Correlation-based multi-view representation learning for information fusion

Wu, Zuobin

2020

Wu, Z. (2020). Correlation-based multi-view representation learning for information fusion.
Doctoral thesis, Nanyang Technological University, Singapore

<https://hdl.handle.net/10356/144264>

<https://doi.org/10.32657/10356/144264>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0
International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 23:11:01 SGT

Correlation-Based Multi-View Representation Learning for Information Fusion

Zuobin Wu

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

28/08/2020

.....

Date

A handwritten signature in black ink, appearing to read 'Zuobin Wu', with a long horizontal stroke extending to the right.

Zuobin Wu

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

28/08/2020

.....

Date

kezhi mao

.....

Prof. Kezhi Mao

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

28/08/2020

.....

Date

A handwritten signature in black ink, appearing to read 'Zuobin Wu', with a long horizontal stroke extending to the left.

Zuobin Wu

Acknowledgements

First of all, it is my pleasure to express my sincere thanks and great gratitude to my supervisor, Professor Kezhi Mao, who has given me an exciting research opportunity to pursue my PhD. I would like to express thanks to his insightful guidance, advice, discussions and experience. His expertise and insights in machine learning let me uncover the nature behind the research problems. Prof. Mao has spent a large amount of time discussing research ideas and methodologies with me and help me conduct solid research work. His encouragement and moral support are the boost for me during my PhD study. I believe the training I have received from Prof. Mao will be the treasure of my whole life.

I am also grateful to Dr. Wenyin Tang, Dr. Feng Yang, Dr. Wing Yee Sit, Dr. Xuefeng Yang, Dr. Rui Zhao and Dr. Dongzhe Wang. As my seniors, they have given me many valuable suggestions in both research and life. I would also like to thank all my friends in the Internet of Things Laboratory and Singapore for their encouragement and help.

Last but not least, I would like to express my eternal gratitude to my family members for their unconditional love. They are the best in my life!

Acronym List

LDA		Linear Discriminative Analysis
SC		Sparse Coding
CNN		Convolutional Neural Network
CCA		Canonical Correlation Analysis
ILSVRC		Large Scale Visual Recognition Challenge
VOC		Visual Object Classes
TF-IDF		Term Frequency-Inverse Document Frequency
DNN		Deep Neural Network
IMCCA		Informative Multiset Canonical Correlation Analysis
MDCA		Multiset Discriminant Correlation Analysis
PCA		Principal Component Analysis
MSCCA		Multi-view Supervised Canonical Correlation Analysis
JSR		Joint Sparse Representation
IEDCA		Intra-Class and Extra-Class Discriminative Correlation Analysis
KNN		K-nearest Neighbor
SGD		Stochastic Gradient Descent
MLP		Multilayer Perceptron
BP		Backpropagation
VSM		Vector Space Model
RQA		Recurrence Quantification Analysis
LAM		Laminarity
REC		Recurrence

SVM		Support Vector Machine
ReLU		Rectified Linear Unit
VGG		Visual Geometry Group
AP		Average Precision
mAP		mean Average Precision
t-SNE		t-Distributed Stochastic Neighbor Embedding
mRMR		minimum Redundancy Maximum Relevance
RBF		Radial Basis Function
MFCC		Mel-frequency Cepstrum
Ncut		Normalized Cut
DCT		Discrete Cosine Transform

Notations and Variables

C	Covariance matrix
x_l^i	Feature vector of l^{th} sample in the i^{th} class
$w(i, l)$	Weight corresponding to x_l^i
$\widehat{NN}_k$	k -th nearest neighbor
Λ	Diagonal matrix
I	Mutual information
ϕ	Feature map from a set to a Hilbert space
X	Sample feature set
Var	Variance
S_b	Between-class scatter matrix
S_w	Within-class scatter matrix
$\mathbb{R}$	Real numbers
$\mathbb{N}$	Natural numbers
ρ	Correlation coefficient
L	Label matrix
η	Learning rate
$\ \cdot\ _F$	Frobenius norm
$\ \cdot\ _1$	Operator norm corresponding to ℓ_1 norm

Abstract

With the advancement of information technology, a large amount of data are generated from different sources. A lot of pattern recognition systems employ a fusion of multiple sources of information to enhance learning performance. The information is collected from various domains or obtained from diverse feature extractors with heterogeneous properties. Each information can be referred to as a particular view. Single-view learning may not be sufficient to describe the underlying distribution of information. A conventional way of using multi-view information is concatenating them into one single compact information to adapt to the learning setting. However, this concatenation may not be the best way to integrate different sources of data. In contrast to single-view learning, multi-view representation learning introduces one function to model a particular view and jointly optimizes all the functions to exploit the redundant views of the same input data to improve learning performance. Although there exist various multi-view representation learning methods for information fusion, how to learn a good association among multi-view data is still an unsolved problem. In this thesis, we try to solve this problem from the perspective of correlation-based multi-view representation learning and feature fusion strategies because the correlation structure is important in multi-view representation learning.

The motivation of correlation-based multi-view representation learning is introduced in Chapter 1. Various correlation-based multi-view representation learning and deep representation learning models are reviewed in Chapter 2. Three correlation-based multi-view representation learning models are proposed in Chapter 3, Chapter 4 and Chapter 5, respectively. Chapter 6 elaborates a feature regrouping algorithm for correlation-based multi-view representation learning. These models are proposed to solve the problems in the existing correlation-based multi-view representation learning methods.

Traditional correlation-based multi-view representation learning approach may not preserve the complete and effective discriminative structure underlying multi-view

representations. In Chapter 3, we propose intra-class and extra-class discriminative correlation analysis (IEDCA) to provide more discriminative and complete representations through correlation-based multi-view representation learning. IEDCA explores both the pair-wise correlation like canonical correlation analysis (CCA) based feature fusion approaches and the correlation across different features within the same class.

Most representation generated by correlation-based multi-view representation learning may still contain redundancies due to the correlation criterion. In Chapter 4, to explore the correlation underlying different views and jointly optimize the learned representation to remove the irrelevant redundancies brought by the correlation-based multi-view representation approach, we incorporate the proposed IEDCA in Chapter 3 into minimum Redundancy Maximum Relevance (mRMR) and further improve the learning performance.

The nonlinear representation produced by kernelized correlation-based multi-view representation learning approach is limited by a fixed kernel. Moreover, the training time scales poorly with the size of the training data. In Chapter 5, we combine the flexibility of deep representation learning with the canonical correlation analysis from multiple views of data and propose deep supervised multi-view CCA (DMCCA). DMCCA can learn flexible nonlinear representations by passing multi-view data through multiple stacked layers of nonlinear transformation.

In most correlation-based multi-view representation learning approach, the natural groupings of features are directly used. The complementary and mutual information existed in multi-view representations may not be fully utilized. In Chapter 6, we propose a feature regrouping approach for correlation-based multi-view representation learning to reduce the variances of learned representations and enhance the learning performance.

In Chapter 7, we have the summary and future directions of this PhD work.

Contents

Acknowledgements	i
Acronym List	iii
Notations and Variables	v
Abstract	vi
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Overview	1
1.2 Contributions and Thesis Outline	5
2 Literature Review	9
2.1 Correlation-Based Multi-View Representation Learning	9
2.1.1 Feature Fusion by Canonical Correlation Analysis	10
2.1.2 Sparse Canonical Correlation Analysis	11
2.1.3 Discriminative Canonical Correlation Analysis	13
2.1.4 Kernel Canonical Correlation Analysis	14
2.1.5 Deep Canonical Correlation Analysis	16
2.2 Deep Feature Representation	18
2.3 Connections of Proposed Approaches with Related Work	22
3 Effective Feature Representation Learning for Pattern Classification Based on Intra-Class and Extra-Class Discriminative Correlation Analysis	25
3.1 Introduction and Motivation	26
3.2 Intra-Class and Extra-Class Discriminative Correlation Analysis . .	27
3.2.1 Intra-Class and Extra-Class Discriminative Analysis	27
3.2.2 Intra-Class Correlation Analysis	29

3.2.3	Multiset Intra-Class and Extra-Class Discriminative Correlation Analysis	31
3.3	Experimental Results	32
3.3.1	Unimodal Feature Fusion	33
3.3.2	Multimodal Feature Fusion: SDUMLA-HMT Database	35
3.3.3	Effect of Dimensions Retained for Correlation	39
3.4	Conclusions	40
4	Enhanced Feature Representation Learning through Irrelevant Redundancy Elimination in Intra-Class and Extra-Class Discriminative Correlation Analysis	41
4.1	Introduction	42
4.1.1	Motivation and Overview	42
4.1.2	Minimum Redundancy Maximum Relevance	45
4.2	Enhanced Feature Fusion Through Irrelevant Redundancy Elimination in Intra-Class and Extra-Class Discriminative Correlation Analysis	46
4.2.1	Analysis of CCA-based Representation Learning	46
4.2.2	Kernelized Intra-Class Correlation Analysis	49
4.2.3	Incorporation of Intra-Class Correlation Analysis into mRMR	50
4.2.4	IEDCA-IRE Algorithm	52
4.2.5	Computational Complexity Analysis	52
4.3	Experiments	53
4.3.1	Unimodal Feature Fusion	53
4.3.2	Multimodal Feature Fusion: SDUMLA-HMT Database	56
4.3.3	Deep Feature Fusion	58
4.3.4	Fusion of Deep Feature and Handcrafted Feature	63
4.3.5	Effect of Nearest Neighbor Parameter k	66
4.4	Conclusions	71
5	Deep Supervised Multi-View Canonical Correlation Analysis	73
5.1	Introduction	74
5.1.1	Motivation	74
5.1.2	Multi-View Canonical Correlation Analysis	74
5.2	Deep Supervised Multi-View Canonical Correlation Analysis	76
5.2.1	Derivation of DMCCA Gradient	78
5.3	Experiments and Discussions	82
5.3.1	Multimodal feature fusion: Phonetic Recognition on XRMB database	82
5.3.2	Different Fingerprint Fusion	84
5.3.3	Unimodal Feature Fusion: AR Database	86
5.4	Conclusions	87

6	Feature Regrouping for CCA-Based Feature Fusion and Extraction Through Normalized Cut	89
6.1	Introduction	89
6.1.1	Motivation	90
6.1.2	Normalized Cut	92
6.2	Feature Regrouping for CCA-based Feature Fusion and Extraction through Normalized Cut	93
6.2.1	Feature Regrouping Through Normalized Cut	94
6.2.2	Optimization of Feature Regrouping Via Normalized Cut	97
6.3	Experiments	99
6.3.1	Unimodal Feature Fusion	99
6.3.2	Multimodal Feature Fusion: SDUMLA-HMT Database	101
6.3.3	Deep Feature Regrouping	104
6.4	Conclusions	107
7	Conclusions and Future Works	109
7.1	Conclusions	109
7.2	Future Works	113
A	Analysis of IEDCA Fused Feature	115
	Author's Publications	119
	Bibliography	121

List of Figures

1.1	The ground truth and different views of synthetic data. (a) Ground Truth (b) View 1 (c) View 2 (d) View 3	2
2.1	An illustration of AlexNet with five convolutional layers and three fully connected (FC) layers.	19
2.2	An illustration of VGG with five convolutional layers and three fully connected (FC) layers.	19
2.3	An illustration of the basic residual block in ResNet.	20
3.1	The overview of intra-class and extra-class discriminative correlation analysis. The solid and hollow shapes denote data points represented by feature set X and Y, respectively. Circles, squares, and triangles denote different classes. The solid lines under intra-class correlation analysis indicate the correlation between the paired feature in each class, while the dashed lines indicate the correlation across the different features in each class.	28
3.2	Traditional correlation analysis VS Intra-class correlation analysis.	29
3.3	Multiset Intra-Class and Extra-Class Discriminative Correlation Analysis for five sample sets.	32
3.4	Iris preprocessing. (a) Sample image in SDUMLA-HMT database. (b) Valid region of iris after segmentation. (c) Generated binary iris template.	37
3.5	Fingerprint preprocessing. (a) Sample fingerprint in SDUMLA-HMT database. (b) Enhanced fingerprint. (c) Detected core point and region of interest.	38
4.1	(a) t-SNE on original features of SDUMLA-HMT database (b) t-SNE on the fused features through intra-class and extra-class discriminative correlation analysis of SDUMLA-HMT database. Each colour represents a different class.	43
4.2	An illustration of the relationship for variables X_1 , X_2 , X_3 and C, C is the target class.	48
4.3	Iris preprocessing. (a) Sample image in SDUMLA-HMT database. (b) Valid region of iris after segmentation. (c) Generated binary iris template.	59

4.4	Fingerprint preprocessing. (a) Sample fingerprint in SDUMLA-HMT database. (b) Enhanced fingerprint. (c) Detected core point and region of interest.	60
4.5	The pipeline of the application of redundancy without relevance eliminated IEDAC in deep learning. The raw images are preprocessed and forwarded into each deep neural network model: VGG, ResNet and AlexNet. Feature extraction is done on the last layer of the three models. These vectors are fed into the redundancy without relevance eliminated IEDCA, concatenated, and forwarded into softmax for classification.	61
4.6	Performance on D-case dataset, SDUMLA-HMT database, PASCAL VOC 2007 and Stack Overflow. The figures (a), (b), (c), (d) provide the classification accuracy of D-case dataset, SDUMLA-HMT database and Stack Overflow, as well as the mean average precision of PASCAL VOC 2007. The figures (e), (f), (g), (h) provide the redundancy rate.	69
4.7	The recognition performance versus feature dimension under a small range of k (a) AR database (b) PASCAL VOC-2007 database. . . .	70
4.8	The recognition performance versus different k on PASCAL VOC-2007 database.	71
5.1	A schematic of DMCCA, consisting of deep networks for K views. The networks are jointly trained with label view so that the output layers of the networks are maximally correlated.	76
5.2	Confusion matrix for speaker-dependent DCCA.	84
5.3	Confusion matrix for speaker-dependent DMCCA.	85
6.1	Visualization of correlation matrices in SDUMLA-HMT database before and after performing feature regrouping for CCA-based feature fusion and extraction through normalized cut (FR-NC), elements with zero values are represented in dark colour, elements with higher values are represented in light colour. (a) the correlation matrix of the natural feature groups. (b) the correlation matrix of the regrouped features using FR-NC.	91
6.2	The flowchart of feature regrouping for CCA-based feature fusion and extraction through normalized cut.	95
6.3	The whole process of the proposed feature regrouping for multimodal CCA-based feature fusion in SDUMLA-HMT database.	103

