extracted features in two views become correlated. Nevertheless, the best single view accuracy (face modality) of CCA, KCCA, and Co-LPP out-performs the score-level fusion accuracy in PCA and LPP, which demonstrates the effectiveness of multi-view dimensionality reduction in exploring multiple information sources. The performance of KCCA is slightly better than CCA, but kernel parameters need careful tuning. Finally, the proposed Co-LPP algorithm out-performs the closest-performing KCCA approach with a significant margin. In both experiments, even using the weaker single view (speech in audio-visual retrieval and pixel in face retrieval), better retrieval accuracy is obtained than the fusion scheme in KCCA. The score-level fusion scheme in Co-LPP subspaces obtain near perfect retrieval performance ($Precision = Recall > 99\%$) in the audio-visual retrieval experiment. Figure 6.3 (a) and (b) show the variation in retrieval accuracy and agreement scores (objective function defined in Equation 6.8) as a function of the number of iterations, for audio-visual speaker retrieval and face retrieval experiments respectively. In both experiments, the agreement score is seen to stabilize after approximately 5 and 15 iterations respectively.



(a) Audio-visual retrieval



(b) Face retrieval

Figure 6.3: Retrieval accuracy and agreement score as a function of number of co-training iterations for audio-visual speaker retrieval (a) and face retrieval (b) experiments. In each figure, red and blue curves indicate the retrieval accuracy with each single feature, the green curve indicates the retrieval accuracy with score level fusion, and the black dashed curve indicates the agreement score.

For the face retrieval experiment, we show in Figure 6.4 the first 10 retrieved faces while using a screaming face image as a query. The first two rows of images represent the retrieved images in PCA subspace of pixel intensity and LBP features respectively, while the last row represents the Co-LPP retrieval result in the fusion scheme. Incorrectly retrieved images are indicated in red color. We see that in PCA subspace, both features have a tendency to retrieve faces with the same screaming expression, since inter-class variation is more significant than intra-class variation in this case. However, the errors in the two views are different. The proposed Co-LPP algorithm efficiently reduces the intra-class variation by

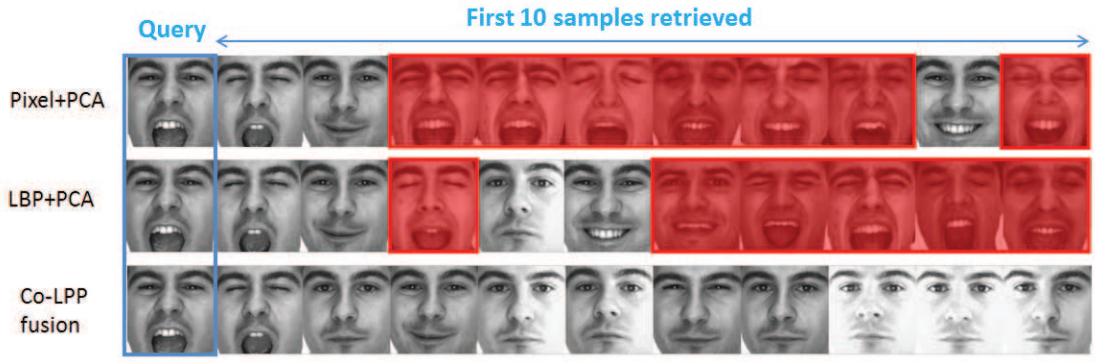
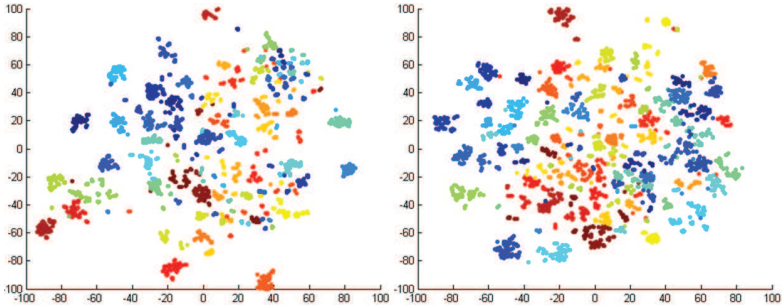


Figure 6.4: Comparison of retrieval results using only pixel features (first row), LBP features (second row) and proposed Co-LPP retrieval scheme. Faces in red indicate incorrect retrieved images.

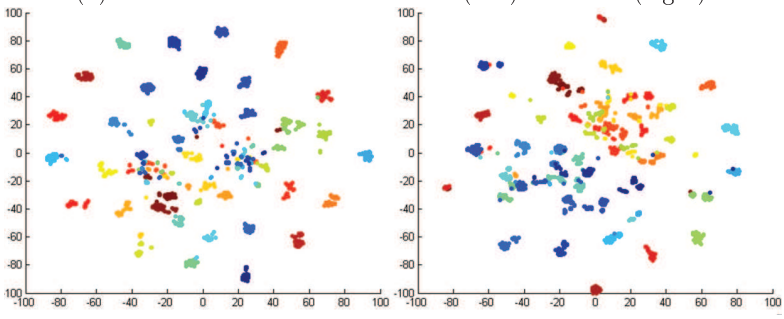
requiring a consensus data structure between the two views.

For audio-visual speaker retrieval experiments, all 2400 samples in the audio-visual speaker retrieval experiment are projected in the PCA, LPP, CCA, and Co-LPP subspaces and their embeddings are visualized in the 2-D plane through the application of t-SNE, which is shown in Figure 6.5(a). In all cases, samples belonging to different classes are represented by different colours. In PCA subspace (Figure 6.5(a)), the sample distribution is especially noisy and explains poor retrieval performance in this case. In LPP subspace (Figure 6.5(b)), some classes form compact clusters while other classes are mixed together. In CCA subspace (Figure 6.5(c)), the mixing of different classes is significantly reduced, yet intra-class scattering is still relatively large. This observation confirms the analysis in [Chaudhuri et al., 2009], CCA is able to maximize the scattering of the centroids of each underlying classes (inter-class scattering), but is not able to minimize the intra-class scattering. Finally, in the proposed Co-LPP subspaces (Figure 6.5(d)), same-class samples are well located in compact clusters with a much higher between-class separation, thereby illustrating its superior performance. In other words, Co-LPP algorithm has high discriminative power despite its unsupervised nature.

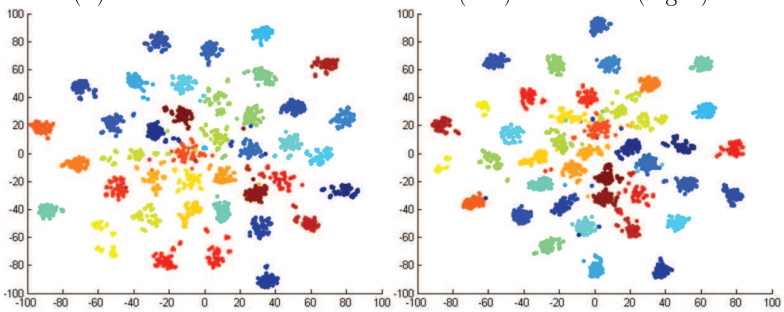
Besides its application in multi-view data retrieval, using the proposed Co-LPP algorithm is also well suited to clustering of multi-view data. Still using the same database as in the retrieval experiment, we performed k-means clustering on the 2400 samples in PCA, LPP, CCA and Co-LPP subspaces. The best clustering accuracy is achieved in Co-LPP subspace, which could be expected from the observation of data structures in Figure 6.5.



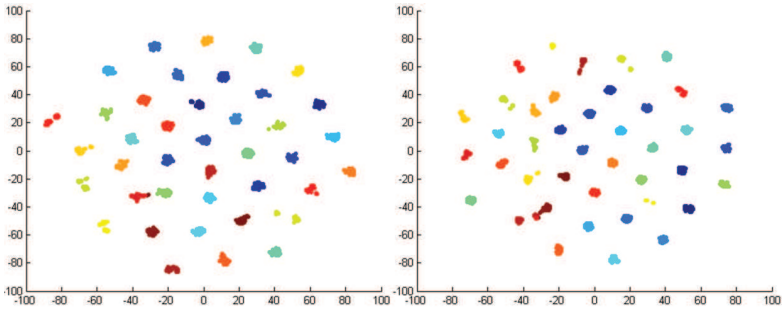
(a) data structure in PCA for face(left) and voice(right)



(b) data structure in LPP for face(left) and voice (right)



(c) data structure in CCA for face(left) and voice (right)



(d) data structure in Co-LPP for face(left) and voice (right).

Figure 6.5: 2-D t-SNE visualizations of data structures for PCA, LPP, CCA, and Co-LPP subspaces. Different subjects/classes are represented by different colors.

6.5 Summary

This chapter proposes a new unsupervised multi-view dimensionality reduction algorithm. Given data with multiple representations, the proposed algorithm aims to learn a subspace projection for each view such that the local data structure in each subspace is in maximal agreement across each view. We show that this objective can be achieved by a graph-based co-training process of LPP. The method is unsupervised, leading to the potential to avoid expensive and time-consuming manual labelling in scenarios where labelled data is scarce, and has potential applications in multi-modal biometric data retrieval.

Chapter 7

Conclusion

In this chapter, we elaborate the main achievements of this thesis and discuss the future perspectives.

7.1 Contributions

Biometric data is often represented by high-dimensional feature vectors which contain significant inter-session variation. Efficient dimensionality reduction techniques are thus needed in order to to extract class-discriminative, low-dimensional features and to attenuate unwanted variations which is redundant to recognition. Such discriminative dimensionality reduction techniques generally follow a supervised learning scheme, in which a subspace projection P is learnt using feature-label pairs $< X, Y >$. However, the acquisition of labelled training data needs expensive human manual labelling and in biometric systems, labelled training data is generally limited in quantity and often does not reliably represent the inter-session variation encountered in test data. The limited size of labelled training set often leads to biased projection matrices and degraded recognition performance.

This thesis proposes to use multi-view dimensionality reduction (MVDR) which aims to extract discriminative features in multi-modal biometric systems, where different modalities are regarded as different views of the same data. Instead of training on feature-label pairs $< X, Y >$, MVDR projections are trained on feature-feature pairs $< X^{(1)}, X^{(2)} >$ where label information is not required. Since unlabelled data is easier to acquire in large

quantities, and because of the natural co-existence of multiple views in multi-modal biometric problems, discriminant, low-dimensional subspaces can be learnt using the proposed MVDR approaches in a largely unsupervised manner.