List of Tables

3.1	Classification accuracies (%) by SVM for different features of D-case database.	35
3.2	Classification accuracies (%) for feature fusion of MFCC, RQA and DWT in D-case database.	36
3.3	Classification accuracies (%) by SVM for individual modalities of SDUMLA-HMT database.	37
3.4	Classification accuracies (%) of multimodal fusion in SDUMLA-HMT database.	39
3.5	Average classification accuracies (%) for the different number of dimensions kept in discriminative analysis in AR database.	40
4.1	The mutual information and corresponding area.	49
4.2	Classification accuracies (%) by SVM for different features of D-case database.	56
4.3	Classification accuracies (%) for feature fusion of MFCC, RQA and DWT in D-case database.	57
4.4	Classification accuracies (%) by SVM for individual modalities of SDUMLA-HMT database.	58
4.5	Classification accuracies (%) of multimodal fusion in SDUMLA-HMT database.	58
4.6	Mean average precision (%) for the fusion of AlexNet (A), VGG-16 (V) and ResNet-50 (R) in PASCAL VOC-2007 database.	63
4.7	Classification results (%) for deep feature fusion of AlexNet, VGG and ResNet in VOC-2007 database.	64
4.8	Classification accuracies (%) in Stack Overflow dataset.	66
5.1	Classification accuracies (%) in XRMB database.	83
5.2	Individual classification accuracies (%) for fingerprint fusion on SDUMLA-HMT database.	86
5.3	Classification accuracies (%) for fingerprint fusion on SDUMLA-HMT database.	86
5.4	Classification accuracies (%) in AR database.	87
6.1	Classification accuracies (%) by SVM for different features of D-case database.	101

6.2	Classification accuracies (%) for feature fusion of regrouped features in D-case database.	102
6.3	Classification accuracies (%) by SVM for individual modalities of SDUMLA-HMT database.	104
6.4	Classification accuracies (%) of multimodal fusion of regrouped features in SDUMLA-HMT database.	105
6.5	Mean average precision (%) for the fusion of regrouped deep features from VGG-16 in PASCAL VOC-2007 database.	106
6.6	Classification results (%) for regrouped deep feature fusion (NO. of groups =30) from VGG in VOC-2007 database.	106
7.1	Classification accuracy (%) comparison of proposed methods over existing correlation-based multi-view representation methods on SDUMLA-HMT database.	111

Chapter 1

Introduction

1.1 Overview

With the advancement of information technology, a large amount of data are generated from different sources. A lot of pattern recognition systems employ the fusion of these multiple sources of information to provide enhancements in data authenticity for improved detection, reliability, confidence and reduction in data ambiguity [1]. Information fusion has been used in a variety of applications, such as odor classification [2], multimodal medical image analysis [3], image annotation [4], person re-identification [5], etc. The information is collected from various domains or obtained from diverse feature extractors and they have heterogeneous properties. Each information is referred to as a particular view. Single-view learning may not be sufficient to describe the underlying distribution of information. For example, as shown in Figure 1.1, each view may not fully represent the ground truth of data. A conventional way of using multi-view information is concatenating them into one single compact information to adapt to the learning setting. However, this concatenation may not be the best way to integrate different sources of data.

In contrast to single-view learning, multi-view representation (or feature) learning introduces one function to model a particular view and jointly optimizes all the functions to exploit the redundant views of the same input data to improve learning performance. This learning mechanism has attracted lots of attention with the increase of multi-view data in real applications where samples are depicted with multi-modal measurements of an underlying signal, such as audio and

FIGURE 1.1: The ground truth and different views of synthetic data. (a) Ground Truth (b) View 1 (c) View 2 (d) View 3

articulation, audio and video, image and text, or different information extracted from the same source. Multi-view data usually contains complementary information and multi-view representation learning exploits this point and tries to learn more comprehensive representations. Moreover, multi-view representation learning can preserve the discriminative structure underlying different information and provide a complete description of the intrinsic pattern characteristics [6]. As the performance of machine learning approaches is heavily influenced by the data representation, multi-view representation learning has received a lot of attention and become a very promising topic.

In recent years, deep learning has shown its superiority on a variety of tasks such as speech recognition[7], text classification [8], and object recognition [9]. Besides the well-known deep neural network models such as VGGNet [10], ResNet [11] and AlexNet [12], there are plenty of variants designed for different tasks with various depth, kernel size, loss function, activation function, number of hidden units, types of connections, etc. Since different networks are trained for different tasks and

objective functions, the representations taken from some networks may be more robust to certain variations, for example, illumination variation, than others, while some can capture more local information. There is no one universal network that can perform the best in all situations. Deep learning-based multi-view representation learning can reap the complementary information and enhance performance. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13], top performance is achieved by using an ensemble of different networks. There is a need of combining different deep neural network models by multi-view representation learning. Ding and Tao [14] concatenated complementary facial features extracted from a set of convolutional neural networks into a high-dimensional fused feature vector for face recognition. Liu et al. [15] integrated spatial and spectral information in deep multiple instance learning for multispectral and panchromatic imagery classification. Poria et al. [16] integrated multiple knowledge sources via feature fusion of audio, visual and text for sentiment analysis. Ma et al. [17] explored a multi-feature fusion deep networks based on autoencoder. Tang et al. [18] proposed a GoogLeNet based multi-stage feature fusion for scene recognition. Li et al. [19] exploited the multimodal fusion of convolutional neural networks for robust RGB-T tracking. The deep architectures trained on large datasets can adapt to a specific domain using multi-view measurements of the signal, for example, cross-media retrieval. Due to the powerful representation ability with multiple levels of abstraction, deep learning-based multi-view representation learning has received a lot of attention. From the perspective of neural network, multi-view representation learning learns the mid-level features from each view, and integrates them into a single compact representation. Some of the representative examples are multi-view convolutional neural network [20], multi-modal recurrent neural network [21] and multi-modal autoencoder [22].

Canonical correlation analysis (CCA) [23] is a representative approach in the studies of multi-view representation learning. It is widely used to find correlation structures in multi-view data. It is proved that CCA can find the right latent structure under certain hidden state model [24]. CCA finds linear projections of two views of the data that are maximally correlated. An appealing property of correlation-based multi-view representation learning is that, if there is noise in either view that is uncorrelated with the other view, the learned representations would not contain the noise in the uncorrelated dimensions. Another appealing property is that correlated representations might partially reconstruct the information in the

second view, if it is unavailable at test time. In machine learning, CCA has been successfully used as a dimension reduction method in multi-view datasets. For example, in [25], CCA is used in image and text to find common structures of these two different views. In [26], CCA is performed on a large unlabeled dataset to generate low dimensional features to a regression problem where the size of the labeled dataset is small. In [27], CCA is implemented between words and their context on several large corpora to generate low dimensional vector representations of words which capture useful semantic features. Multiset canonical correlation analysis is proposed as an extension of CCA to deal with multiple sets of features [28, 29]. However, such extension cannot infer the presence of latent correlations when the sample size is less than the feature dimension. Nicholas [30] proposed an informative multiset canonical correlation analysis (IMCCA) to solve this problem. IMCCA forms the block-structured matrix using a smaller subset of singular vectors. Since multiset canonical correlation analysis is essentially an unsupervised learning method, discriminant information may be overlooked when features are projected into canonical subspaces. The kernel extension of correlation-based multi-view representation learning allows the learning of nonlinear representations.

Despite different correlation-based multi-view representation learning methods for information fusion, how to learn a good association among multi-view data is still an unsolved problem. In this PhD thesis, we try to solve this problem from the perspective of correlation-based multi-view representation learning and feature fusion strategies. Since the correlation among multi-view data is the focus in our work, the proposed methods should be able to:

- Learn a more discriminative representation through correlation-based multi-view representation learning to preserve the complete and effective discriminative structure underlying multi-view representations;
- Explore the correlation underlying different views and jointly optimize the learned representation to remove the irrelevant redundancies brought by CCA-based representation learning approach and improve the learning performance;
- Learn a shared complex nonlinear representation from multiple views of data and simultaneously learn complex nonlinear transformations from each view to this shared space in which the resulting transformations are maximally

informative of each other. Combine the flexibility of deep representation learning with the canonical correlation analysis from multiple views of data;

- Fully utilize the complementary and mutual information existed in correlation-based multi-view representation learning, reduce the variances and enhance the performance of correlation-based multi-view representation learning.

To fulfill the above-mentioned objectives, we present our work and organize the outline of the thesis in the following section.

1.2 Contributions and Thesis Outline

The main contribution of this thesis is to address the limitation of correlation-based multi-view representation learning. To begin with, we analyze Canonical correlation analysis (CCA), which is the prerequisite of the methods presented in this thesis, and propose intra-class and extra-class discriminative correlation analysis (IEDCA). By exploring both the pairs-wise correlation like CCA-based representation learning approaches and the correlation across different features within the same class, IEDCA provides a more complete and effective discriminative correlation-based multi-view representation learning approach. We further analyze the CCA-based representation learning approach and find inevitable redundancies brought by the correlation criterion in the CCA-based representation learning approach. We propose an enhanced correlation-based multi-view representation learning approach. This method classifies the redundancies existed in CCA-based representation learning into relevant and irrelevant redundancies. The enhanced approach can retain only the relevant redundancies and provide an optimal correlation-based representation by removing irrelevant redundancies. By combining deep representation learning and Canonical correlation analysis, we propose deep supervised multi-view canonical correlation analysis. This correlation-based multi-view representation learning approach can capture high-level associations among different views of data and generate more abstract representations. Moreover, we find that feature representation regrouping can also benefit the performance of correlation-based multi-view representation learning. In this thesis, we also propose a feature representation regrouping algorithm for correlation-based

multi-view representation learning to help reduce the variances in the learned representation.

The related work in the correlation-based multi-view representation learning are reviewed in Chapter 2. Three correlation-based multi-view representation learning models are proposed in Chapter 3, Chapter 4 and Chapter 5, respectively. Chapter 6 elaborates a feature regrouping algorithm for correlation-based multi-view representation learning. Chapter 7 provides brief conclusions and discussions of this thesis and discusses future directions. The details of the proposed approaches in this thesis are discussed as follows:

- *Chapter 3: Effective Feature Representation Learning for Pattern Classification Based on Intra-Class and Extra-Class Discriminative Correlation Analysis:*

The method presented in this chapter is our first attempt to explore correlation-based multi-view representation learning for pattern classification tasks. In contrast to traditional CCA-based feature representation learning approaches, which only consider one-to-one pairs-wise correlation, we consider both the pairs-wise correlation and the correlation across different features within the same class. In this way, class separability is enhanced, and the proposed correlation-based representation learning method can generate more discriminative representations. In correlation-based multi-view representation learning, the dimensionality of the representations from each view (each feature set) is usually first reduced using principal component analysis (PCA), linear discriminant analysis (LDA), etc. to ensure non-singularity and invertibility of covariance matrices. One issue with this standard CCA-based representation learning is that the reduced feature sets generated by PCA or LDA may neglect certain correlation information among different feature sets which is useful for correlation analysis, and this in turn may degrade the following classification performance. To address this issue, our proposed correlation-based representation learning method can retain enough dimensions of feature for the correlation analysis and preserve the complete and effective discriminative structure underlying multi-view representations. In each feature set, class separability is enhanced by the discriminative analysis through the intra-class and extra-class nearest neighbor of each sample in the feature set. The effectiveness of our proposed method is demonstrated through both the pairs-wise

correlation and the correlation across different features within the same class for improved pattern classification. The work in this Chapter is published in [31].

- *Chapter 4: Enhanced Feature Representation Learning through Irrelevant Redundancy Elimination in Intra-Class and Extra-Class Discriminative Correlation Analysis:*

Correlation-based multi-view representation learning explores the mutual relationships of multiple views of data. However, most representation generated by correlation-based multi-view representation learning may still contain redundancies due to the correlation criterion. These redundancies may be relevant or irrelevant to class labels. The irrelevant redundancies may degrade the pattern recognition performance, while the relevant redundancies can make the pattern recognition system more robust. To address this potential limitation, we propose an enhanced feature representation learning scheme through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis (IEDCA-IRE). IEDCA-IRE explores the intra-class correlation including both the pairs-wise correlation like CCA-based feature fusion approaches and the correlation across different features within the same class. By incorporating kernelized IEDCA into minimum Redundancy Maximum Relevance (mRMR) criterion, only the relevant redundancy is retained in the fused feature. This part of work is published in [32].

- *Chapter 5: Deep Supervised Multi-View Canonical Correlation Analysis:*

Traditional CCA-based multi-view representation learning tries to find linear projections of two views of the data that are maximally correlated. Its kernel extension allows the learning of nonlinear representations. However, the learned representation by kernel CCA is limited by the fixed kernel. Moreover, the training time of kernel CCA scales poorly with the size of training data as it is a nonparametric method. Although Canonical Correlation Analysis and the kernel extension can effectively model the relationship between two sets of variables, they are limited in dealing with high-level associations among multi-view data. With the development of deep neural networks, we propose deep supervised multi-view canonical correlation analysis (DMCCA) to solve this problem, with a strategy of learning a joint representation that

is associated with multiple views of data at a higher level after learning several layers of view-specific representations from lower layers. DMCCA tries to learn the shared complex nonlinear representation from multiple views of data and simultaneously learn complex nonlinear transformations from each view to this shared space in which the resulting transformations are maximally informative to each other. It combines canonical correlation analysis from multiple views of data with the flexibility of deep representation learning. The label information is leveraged for learning better representations and embedded as a view into the whole networks at the training stage. During training, the label information and the network structure of other views of data are jointly trained to update the parameters of the whole networks and maximize the total correlation simultaneously. This part of work is in [33].

- *Chapter 6: Feature Regrouping for CCA-Based Feature Fusion and Extraction Through Normalized Cut:*

In traditional CCA-based feature fusion and extraction, the natural groupings of features are directly used. It is still unclear whether the natural groupings of features are optimal for CCA-based fusion. In this chapter, we propose a feature regrouping algorithm for CCA-based feature fusion and extraction through normalized cut (FR-NC). Feature correlation analysis is incorporated into normalized cut, in which the intra-group correlation is maximized, and the extra-group correlation is minimized simultaneously. CCA-based feature fusion is performed on the regrouped features. The proposed feature regrouping algorithm aims to provide enhanced fused features for pattern classification. The work in this Chapter is published in [34].

Chapter 2

Literature Review

In this chapter, the state-of-the-art work that is related to our proposed methods are reviewed from the perspective of correlation-based multi-view representation learning and deep feature representation methods. In Section 2.1, we start our review from classical canonical correlation analysis (CCA), sparse CCA, discriminative CCA, kernel CCA and deep CCA. Then we discuss the state-of-the-art deep feature representations learning methods in Section 2.2. Many of these correlation-based multi-view representation learning approaches and deep feature representations learning methods are exploited in our proposed methods in Chapter 3, Chapter 4, Chapter 5 and Chapter 6.

2.1 Correlation-Based Multi-View Representation Learning

Correlation-based multi-view representation learning is a representative technique in the studies of multi-view representation learning with an objective to maximize the correlations of variables among multiple views by CCA. In this section, we will review correlation-based multi-view representation learning methods: CCA, sparse CCA, discriminative CCA, kernel CCA, and deep CCA.

2.1.1 Feature Fusion by Canonical Correlation Analysis

Canonical Correlation Analysis [23] increasingly becomes popular since it can effectively model the relationship among different sets of variables. From the perspective of multi-view representation learning, Canonical Correlation Analysis (CCA) computes a shared embedding of different sets of variables by maximizing the correlations among these different sets. To be more specific, CCA is widely used in the tasks of multi-view representation learning to generate low-dimensional representations [35, 36]. It has a variety of applications such as dimensionality reduction [37, 38], clustering [39, 40], discriminant learning [41, 42], word embeddings [43], and regression [44, 45].

Suppose there are two feature sets from the same n objects denoted by $X = [x_1, x_2, \dots, x_n] \in R^{p \times n}$ and $Y = [y_1, y_2, \dots, y_n] \in R^{q \times n}$, respectively, where x_i (or y_i) can be treated as a sample. CCA is used to find optimal pairs of correlation projection matrices W_x and W_y so that $W_x^T X$ and $W_y^T Y$ have maximum correlation with each other. The following correlation function is maximized to find W_x and W_y :

$$\text{corr}(U, V) = \frac{\text{cov}(U, V)}{\text{var}(U) \cdot \text{var}(V)} \quad (2.1)$$

where $U = W_x^T X$, $V = W_y^T Y$, $\text{cov}(U, V) = W_x^T C_{xy} W_y$, $\text{var}(U) = W_x^T C_{xx} W_x$ and $\text{var}(V) = W_y^T C_{yy} W_y$. $C_{xy} \in R^{p \times q}$ is the cross-set covariance matrix and it is defined as:

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)^T \quad (2.2)$$

where $\mu_x = \frac{1}{n} \sum_{i=1}^n x_i$ and $\mu_y = \frac{1}{n} \sum_{i=1}^n y_i$ are the means of the two views, respectively. $C_{xx} \in R^{p \times p}$ and $C_{yy} \in R^{q \times q}$ are the within-set covariance matrices and can be obtained similarly. As the correlation is invariant to scaling, the maximization of the above-mentioned criterion can be solved by maximizing $\text{cov}(U, V)$ with constraints $\text{var}(U) = \text{var}(V) = 1$. W_x and W_y are obtained through solving the following generalized eigenvalue problem [46]:

$$\begin{cases} C_{xy}C_{yy}^{-1}C_{xy}^TW_x = \lambda^2C_{xx}W_x \\ C_{xy}^TC_{xx}^{-1}C_{xy}W_y = \lambda^2C_{yy}W_y \end{cases} \quad (2.3)$$

The correlation projection matrices $W_x \in R^{p \times d}$ and $W_y \in R^{q \times d}$ correspond to top d eigenvalues. The projected correlation features can be extracted in the form of W_x^TX and W_y^TY , and concatenated into a fused feature vector.

There also exist other different ways defining the canonical correlations for two matrices besides the above definition. They are proved to be equivalent [47]. In [28], Kettenring demonstrates CCA is equivalent to a constrained least-square problem. In [47], Golub provides a classical two-step algorithm to perform CCA. QR decomposition is firstly conducted on the whitened data matrices, followed by an SVD of the whitened covariance matrix. However, this two-step algorithm becomes extremely slow with the big data matrices. To solve this problem, Avron [48] proposes a fast CCA algorithm. The idea is inspired by subsampled randomized Walsh-Hadamard transform [49]. In fast CCA, the subsampled small proportion of training data are used to approximate matrix product. Moreover, Lu [24] considers sparse design matrices and comes up with an iterative regression algorithm to calculate the canonical correlations for large scale data matrices.

Although CCA can perform multi-view feature learning and it is extensively used in a variety of fields, there still exist some limitations. For example, traditional CCA neglects the non-linearity which often exists in multi-view data. To overcome the limitations of CCA, a lot of CCA-based algorithms have been proposed. In the following sections, we will review some typical extensions such as sparse CCA, discriminative CCA, kernel CCA, and deep CCA.

2.1.2 Sparse Canonical Correlation Analysis

Sparse representation has also proven to be a powerful tool in pattern recognition. It assumes that there exists a linear combination of training samples to sufficiently represent a query sample. In sparse representation, the coefficients of the linear representation are computed, followed by the calculation of the reconstruction residuals. The query sample will be classified according to the minimum reconstruction residual. Joint sparse representation (JSR) [50] is regarded as a

multimodal feature fusion approach, where multiple dictionaries are created from training samples of each modality. The objective of JSR is to find joint sparse representation coefficients that have the sparsity patterns in common. The query sample is then reconstructed by the multimodal training samples belonging to the same class. Multimodal task-driven dictionary learning [51] is developed to enforce the collaboration of multiple sources of information.

Sparse Canonical Correlation Analysis (sparse CCA) has also attracted a lot of attention in the field of multi-view representation learning. The sparsity is explored in CCA due to the following aspects. First, the sparsity can account for predicted results. Second, regularization and stability play important roles in sparsity. Appropriate regularization can help eliminate the influence of noisy data and reduce the sensitivity of CCA when there are only limited observations. Additionally, sparse CCA can be formulated into a subset selection to reduce dimension and provide a stable solution.

Most sparse CCA approaches are based on LASSO [52]. LASSO is a shrinkage and selection method for linear regression. In [36], Sun proposes a sparse CCA through formulating CCA into a least square problem. This method is used in the application of multi-label classification. It can directly compute canonical correlations through the least angle regression algorithm. Moreover, this least square solution can facilitate the incorporation of unlabeled data into CCA and help capture the local geometry of the data. In [53], graph Laplacian is embedded to deal with the unlabeled data.

The advance of sparse CCA is closely related to the development of sparse PCA [54]. Classical solutions to sparse PCA lead to generalized eigenvalue problem. Similar approaches are extended to solve sparse CCA [55, 56]. For example, Torres et al. [55] extend the approach for solving sparse eigenvalue problem and develop a sparse CCA algorithm through difference of convex functions. In [56], Wiesel proposes a greedy backward algorithm to solve the problem of sparse CCA through bounding the canonical correlations during each stage. In [57], Witten et al. use penalized matrix decomposition to solve the covariance matrix C_{xy} . This method leads to a penalized sparse CCA. Similarly, by extending penalized CCA with structured sparsity, structured sparse CCA has been proposed in [58].

2.1.3 Discriminative Canonical Correlation Analysis

The class-label information is not utilized in the formulation of CCA. To improve CCA for obtaining discriminative features, Gatto et al. [59] and Sun et al. [60] proposed discriminative CCA (DCA) by combining the advantages of CCA and linear discriminant analysis (LDA). These two DCA methods are closely related, and they use two views of the same underlying semantics together with the class-label information to extracted features. DCA utilizes class-label information through the maximization of the correlation between feature vectors within the same class and provides more discriminative features. The optimization of DCA is usually written as:

$$\max_{w_x, w_y} (w_x^T C_w w_y - \eta \cdot w_x^T C_b w_y) \quad (2.4)$$

with the constraints that $w_x^T X X^T w_x = 1$, $w_y^T Y Y^T w_y = 1$. C_w and C_b denote the within-class correlations and the between-class correlations, respectively, such that

$$C_w = \sum_{i=1}^c \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} (x_k^i) (y_l^i)^T \quad (2.5)$$

$$C_b = \sum_{i=1}^c \sum_{j=1, j \neq i}^c \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (x_k^i) (y_l^j)^T \quad (2.6)$$

where c is the number of classes, n_i is the number of instances that belong to class i , and x_j^i denotes the j -th example in class i . The solution of (2.4) is similar to CCA. DCA has the limitation of extracting fewer features than the number of classes because the rank of C_w is $c - 1$ when the number of features for X and Y is greater than c (high-dimensional data). Although DCA can provide more discriminative features by using class-label information in CCA, it is proved that the projective directions of DCA are equal to the directions of LDA with orthogonal constraints [61]. They are essentially supervised dimensionality reduction methods. Moreover, DCA has the same generalization problem as CCA since they have the sample covariance matrices.

2.1.4 Kernel Canonical Correlation Analysis

Canonical Correlation Analysis is a linear multi-view representation learning method. However, in many real-world applications, the relationship among multiple views of data is often nonlinear. It is very difficult for the linear CCA to capture all underlying relationships of multi-view data [62]. With the advent of kernel trick, kernelized Canonical Correlation Analysis (KCCA) [63] provides a solution to capture the underlying nonlinear relationships. KCCA can be considered as a nonlinear extension of CCA, and it has been widely used in many applications such as acoustic feature learning [64], multi-view clustering [65], cross-media information retrieval [66] and independent component analysis [67].

The main idea behind KCCA is by projecting X' and Y' into a higher dimensional space:

$$\phi : x' \rightarrow \phi_{x'}(x'), y' \rightarrow \phi_{y'}(y') \quad (2.7)$$

correlation analysis is performed in the above higher dimensional space. In this higher dimensional space, it is difficult to conduct normal calculation. To avoid the calculation of complex nonlinear function, the kernel method constructs an inner product kernel function by:

$$k(x', z') = \langle \phi(x'), \phi(z') \rangle \quad (2.8)$$

where $x', z' \in X'$. The maximization of the criterion (2.1) remain the same, where U and V become:

$$U = W_{\phi(x')}^T \phi_{x'}(X') \quad (2.9)$$

$$V = W_{\phi(y')}^T \phi_{y'}(Y') \quad (2.10)$$

The directions of $W_{\phi(x')}$ and $W_{\phi(y')}$ can be expressed as the projection of X' and Y' on W_x and W_y :

$$W_{\phi(x')} = \phi_{x'}(X') W_x \quad (2.11)$$

$$W_{\phi(y')} = \phi_{y'}(Y') W_y \quad (2.12)$$

Substitute (2.11) and (2.12) in (2.1), let $K_x = (\phi_{x'}(X'))^T \phi_{x'}(X')$ and $K_y = (\phi_{y'}(Y'))^T \phi_{y'}(Y')$, $\text{cov}(U, V)$, $\text{var}(U)$ and $\text{var}(V)$ become:

$$\text{cov}(U, V) = W_x^T K_x K_y W_y \quad (2.13)$$

$$\text{var}(U) = W_x^T K_x^2 W_x \quad (2.14)$$

$$\text{var}(V) = W_y^T K_y^2 W_y \quad (2.15)$$

Similarly, maximization of (2.1) can be performed by Lagrange multiplier on the covariance between U and V subject to $\text{var}(U) = \text{var}(V) = 1$. The transformation matrices, W_x and W_y can be obtained through solving a generalized eigenvalue problem with the corresponding first s largest eigenvalues:

$$W_x = [a_1, a_2, \dots, a_s] \quad (2.16)$$

$$W_y = [b_1, b_2, \dots, b_s] \quad (2.17)$$

$$W_{\phi(x')}^T \phi_{x'}(X') = [a_1, a_2, \dots, a_s]^T K_x^T \quad (2.18)$$

$$W_{\phi(y')}^T \phi_{y'}(Y') = [b_1, b_2, \dots, b_s]^T K_y^T \quad (2.19)$$

where a_i and b_i indicate the directions of the i -th pair of canonical projective vectors. To avoid the computational problems and degeneracy caused by large training sets, incomplete Cholesky decomposition (ICD) is adopted for the dimension reduction of the kernel matrices.

In KCCA, the statistical property of KCCA has received a lot of attention and it is mainly from the following aspects [67, 68]. In [69], the statistical convergence of KCCA is proved by Fukumizu et al. by providing rates for regularization parameters. In [70], a more detailed theoretical analysis of KCCA is provided by Haroon and Shawe-Taylor. Moreover, through regression formulation, they also provide a

finite sample statistical analysis for KCCA. Later in [71], a convergence rate analysis of KCCA is provided by Cai and Sun under an approximation assumption. However, how to choose appropriate regularization parameters remains an open problem.

Similar to PCA, there is a closed-form solution for KCCA through the eigenvalue system. However, the computational complexity becomes extremely large with the increase of training data. Therefore, a variety of approximation methods have been proposed through the construction of low-rank approximation of kernel matrices, such as block incremental SVD [72, 73], partial Gram-Schmidt orthogonalization [46], and Cholesky decomposition [67]. It is worth noting that Nystrom’s method [74] is used to speed up kernel machines [75, 76]. This method can be divided into two stages. The first stage is eigendecomposition to a lower-dimensional space, followed by the expansion of the results back to the original space.

2.1.5 Deep Canonical Correlation Analysis

The objective function in Canonical Correlation Analysis can be integrated with neural networks to capture high-level associations among multi-view data. For example, Becker and Hinton [77] propose a multi-layer nonlinear canonical correlation through the maximization of the covariation between two neural networks. They assume different parts of the perceptual input share the cause. Later, by the maximization of the mutual information among different networks, Becker [78] extracts higher-level features from the coherence inputs.

In [79, 80], Lai and Fyfe propose a neural network implementation of CCA. Instead of maximizing the canonical correlation, they only maximize the correlation among the network outputs from different views. Moreover, by using three feedforward neural networks, Hsieh [81] develops a nonlinear CCA approach. These three networks play different roles in this approach. The first one tries to maximize the canonical correlation of the other two neural network neurons). The other two try to project the canonical variates of the original variables.