According to different functionalities of biometric systems, namely recognition (including identification and verification), clustering, and retrieval, we propose three MVDR frameworks which meet the requirements for each functionality.

1. **MVDR by incremental co-training**

   This framework is designed for biometric recognition problems. We assume that a small quantity of labelled data is available during an enrolment session while a larger pool of unlabelled data can be acquired during a period of normal system use. Following a typical co-training procedure, Linear Discriminant Analysis (LDA) projections initially weakly learnt with the small set of labelled data are incrementally re-learnt with automatically labelled data from the unlabelled dataset. This algorithm is referred to as Co-LDA and is applied to the audio-visual person recognition problem. Experimental results on both identification and verification tasks show significant improvements in performance and demonstrate the effectiveness of Co-LDA. In order to deal with out-of-class samples existed in unlabelled dataset, we also provided an extension of the Co-LDA algorithm by incorporating a Sparse Representation Classifier (SRC).

2. **MVDR by subspace clustering agreement**

   This framework is designed for clustering high-dimensional, multi-view data, e.g. facial-vocal biometric data in videos. The framework combines the simplicity of k-means clustering and LDA within a co-training scheme which exploits labels learned automatically in one view to learn discriminative subspaces in another, and this new algorithm is referred to as CoKmLDA. We also present a theoretical treatment which shows how LDA projections learned from samples with random label noise are equivalent to those learned with entirely clean labels. In essence, CoKmLDA algorithm learns a subspace for each view such that the clustering structure is in maximum agreement across each view. We also provide an extension of the two-view CoKmLDA to more than two views. The effectiveness of the proposed algorithm is demonstrated empirically with an audio-visual speaker clustering experiment. Significant improvements over alternative multi-view clustering approaches are reported. The CoKmLDA algorithm is also tested on other multi-view clustering problems such as text clustering and image clustering.

3. **MVDR by subspace graph agreement**

This framework is designed for biometric data retrieval problems, where the purpose is to return from a database a set of samples similar to a given query sample where, in this thesis and the context of biometrics, similarity infers the same subject identity. The similarity relationship between samples in a dataset can be represented by a similarity graph, thus this framework aims to learn a subspace projection for each view such that the difference between the similarity graphs built on the projected samples in each view is minimized. We have shown that this objective can be achieved by a graph-based co-training process of Locality Preserving Projections (LPP), and the new algorithm is referred to as Co-LPP. The effectiveness of the proposed algorithm is validated by audio-visual speaker retrieval experiment and a face retrieval experiment with two different facial features. The retrieval performance of the proposed Co-LPP algorithm out-performs other state-of-the-art MVDR methods such as CCA and KCCA by a significant margin.

The three MVDR frameworks proposed in this thesis share the same spirit: all methods aim to learn a projection for each view such that a certain form of agreement is attained in the subspaces across different views. The definition of such an agreement is, however, different according to the different functionality of the framework. The first MVDR framework is designed for biometric verification and identification, which is a classification problem. In this case the solution involves the agreement between classifier predictions applied to each view in each view. The second MVDR framework is designed for clustering problems. Here the solution involves the agreement between clustering results in different views. The third MVDR framework is for retrieval problems. This solution involves the agreement between the k-nearest-neighbour (kNN) graph build in the subspaces of each view which infers that, given each sample as a query, its kNN retrieval results should be similar in each view.

To summarize, the three MVDR framework can be unified into one general framework for multi-view dimensionality reduction through subspace agreement. *We regard this novel concept of subspace agreement to be the primary contribution of this thesis.*

7.2 Future Work

The work presented in this thesis can be improve in future in terms of both theoretical and practical aspects.

From a theory perspective, a mathematical treatment of the convergence of the proposed MVDR framework is needed. All the three MVDR frameworks make use of the idea of co-training to achieve the agreement between different views. However, co-training itself is

an heuristic approach for which no strict proof of convergence has been given. Although divergence has never been observed in our experiments, a mathematical demonstration of system convergence will lead to more confidence in the proposed algorithms. Moreover, co-training is one option to obtain the objective of subspace agreement, but it is maybe not the only solution. For example, Kumar et al. [Kumar and Daumé III, 2011] first used heuristic co-training in multi-view spectral clustering problem, but in their following work [Kumar et al., 2011], it is show that co-regulation can solve the same problem with a closed-form solution where convergence is guaranteed. Similar efforts could extend the work presented in this thesis by looking for a closed-form solution to the objective function of subspace agreement.

From a practical perspective, the proposed MVDR methods can be extended to solve more problems besides biometrics, e.g. content-base image retrieval (CBIR). Some challenges in CBIR problems are similar to multi-modal biometric problems. For example, images are often represented by high-dimensional feature vectors, and dimensionality reduction is needed. Image retrieval problems can be multi-view, since images have different representations in color, texture and associated text, which could be regarded as different views. Moreover, manual labelling is also expensive so labelled data can be scarce. The new MVDR approaches presented in this thesis present potential solutions to all these challenges.

Appendix A

Semi-supervised Face Recognition with LDA Self-training

This thesis mainly is focused on multi-view dimensionality reduction (MVDR) for multi-modal biometrics. Some of our early work, however, involves semi-supervised dimensionality for single-model biometric system, face recognition in particular. For example, one of our publications in International Conference on Image Processing (ICIP) 2011 deals with a semi-supervised, self-training LDA-based face recognition system. We show that, given only one single face modality and a single classifier, self-training can be applied to augment a manually labelled training set with new data from an unlabelled set, in order to improve the recognition performance of a face recognition system. Since this work is not directly related to MVDR, we would like to presented it in an appendix chapter.

A.1 Introduction

For more than a decade automatic face recognition (AFR) has been one of the most active research topics in computer vision, machine learning and biometrics. In addition to established applications in access control, surveillance and general security, relatively new applications in digital content structuring, search and retrieval are fast gaining popularity. For example, Google's Picasa application utilizes AFR to label faces detected within a photograph so that queries can be performed to return all the pictures containing a particular person. The extension of such algorithms to the wider Internet has already been reported [Kumar et al., 2008].

Many practical AFR applications are characterized by the weak training of templates or models involving only a small number of labelled training data. In these cases AFR performance is generally not robust to inter-session variation in illumination, occlusion, pose and expression since such variation is not well represented in the template or model. Meanwhile, a large pool of unlabeled auxiliary data is generally easily obtained since its collection does not entail costly manual labelling. Images acquired during testing and general operation may be more representative of inter-session variation and may be used to enhance the template or model via appropriate adaptive or self-training approaches. By iteratively augmenting the training set with more and more images, inter-session variation may be incorporated into the template or model and thus better performance can be expected.

Semi-supervised learning refers to a general class of machine learning techniques that make use of both labelled and unlabelled data for training, typically a small amount of labelled data and a larger amount of unlabelled data [Zhu, 2005]. Roli and Marcialis [Roli and Marcialis, 2006] proposed an original semi-supervised face recognition algorithm whereby a PCA-based classifier is initially weakly trained with a small number of manually labelled examples before it is used to classify unlabelled auxiliary data to augment the training set. In related work, also applied to PCA-based classifiers, Roli [Roli, 2005] proposed a variation in which 3 independent classifiers were used. In this work unlabelled auxiliary data are added to augment the labelled dataset only if more than two classifiers agree on the classification result. Neither of the approaches, however, embraces the discriminant power of linear discriminant analysis (LDA). LDA is one of the most popular linear projection techniques for feature extraction, and it is a powerful tool for face recognition when sufficient and representative training examples are available [Belhumeur et al., 1997]. Over-fitting can occur, however, when the training data is limited and in this case performance can be drastically reduced [Martinez and Kak, 2001]. To this end, Cai et al. [Cai et al., 2007] proposed a semi-supervised LDA (SDA) approach which aims to discover the geometrical structure of the data manifold from the unlabeled data but this work did not consider self-training.

In this appendix chapter, we propose a new semi-supervised face recognition approach based on LDA and self-training. In contrast to the work in [Cai et al., 2007], the principal objective is to use automatically labelled, auxiliary data to improve the performance of a classifier that is weakly trained on a small amount of manually labelled data. To our knowledge, it is the first work to couple semi-supervised self-training with an LDA-based approach to face recognition.

The remainder of this appendix chapter is organized as follows. The new LDA self-training algorithm in described in Section A.2. Experiments and results are detailed in

Section A.4 before our conclusions are presented in Section A.5.

A.2 LDA Self-training Algorithm

Here we describe our baseline LDA-based AFR system and then a semi-supervised variant based on self-training.

A.2.1 Baseline system

Linear subspace analysis has been used for AFR over many years and is now a well-known simple, efficient and proven approach. LDA is a supervised algorithm which, according to an optimised projection W_{opt}, projects data vectors x_i in a new space where the ratio between the inter-class (or between, S_B) and intra-class (or within, S_W) scatter is maximized. S_W and S_B are determined according to:

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{l_j} (x_i^j - \mu_j)(x_i^j - \mu_j)^T, \tag{A.1}$$

$$S_B = \sum_{j=1}^{c} l_j (\mu_j - \mu)(\mu_j - \mu)^T, \tag{A.2}$$

where x_i^j is the i-th sample of class j, μ_j is the mean of class j, c is the number of classes, and l_j is the number of samples in class j. The global mean, subsuming all classes, is denoted by μ. We define the total scatter according to:

$$S_T = \sum_{i=1}^{l} (x_i^j - \mu)(x_i^j - \mu)^T, \tag{A.3}$$

where l is the total number of samples such that $S_T = S_B + S_W$. W_{opt} is obtained according to the objective function:

$$W_{opt} = \arg\max_W \frac{W_T S_B W}{W_T S_T W} = [w_1, \ldots, w_m], \tag{A.4}$$

where $\{w_i | i = 1, \ldots, m\}$ are the eigenvectors of S_B and S_T which correspond to the m largest generalized eigenvalues according to:

$$S_B w_i = \lambda_i S_T W_i, i = 1, \ldots, m. \tag{A.5}$$

Note that there are at most $c - 1$ nonzero generalized eigenvalues, so m is upper-bounded by $c-1$. Since S_W is often singular it is common to first apply principal component analysis

(PCA) to reduce the t-dimensional image vector to a g?dimensional vector, $wheret > g > c - 1$, before LDA is used to obtain (c?1)-dimensional vectors.