Canonical Correlation Analysis is only partially involved in the above-mentioned CCA-based neural network models. The full integration of CCA and deep neural network, which is referred to deep CCA, has recently been proposed in [82].

Inspired by the development of deep learning, Andrew et al. [82] integrate CCA and deep neural network models. Consequently, the learned deep nonlinear representations of two views, X and Y , are maximally correlated with each other. In deep CCA, multiple stacked layers of nonlinear mappings are used to learn the two-view representations. Let us assume a network with d layers. In deep CCA, it first learns representation from $f_x(x) = h_{W_x, b_x}(x)$ with parameters $(W_x, b_x) = (W_x^1, b_x^1, W_x^2, b_x^2, \dots, W_x^d, b_x^d)$, where $W_x^l(ij)$ represents the parameters connecting unit i in layer l , and unit j in layer $l + 1$. Moreover, $b_x^l(j)$ represents the bias associated with unit j in layer $l + 1$. For a sample from the second view, its representation $f_y(y)$ is computed similarly with different parameters (W_y, b_y) . The objective of deep CCA is to jointly learn these parameters for both views such that the correlation, $\text{corr}(f_x(X), f_y(Y))$, is as high as possible. Let θ_x represent the vector of all parameters (W_x, b_x) of the first view and similarly for θ_y , then

$$\underset{(\theta_x, \theta_y)}{\text{argmax}} \text{corr}(f_x(X; \theta_x), f_y(Y; \theta_y)) \quad (2.20)$$

The parameters are estimated through gradient-based methods during the training of deep neural networks. For example, the parameters (θ_x^*, θ_y^*) are estimated through the training data using the gradient of the correlation objective, with the batch-based algorithms such as L-BFGS in [82] and stochastic optimization with mini-batches [83].

Deep CCA has been extensively used in representation learning tasks where two views of data are provided. In [84], a joint latent space for the task of image caption is learned through deep CCA with GPU implementation. In [83], the deep nonlinear embeddings of two languages are learned through deep CCA to exploit multilingual context.

Deep canonically correlated autoencoder [85] has also been recently developed by combining the advantages of autoencoder-based methods and deep CCA. This model consists of two autoencoders. In deep canonically correlated autoencoder, both the canonical correlation between the learned bottleneck representations and the reconstruction error of the autoencoders are optimized. In this way, the model can offer a trade-off between the information which is captured in the embedding within each view, and the information in the relationship across different views.

2.2 Deep Feature Representation

Deep learning [86] has become one of the most widely used representation learning methods nowadays. It explores multiple levels of representation. The higher level feature representation can provide more abstract information. The mid-level representations are transformed from the lower-level representation in a hierarchical architecture. Same as traditional machine learning approaches, deep learning methods can be categorized into unsupervised learning, semi-supervised learning and supervised learning. In this thesis, we mainly focus on supervised learning approaches and apply in acoustic scene classification, multimodal phonetic recognition, multimodal biometric recognition, image classification and text classification. Inspired by the mechanism of computation in the human brain, deep neural network imitates the biological neural system and conducts in a type of feed-forward neural network with many computational neurons. Deep neural networks (DNNs) are generally trained with gradient-based optimization methods such as back-propagation [87] and stochastic Gradient Descent [88]. Convolutional neural network (CNN) is a representative example of deep neural networks. It was firstly proposed by Fukushima back in 1988 [89] which has not been widely spread until 1990s. Starting from 2000s, CNN has been successfully used in many pattern recognition tasks such as speech recognition [7], image recognition [9, 90–92], text mining [8], object detection [93, 94] and image segmentation [95, 96].

AlexNet [12] is a typical CNN. It contains convolutional layers wrapped with pooling steps and followed with fully connected layers. These are also known as the dense layers in multilayer neural network. Figure 2.1 illustrates an AlexNet framework with five convolutional layers (conv) and three fully-connected layers (FC). Since the convolutional neural network is usually trained using a huge amount of data, the learning of CNNs is extremely computationally expensive. For instance, an AlexNet with eight layers [12] has 650,000 neurons. This will result in 60 million parameters. It takes more than one week to train through backpropagation using GPUs on ImageNet ILSVRC [97] dataset.

Starting with AlexNet, CNN-based architecture develops rapidly. To solve the expensive computational problem, ZFNet [98] was proposed through reducing the parameters of the network in AlexNet. Global average pooling is used to replace

FIGURE 2.1: An illustration of AlexNet with five convolutional layers and three fully connected (FC) layers.

FIGURE 2.2: An illustration of VGG with five convolutional layers and three fully connected (FC) layers.

FIGURE 2.3: An illustration of the basic residual block in ResNet.

fully connected layers. In this way, the parameters of the network are significantly reduced. Meanwhile, the overfitting issue is avoided. Following AlexNet, VGG [10, 99] was a series of AlexNet-based convolutional neural network. Figure 2.2 demonstrates a VGG framework with five convolutional layers and three fully-connected layers. VGG demonstrates that the depth of convolutional neural network architecture is an intractable factor to enhance the generalization and performance in pattern recognition tasks.

Another framework to solve the expensive computational problem is GoogLeNet [100]. In GoogLeNet, inception layers are incorporated to reduce the expensive computational complexity aroused in CNN architecture. Residual network architecture ResNet [11] was proposed as an ultra-deep neural network. By using the residual connections which include the input to the layer in addition to nonlinear transform, ResNet is able to be as deep as to 1202 layers in a single end-to-end architecture. In practice, residual mapping is often easier to optimize. The residual module of ResNet is illustrated in Figure 2.3. ResNet follows VGG's full 3×3 convolutional layer design. The residual block has two 3×3 convolutional layers with the same number of output channels. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function. GoogLeNet uses four blocks made up of Inception blocks. However, ResNet uses four modules made up of residual blocks, each of which uses several residual blocks with the same number of output channels. There are also some recently proposed networks which continuously improved the state-of-the-art pattern classification performance, such as CapsuleNet [101], DenseNet [102], etc. In the following chapters, many aforementioned CNN baselines are implemented as deep feature representation extractors in our proposed methods.

The well-trained CNN usually has good adaptability of parameters transferring. They can be applied to a specific task with appropriate fine-tuning. For example, Oquab et al. [103], Caffe [104] and DeCAF [105] apply the deep representations obtained from the pre-trained models and re-trained the adaptation layers for a different task. The activations of the penultimate layer of CNN can be regarded as powerful CNN feature vectors [106].

2.3 Connections of Proposed Approaches with Related Work

In spite of different correlation-based multi-view representation learning methods for information fusion, how to learn a good association among multi-view data is still an unsolved problem. In the following chapters, we try to solve this problem from the perspective of correlation-based multi-view representation learning and feature fusion strategies. The connection of our proposed approaches in the following chapters and their related work discussed in this literature review section are demonstrated as follows. The discussion will follow the order of the subsections with respect to their corresponding related work in this thesis. We solve the issues existed in the current correlation-based multi-view representation learning approach from four perspectives.

First, traditional correlation-based multi-view representation learning approach may not preserve the complete and effective discriminative structure underlying multi-view representations. In Chapter 3, we propose intra-class and extra-class discriminative correlation analysis (IEDCA) to provide more discriminative and complete representations through correlation-based multi-view representation learning. In Chapter 3, IEDCA explores both the pair-wise correlation like CCA-based feature fusion approaches and the correlation across different features within the same class. Therefore, it is highly related to the existing methods discussed in Section 2.1.1.

Second, most representation generated by correlation-based multi-view representation learning may still contain redundancies due to the correlation criterion. In Chapter 4, to explore the correlation underlying different views and jointly optimize the learned representation to remove the irrelevant redundancies brought by the correlation-based multi-view representation approach, we incorporate the proposed IEDCA in Chapter 3 into minimum Redundancy Maximum Relevance (mRMR) and further improve the learning performance.

Third, the nonlinear representation produced by the kernelized correlation-based multi-view representation learning approach is limited by a fixed kernel. Moreover, the training time scales poorly with the size of the training data. In Chapter 5, we

combine the flexibility of deep representation learning with the canonical correlation analysis from multiple views of data and propose deep supervised multi-view CCA (DMCCA). DMCCA can learn flexible nonlinear representations by passing multi-view data through multiple stacked layers of nonlinear transformation. This work draws inspiration from deep CCA in Section 2.1.5.

Fourth, in most correlation-based multi-view representation learning approach, the natural groupings of features are directly used. The complementary and mutual information existed in multi-view representations may not be fully utilized. In Chapter 6, we propose a feature regrouping approach for correlation-based multi-view representation learning to reduce the variances of learned representations and enhance the learning performance.

In addition, The deep feature representations mentioned in Section 2.2 are extensively used as the baseline representations in Chapter 4 and Chapter 6.

Chapter 3

Effective Feature Representation Learning for Pattern Classification Based on Intra-Class and Extra-Class Discriminative Correlation Analysis

In this chapter, we demonstrate one of our proposed correlation-based multi-view representation learning frameworks which considers both the pairs-wise correlation and the correlation across different features within the same class. The proposed correlation-based multi-view representation learning framework can preserve the complete and effective discriminative structure underlying multi-view representations. The effectiveness of our proposed model is demonstrated through several classification tasks on both the object-centric datasets and the scene-centric datasets.

The rest of this chapter is organized as follows. The motivation and overview of this work are provided in Section 3.1. The details of the proposed Intra-Class and Extra-Class Discriminative Correlation Analysis and its extension to deal with

multiple views of data are elaborated in Section 3.2. In Section 3.3 and Section 3.4, we demonstrate the experimental results and provide conclusions.

3.1 Introduction and Motivation

Correlation-based multi-view representation learning encounters certain limitations. If the dimension of any view exceeds the number of samples ($p > n$ or $q > n$), covariance matrices C_{xx} and C_{yy} become singular and non-invertible. This is often referred to small sample problem. To overcome this problem, one solution is to reduce the feature dimension using principal component analysis (PCA) before performing CCA. However, the negligence of class label information in PCA may deteriorate the classification performance. In this sense, linear discriminant analysis (LDA) may be a better choice.

Linear discriminant analysis is a classical supervised technique for dimension reduction and continues to be a popular tool. For a problem of c classes, LDA projects the data into a $(c - 1)$ -dimension space through maximizing between-class scatter matrix and minimizing within-class scatter matrix. The between-class and within-class scatter matrices are given respectively as follows:

$$S_b = \sum_{i=1}^c n_i (m_i - m) (m_i - m)^T \quad (3.1)$$

$$S_w = \sum_{i=1}^c \sum_{j=1}^{n_i} (x_j - m_i) (x_j - m_i)^T \quad (3.2)$$

where n_i is the number of samples for class i , m_i is the mean vector of class i , and $m = \frac{1}{n} \sum_{i=1}^n x_i$ is the mean vector of the whole training set. Through maximizing the trace ratio of between-class scatter matrix and within-class scatter matrix, the projection matrix is obtained as:

$$W = \operatorname{argmax}_W \frac{\operatorname{tr}(W^T S_b W)}{\operatorname{tr}(W^T S_w W)} \quad (3.3)$$

where $tr()$ represents the trace of a matrix. Instead of solving Eq. (3.3) directly, the trace differential solution is usually performed as [107]:

$$W = \operatorname{argmax}_W \operatorname{tr} (W^T (S_b - S_w) W) \quad (3.4)$$

The transformation matrix W corresponds to the top d eigenvalues of the above equation. Linear Discriminant Analysis (LDA) can enhance class separability by using class label information. However, the discriminating structure from LDA may not be inherited by CCA since it is unsupervised. The utility of CCA in fusing heterogeneous data for classification is limited since it cannot guarantee that learned correlated features are discriminative. Supervised CCA methods [108, 109] have been proposed to preserve the discriminating structure and can be applied together with LDA. But LDA reduces the original feature dimension to at most $(c - 1)$ discriminant features, where c is the number of classes. This may neglect some information which is important for the subsequent correlation analysis and eventually degrade the performance. Therefore, we propose to seek transformations that can retain enough dimension of features for class separation under each set of features, as well as inherit the discriminating structure in the learned correlation features.

3.2 Intra-Class and Extra-Class Discriminative Correlation Analysis

The basic idea of Intra-Class and Extra-Class Discriminative Correlation Analysis (IEDCA) is illustrated in Figure 3.1. IEDCA attempts to find transforms where classes are separated within each feature set while the discriminating structure is inherited in the correlation analysis. It consists of two stages, a discriminative analysis followed by correlation analysis.

3.2.1 Intra-Class and Extra-Class Discriminative Analysis

Assume the samples are from c separate classes. Let x_l^i denotes the feature vector of l^{th} sample in the i^{th} class. $m_e(x_l^i)$ denotes the extra- class k-nearest neighbor

FIGURE 3.1: The overview of intra-class and extra-class discriminative correlation analysis. The solid and hollow shapes denote data points represented by feature set X and Y, respectively. Circles, squares, and triangles denote different classes. The solid lines under intra-class correlation analysis indicate the correlation between the paired feature in each class, while the dashed lines indicate the correlation across the different features in each class.

(KNN) mean. $w(i, l)$ denotes the weight corresponding to x_l^i . The between-class scatter matrix is thus defined as:

$$S_b^n = \sum_{i=1}^c \sum_{l=1}^{n_i} w(i, l) (x_l^i - m_e(x_l^i)) (x_l^i - m_e(x_l^i))^T \quad (3.5)$$

where the extra-class KNN mean $m_e(x_l^i)$ is defined as:

$$m_e(x_l^i) = \frac{1}{k} \sum_{p=1}^k \widehat{NN}_p(x_l^i)_{extra} \quad (3.6)$$

$\widehat{NN}_p(x_l^i)_{extra}$ is the p -th nearest neighbor from extra-class to x_l^i , and weight $w(i, l)$ is defined as:

$$w(i, l) = \frac{\min \left\{ d(x_l^i, \widehat{NN}_k(x_l^i)_{intra}), d(x_l^i, \widehat{NN}_k(x_l^i)_{extra}) \right\}}{d(x_l^i, \widehat{NN}_k(x_l^i)_{intra}) + d(x_l^i, \widehat{NN}_k(x_l^i)_{extra})} \quad (3.7)$$

$d(v_1, v_2)$ denotes the Euclidean distance between vector v_1 and v_2 . $\widehat{NN}_k(x_l^i)_{intra}$ is the k -th nearest neighbor from intra-class (class i) to x_l^i . The weighting function emphasizes the boundary information which is important for pattern classification. The between-class scatter matrix, S_b^n can be diagonalized as follows:

$$Q^{-1} S_b^n Q = \Lambda \quad (3.8)$$

FIGURE 3.2: Traditional correlation analysis VS Intra-class correlation analysis.

where Q denotes the right eigenvectors and Λ denotes the diagonal matrix of descending eigenvalues of S_b^n , respectively. By choosing top r eigenvalues and their corresponding eigenvectors ($r \geq c - 1$):

$$Q_{(r \times p)}^{-1} S_b^n Q_{(p \times r)} = \Lambda_{(r \times r)} \quad (3.9)$$

where p is the number of features. Instead of reducing the dimension directly to $c - 1$, the reasonable dimension of feature is kept before correlation analysis:

$$X'_{r \times n} = Q_{(r \times p)}^T X_{(p \times n)} \quad (3.10)$$

Similarly, Y can be transformed to Y' . In this way, more information on the classification structure is preserved.

3.2.2 Intra-Class Correlation Analysis

Now X and Y are transformed to X' and Y' , respectively. Traditional correlation analysis only considers one-to-one correlation. The class structure which contains discriminative information is neglected. Intra-class correlation explores not only one-to-one correlation but also the correlation across different features within the same class. Figure 3.2 demonstrates the difference between traditional correlation analysis VS intra-class correlation analysis.

To utilize intra-class correlation, we introduce a class matrix $L \in R^{n \times c}$ (c denotes the number of the classes, while n denotes the number of training samples). Each

row of the class matrix L represents the sample label. For instance, if there are three classes ($c = 3$), $[1 \ 0 \ 0]$, $[0 \ 1 \ 0]$ and $[0 \ 0 \ 1]$ denote class 1, class 2 and class 3, respectively. The covariance matrix between the transformed feature sets with class correlation is:

$$C_L = (X'HL)(Y'HL)^T \quad (3.11)$$

where the centering matrix $H \in R^{n \times n}$ is defined as $H = I - n^{-1}ee^T$, and e is a vector of all ones. To obtain the transformation that maximizes both the pair-wise correlation between different features and the correlation across different features within each class and meanwhile preserves the class structure in the learned correlation features, we need to maximize the following criterion:

$$\text{corr}(U, V) = \frac{\text{cov}(UL, VL)}{\text{var}(U) \cdot \text{var}(V)} \quad (3.12)$$

where $U = W_x^T X' H$, $V = W_y^T Y' H$, $\text{cov}(UL, VL) = W_x^T C_L W_y$. It can be easily verified that $\text{var}(U) = W_x^T C_{xx} W_x$ and $\text{var}(V) = W_y^T C_{yy} W_y$:

$$\begin{aligned} \text{var}(U) &= UU^T \\ &= (W_x^T X' H)(W_x^T X' H)^T \\ &= W_x^T (X' H)(X' H)^T W_x \\ &= W_x^T [X' (I - n^{-1}ee^T)] [X' (I - n^{-1}ee^T)]^T W_x \\ &= W_x^T (X' - \mu_x)(X' - \mu_x)^T W_x \\ &= W_x^T C_{xx} W_x \end{aligned} \quad (3.13)$$

Maximization of Eq. (3.12) can be performed by Lagrange multiplier on the covariance between UL and VL subject to $\text{var}(U) = \text{var}(V) = 1$. The transformation matrices, W_x and W_y , are obtained through solving the following generalized eigenvalue equation:

$$\begin{cases} C_L C_{yy}^{-1} C_L^T W_x = \lambda^2 C_{xx} W_x \\ C_L^T C_{xx}^{-1} C_L W_y = \lambda^2 C_{yy} W_y \end{cases} \quad (3.14)$$

Since C_{xx} and C_{yy} are symmetric positive definite, we can decompose them as:

$$C_{xx} = R_{xx} R_{xx}^T \quad (3.15)$$

where R_{xx} is a lower triangular matrix. Let $A_x = R_{xx}^T W_x$, the first part in Eq. (3.14) becomes:

$$(R_{xx})^{-1} C_L C_{yy}^{-1} C_L^T (R_{xx}^T)^{-1} A_x = \lambda^2 A_x \quad (3.16)$$

which is a symmetric standard eigenproblem. W_x can be obtained as $(R_{xx}^T)^{-1} A_x$, W_y can be obtained as $C_{yy}^{-1} C_L^T W_x / \lambda$. Since the number of nonzero generalized eigenvalues is at most $c - 1$, the upper limit of W_x and W_y is $\min(c - 1, \text{rank}(X'), \text{rank}(Y'))$. The whole transformation matrices are $W_x^T Q_x^T$ and $W_y^T Q_y^T$, respectively. Eventually, the transformed features are concatenated together into a fused feature vector.

The process of IEDCA is summarized below.

Algorithm 3.1 IEDCA Algorithm

Input: training data matrix, X and Y , label matrix L .

Output: Transformation matrix $W_x^T Q_x^T$ and $W_y^T Q_y^T$.

1. Construct a between-class scatter matrix through the nearest neighbor from both intra-class and extra-class.
 2. Diagonalize the between-class scatter matrix.
 - $Q \leftarrow$ eigenvectors of Eq. (3.9) corresponding to top r eigenvalues.
 3. $X' \leftarrow Q_x^T X$. Similarly, Y is transformed into Y' .
 4. $H \leftarrow I - n^{-1} e e^T$
 5. $C_L \leftarrow (X' H L)(Y' H L)^T$
 6. Compute basis: $W_x, W_y \leftarrow$ eigenvectors of Eq. (3.14) corresponding to maximum $c - 1$ nonzero generalized eigenvalues when seeking the maximization of correlation UL and VL subject to $\text{var}(U) = \text{var}(V) = 1$ as in Eq. (3.12). The entire transformation matrices are $W_x^T Q_x^T$ and $W_y^T Q_y^T$, respectively.
 7. Concatenate transformed features.
-

3.2.3 Multiset Intra-Class and Extra-Class Discriminative Correlation Analysis

In this section, the proposed intra-class and extra-class discriminative correlation analysis is extended to multiple sets of features. Inspired by sequential forward selection method [110], we develop the following procedure for multiset IEDCA:

FIGURE 3.3: Multiset Intra-Class and Extra-Class Discriminative Correlation Analysis for five sample sets.

- 1) First, all sets of features are ranked in the decreasing order based on their classification accuracy. The first set of feature which has the highest classification accuracy is selected among all the sets.
- 2) Second, the next set of features is selected among all the unselected sets features which leads to the highest classification accuracy when fusing with the selected set in the previous step.
- 3) Third, the fused feature is considered as a new set of features. Repeat the above process until enough sets of features are selected, or until the classification accuracy is good enough.

Figure 3.3 demonstrates the Multiset Intra-Class and Extra-Class Discriminative Correlation Analysis for five sample sets.

3.3 Experimental Results

To test the effectiveness of our proposed model, experiments are performed on D-case dataset [111], SDUMLA-HMT database [112] and AR database [113] respectively. Experiments are designed to fuse a variety of features from single modality as well as multiple modalities. Part A demonstrates the experiment of unimodal feature fusion, while Part B shows the experiment of multimodal feature fusion. Part C evaluates the effect of different dimensions kept in IEDCA.

3.3.1 Unimodal Feature Fusion

In this experiment, the effectiveness of IEDCA is demonstrated in the case of fusing various features from single modality using D-case dataset [111]. The D-case dataset is provided by IEEE AASP challenge involving acoustic scene classification and event detection. Here, we use the scene classification dataset to evaluate our proposed model. The scene classification dataset consists of ten different scenes: tube, subway station, supermarket, restaurant, quiet street, park, open-air market, office, busy street and bus. For each scene, ten audio clips are provided and each of them is 30 seconds long. All the audio clips are sampled at 44.1 kHz.

In our experiment, three types of feature extraction methods were employed. For each audio clips, the first type of features, Mel-frequency cepstral coefficients (MFCC) were extracted. The frame window is set to 25 ms, and the step between successive windows is 10 ms. The frequency range is limited to 0-900 Hz. The periodogram is calculated for each frame. The Mel filter bank is applied, followed by taking the DCT of the calculated energies.

Recurrence Quantification Analysis (RQA) is the second type of feature extracted. The idea of RQA is to quantify the patterns that emerge in the recurrence plots by certain metrics. The details can be found in [114]. RQA feature is a combination of the following metrics:

- Recurrence (REC), which represents the percentage of points in the threshold plot. It will be high for self-similar sound when the radius in RQA is fixed.
- Determinism, which denotes the proportion of points in the diagonal lines of RQA. It is used to identify periodic sounds.
- Laminarity (LAM), which represents the proportion of points in forming the vertical lines. It is used to identify stationary segments of sounds. Maximum length (V max) is used to characterize durations of stationary periods.
- The average diagonal length (LEN) and longest diagonal size (Lmax), which are used to characterize periodic repetitions.
- Trapping time, which represents the average vertical line length, while Entropy is the Shannon entropy of diagonal line length.

Texture windows of 40 MFCC frames representing 400ms of audio were used in computing the recurrence plots. In this way, the temporal evolution of the features, which is averaged to obtain a document level representation, is captured. A cosine similarity of 0.03 between MFCC frames is adopted.

Discrete Wavelet Transform (DWT) is the third type of features employed in this experiment. DWT provides frequency and time resolution which closely matches the human ear. A six-level wavelet decomposition using the Daubechies4 (D4) wavelet was used. For each level, both the mean and the variance of the wavelet coefficients over each scene segment were computed and used as features.

We performed five-fold cross-validation for ten runs and averaged the recognition results. Several state-of-the-art feature fusion algorithms are compared with our proposed feature fusion algorithm: serial fusion, parallel fusion, conventional CCA, MDCA, and Multi-view Supervised CCA (MSCCA). In the CCA-based technique, dimension reduction such as PCA and LDA is applied to overcome the small sample size problem. SVM is chosen as the classifier. When there are more than two feature sets, the method of parallel fusion cannot be applied and the conventional CCA is replaced by multiset-CCA.

Table 3.1 shows the classification accuracy of the individual type of features, while Table 3.2 shows the classification accuracy of feature fusion. As demonstrated in the results, serial fusion does not always work. In the case of DWT, the serial-fused feature is less discriminative than the individual feature. The parallel-fused feature is also less discriminative than the individual feature in the cases of MFCC, RQA, and DWT, and it can only be applied to the fusion of two feature sets. Most CCA-based feature fusion results are good since they maximize the correlation between different features. Supervised CCA-based feature fusion techniques perform even better due to the incorporation of class information into correlation analysis. Our proposed IEDCA not only inherits the class information in the correlation analysis but also retains enough dimensions of feature for correlation analysis. Therefore, IEDCA provides a more discriminative feature fusion scheme. Experiment results show that our proposed IEDCA outperforms serial, parallel, CCA-based and supervised CCA- based feature fusion methods.

TABLE 3.1: Classification accuracies (%) by SVM for different features of D-case database.

<i>Feature Type</i>	<i>Classification Accuracy</i>
MFCC (Channel 1)	65.02
MFCC (Channel 2)	64.66
RQA (Channel 1)	55.38
RQA (Channel 2)	54.97
DWT (Channel 1)	64.84
DWT (Channel 2)	56.23

3.3.2 Multimodal Feature Fusion: SDUMLA-HMT Database

In this experiment, SDUMLA-HMT database is used. This database is a comprehensive collection of multiple modalities: finger vein, face, iris, gait and fingerprint from 106 individuals. Experimental settings follow those in [51]. Irises of both eyes and fingerprints from the thumb and index of both hands are used.

Segmentation was performed on iris images using the technique in [51]. The valid part of the iris is illustrated in Figure 3.4(b). The segmented area is normalized and a binary iris template is generated from the segmented iris as shown in Figure 3.4(c). Log-Gabor features are extracted from each binary iris template.

For fingerprints, they were enhanced following the method in [115]. After enhancement, the core point was detected and convolution was performed on the enhanced fingerprints using Gabor filters which have eight orientations. Around the detected core point, circular tessellations consisting of 15 concentric bands were extracted. The entire preprocessing is shown in Figure 3.5.

We performed five-fold cross-validation for ten runs and averaged the recognition results. Our proposed feature fusion algorithm is compared with other feature fusion algorithms as in the previous experiment. Besides, we also compare with Joint Sparse Representation (JSR) and SMDL [51] as mentioned in Chapter 2, are used in combination with a sparse representation classifier. JSR and SMDL are not included in experiment A since they are used for multimodal fusion.

Table 3.3 shows the classification accuracy for each modality, while Table 3.4 provides the classification accuracy for multimodal fusion. Experimental results demonstrate our IEDCA outperforms other fusion methods.

TABLE 3.2: Classification accuracies (%) for feature fusion of MFCC, RQA and DWT in D-case database.

<i>Method</i>	<i>Feature Type</i>	<i>(Channel 1 and 2)</i>					
		<i>MFCC</i>	<i>RQA</i>	<i>DWT</i>	<i>MFCC+RQA</i>	<i>MFCC+DWT</i>	<i>All Six Feature Sets</i>
Serial + PCA + SVM		65.74	55.82	62.61	65.49	65.91	61.94
Serial + LDA + SVM		66.54	56.50	63.36	66.17	66.72	63.51
Parallel + PCA + SVM		63.67	54.22	62.05	-	-	-
Parallel + LDA + SVM		65.01	55.76	62.51	-	-	-
PCA + CGA/MCCA + SVM		67.13	56.68	63.47	67.28	67.03	65.28
LDA + CGA/MCCA + SVM		67.49	57.42	64.79	67.91	67.48	69.36
MSCGA + SVM		67.51	58.18	65.18	68.77	68.53	70.76
MDCGA + SVM		67.62	60.51	64.91	72.95	68.91	74.95
Proposed Method + SVM		67.83	61.28	65.32	74.56	69.63	76.72

FIGURE 3.4: Iris preprocessing. (a) Sample image in SDUMLA-HMT database. (b) Valid region of iris after segmentation. (c) Generated binary iris template.