This is the well-known Fisherface algorithm [Belhumeur et al., 1997] which generally outperforms the Eigenface approach [Turk and Pentland, 1991] when sufficient quantities of labelled data are available. When the quantity of data is low S_w in particular can be noisy which leads to unreliable projections and poor performance [Cai et al., 2007].

A.3 LDA self-training algorithm

A possible solution to deal with insufficient training examples involves semi-supervised learning, which learns from both labelled and unlabelled examples. The semi-supervised PCA-based self-training AFR algorithm proposed in [Roli and Marcialis, 2006] is applied to improve classifiers that are weakly trained using a small labelled dataset $\mathbf{D}_l$. This classifier is then used to automatically label an auxiliary dataset $\mathbf{D}_u$. A fraction of the data with which the system is most confident is then reassigned to $\mathbf{D}_l$ and the classifier is re-trained using the augmented dataset. When repeated iteratively the labelled dataset is steadily enlarged and thus the recogniser is potentially more robust.

However, PCA is an unsupervised approach to dimension reduction. Self-training approaches can thus only help to update the templates for each subject rather than to improve the PCA projection itself. With LDA, in contrast, automatically labelled data not only serve to update templates, but also to increase the amount of data for learning and hence to improve the projection. In this appendix chapter, we demonstrate how a standard LDA-based AFR system can be enhanced through the power of self-learning.

The algorithm is summarized in Algorithm 6. The input to the system is a labelled dataset $\mathbf{D}_l$ and a larger unlabelled auxiliary dataset $\mathbf{D}_u$. First a supervised Fisherface algorithm is applied to reduce the t?dimensional image vectors to a g-dimensional vector through PCA and then to a $(c-1)$-dimensional vector through LDA. A template is calculated for each class by calculating the projected mean. The set of unlabelled samples $\mathbf{D}_u$ is then automatically assigned the label of its nearest template, using the Euclidean distance.

Then, for each class, the single example which is nearest to the corresponding template is removed from $\mathbf{D}_u$ and added to the labelled set $\mathbf{D}_l$. If, for any given class, there are no corresponding examples in $\mathbf{D}_u$ then the corresponding labelled set in $\mathbf{D}_l$ is left unchanged. The PCA and LDA projections are relearned and the templates are recalculated. The process is repeated iteratively until $\mathbf{D}_u$ is empty. A less conservative strategy can also be used whereby, upon each iteration, more than one automatically labelled example is added

Algorithm 6 LDA Self-training

Input:
- $\mathbf{D}_l$, a set of labelled examples from c classes;
- $\mathbf{D}_u$, a set of unlabelled examples;
- g, PCA inter-media space dimension in Fisherface algorithm;

Output:
- Projection matrices P_{pca}, P_{lda} and c updated templates.

Initialize:
- Use PCA to project $\mathbf{D}_l$ into a g-dimension inter-space P_{pca}, then use LDA to further project into $c-1$ dimension feature space P_{lda};
- A template is created by calculating the projected mean of each class;

Iterative LDA self-training:
- Project the $\mathbf{D}_u$ into the PCA inter-space then into the LDA feature space;
- Label each example in $\mathbf{D}_u$ according to the nearest template;
- For each class, the n examples closest to the template are removed from $\mathbf{D}_u$ and added to $\mathbf{D}_l$;
- Update P_{pca} and P_{lda} projections with new $\mathbf{D}_l$, and also templates;
- Iterate until $\mathbf{D}_u$ is empty;

to the training data for each class. This results in a faster algorithm but one which does not capitalise on all the additional training data when each individual sample is selected. Improved computational efficiency thus comes at the cost of reduced performance. The algorithm can work both in a transductive or semi-supervised configuration. A transductive configuration refers to the situation where both the training and testing set are available in the learning process, which reflects an application similar to the automatic labelling of photos in a digital album; A semi-supervised configuration refers to the situation where the testing set is not available during the learning process, and reflects a video security application, for example.

Finally we note that, to avoid S_B and S_T being identical, the LDA algorithm needs at least 2 initial training examples per class. When only a single labelled image is available this restriction can be easily overcome by acquiring a second image through Eigenface recognition, so that LDA may then be applied in the normal way.

A.4 Experimental Results

In this section we report experiments that aim to assess the LDA self-training algorithm and to compare its performance to that of other semi-supervised learning methods. Our experiments were performed on three standard, independent datasets: the Olivetti Research

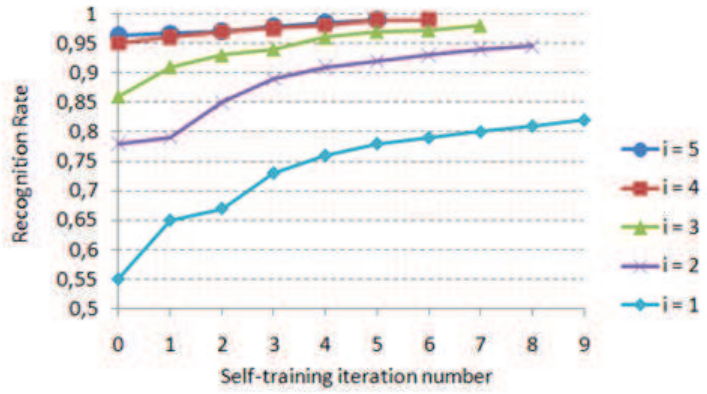


Figure A.1: Recognition rate as a function of the number of self-training iterations.

Lab (ORL) database [1], the AR database [Martinez, 1998] and the CMU PIE database [Sim et al., 2002].

Experiments with the ORL database were performed with a transductive configuration while those with the AR database were performed in a semi-supervised configuration. The aim is to show that our method is beneficial in both cases. Experiments conducted with the CMU PIE database relate to single training images. Here we aim to show the benefit of our algorithm over that reported in [Cai et al., 2007] which was assessed on the same database.

The PCA inter-space dimension g was seen to have a strong influence on performance but behaviour was observed to be consistent across the three different datasets. For all experiments reported here g was set equal to 1.5 times the number of classes (persons).

A.4.1 Tranductive configuration

The ORL database contains images from 40 subjects with 10 images per subject, including pose and expression variations. Original images contain 92×112 pixels but, for computational efficiency, all images were down sampled to 23×28 pixels. Results reported below indicate that our algorithm works well with such low-resolution images.

For any single trial, a template is derived for each subject using between $i = 1$ to 5 labelled training images which are randomly selected according to the ground-truth reference. The remaining images are used either as unlabelled examples for self-training or as test data. Figure A.1 shows the average recognition rate observed from 20 trials. The

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

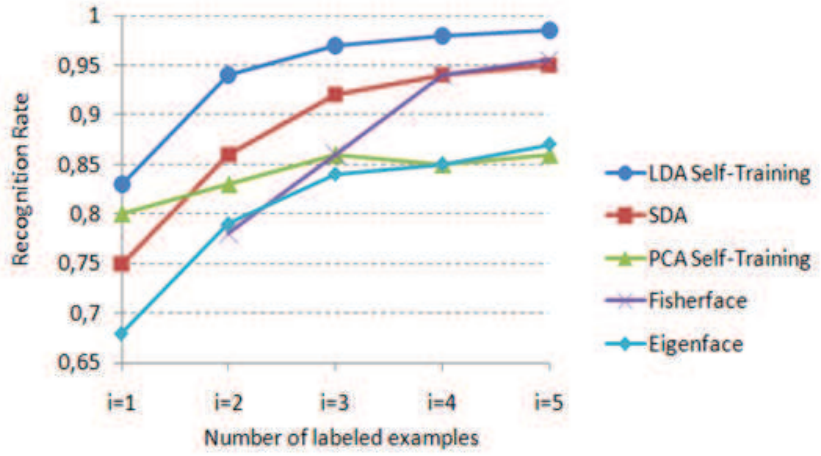


Figure A.2: Recognition rate comparison of ORL database.

horizontal axis represents the number of self-training iterations, while the vertical axis is the recognition rate. Profiles are illustrated for each value of i and confirm that recognition accuracy increases when a greater number of images is used for training (55% for i=1 and 96% for i=5, without self-training). All profiles are further shown to rise as more training images are acquired through self-training (55% without self-training cf. 82% with 9 iterations, for $i = 1$). Note that the maximum number of self-training iterations decreases with increasing i since there are then fewer unlabelled images available.

Figure A.2 illustrates comparative results for alternative semi-supervised AFR algorithms, namely PCA-self training [Roli and Marcialis, 2006], semi-supervised discriminant analysis (SDA) [Cai et al., 2007] in addition to profiles for supervised Eigenface [Turk and Pentland, 1991] and Fisherface [5] algorithms. All systems are our own implementations except for the SDA algorithm which comes from the source code provided by the authors of [Cai et al., 2007]. In all cases results are averaged over 20 trials. Results show that LDA self-training outperforms all other algorithms by a significant margin and serve to demonstrate the merit in combining a discriminant classifier with self-training.

A.4.2 Semi-supervised configuration

The second set of experiments aim to evaluate the self-training LDA algorithm in a semi-supervised configuration, where test data is not available during the learning process. Here experiments where performed on the AR database which contains over 4,000 face images

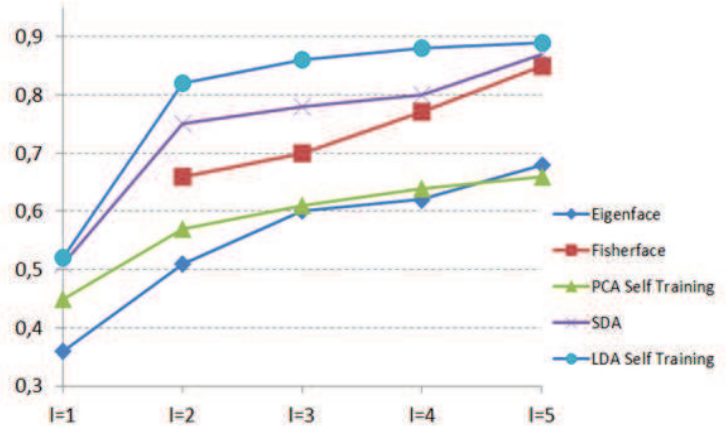


Figure A.3: Recognition rate comparison of AR face database.

from 126 people, and includes expression, illumination and occlusion variations. We first purged the dataset of occluded images and randomly selected 100 subjects (50 male, 50 female). The resulting subset contained 14 images per subject. All images were manually cropped to focus on the face and resized to 32×32 pixels. 3 images per subject were randomly selected as test images. For any one trial $i = 1$ to 5 images were labelled according to the ground-truth reference and used for template learning. The others are used as unlabelled images for self-training. Results for the five different algorithms are illustrated in Figure A.3 and again show that the self-training algorithm outperforms the other algorithms.