TABLE 3.3: Classification accuracies (%) by SVM for individual modalities of SDUMLA-HMT database.

<i>Modality</i>	<i>Classification Accuracy</i>
Left Iris	61.05
Right Iris	55.79
Left Thumb Fingerprint	76.84
Left Index Fingerprint	88.53
Right Thumb Fingerprint	81.05
Right Index Fingerprint	89.58

FIGURE 3.5: Fingerprint preprocessing. (a) Sample fingerprint in SDUMLA-HMT database. (b) Enhanced fingerprint. (c) Detected core point and region of interest.

TABLE 3.4: Classification accuracies (%) of multimodal fusion in SDUMLA-HMT database.

<i>Method</i>	<i>Modality</i>	<i>Two Irises</i>	<i>Four Fingerprints</i>	<i>All Six Modalities</i>
Serial + PCA + SVM		66.83	89.94	90.53
Serial + LDA + SVM		73.68	90.21	91.58
Parallel + PCA + SVM		72.42	-	-
Parallel + LDA + SVM		76.21	-	-
PCA + CCA/MCCA + SVM		80.32	92.25	92.63
LDA + CCA/MCCA + SVM		81.05	92.53	93.68
MSCCA + SVM		84.47	94.16	95.79
MDCA + SVM		85.42	94.42	97.89
JSR		80.29	93.69	94.74
SMDL		84.18	93.71	96.84
Proposed Method + SVM		86.37	95.58	98.95

3.3.3 Effect of Dimensions Retained for Correlation

In this section, the effect of different dimensions retained for correlation analysis is evaluated on AR database [113]. This database contains images from 100 subjects under various illuminations and facial expressions. The subset of AR database without occlusion is chosen. In the experiment, seven images from each subject are randomly chosen for training, while the remaining are used for testing. Ten independent recognition tests are conducted.

Gabor features and Histogram of Oriented Gradients (HOG) are extracted. Convolution is performed on each image with a set of 40 Gabor filters with eight orientations and five scales, followed by a down-sampling factor of five. HOG features are extracted with nine orientations in 5×5 cells. UOCTTI variant [115] is adopted in calculating directed and undirected gradients of HOG. The result is projected down to 31 dimensions, which consist of 27-dimension orientation channels and 4-dimension gradient energy. Since we only explore the number of dimensions kept in IEDCA, a simple nearest neighbor classifier would be enough. Table 3.5 shows the average classification accuracies for various numbers of dimensions kept in the discriminative analysis, obtained by the individual feature and the fused feature.

As mentioned in Section 3.1.1, to solve small sample problem, the dimension of feature usually reduces to at most $c - 1$ (c is the number of classes), followed by correlation analysis. In this experiment, $c = 100$, the feature dimension is reduced to 99 and the average classification accuracies achieved by Gabor features and HOG features are 88.57% and 77.86%, respectively. The average classification accuracy

TABLE 3.5: Average classification accuracies (%) for the different number of dimensions kept in discriminative analysis in AR database.

<i>Number of dimensions kept in discriminative analysis</i>	<i>Gabor</i>	<i>HOG</i>	<i>Gabor + HOG</i>
190	84.14	77.86	98.29
180	84.49	77.86	98.29
160	85.57	77.86	97.71
150	86.00	77.86	97.43
99	88.57	77.86	96.29

achieved by IEDCA fusion of Gabor and HOG features is 96.29%. In this case, the dimension of feature is reduced to c-1, and certain information that is important for the correlation analysis between different features is lost. This will, in turn, affect the eventual performance.

In IEDCA, dimension reduction of features for correlation analysis is not restricted to c-1. In Table 3.5, with the increase of dimensions kept in discrimination analysis (from 99 to 190), the average classification accuracy achieved by Gabor features drops from 88.57% to 84.14%. This is because part of the information is redundant when we increase dimensions kept in discrimination analysis. Meanwhile, the average classification accuracy achieved by HOG stays at 77.86%. The increase of dimensions kept for HOG does not affect the recognition. Average classification accuracy achieved by IEDCA fusion of Gabor features and HOG features improves from 96.29% to 98.29% with the increase of dimensions kept in discrimination analysis. This is because part of redundant information underlying individual features contains correlation information which is important for feature fusion.

3.4 Conclusions

In this chapter, we propose a feature fusion scheme named intra-class and extra-class discriminative correlation analysis (IEDCA). The proposed IEDCA inherits the discriminating structure in the learned correlation features by incorporating the class information and fully explores the intra-class correlation including both one-to-one pair-wise correlation and the correlation between different features within the same class. Moreover, more information is kept in the discriminative analysis to enhance the performance of learned correlation features. Various experiments ranging from the fusion of different features of single modality to multimodal feature fusion demonstrate the effectiveness of our proposed approach.

Chapter 4

Enhanced Feature Representation Learning through Irrelevant Redundancy Elimination in Intra-Class and Extra-Class Discriminative Correlation Analysis

Most CCA fused features may still contain redundancies due to the correlation criterion. The correlation-based multi-view representation learning framework we proposed in Chapter 3 also has this redundancy problem. These redundancies may be relevant or irrelevant to class labels. The irrelevant redundancies may degrade the pattern recognition performance, while the relevant redundancies can make the pattern recognition system more robust. In this chapter, we propose an enhanced feature fusion scheme through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis (IEDCA-IRE) addressing this issue. IEDCA-IRE explores the intra-class correlation including both the pairs-wise correlation like CCA-based feature fusion approaches and the correlation across different features within the same class. By incorporating kernelized IEDCA into minimum Redundancy Maximum Relevance (mRMR) criterion, only the relevant redundancy is retained in the fused feature. Our proposed IEDCA-IRE can be

used in unimodal feature fusion, multimodal feature fusion, the fusion of deep features extracted from different deep neural network models as well as the fusion of deep features and handcrafted features. Extensive experiments have proved its effectiveness.

4.1 Introduction

In this section, we discuss the motivation of our proposed methodology and the related minimum Redundancy Maximum Relevance (mRMR) method inspiring our work.

4.1.1 Motivation and Overview

Supervised CCA-based representation fusion approaches can produce a compact representation ensuring that the information captured from different feature sets are maximally correlated and the effective discriminative information of multiple feature sets is retained. Our previous representation fusion framework falls into the category of supervised CCA-based representation fusion approach dealing with the data from multiple views.

Let's have a brief review of our previous representation fusion framework Intra-Class and Extra-Class Discriminative Correlation Analysis (IEDCA) [31]. For the details of IEDCA, please refer to Chapter 3. The basic idea of Intra-Class and Extra-Class Discriminative Correlation Analysis (IEDCA) is to find transforms where classes are separated within each feature set while the discriminating structure is inherited in the correlation analysis. It consists of two stages, a discriminative analysis followed by correlation analysis. The advantage of IEDCA is illustrated in Fig.4.1 using t-distributed stochastic neighbor embedding(t-SNE) [116] on SDUMLA-HMT database [112]. Fig.4.1 shows the fused features through IEDCA enhances the separation among classes in t-SNE projection (Fig.4.1 (b)).

However, the redundant information therein is only partially eliminated. This is because, in CCA, the mutual relationships of the input feature sets are maximized through correlation criterion. Such mutual information of each canonical

(a)

(b)

FIGURE 4.1: (a) t-SNE on original features of SDUMLA-HMT database (b) t-SNE on the fused features through intra-class and extra-class discriminative correlation analysis of SDUMLA-HMT database. Each colour represents a different class.

projected pairs contains certain redundancy. On the other hand, redundancy provides robustness in the pattern recognition system [117]. There exists a question that how we can enhance the CCA-fused representation and only keep the essential redundancy. From machine learning perspective, it is important to find the most characterizing features to minimize the classification error. Inspired by the idea that such minimization usually requires the maximum of relevance and the minimum of redundancy simultaneously, we classify the redundancy existed in the representation into relevant and irrelevant redundancy. To keep the relevant redundancy and remove the irrelevant redundancy during feature fusion, we propose

a feature fusion scheme called Enhanced Feature Fusion through Irrelevant Redundancy Elimination in Intra-Class and Extra-Class Discriminative Correlation Analysis (IEDCA-IRE). The irrelevant redundancy generated by the maximization criterion of the correlation analysis is eliminated by incorporating kernelized intra-class correlation analysis into minimum Redundancy Maximum Relevance (mRMR) criterion. In each feature set, class separability is enhanced by the discriminative analysis through the intra-class and extra-class nearest neighbor of each sample in the feature set. IEDCA-IRE also explores intra-class correlation, including both the pair-wise correlation like CCA-based feature fusion approaches and the correlation across different features within the same class. The discriminating structure is inherited by incorporating class matrix. The proposed method overcomes the problem of small sample size encountered in CCA and is extended to deal with multiset features by sequential forward selection in IEDCA-IRE. This enhanced approach contains the following aspects:

- To obtain an enhanced fused representation, we analyze the fused representation generated by IEDCA. By optimizing and combining different sets of features through IEDCA, the effective discriminative information from multiple feature sets is kept. However, the fused features may still contain redundancies due to the correlation criterion. These redundancies may be relevant or irrelevant to class labels. We classify the redundancies into relevant redundancies and irrelevant redundancies. The irrelevant redundancies may degrade the pattern recognition performance, while the relevant redundancies can make the pattern recognition system more robust. Since IEDCA is essentially a CCA-based method, other CCA-based feature fusion schemes have this problem in common. The details of the analysis are provided in Section 4.2.1.
- To keep only the relevant redundancies and remove the irrelevant redundancies in the feature fusion. We incorporate kernelized intra-class correlation analysis into mRMR. We use the transformed features by the proposed kernelized intra-class correlation analysis between the feature and the target class instead of using the features directly to calculate the mutual information in mRMR. In this way, the redundancy estimated in mRMR is only the relevant redundancy.

- To capture the nonlinear relationship which often exists between the feature and the class label, we kernelized IEDCA. This is used for irrelevant redundancy elimination.

4.1.2 Minimum Redundancy Maximum Relevance

Minimum Redundancy Maximum Relevance (mRMR) [118] is a classical method to find a subset of features in which the relevance between the features and the target class is maximized and the redundancy among the features is minimized based on mutual information. Assume the input features are $\{x_j \in S_m, j = 1, \dots, m\}$. In mRMR, the mutual information (MI) between the candidate feature and the target class label is firstly calculated as relevance:

$$\max D(S_m, c), \quad D = \frac{1}{m} \sum_{x_j \in S_m} I(x_j; c) \quad (4.1)$$

where $I(x_j; c)$ denotes the MI between x_j and c . Then the average MI between the candidate feature and the selected features is computed as redundancy term:

$$\min R(S_m), \quad R = \frac{1}{m^2} \sum_{x_j, x_i \in S_m} I(x_i; x_j) \quad (4.2)$$

The combination of relevance and redundancy is either in difference form as:

$$\max \phi(D; R), \quad \phi = D - R \quad (4.3)$$

or in quotient form as:

$$\max \phi(D; R), \quad \phi = D/R \quad (4.4)$$

Suppose we have already selected $m-1$ features in S_{m-1} . mRMR selects the coming features as:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, c) - \frac{1}{(m-1)} \sum_{x_i \in S_{m-1}} I(x_j, x_i) \right] \quad (4.5)$$

or in quotient form as:

$$\max_{x_j \in X - S_{m-1}} \left[\frac{(m-1)I(x_j, c)}{\sum_{x_i \in S_{m-1}} I(x_j, x_i)} \right] \quad (4.6)$$

where S_{m-1} is the existing selected feature set with $m-1$ features, c is the target class, I calculates the mutual information.

4.2 Enhanced Feature Fusion Through Irrelevant Redundancy Elimination in Intra-Class and Extra-Class Discriminative Correlation Analysis

In this section, we firstly analyze CCA-based feature fusion and then kernelized intra-class correlation analysis. We propose the enhanced feature fusion scheme through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis (IEDCA-IRE) by incorporating kernelized intra-class correlation analysis into mRMR.

4.2.1 Analysis of CCA-based Representation Learning

CCA-based feature fusion approach provides a compact representation, where the information captured from different feature sets are maximally correlated and the effective discriminative information of multiple feature sets is retained. However, the redundant information is only partially eliminated. In the following, we show that in the fused features generated by IEDCA, different pairs of the transformed features by IEDAC are uncorrelated and thus nonredundant. The details are provided in Appendix A.

$$\text{cov}(x_i^*, x_j^*) = a_i^T C_{xx} a_j = 0 \quad (4.7)$$

$$\text{cov}(y_i^*, y_j^*) = b_i^T C_{yy} b_j = 0 \quad (4.8)$$

$$\text{cov}(x_i^*, y_j^*) = a_i^T C_L b_j = 0 \quad (4.9)$$

$$\text{cov}(x_i^*, y_i^*) = a_i^T C_L b_i = \lambda_i \quad (4.10)$$

$$\text{cov}(x_i^* + y_i^*, x_j^* + y_j^*) = \text{cov}(x_i^*, x_j^*) + 2\text{cov}(x_i^*, y_j^*) + \text{cov}(y_i^*, y_j^*) = 0 \quad (4.11)$$

Since the correlation criterion is to maximize the correlation between each pair through the calculation of λ_i in (4.10), the mutual information in each pair may contain both the relevant and irrelevant redundancy. The irrelevant redundancy may degrade the performance of the pattern recognition task. Minimum Redundancy Maximum Relevance (mRMR) [118] can eliminate the redundancy through mutual information. In mRMR, the mutual information (MI) between the candidate feature and the target class label is firstly calculated as the relevance measure. Then the average MI between the candidate feature and the selected features is computed as a redundancy measure. The maximization of relevance and minimization of redundancy is either in difference form as in (4.5) or in quotient form as in (4.6).