A.4.3 Single training image test

The CMU PIE face database contains 68 subjects with 41,368 face images captured with varying pose, illumination and expressions. Each image contains 32×32 pixels. For all experiments reported here we used only frontal pose images which correspond to 43 per subject from which 30 were randomly selected as training data. For any single trial, a single training image is randomly selected for each subject and the remaining 29 images are left unlabelled and are pooled for subsequent self-training. As before results are averaged over 20 trials. From the results illustrated in Table 2 we can see that although the LDA self-training algorithm exhibits larger standard deviation among different trials, it nevertheless achieves the best performance among all the other algorithms, with a significant margin.

Table A.1: Recognition rate on CMU PIE database. The number in parentheses indicates the standard variation.

Accuray (std.)	Unlabeled Set	Test Set
Eigenface [Turk and Pentland, 1991]	25.3(1.7)	25.3(1.6)
Laplacianface [He et al., 2005]	56.1(2.3)	56.4(2.4)
Consistency [Zhou et al., 2004]	52.0(1.8)	–
LapSVM [Belkin et al., 2006]	56.5(1.6)	56.9(2.6)
LapRLS [Belkin et al., 2006]	57.5(1.6)	57.9(2.6)
SDA [Cai et al., 2007]	59.0(2.0)	59.5(2.7)
LDA self-training	**84.5(9.5)**	**71.3(6.5)**

A.5 Conclusion

This chapter presents a new semi-supervised face recognition algorithm based on LDA self-training. Despite its simplicity it successfully exploits both labelled and un-labelled data for template learning and delivers superior performance than existing approaches. Training data is augmented with automatically labelled, auxiliary data that is often easily obtained without the cost of manual labelling. Experiments on three independent datasets show that the new algorithm is robust to variations in illumination, pose and expression and that it outperforms related approaches in both transductive and semi-supervised configurations. These observations indicate that the new self-training algorithm is successful in overcoming the over-fitting problems which typify LDA-based approaches to automatic face recognition and that they warrant further attention.

Bibliography

T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006.

X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):356–370, 2012.

C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5): 1505–1512, 2006.

M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464, 2002.

S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proceedings of 19th International Conference on Machine Learning (ICML)*, 2002.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computing*, 12(10):2385–2404, Oct. 2000. ISSN 0899-7667. doi: 10.1162/089976600300014980. URL `http://dx.doi.org/10.1162/089976600300014980`.

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

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework

for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

R. Bellman. *Adaptive control processes: a guided tour*, volume 4. Princeton university press Princeton, 1961.

L. Besacier, J. Bonastre, and C. Fredouille. Localization and selection of speaker-specific information with statistical modeling. *Speech Communication*, 31(2):89–106, 2000.

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is âĂIJnearest neighborâĂİ meaningful? In *Database TheoryâĂŤICDTâĂŹ99*, pages 217–235. Springer, 1999.

H. S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, A. Noore, and A. Ross. On co-training online biometric classifiers. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7. IEEE, 2011.

M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of sift features for face authentication. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 35–35. IEEE, 2006.

S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the IEEE International Conference on Data Mining*, volume 36, 2004.

M. B. Blaschko and C. H. Lampert. Correlational spectral clustering. In *In CVPR ?08*, 2008.

M. B. Blaschko, J. A. Shelton, A. Bartels, C. H. Lampert, and A. Gretton. Semi-supervised kernel canonical correlation analysis with application to human fmri. *Pattern Recognition Letters*, 32(11):1572–1583, 2011.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998a.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998b.

K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer Vision and Image Understanding*, 101 (1):1–15, 2006.

D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7. IEEE, 2007.

W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *Signal Processing Letters, IEEE*, 13(5):308–311, 2006.

C.-H. Chan, J. Kittler, and K. Messer. *Multi-scale local binary pattern histograms for face recognition*. Springer, 2007.

K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 129–136. ACM, 2009.

X. Chen, P. J. Flynn, and K. W. Bowyer. Ir and visible light face recognition. *Computer Vision and Image Understanding*, 99(3):332–358, 2005.

C. Chibelushi, F. Deravi, and J. Mason. Audio-visual person recognition: An evaluation of data fusion strategies. In *Security and Detection, 1997. ECOS 97., European Conference on*, pages 26–30. IET, 1997.

R. Cucchiara. Multimedia surveillance systems. In *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 3–10. ACM, 2005.

V. R. de Sa. Spectral clustering with two views. In *Proceedings of Workshop of Learning with Multiple Views*, 2005.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

K. Delac, M. Grgic, and S. Grgic. Independent comparative study of pca, ica, and lda on the feret data set. *International Journal of Imaging Systems and Technology*, 15(5): 252–260, 2005.

I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2):143–175, Jan. 2001. ISSN 0885-6125. doi: 10.1023/A:1007612920971. URL `http://dx.doi.org/10.1023/A:1007612920971`.

T. Diethe, D. R. Hardoon, and J. Shawe-Taylor. Multiview fisher discriminant analysis. In *NIPS workshop on learning from multiple sources*, 2008.

C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *In International Conference on Machine Learning*, pages 84–405. Academic Press, 2007.

G. Evangelopoulos and P. Maragos. Speech event detection using multiband modulation energy. In *Proc. Interspeech*, pages 685–688, 2005.

D. P. Foster, S. M. Kakade, and T. Zhang. Multi-view dimensionality reduction via canonical correlation analysis. Technical report, Technical Report TR-2008-4, TTI-Chicago, 2008.

D. Gafurov. A survey of biometric gait recognition: Approaches, security and challenges. In *Annual Norwegian Computer Science Conference*, pages 19–21, 2007.

D. Greene and P. Cunningham. Producing accurate interpretable clusters from high-dimensional data. *Knowledge Discovery in Databases: PKDD 2005*, pages 486–494, 2005.

D. Greene and P. Cunningham. A matrix factorization approach for integrating multiple data views. *Machine Learning and Knowledge Discovery in Databases*, pages 423–438, 2009.

Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang. Sparse unsupervised dimensionality reduction for multiple view data. 2012.

D. R. Hardoon and J. Shawe-Taylor. Kcca for different level precision in content-based image retrieval. In *Proceedings of Third International Workshop on Content-Based Multimedia Indexing, IRISA, Rennes, France*, 2003.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):328–340, 2005.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

C. Hou, C. Zhang, Y. Wu, and F. Nie. Multiple view semi-supervised dimensionality reduction. *Pattern Recognition*, 43(3):720–730, 2010.

M. Huijbregts and D. van Leeuwen. Towards automatic speaker retrieval for large multimedia archives. In *Proceedings of the 3rd international workshop on Automated information extraction in media production*, pages 15–20. ACM, 2010.

A. K. Jain and A. Ross. Multibiometric systems. *Communications of the ACM*, 47(1): 34–40, 2004.

A. K. Jain, R. M. Bolle, and S. Pankanti. *Biometrics: personal identification in networked society*. Springer, 1999.

A. K. Jain, B. Klare, and U. Park. Face matching and retrieval in forensics applications. *IEEE MultiMedia*, 19(1):20, 2012.

H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved lbp under bayesian framework. In *Image and Graphics, 2004. Proceedings. Third International Conference on*, pages 306–309. IEEE, 2004.

I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2005.

P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1435–1447, 2007.

J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239, 1998.

J. Kittler, Y. Li, and J. Matas. On matching scores for lda-based face verification. In *Proceedings of British Machine Vision Conference*, pages 42–51, 2000.

B. Kotnik, Z. Kacic, and B. Horvat. A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. In *Proc. 7th EUROSPEECH, Aalborg, Denmark*, pages 197–200, 2001.

H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1–58, 2009.

A. Kumar and H. Daumé III. A co-training approach for multi-view spectral clustering. In *International Conference on Machine Learning*, 2011.

A. Kumar, P. Rai, and H. Daumé III. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems*, 24:1413–1421, 2011.

N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *Computer Vision–ECCV 2008*, pages 340–353. Springer, 2008.

P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000.

H. Lee, Y. Chung, J. Kim, and D. Park. Face image retrieval using sparse representation classifier with gabor-lbp histogram. In *Information Security Applications*, pages 273–280. Springer, 2011.

J. Lee, B. Moghaddam, H. Pfister, and R. Machiraju. Finding optimal views for 3d face shape modeling. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 31–36. IEEE, 2004.

Y. Li and J. Shawe-Taylor. Using kcca for japanese–english cross-language information retrieval and document classification. *Journal of intelligent information systems*, 27(2): 117–133, 2006.

Z. Li, W. Jiang, and H. Meng. Fishervioce: A discriminant subspace framework for speaker recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4522–4525. IEEE, 2010.

B. Long, S. Y. Philip, and Z. M. Zhang. A general model for multiple view unsupervised learning. In *SDM*, pages 822–833, 2008.

X. Lu, Y. Wang, and A. K. Jain. Combining classifiers for face recognition. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–13. IEEE, 2003.

S. Marcel, C. McCool, P. Matějka, T. Ahonen, J. Černockỳ, S. Chakraborty, V. Balasubramanian, S. Panchanathan, C. H. Chan, J. Kittler, et al. On the results of the first mobile biometry (mobio) face and speaker verification evaluation. In *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 210–225. Springer, 2010.

G. L. Marcialis and F. Roli. Fingerprint verification by fusion of optical and capacitive sensors. *Pattern Recognition Letters*, 25(11):1315–1322, 2004.

A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

A. M. Martinez and A. C. Kak. Pca versus lda. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):228–233, 2001.

C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J. Bonastre, P. Tresadern, and T. Cootes. Bi-modal person recognition on a mobile phone: Using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635 –640, july 2012. doi: 10.1109/ICMEW.2012.116.

I. Naseem, R. Togneri, and M. Bennamoun. Sparse representation for speaker identification. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4460–4463. IEEE, 2010.