However, mRMR drives an unbalanced selection process between the relevance term $I(x_j, c)$ and redundancy term $I(x_j, x_i)$. $I(x_j, c)$ measures the relevance of the feature to be added, while the mutual information $I(x_j, x_i)$ only estimates the redundancy between the j -th feature and the existing selected feature. The redundancy is not associated with relevance. The feature with lower relevance may have a higher chance to be selected than the feature with both higher relevance and higher irrelevant redundancy. The problem is illustrated using Fig.4.2. The mutual information is represented by the overlap area of circles and summarized in Table 4.1. Assume $I(x_1, c) > I(x_2, c) > I(x_3, c)$, x_1 is selected first. Suppose $I(x_2, c) = I(x_2, x_1)$, the calculation in mRMR for x_3 can be represented by the overlap area of x_3 and c (CX_3) since $I(x_3, x_1) = 0$. The mRMR calculation for

FIGURE 4.2: An illustration of the relationship for variables X_1 , X_2 , X_3 and C , C is the target class.

x_2 is 0 since $I(x_2, c) = I(x_2, x_1)$. Following the optimizing condition of either (4.5) or (4.6), x_3 is selected second. However, the nonredundant relevance of x_2 is larger than x_3 ($CX_2 > CX_3$). Here, we define the nonredundant relevance as the unique information carried by the variables with the target class. x_2 is supposed to be selected before x_3 . This is because in mRMR, the mutual information $I(x_j, x_i)$ estimates both the relevant redundancy and the irrelevant redundancy. The calculation in mRMR is highly affected by irrelevant redundancy. This makes the nonredundant relevance been overlooked and the selection of features with lower relevance ahead of the feature which has both higher relevance and higher irrelevant redundancy. To overcome this problem, we should only consider the relevant redundancy and balance the relevance term and redundancy term. To associate the redundancy with relevance, we use the transformed features by the correlation analysis between the feature and the target class instead of using the features directly to calculate the mutual information in mRMR. In this way, the redundancy estimated in mRMR is only the relevant redundancy. The relationship between the feature and the target class could be nonlinear, and the kernel method is an effective way to solve the nonlinear problem [119]. Therefore, we propose to incorporate kernelized intra-class correlation analysis into mRMR to eliminate the irrelevance redundancy generated by the maximization criterion of the correlation analysis. In the following sections, we kernelize IEDCA first and then incorporate the kernelized IEDCA into mRMR.

TABLE 4.1: The mutual information and corresponding area.

<i>Term</i>	<i>Corresponding area</i>
$I(x_1, c)$	$CX_1 + CX_1X_2$
$I(x_2, c)$	$CX_2 + CX_1X_2$
$I(x_3, c)$	CX_3
$I(x_2, x_1)$	$CX_1X_2 + X_1X_2$

4.2.2 Kernelized Intra-Class Correlation Analysis

In this subsection, we kernelized our intra-class correlation analysis by first projecting X' and Y' into a higher-dimensional feature space:

$$\phi : x' \rightarrow \phi_{x'}(x'), y' \rightarrow \phi_{y'}(y') \quad (4.12)$$

Intra-class correlation analysis is performed in the above higher dimensional space. In this higher-dimensional space, it is difficult to conduct normal calculation. To avoid the calculation of complex nonlinear function, the kernel method constructs an inner product kernel function by:

$$k(x', z') = \langle \phi(x'), \phi(z') \rangle \quad (4.13)$$

where $x', z' \in X'$. The maximization of the criterion (3.12) remains the same, where U and V become:

$$U = W_{\phi(x')}^T \phi_{x'}(X') H \quad (4.14)$$

$$V = W_{\phi(y')}^T \phi_{y'}(Y') H \quad (4.15)$$

The directions of $W_{\phi(x')}$ and $W_{\phi(y')}$ can be expressed as the projection of X' and Y' on W_x and W_y :

$$W_{\phi(x')} = \phi_{x'}(X') W_x \quad (4.16)$$

$$W_{\phi(y')} = \phi_{y'}(Y') W_y \quad (4.17)$$

Substitute (4.16) and (4.17) in (3.12), let $K_x = (\phi_{x'}(X'))^T \phi_{x'}(X')$ and $K_y = (\phi_{y'}(Y'))^T \phi_{y'}(Y')$, $cov(UL, VL)$, $var(U)$ and $var(V)$ become:

$$cov(UL, VL) = W_x^T K_x (HL) (HL)^T K_y W_y \quad (4.18)$$

$$var(U) = W_x^T K_x^2 W_x \quad (4.19)$$

$$var(V) = W_y^T K_y^2 W_y \quad (4.20)$$

Similarly, maximization of (3.12) can be performed by Lagrange multiplier on the covariance between UL and VL subject to $var(U) = var(V) = 1$. The transformation matrices, W_x and W_y can be obtained through solving a generalized eigenvalue problem with the corresponding first s largest eigenvalue:

$$W_x = [a_1, a_2, \dots, a_s] \quad (4.21)$$

$$W_y = [b_1, b_2, \dots, b_s] \quad (4.22)$$

$$W_{\phi(x')}^T \phi_{x'}(X') = [a_1, a_2, \dots, a_s]^T K_x^T \quad (4.23)$$

$$W_{\phi(y')}^T \phi_{y'}(Y') = [b_1, b_2, \dots, b_s]^T K_y^T \quad (4.24)$$

To avoid the computational problems and degeneracy caused by large training sets, incomplete Cholesky decomposition (ICD) is adopted for the dimension reduction of the kernel matrices.

4.2.3 Incorporation of Intra-Class Correlation Analysis into mRMR

By combining different sets of features through IEDCA, the effective discriminative information of multiple sets of features is kept. However, the redundant information is only partially eliminated as we analyzed. The mutual information of each canonical projected pairs contains irrelevant redundancy. It is helpful to correlate

redundancy components obtained by kernelized IEDCA with predictor variables C to extract the relevant redundancy in the features. To calculate only the relevant redundancy and remove the irrelevant redundancy, we use the transformed features by the proposed intra-class correlation analysis between the feature and the target class instead of using the features directly to calculate the mutual information in mRMR.

To capture the nonlinear relationship that often exists between the feature and target class, kernelized intra-class correlation analysis is used to maximize the correlation between the features transformed by IEDCA and the corresponding target class. The feature transformed by the kernelized intra-class correlation analysis is used to calculate the mutual information in mRMR. In this way, the redundancy estimated in mRMR is only the relevant redundancy:

$$\max_{x_j \in X - S_{m-1}} \left[\frac{(m-1) \sum_{d=1}^s \lambda_d^2 I(a_d K_j^T, c)}{\sum_{x_i \in S_{m-1}} \sum_{d,f=1}^s \lambda_d \lambda_f I(a_d K_j^T, b_f K_i^T)} \right] \quad (4.25)$$

$a_d K_j^T$ indicate the d -th dimension transformed feature by kernelized intra-class correlation analysis between feature j and the corresponding target class. λ_d is the d -th correlation coefficient between feature j and corresponding target class. We use the mutual information quotient scheme since the mutual information quotient scheme outperforms mutual information difference scheme, especially for discrete data [120].

The numerator of (4.25) calculates the relevance between feature j and corresponding target class. By kernelized intra-class correlation analysis, each feature is transformed into a higher dimension. The relevance term is summed over all the dimensions transformed by the kernelized intra-class correlation analysis. By multiplying the square of the correlation coefficient, we provide different weights to the different dimensions of the transformed features according to their relation with target class and combine them with consideration of the variance from covariates. By summing up all the dimension pairs of the transformed features by kernelized intra-class correlation analysis between each feature and corresponding target class, only the relevant redundancy is considered in the denominator of (4.25).

4.2.4 IEDCA-IRE Algorithm

The process of enhanced feature fusion through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis (IEDCA-IRE) consists of two stages. First, multisets of features are transformed by IEDAC and concatenated into a fused feature vector. Second, by incorporating intra-class correlation analysis into mRMR, the irrelevant redundancy of the fused feature vector is eliminated. The whole algorithm is summarized below.

Algorithm 4.1 IEDCA-IRE Algorithm

Input: training data matrix, X and Y , label matrix L .

Output: irrelevant redundancy eliminated fused feature vector.

1. Construct a between-class scatter matrix through the nearest neighbor from both intra-class and extra-class.
 2. Diagonalize the between-class scatter matrix $Q \leftarrow$ eigenvectors of (3.9) corresponding to top r eigenvalues.
 3. $X' \leftarrow Q_x^T X$. Similarly, Y is transformed into Y' .
 4. $H \leftarrow I - n^{-1} \mathbf{e} \mathbf{e}^T$
 5. $C_L \leftarrow (X' H L)(Y' H L)^T$
 6. Compute basis: $W_x, W_y \leftarrow$ eigenvectors corresponding to nonzero generalized eigenvalues when seeking the maximization of correlation UL and VL subject to $\text{var}(U) = \text{var}(V) = 1$. The entire transformation matrices are $W_x^T Q_x^T$ and $W_y^T Q_y^T$, respectively.
 7. Concatenate transformed features into a fused feature vector.
 8. By incorporating kernelized intra-class correlation analysis into mRMR, the irrelevant redundancy in the fused feature vector is eliminated through (4.25)
-

4.2.5 Computational Complexity Analysis

Let us assume $p > q$. For the first stage (step1-3), the main computational burden lies in diagonalizing the between-class scatter matrix. It is known that the computational complexity is $O(p^3)$ which is similar to Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). For the second stage (step 4-6), the computational complexity to calculate cross-set covariance matrix C_{xy} is $O(pqn)$, whereas for within-set covariance matrices C_{xx} and C_{yy} is $O(p^2n + q^2n)$. In step 6, the eigenvalues with corresponding eigenvectors can be calculated with

complexity $O(p^3 + q^3)$. Hence, the total computational complexity of the second stage is $O(pqn + p^2n + q^2n + p^3 + q^3) = O(p^3)$. For the third stage (step 7-8), kernelized intra-class correlation analysis has a computational complexity of $O(\Omega^3)$, where Ω is the number of features for the fused feature vector in step 7. Let Φ denote the set of irrelevant redundancy eliminated features we are seeking. In step 8, the computational complexity is $O(|\Phi|\Omega)$. Hence, the overall computational complexity of the proposed algorithm is $2O(p^3) + O(|\Phi|\Omega)$. CCA has a computational complexity of $O(p^3)$. The difference between CCA and MSCCA is that in MSCCA within-class scatter matrices replace the covariance matrices. Therefore, the computational complexities of CCA and MSCCA are roughly the same.

4.3 Experiments

To test the effectiveness of our proposed IEDCA-IRE, experiments are performed on D-case dataset [111], SDUMLA-HMT database [112], AR database [113], PASCAL VOC 2007 [121] and Stack Overflow [122] respectively. Experiments are designed to fuse a variety of features from single modality as well as multiple modalities. Moreover, we also evaluate our proposed model in the fusion of deep features extracted from different deep neural network models as well as the fusion of deep features and handcrafted features. Section 4.3.1 demonstrates the experiment of unimodal feature fusion, while Section 4.3.2 shows the experiment of multimodal feature fusion. Section 4.3.3 evaluates the effect of different dimensions retained for the correlation analysis in IEDCA-IRE. Section 4.3.4 demonstrates the experiment of deep feature fusion from different deep neural networks. Section 4.3.5 shows the experiment of fusion of deep features and handcrafted features.

4.3.1 Unimodal Feature Fusion

In this experiment, the effectiveness of IEDCA-IRE is demonstrated in the case of fusing various features from single modality using D-case dataset [111]. D-case dataset is provided by IEEE AASP challenge involving acoustic scene classification and event detection. Here, we use the scene classification dataset to evaluate our proposed model. The scene classification dataset consists of ten different scenes: tube, subway station, supermarket, restaurant, quiet street, park, open-air market,

office, busy street and bus. For each scene, ten audio clips are provided and each of them is 30 seconds long. All the audio clips are sampled at 44.1 kHz. Three types of feature extraction methods were employed. For each audio clips, the first type of features, Mel-frequency cepstral coefficients (MFCC) were extracted with a dimension of 13. The frame window is set to 25 ms, and the step between successive windows is 10 ms. The frequency range is limited to 0-900 Hz. The periodogram is calculated for each frame. Mel filter bank is applied, followed by taking DCT of the calculated energies. Recurrence Quantification Analysis (RQA) is the second type of feature extracted with a dimension of 11. The basic idea of RQA is to quantify the patterns emerging in the recurrence plots by certain metrics. The details can be found in [114]. RQA feature is a combination of the following metrics:

- Recurrence (REC), which represents the percentage of points in the threshold plot. It will be high for self-similar sound when the radius in RQA is fixed.
- Determinism, which denotes the proportion of points in the diagonal lines of RQA. It is used to identify periodic sounds.
- Laminarity (LAM), which represents the proportion of points in forming the vertical lines. It is used to identify stationary segments of sounds. Maximum length (V max) is used to characterize durations of stationary periods.
- The average diagonal length (LEN) and longest diagonal size (Lmax), which are used to characterize periodic repetitions.
- Trapping time, which represents the average vertical line length, while Entropy is the Shannon entropy of diagonal line length.

Texture windows of 40 MFCC frames representing 400ms of audio were used in computing the recurrence plots. In this way, the temporal evolution of the features, which is averaged to obtain a document level representation, is captured. Cosine similarity of 0.03 between MFCC frames is adopted. Discrete Wavelet Transform (DWT) is the third type of features with a dimension of 255 employed in this experiment. DWT provides frequency and time resolution which closely matches the human ear. A six-level wavelet decomposition using the Daubechies4 (D4) wavelet was used. For each level, both the mean and the variance of the wavelet coefficients over each scene segment were computed and used as features.