A. V. Nefian, L. H. Liang, T. Fu, and X. X. Liu. A bayesian approach to audio-visual speaker identification. In *Audio-and Video-Based Biometric Person Authentication*, pages 761–769. Springer, 2003.

C. A. Nelson. The development and neural bases of face recognition. *Infant and child development*, 10(1-2):3–18, 2001.

A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

F. Nie, D. Xu, X. Li, and S. Xiang. Semisupervised dimensionality reduction and classification through virtual label regression. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(3):675–685, 2011.

X. Niyogi. Locality preserving projections. In *Neural information processing systems*, volume 16, page 153, 2004.

T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

N. Poh, R. Wong, J. Kittler, and F. Roli. Challenges and research directions for adaptive biometric recognition systems. In *Advances in Biometrics*, pages 753–764. Springer, 2009.

A. Rattani, B. Freni, G. L. Marcialis, and F. Roli. Template update methods in adaptive biometric systems: a critical review. In *Advances in Biometrics*, pages 847–856. Springer, 2009.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, page 2000, 2000a.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000b.

F. Roli. Semi-supervised multiple classifier systems: Background and research directions. In *Multiple Classifier Systems*, pages 1–11. Springer, 2005.

F. Roli and G. L. Marcialis. Semi-supervised pca-based face recognition using self-training. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 560–568. Springer, 2006.

F. Roli, L. Didaci, and G. L. Marcialis. Template co-update in multimodal biometric systems. In *Advances in Biometrics*, pages 1194–1202. Springer, 2007.

A. Ross, K. Nandakumar, and A. K. Jain. Introduction to multibiometrics. In *Handbook of Biometrics*, pages 271–292. Springer, 2008.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao. Audiovisual celebrity recognition in unconstrained web videos. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1977–1980. IEEE, 2009.

M. Sargm, E. Erzin, Y. Yemez, and A. M. Tekalp. Multimodal speaker identification using canonical correlation analysis. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

B. Scholkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *Advances in kernel methods-support vector learning*. Citeseer, 1999.

J. W. Shin, J.-H. Chang, and N. S. Kim. Voice activity detection based on statistical models and machine learning approaches. *Computer Speech & Language*, 24(3):515–530, 2010.

T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002.

K. Steiglitz and C. H. Papadimitriou. Combinatorial optimization: algorithms and complexity. *Prentice Hall, New Jersey., UV Vazirani (1984). On two geometric problems related to the travelling salesman problem. J. Algorithms*, 5:231–246, 1982.

A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

H. Tang, S. Chu, M. Hasegawa-Johnson, and T. Huang. Partially supervised speaker clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(5):959–971, May 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.174. URL `http://dx.doi.org/10.1109/TPAMI.2011.174`.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.

M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

F. Wang, C. Zhang, H. C. Shen, and J. Wang. Semi-supervised classification using linear neighborhood propagation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 160–167. IEEE, 2006.

H. Wang, C. Ding, and H. Huang. Multi-label linear discriminant analysis. In *Computer Vision–ECCV 2010*, pages 126–139. Springer, 2010.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

T. Xia, D. Tao, T. Mei, and Y. Zhang. Multiview spectral embedding. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(6):1438–1446, 2010.

M. Yamada, M. Sugiyama, and T. Matsui. Semi-supervised speaker identification under covariate shift. *Signal Processing*, 90(8):2353–2361, 2010.

M. Yang, Y. Yang, and Z. Wu. A pitch-based rapid speech segmentation for speaker indexing. In *Multimedia, Seventh IEEE International Symposium on*, pages 6–pp. IEEE, 2005.

J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. *Advances in Neural Information Processing Systems*, 20:1649–1656, 2007.

W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 786–791. IEEE, 2005.

X. Zhao, N. Evans, and J.-L. Dugelay. A co-training approach to automatic face recognition. *EUSIPCO 2011*, 2011.

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16(16):321–328, 2004.

X. Zhu. Semi-supervised learning literature survey. 2005.

Réduction multivue de la dimensionnalité pour la biométrie multimodale

--un résumé Français de la thèse

Contexte

La biométrie se réfère à la reconnaissance des l'hommes par leurs traits physiques ou comportementaux. Dans la vie de tous les jours, il y a beaucoup de problèmes concernant les identités personnelles. Les innovations récentes dans les systèmes biométriques apportent une solution plus simple, plus rapide et plus sûr. Par exemple, les systèmes biométriques sont largement utilisés dans le contrôle d'accès, soit l'accès physique à une ressource spécifique, l'emplacement ou le territoire (contrôle d'accès à un bâtiment, le contrôle des frontières pour l'immigration, etc), ou un accès virtuel à un réseau informatique, compte bancaire en ligne, par exemple. Dans ces applications, les traits biométriques comme le visage, l'iris, les empreintes digitales et la voix peuvent être utilisés pour remplacer (ou compléter) les mots de passe ou les cartes d'identité, qui pourrait être soit oubliés ou volés. Les systèmes de biométrie basés sur la reconnaissance de visage ou du démarche pourraient êtres utilisé pour identifier les individus (par exemple, les criminels) dans les systèmes de vidéo-surveillance, car ils n'ont pqs besoin de la coopération minimale de l'utilisateur. Les informations sur le visage et la voix de l'homme peuvent aussi contribuer à la gestion de données multimédia, afin de rendre la récupération ou l'indexation des fichiers multimédias plus précis et plus efficace.

Quelles que soient les applications, du point de vue informatique, la biométrie est un problème de reconnaissance de formes. Les systèmes biométriques contiennent généralement deux modules, l'extraction de caractéristiques et la comparaison (ou classification). Dans le module d'extraction de caractéristiques, des échantillons biométriques sont représentés par des fonctions numériques qui peuvent être traité par des programmes informatiques; dans le module de comparaison ou de classification, lq caractéristique extraite à partir d'un échantillon de test est comparée à une ou plusieurs caractéristiques obtenues à partir des

échantillons d'enrolement (connus en tant que modèle) pour déterminer si l'échantillon du test ont l'identité déclarée (mode de vérification) ou à quelle l'identité enregistrée l'échantillon de test appartient. Dans la plupart des systèmes biométriques de l'état de l'art, les données biométrique sont souvent représentés par des vecteurs de grande dimensionalité (par exemple, les local binary pattern (LBP) pour la reconnaissance de visage et les Gaussian Mixture Models (GMM) supervecteurs pour la reconnaissance du locuteur. La dimensionnalité d'éléments biométriques génèrent un stockage lourd et des calculs informatiques important, et plus grave encore, la soi-disante *malédiction de dimensionnalité*, peuvent avoir un impact sur la performance de la reconnaissance dans le module de comparaison ou classification suivant.

Les difficultés liées à la dimensionalité sont généralement surmontés grâce à l'application de techniques de réduction de la dimensionnalité (DR), qui cherchent une representation de plus faible dimensionnelle des donnée. Selon que si l'information sur l'étiquette est nécessaire ou non, les techniques DR peuvent être classées dans deux catégories: méthodes supervisées et celles non supervisées. Un dilemme se pose sur le compromis entre la disponibilité de l'information sur l'étiquette et le pouvoir discriminant des caractéristiques de faibles dimensions extraites. Les méthodes supervisées telles que Linear Discriminant Analysis (LDA) ont une puissance discriminante importante, mais ils ont besoin de grandes quantités de données d'apprentissage étiquetés manuellement. Les méthodes non supervisées telles que Principle Component Analysis (PCA) n'ont pas besoin d'étiquettes de classe, mais n'ont généralement que peu de pouvoir discriminant. Dans les systèmes d'identification et de vérification biométriques, les données étiquetées manuellement sont normalement en nombre limité, mais une grande quantité de données non étiquetées peut être facilement acquise lors de l'utilisation normale du système.

Dans la biométrie multimodale, différentes modalités biométriques peuvent former différents entrés des algorithmes de classification. Les systèmes biométriques multimodaux peuvent obtenir de multiples ensembles d'informations de la même modalité (2D+3D dans la reconnaissance de visage) ou des information de différentes modalités biométriques (système

biométrique avec le visage et la voix). La fusion des modalités reste un problème difficile et est généralement traitée de manière isolée à celui de dimensionalité élevée.

Cette thèse aborde le problème de la dimensionnalité élevée et le problème de la fusion multimodale dans un cadre unifié. En vertu d'un paramètre biométrique multi-modale et les données non étiquetées abondantes données, nous cherchons à extraire des caractéristiques discriminatoires de multiples modalités d'une manière non supervisée.

Contributions

Dans cette section, nous résumons brièvement le contenu de la thèse et les contributions.

Les systèmes biométriques multimodaux utilisent deux ou plusieurs modalités individuelles pour améliorer la précision de la reconnaissance des méthodes uni-modaux classiques. Dans un système biométrique bimodale qui emploie deux modalités différentes, des échantillons de données peuvent être représentées par des caractéristiques appariées $<\mathbf{X}^{(1)},\mathbf{X}^{(2)}>$ d' un sujet et de l'identité $\mathbf{Y}$ en tant que variable de référence. L'etat de l'art des systèmes biométriques utilisent souvent des caractéristiques de grande dimensionalité, ainsi les techniaues de réduction de dimensionnalité (RD) sont souvent appliquées pour atténuer le soi-disant problème de *malédiction de dimensionnalité* dans l'étape de classification aui suit. Cette thèse présente une étude de la RD s'approche de la biométrie multimodale. Communément appelée réduction multi-vues de la dimensionnalité (RMVD), ce domaine a suscité un intérêt considérable en recherche au cours des dernières années. La plupart des algorithmes de RMVD existants sont basés sur l'analyse de corrélation canonique (CCA) et ses variantes. Ces algorithmes visent généralement à apprendre deux projections $\mathbf{P}^{(1)}$ et $\mathbf{P}^{(2)}$ tels que les échantillons projetés $\mathbf{P}^{(1)}\mathbf{X}^{(1)}$ et $\mathbf{P}^{(2)}\mathbf{X}^{(2)}$ sont en corrélation maximum. Lorsqu'il est appliqué à des fonctions paires, le principal avantage des méthodes RMVD sur RD d'une seule vue est qu'une des vue peut être considérée comme étiquettes faibles pour l'autre. En conséquence, les caractéristiques discriminantes peuvent ainsi être extraites même si les échantillons étiquetées sont soit limités en nombre soit totalement absents. Contrairement aux approches précédentes, le nouveau travail présenté dans cette thèse aborde le problème de MVDR sous

un angle différent. Inspiré par la méthode innovqtrice d'apprentissage semi-supervisé, cette thèse présente un nouveau concept d'*accord de structure de sous-espace*. L'idée principale consiste projections d'apprentissage $\mathbf{P}^{(1)}$ et $\mathbf{P}^{(2)}$, de sorte que la structure de données d'échantillons prévus $\mathbf{P}^{(1)}\mathbf{X}^{(1)}$ et $\mathbf{P}^{(2)}\mathbf{X}^{(2)}$ est aussi proche que possible. Selon les différentes définitions de *structure de données*, et pour des applications diffirentes, à savoir la classification semi-supervisée, classification non supervisée et la récupération, nous proposons trois approches différentes RMVD, qui sont décrits ci-après.

La première approche est une extension directe du *co-apprendissage* supplémentaire aux problèmes de MVDR semi- supervisés par la co-apprendissage de l'analyse discriminante linéaire (LDA) projective. L'algorithme co-LDA a été publié dans les actes de la International Conference on Multimédia and Exposition (ICME) en 2012. L'entrée comporte un petit ensemble de deux-vue, données étiquetées { $\mathbf{X}_L^{(1)}$, $\mathbf{X}_L^{(2)}$, $\mathbf{Y}$} et un plus grand nombre de données nonétiquetées { $\mathbf{X}_U^{(1)}$, $\mathbf{X}_U^{(2)}$ }. Alors que pour les données non-étiquetées, la taille est plus représentatif de la distribution des données sous-jacente. Projections LDA $\mathbf{P}^{(1)}$ et $\mathbf{P}^{(2)}$ sont initialement appris sur chaque vue de l'ensemble de la formation marqué. L'ensemble non-étiqueté est ensuite projeté dans les mêmes sous-espaces, et les échantillons sont affectés qux étiquettes selon un classificateur du plus proche centre de gravité. Pour chaque point de vue, le sous-ensemble des échantillons non marqués qui sont classifié avec le plus de confiance sont retirés de l' ensemble non-étiquetéet ajouté à l'ensembleétiqueté. Les projections et les classificateurs LDA sont ensuite réitérés. Itération de la procédure continue jusqu'à ce que l'ensemble non étiquetéest vide. Le nouvel algorithme est un succès dans l'utilisation des données non-étiquetées pour éviter une coupe près du corps à l'ensemble de données étiquetées à la main plus faible. Lors d'une experience sur la base de données bimodale MOBIO, l'algorithme Co-LDA proposé soulève un taux d'identification de base du 71 % à 99 %, tandis que dans une tâche de vérification du taux d'erreur égal (EER) est réduite de 16% à moins de 1%. Dans le prolongement de ce travail qui a été publié dans les actes de la International Conference on Acoustics Speech and Signal Processing (ICASSP) 2013, nous montrons que'une Représentation Eparce Classificateur (SRC) pourrait être utilisée pour rejeter les échantillons hors- classe qui appartiennent à aucune des classes enregistrées. Dans

des travaux connexes publiés dans les actes de la International Conference on Image Processing (ICIP) 2011 et de la European Conference on Signal Processing (EUSIPCO) 2011, nous avons également proposé une version self-training de l'algorithme qui pourrait être appliquée aux systèmes mono-modal.

L'algorithme de co-apprendissage standard est semi-supervisé et nécessite certaines données étiquetées pour l'initialisation. Pour les problèmes de clustering qui est purement non-supervisé, nous avons proposé un algorithme multi-vue de regroupement sous-espace qui est basé sur une hypothèse d'accord multi-vues de regroupement. Nous considérons le problème de regrouper des donées deux-avis, de grande $<\mathbf{X}^{(1)},\mathbf{X}^{(2)}>$ dans k groupes, et des échantillons de la même classe devraient se placer dans la même cluster. Puisque $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ sont des représentations différentes de la même sous-jacent classe $\mathbf{Y}$, dans des conditions idéales, les résultats de regroupement devrait être identiques indépendamment de la vue utilisée pour le regroupement. Cependant, il est peu probable si le regroupement est effectué dans les espaces de caractéristiques originales $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$, car ils sont corrompus par différents variations intra-classe. Cette thèse présente une nouvelle approche de regroupement multi-veu sous-espace qui cherchent des projections $\mathbf{P}^{(1)}$ et $\mathbf{P}^{(2)}$ tels que les résultats de regroupement sont maximalment d'accord dans les sous-espaces de chaque vue. Nous montrons que cet objectif peut être obtenu en combinant la simplicité du k-means et l'analyse discriminante linéaire (LDA) au sein d'un système de co-apprendissage. Le nouvel algorithme exploite les indicateurs de munitions obtenues à partir de k-means dans une vue pour apprendre sous-espaces discriminants dans une autre. L'algorithme est appelé CoKmLDA. Nous montrons mathématiquement que les projections LDA apprises à partir d'échantillons de bruit de l'étiquette aléatoire sont probabiliste équivalentes à celles appris avec étiquettes propres et que l'étiquetage intra-vue, ou co-apprendissage, est efficace dans la correction des échantillons des étiquettes erronées. De plus, l'algorithme ne nécessite pas l' optimisation de tous les hyperparamètres. L'efficacité de l'algorithme proposé est démontré non seulement en enceintes regroupement d'expériences bimodale, voix visage de base de données MOBIO, mais aussi dans les tâches de plus générale telles que le regroupement de chiffres manuscrits et documents de texte regroupement. L'amélioration significative par rapport aux de

regroupement du vues multiples des approches alternatives telles que le CCA et le regroupement co-spectrale sont indiquées. Ce travail a été présenté à un numéro spécial sur l'apprentissage non supervisé et supervisé dans Pattern Recognition Letters.

L'algorithme proposé CoKmLDA est adapté pour des problèmes de regroupement, mais n'est pas bien adapté d'autre problèmes des apprentissage non supervisé, tels que l'extraction, car elle doit connaître le nombre de classes en tant que paramètre d'entrée. Dans cette thèse, nous avons proposé en outre algorithme multi-veu de réduction de la dimension des problèmes de récupération, basée sur des graphes de similarité. Les méthodes de réduction de la dimensionnalité basées sur les graphes ont récemment émergé comme un outil puissant pour l'analyse des données de grande dimension. L'exemples d'algorithmes comprennent des méthodes non linéaires telles que Isomap, Embedding linéaire locale (LLE), eigenmaps de Laplace et méthodes linéaires tels que Localité Préserver projection (LPP). Ces méthodes commencent avec la construction d'un graphe de similarité **S** dans lequel les noeuds représentent des échantillons de données tandis que les bords s_{ij} représentent la mesure de similarité entre la i^{eme} et j^{eme} échantillon. Les méthodes de réduction de la dimensionnalité fondées sur les graphiques sont en grande partie pour révéler la structure de variété de basse dimension des données d'origine, mais ne sont pas capable d'extraire des caractéristiques discriminantes spécifiques à la classe en raison de leur nature non supervisée. En raison des variations importantes intra-classes, s_{ij} measuré pourrait être très faible mesurée dans les espaces d'origine, même si l'échantillon i et j appartiennent à la même classe. L'estimation fiable de la similitude va influencer les projections et donc conduire à des sous-espaces sous-optimales. Nous considérons la réduction de la dimensionnalité à base de graphes dans un cadre de deux vues, où des échantillons de données peuvent être à nouveau représentés sous la forme de $<\mathbf{X}^{(1)},\mathbf{X}^{(2)}>$ et les deux points de vue présenter une certain niveau d'indépendance conditionnelle, comme c'est souvent le cas en la biométrie. Si la matrice de similitude $\mathbf{S^{(1)}}$ et $\mathbf{S^{(2)}}$ sont construits avec des $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ respectivement, alors qu'ils sont censés être différent puisque $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ contient différentes variations intra-classes. Supposons qu'il existe des projections optimales $\mathbf{P_{opt}^{(1)}}$ et $\mathbf{P_{opt}^{(2)}}$ telles que, dans les deux sous-espaces projetés, des échantillons de même classe sont situées à proximité l'un de l'autre

et les échantillons de différents classe sont situés loin l'un de l'autre, et $\mathbf{S}^{(1)}$ et \$ $\mathbf{S}^{(2)}$ sont construits avec les échantillons projetés $\mathbf{P}_{opt}^{(1)}\mathbf{X}^{(1)}$ et $\mathbf{P}_{opt}^{(2)}\mathbf{X}^{(2)}$, $\mathbf{S}^{(1)}$ et $\mathbf{S}^{(2)}$ devraient être similaires. Avec cette logique, nous proposons d' approcher $\mathbf{P}_{opt}^{(1)}$ et $\mathbf{P}_{opt}^{(2)}$ en trouvant $\mathbf{P}^{(1)}$ et $\mathbf{P}^{(2)}$ qui minimisent la différence entre $\mathbf{S}^{(1)}$ et $\mathbf{S}^{(2)}$. Cet objectif pourrait être atteint grâce aux co-apprendissage basé sur les graphes de la LPP, et cette thèse includ une telle approche appelée Co-LPP. Co-LPP est idéal pour l'apprentissage de métrique pour des problèmes de retrieval, et son efficacité est démontrée par des expériences sur la recherche de personne audiovisuel à partir de audio-vidéos et de la recherche de visage humain avec de multiples traits du visage. Ce travail a été publié dans IEEE International Workshop on Information Forensics and Security (WIFS), 2013. L'algorithme d'accord de sous-espace graphique est très flexible et peut être utilisé pour étendre d'autres methodes mono-vue de réduction de dimensionalité à un réglage multi- vues.

En résumé, les contributions de cette thèse sont les suivantses:

- Un état de l'art des algorithmes RMVD de l'état de l'art ;
- Un nouveau concept de RMVD: accord de la structure de donées dans sous-espace;
- Trois nouveaux algorithmes de MVDR basée sur des définitions différentes de l'accord de la structure dans les sous-espace;
- L'application des algorithmes proposés à la classification semi-supervisée, la classification non supervisée, et les problèmes de récupération de données biométriques, en particulier dans un contexte de la reconnaissance de personne en audio et vidéo;
- L'application des algorithmes proposés à des problèmes plus larges de reconnaissance des formes pour les données non biométriques, tels que l'image et le regroupement de texte et la recherche.

Une revue de publication dans la thèse

Le travail présenté dans cette thèse a été publié par le candidat dans les conférences et les

revues suivantes:

[ICIP2011] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Semi-supervised face recognition using LDA self-training", in Proceedings of IEEE International Conference on Image Processing (ICIP), September, 2011.

[EUSIPCO2011] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "A co-training approach to semi-supervised automatic face recognition", in Proceedings of European Signal Processing Conference (EUSIPCO), September, 2011.

[ICME2012] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Co-LDA: a semi-supervised approach to audio-visual person recognition", in International Conference of Multimedia and Exposition (ICME), July, 2012.

[EUSIPCO2012] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Multi-view Semi-supervised Dimensionality Reduction", in 2012 European Conference on Signal Processing (EUSIPCO), August, 2012.

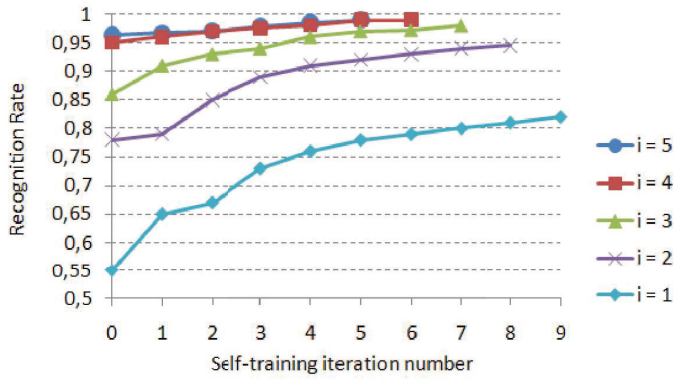
[ICASSP2013] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Open-set semi-supervised audio-visual person identification using co-training LDA and sparse representation classifiers", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

[PRL2013] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "A subspace co-training framework for multi-view clustering", accepted in Pattern Recognition Letters.

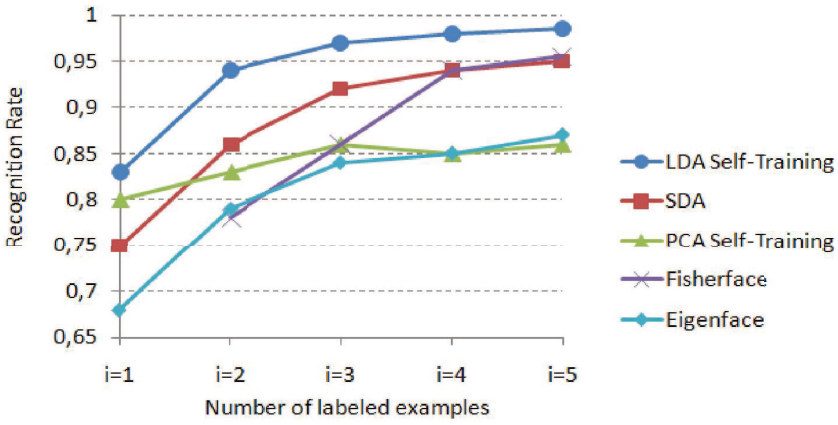
[WIFS2013] Xuran Zhao, Nicholas Evans and Jean-Luc Dugelay, "Unsupervised Multi-view Dimensionality Reduction with Application to Audio-Visual Speaker Retrieval", accepted in IEEE International Workshop on Information Forensics and Security (WIFS),2013

[ICIP2011] Reconnaissance de visage semi-supervisé en utilisant LDA auto-apprentissage

Des algorithmes de reconnaissance de visage basés sur Linear Discriminant Analysis (LDA) donnent généralement une performance satisfaisante, mais ont besoin d'un nombre important d'échantillons afin d'apprendre des projections fiables. Dans de nombreuses applications pratiques en reconnaissance de visage, il y a seulement un petit nombre d'images de visages annotées et dans ce cas les algorithmes basés sur LDA donnent généralement de mauvaises performance. Les contributions dans ce travail concentrent sur un nouvel algorithme semi-supervisé basé sur LDA qui est utilisé pour augmenter d'un ensemble d'apprentissage étiqueté manuellement avec de nouvelles données provenant d'un groupe auxiliaire non annotée et donc d'améliorer les performances de reconnaissance. Sans le coût de l'étiquetage manuel ces données auxiliaires est souvent facilement acquis mais ne sont pas normalement pas utiles pour l'apprentissage. Nous rapportons expériences en reconnaissance de visage sur 3 bases de données indépendantes qui démontrent une amélioration constante des systèmes supervisés. La performance de notre algorithme est également montré une supériorité d'autres algorithmes semi-supervisé.



Performance de reconnaissance en fonction du nombre d'itérations d'auto-apprendissage



Comparaison de précisions de reconnaissance sur la base de données ORL

[EUSIPCO2011] Une approche de co-apprentissage au reconnaissance de visages semi-supervisée

La reconnaissance de visage semi-supervisé utilisant à la fois les données annotées et non-annotées a reçu un intérêt considérable au cours des dernières années. Le co-apprendissage est l'un méthodes des plus connus d'apprentissage semi-supervisée, mais son application en reconnaissance de visages reste presque inexploré parce que son hypothèse indépendance de source peut être rarement satisfait entre deux traits du visage. Cependant, même si deux traits du visage sont corrélés, leurs caractéristiques différentes doivent produire une « marge de classification » possible entre deux classifieurs basés sur eux, et par conséquent, il y a la possibilité d'appendissqge mutuelle. Dans cet article, nous présentons un algorithme de reconnaissance de visages semi-supervisé qui applique co-apprendissage sur deux classifieurs basé sur Linear Discriminant Analysis (LDA) et Local Binary Patterns (LBP). Les résultats experimentals montrent que l'algorithme de co-apprendissqge améliore de manière significative la précision de la reconnaissance par rapport aux méthodes supervisées qui utilisent seulement des données annotées, mais aussi démontre la supériorité du co-apprendissqge sur les méthodes d'auto-apprendissage qui utiliser une seule caratéstique .

	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
Baseline LBP	63%	70%	72%	78%	80%
Baseline LDA	30%	57%	67%	78%	84%
LBP self-training	72%	73%	75%	76%	78%
LDA self-training	78%	88%	92%	93%	94%
LBP co-training	80%	82%	84%	85%	85%
LDA co-training	86%	91%	93%	95%	95%

Comparaison des performances avec des nombres différents de exemples annotées par classe

[ICME2012] Co-LDA: une approche semi-supervisé pour la reconnaissance audiovisuel du locuteur

Les modèles de clients utilisés dans les systèmes de reconnaissance de visage et de reconnaissance du locuteur sont généralement appris avec les données annotées acquises dans un petit nombre de sessions d'enrolment. La quantité de données d'apprentissage est rarement suffisante pour représenter la variation qui se produit plus tard au cours des test. De grandes quantités de données spécifiques aux clients peuvent toujours être obtenues, mais la collecte manuelle et annotation song souvent prohibitif. Le co-apprendissage, un paradigme de l'apprentissage semi-supervisé, qui peut exploiter les données non-étiquetées pour améliorer les modèles de clients faiblement appris. Dans cette article, nous proposons un algorithme de co-LDA qui utilise des données éttiquettées et non éttiquettées pour capturer une plus grande variation de l'inter-session et d'apprendre les sous-espaces discriminants dans lequels les exemples de test peuvent être classés avec plus de précision. L'algorithme proposé est naturellement adapté à la reconnaissance de locuteur audiovisuel parce que les caractéristiques biométriques vocales et visuelles répondent intrinsèquement les hypothèses de suffisance et d'indépendance qui garantissent l'efficacité du co-apprentissage. Lors d'un essai sur la base de données MOBIO, le système de co-apprendissage proposé pousse d'un taux d'identification de base de 71% à 99%, tandis que pour une tâche de vérification d'un taux d'erreur égal (EER) est réduite de 18% à environ 1% . À notre connaissance, c'est la première application réussie de co-apprentissage pour des systèmes biométriques audio-visuel.

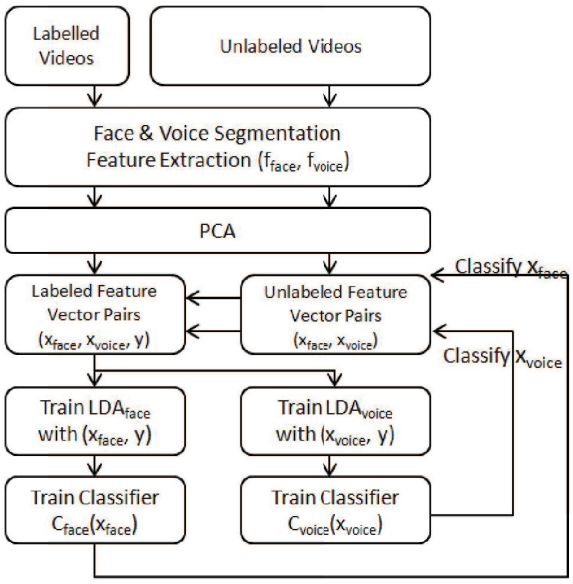
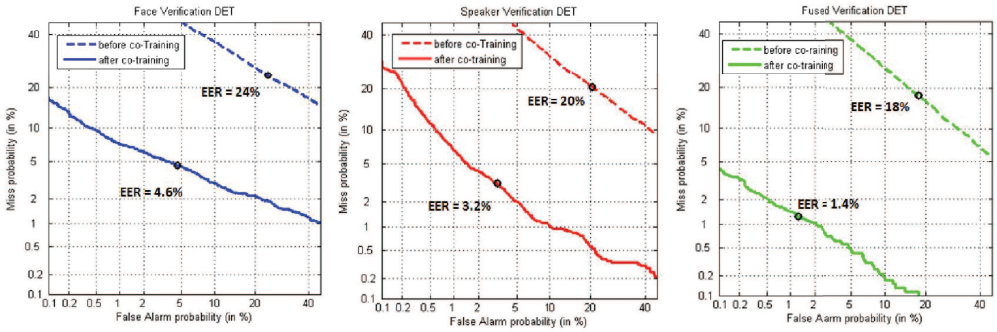


Illustration de l'algorithme de co-LDA,

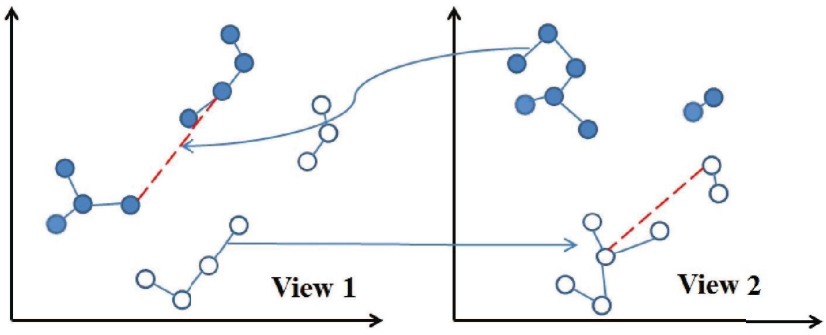


Courbes DET de vérification pour les modelité visage (à gauche), voix (au milieu), et combiné (à droite)

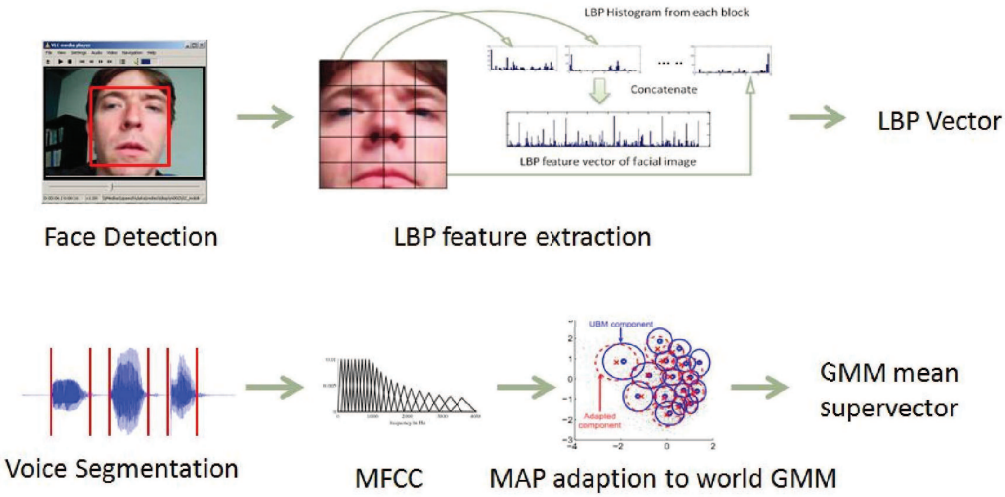
[EUSIPCO2012] Réduction de dimensionnalité multi-vues semi-supervisés pour la reconnaissance du locuteur audio-visuel

Beaucoup de systèmes biométriques utilisent des vecteurs de caractéristiques de grande dimensionalité. Les techniques de réduction de dimensionnalité évitent la supposée *malédiction de dimensionnalité*. Les approches supervisées telles que Linear Discriminant Analysis (LDA) peuvent extraire des caractéristiques discriminantes, mais souffrent de sur-apprendissage quand elles sont utilisés avec de petits ensembles de données

d'apprendissage . Par l'ajout de contraintes de proximités locales, les techniques réduction de la dimensionnalité semi-supervisés peuvent faire usage de données non-étiquetées pour améliorer les performances de la classification . Cet article présente une nouvelle, l'analyse discriminante semi-supervisée multi-vues (APSM), d'un algorithme et de son application en reconnaissance audiovisuel du locuteur. Contrairement aux approches existantes qui utilisent généralement une seule vue , APSM détermine une contrainte de voisinage plus fiable construit conjointement à partir de plusieurs vues des mêmes données . Des résultats expérimentaux sur la base de données MOBIO montrent que notre algorithme non seulement surpassent les méthodes supervisés et non-supervisés, mais il surpasse également les techniques semi-supervisésss sur vue unique.



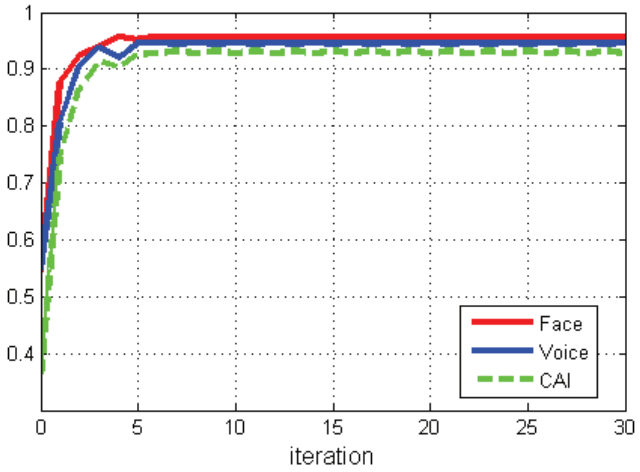
Une illustration de multi-vues contraintes de voisinage



Extraction de caractéristiques pour le visage et la voix

[PRL2013] Un cadre de partitionnement multi-vues par co-apprendissage de sous-espace

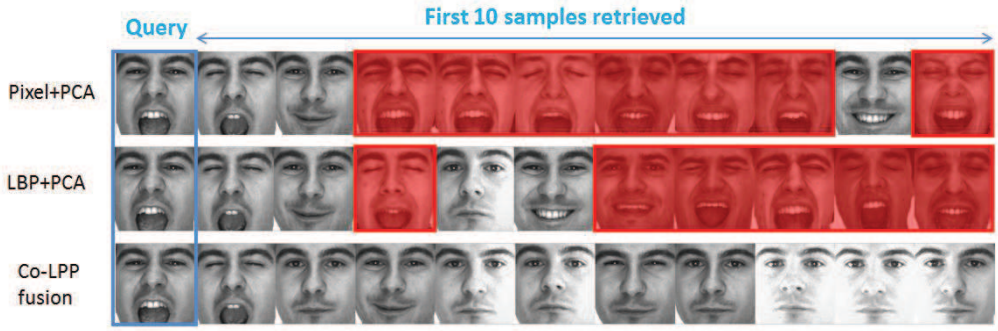
Cet article aborde le problème de la classification non supervisée avec vues multiples pour des données de grande dimension. Nous proposons un nouvel algorithme qui apprend lessous-espaces discriminants dans un mode non-supervisé sur la base de l'hypothèse que le partitionnement fiable devrait être le même dans chaque vue. Ce cadre inclut la simplicité du k-means et Linear Discriminant Analysis (LDA) au sein d'un système de co-apprendissage qui exploitent étiquettes apprises automatiquement dans une vue pour apprendre le sous-espaces dans une autre. Vue un mérite particulier, l'algorithme ne nécessite pas l'optimisation de tous les hyperparamètres. L'efficacité de l'algorithme proposé est démontré dans des test de regroupement audio-visuel, où l'amélioration significative par rapport à d'autres approches de regroupement multi-vues sont reportés. Le nouvel algorithme marche facilement avec des données hors échantillon et peut être étendu à la classification semi-supervisée.



Précision v.s. nombre d'itérations de co-apprendissage

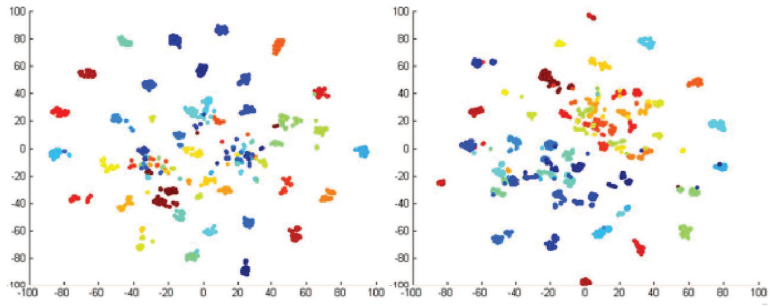
[WIFS2013] Réduction de la dimensionalité multi-vues non-supervisée en recherche biométrique audiovisuel

Cet article présente une nouvelle approche pour la réduction de dimensionalité multi-vues et son application à la recherche des données biométrique multimodales, en particulier audio-visuel. Nous proposons un nouveau concept dénommé *accord multi-vues entre sous-espace*, qui vise à déterminer un sous-espace pour chaque vue qui respecte les relations de similarité entre les points de données dans l'autre vue. L'algorithme proposé est non-supervisé, mais présente des caractéristiques discriminantes et est donc bien adapté aux applications telles que la recherche et le regroupement où les étiquettes de classe sont généralement indisponibles. L'efficacité de l'algorithme proposé est évaluée dans un expérience de recherche audio-visuel de locuteur dans des videos avec la base de données MOBIO. La approche proposée est plus performante que les autres méthods utilisant une vue ou multi-vues avec une marge importante.

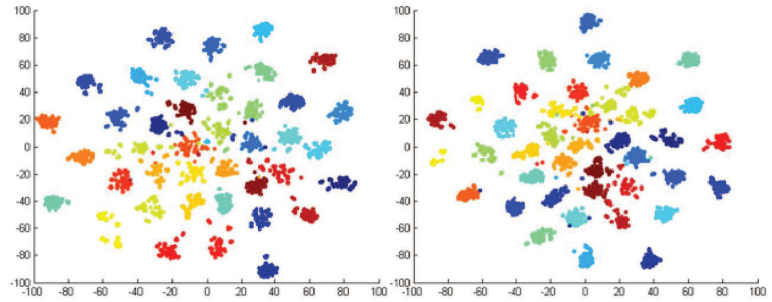


Résultat de la recherche pour deux caractéristiques seules et l'algorithme de co-lpp

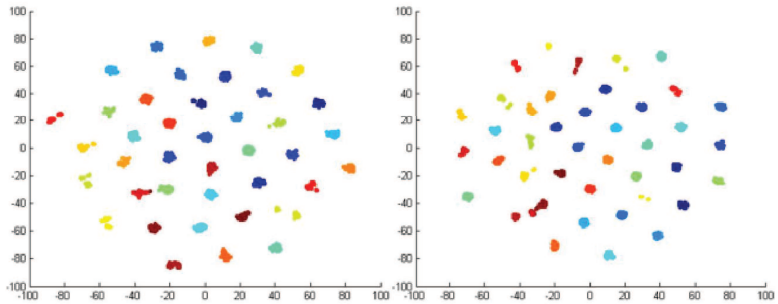
structure de données dans les sous-espace PCA



structure de données dans les sous-espace LPP



structure de données dans les sous-espace CCA



structure de données dans les sous-espace CoLPP