We performed five-fold cross-validation for ten runs and averaged the recognition results. Several state-of-the-art feature fusion algorithms are compared with our proposed feature fusion algorithm: serial fusion, conventional CCA, Multiset Discriminant Correlation Analysis (MDCA), and Multi-view Supervised CCA (MSCCA). In CCA-based technique, dimension reduction, PCA, is applied to overcome the small sample size problem. SVM is chosen as the standard classifier to facilitate fair comparisons. We combine different features for the fusion of Channel 1 and Channel 2. In the case of combining all the three types of features for the fusion of Channel 1 and Channel 2 (totally six feature sets), Channel 1 and Channel 2 each have a combined feature set with dimension of 279 (13+11+255). During the intra-class and extra-class discriminative analysis, the dimension of the combined feature set for Channel 1 and Channel 2 is reduced to 200 by W_x and W_y , respectively. In the subsequent intra-class correlation analysis, Q_x and Q_y are both 200×150 . It results in a fused feature set with a dimension of 300. By incorporating kernelized intra-class correlation analysis into mRMR, we further eliminate the irrelevant redundancy resulting in a dimension of 225. For CCA, PCA is firstly applied to reduce the dimension to 200. W_x and W_y are both 200×160 . This results in a fused feature set with dimension of 320. mRMR is then used to select 225 features for a fair comparison. MSCCA and MDCA result in a fused feature set with dimension of 318 and 310, respectively. Similarly, mRMR is then used to select 225 features. The nearest neighbor parameter k is set to 5. Table 4.2 shows the classification accuracy of the individual type of features, and Table 4.3 shows the classification accuracy of feature fusion. As demonstrated in the results, serial fusion does not always work. Most CCA-based feature fusion results are good since they maximize the correlation between different features. Supervised CCA-based feature fusion techniques perform even better due to the incorporation of class information into correlation analysis. IEDCA-IRE not only inherits the class information in the correlation analysis but also retains enough dimensions of feature for correlation analysis. Therefore, IEDCA-IRE provides a more discriminative feature fusion scheme. The fused feature generated by IEDCA may contain irrelevant redundancy through optimizing the correlation criterion in IEDCA. This may deteriorate the classification performance. By considering only relevant redundancy and eliminating the irrelevant redundancy, experiment results show that our proposed IEDCA-IRE outperforms serial, CCA-based and supervised CCA-based feature fusion methods.

TABLE 4.2: Classification accuracies (%) by SVM for different features of D-case database.

<i>Feature Type</i>	<i>Classification Accuracy</i>
MFCC (Channel 1)	65.02
MFCC (Channel 2)	64.66
RQA (Channel 1)	55.38
RQA (Channel 2)	54.97
DWT (Channel 1)	64.84
DWT (Channel 2)	56.23

4.3.2 Multimodal Feature Fusion: SDUMLA-HMT Database

In this experiment, SDUMLA-HMT database is used. This database is a comprehensive collection of multiple modalities: finger vein, face, iris, gait and fingerprint from 106 individuals. Experimental settings follow [51]. Irises of both eyes and fingerprints from the thumb and index of both hands are used. Segmentation was performed on iris images. The valid part of the iris is illustrated in Fig.4.3(b). The segmented area is normalized, and a binary iris template is generated from the segmented iris as shown in Fig.4.3(c). Log-Gabor features with a dimension of 6000 are extracted from each binary iris template. For fingerprints, they were enhanced following the method in [115]. After enhancement, the core point was detected, and convolution was performed on the enhanced fingerprints using Gabor filters which have eight orientations. Around the detected core point, circular tessellations consisting of 15 concentric bands were extracted. For each fingerprint, the feature dimension is 640. The entire preprocessing is shown in Fig.4.4.

We performed five-fold cross-validation for ten runs and averaged the recognition results. Our proposed feature fusion algorithm is compared with other feature fusion algorithms as in the previous experiment. Besides, we also compare with Joint Sparse Representation (JSR) and SMDL [51], which are used in combination with a sparse representation classifier. In the case of combining all the six modalities, we first perform IEDCA on iris and thumb fingerprint. This generates two 1280-dimensional fused feature sets from the combination of left iris and left thumb fingerprint and the combination of right iris and right thumb fingerprint.

TABLE 4.3: Classification accuracies (%) for feature fusion of MFCC, RQA and DWT in D-case database.

<i>Method</i>	<i>Feature Type</i>	<i>(Channel 1 and 2)</i>						<i>All Six Feature Sets</i>
		<i>MFCC</i>	<i>RQA</i>	<i>DWT</i>	<i>MFCC+RQA</i>	<i>MFCC+DWT</i>	<i>RQA+DWT</i>	
Serial + mRMR CCA/MCCA + mRMR MSCCA + mRMR MDCA + mRMR IEDCA-IRE	Serial + mRMR	67.92	57.85	65.07	69.24	68.29	65.98	70.51
	CCA/MCCA + mRMR	67.54	57.68	65.34	68.38	68.95	65.53	71.06
	MSCCA + mRMR	67.63	58.30	66.86	70.65	70.86	66.94	72.88
	MDCA + mRMR	67.86	60.84	65.98	74.16	70.64	67.21	75.46
	IEDCA-IRE	68.01	61.75	68.14	76.32	71.59	68.57	79.86

TABLE 4.4: Classification accuracies (%) by SVM for individual modalities of SDUMLA-HMT database.

<i>Modality</i>	<i>Classification Accuracy</i>
Left Iris	61.05
Right Iris	55.79
Left Thumb Fingerprint	76.84
Left Index Fingerprint	88.53
Right Thumb Fingerprint	81.05
Right Index Fingerprint	89.58

TABLE 4.5: Classification accuracies (%) of multimodal fusion in SDUMLA-HMT database.

<i>Method \ Modality</i>	<i>Two Irises</i>	<i>Four Fingerprints</i>	<i>All Six Modalities</i>
Serial + mRMR	82.87	91.96	93.64
CCA/MCCA + mRMR	83.92	93.34	95.26
MSCCA + mRMR	86.18	95.22	97.38
MDCA + mRMR	86.35	95.06	98.17
JSR	80.29	93.69	94.74
SMDL	84.18	93.71	96.84
IEDCA-IRE	87.26	96.74	99.51

The subsequent IEDCA is performed on the fused feature set and its corresponding index fingerprint. For example, we perform IEDCA on the 1280-dimensional fused feature set from the combination of left iris and left thumb fingerprint with left index fingerprint. During this process, W_x and W_y are 1280×640 and 640×400 , respectively. Q_x and Q_y are 640×320 and 400×320 , respectively. This results in two 640-dimensional fused feature sets. The eventual IEDCA is performed, and W_x and W_y are both 640×300 . Q_x and Q_y are both 300×170 . The final fused feature set is 340-dimensional. By incorporating kernelized intra-class correlation analysis into mRMR, we further eliminate the irrelevant redundancy resulting in a dimension of 295. The nearest neighbor parameter k is set to 5. Table 4.4 shows the classification accuracy for each modality, while Table 4.5 provides the classification accuracy for multimodal fusion. Experimental results demonstrate our IEDCA-IRE outperforms other fusion methods.

4.3.3 Deep Feature Fusion

In this experiment, we apply our proposed IEDCA-IRE in the fusion of deep heterogeneous features extracted from different state-of-the-art deep neural networks

FIGURE 4.3: Iris preprocessing. (a) Sample image in SDUMLA-HMT database. (b) Valid region of iris after segmentation. (c) Generated binary iris template.

FIGURE 4.4: Fingerprint preprocessing. (a) Sample fingerprint in SDUMLA-HMT database. (b) Enhanced fingerprint. (c) Detected core point and region of interest.

FIGURE 4.5: The pipeline of the application of redundancy without relevance eliminated IEDCA in deep learning. The raw images are preprocessed and forwarded into each deep neural network model: VGG, ResNet and AlexNet. Feature extraction is done on the last layer of the three models. These vectors are fed into the redundancy without relevance eliminated IEDCA, concatenated, and forwarded into softmax for classification.

to exploit the complementary information present in these features for image classification. The application of our proposed IEDCA-IRE in the fusion of three sets of deep features is illustrated in Figure 4.5.

PASCAL VOC-2007 [121] dataset is used. PASCAL VOC-2007 contains about 10,000 images of twenty object classes which have been split into train, validation, and test sets. The three selected pre-trained models are VGG-16 [10], ResNet-50 [11] and AlexNet [12]. They are well-known deep neural networks that perform favorably to the state-of-the-art for classification in the ImageNet Large-Scale Visual Recognition Challenge 2012 (ILSVRC). The activations of the last layer of the network are used as our features, resulting in 4096-dimensional feature vectors.

For image preprocessing, we uniformed them by cropping them into 224×224 since raw images are various in sizes. We first perform IEDCA on two deep feature sets. The subsequent IEDCA is performed on the generated fused feature set and the remaining deep feature set. As an illustration, deep features from AlexNet and

VGG-16 are firstly fused into a 2100-dimensional feature set. Notice that W_x and W_y are 4096×2250 and 4096×2000 , respectively. Q_x and Q_y are 2250×1050 and 2000×1050 , respectively. The fused 2100-dimensional feature set is then integrated with deep features from ResNet-50, generating a 3000-dimensional fused feature set. In this step, W_x and W_y are 2100×2100 and 4096×2400 , respectively. Q_x and Q_y are 2100×1320 and 2400×1320 , respectively. By incorporating kernelized intra-class correlation analysis into mRMR, we further eliminate the irrelevant redundancy of the 2640-dimensional fused feature set. This results in a dimension of 2010. The nearest neighbor parameter k is set to 7. The performance is evaluated by mean Average Precision (mAP) across all classes. The fused features are fed into SoftMax for classification. Our proposed feature fusion algorithm is compared with other feature fusion algorithms as in the previous experiments. The results of the experiment are shown in Table 4.6. Our proposed IEDCA-IRE achieved the best performance in terms of mean average precision (mAP). Table 4.7 lists the results of selected deep neural networks. We can see ResNet-50 reached a classification accuracy of 77.93%. AlexNet and VGG achieved 71.77% and 75.70% respectively. The results were attained on the last layer by SoftMax. Our experiments demonstrate our fusion of three deep neural networks achieved better performance than individual models, with the best classification accuracy of 83.39%. Table 4.6 displays the mean Average Precision of our various fusion models. As demonstrated in the results, serial fusion does not always work. In the case of A+V, A+R and V+R, the serial-fusion model is less discriminative than the individual deep neural network model. Supervised CCA-based feature fusion models perform better due to the incorporation of class information into correlation analysis. The results demonstrate our fusion method outperforms individual deep neural network models. Even the lowest-performing fusion model with two networks (A+V with 79.09%) performs better than the best performance of the individual model (ResNet’s 77.93%). These results indicate our proposed fusion algorithm can improve the classification performance.

As we can see in Table 4.6, the highest mAP increases from 79.09% (for A+V fusion), 79.78% (for A+R fusion), 81.17% (for V+R fusion), and finally 83.39% (for A+V+R fusion); all by SoftMax. Note the highest result of 83.39% is the fusion of the last layer of AlexNet, VGG-16 and ResNet-50. They are all generated by our IEDCA-IRE. Our proposed fusion model not only inherits the class information in the correlation analysis and retains enough dimensions of feature for correlation

TABLE 4.6: Mean average precision (%) for the fusion of AlexNet (A), VGG-16 (V) and ResNet-50 (R) in PASCAL VOC-2007 database.

<i>Method</i> \ <i>Model</i>	<i>A+V</i>	<i>A+R</i>	<i>V+R</i>	<i>A+V+R</i>
Serial + mRMR	70.16	74.86	77.15	78.49
CCA/MCCA + mRMR	74.62	76.01	77.92	78.68
MSCCA + mRMR	75.86	76.58	78.30	80.46
MDCA + mRMR	75.65	77.62	78.96	81.62
IEDCA-IRE	79.09	79.78	81.17	83.39

analysis, but also eliminates the irrelevant redundancy. Therefore, the proposed fusion model provides a more discriminative and effective feature fusion scheme by keeping only the relevant redundancy. Experiment results show that our proposed fusion model outperforms serial, CCA-based and supervised CCA-based feature fusion models. From Table 4.6, it shows the difference ResNet makes in enhancing the performance. The difference between individual ResNet result and VGG-16 is 2.23%, while the difference between individual ResNet result and AlexNet is 6.16%. Also, note the A+V result is 79.09% but the addition of ResNet improves the accuracy to 83.39% (A+V+R), which makes a 4.30% difference. Comparing with all the two-model fusion and three-model fusion, we can see the addition of another model to a two-model fusion enhances the classification performance.

To further illustrate the proposed fusion method in comparison with deep neural networks and other fusion methods, Table 4.7 provides the per-class average precision (AP) as well as the overall mean AP (mAP) of the comparing methods on the VOC-2007 dataset for the fusion of all the three deep neural network models (A+V+R). Our method achieved the best performance of 83.39% in terms of the overall mean average precision (mAP). In each class average precision, our method generates either the highest class average precision or the second-best result.

4.3.4 Fusion of Deep Feature and Handcrafted Feature

In this experiment, we test our algorithm in text classification on Stack Overflow dataset. Stack Overflow is published on the Kaggle website. It contains 3,370,528 samples. We use a subset of it [122] to evaluate our proposed fusion approach on the task of predicting tags from question text. We randomly select 20,000 question titles from 20 different tags. The 20 different tags are wordpress, oracle,

TABLE 4.7: Classification results (%) for deep feature fusion of AlexNet, VGG and ResNet in VOC-2007 database.

Class	Method	AlexNet	VGG	ResNet	A+V+R					
					Serial+mRMR	MCCA+mRMR	MSCCA+mRMR	MDCA+mRMR	IEDCA-IRE
76.40	86.91	86.74	87.16	87.63	88.50	89.37	91.96
75.17	79.84	82.52	83.79	84.81	84.97	84.81	86.72
75.32	82.17	85.16	86.45	88.91	89.48	88.96	90.13
69.89	73.46	76.92	83.16	84.62	86.59	87.91	87.86
44.81	47.83	49.59	49.97	50.38	53.41	54.64	58.52
79.29	82.98	85.57	85.23	85.64	86.38	87.03	88.95
81.76	86.59	88.43	89.44	88.85	90.53	91.28	93.72
82.16	86.73	86.46	88.87	89.57	90.07	91.00	91.66
57.35	59.87	61.93	90.94	57.16	61.96	64.32	66.54
75.32	79.64	80.53	81.13	77.38	75.89	83.64	85.84
65.78	68.86	71.85	74.79	69.23	76.52	74.83	76.48
80.67	83.97	85.31	86.21	86.94	87.42	88.54	90.79
81.26	84.18	85.42	86.55	87.15	87.84	88.37	89.96
79.39	81.84	83.75	81.25	82.91	83.46	83.68	84.29
86.52	89.08	91.37	92.91	90.38	92.27	94.84	95.81
43.35	47.09	52.98	43.86	50.42	54.93	56.00	55.86
69.49	72.68	75.71	76.33	80.56	80.35	84.24	86.67
56.65	59.73	62.16	64.15	62.24	68.91	69.48	69.32
86.41	89.65	92.47	93.06	93.59	94.16	93.27	96.35
38.46	70.91	73.76	74.52	75.14	75.63	76.15	80.43	
mAP		71.77	75.70	77.93	78.49	78.68	80.46	81.62	83.39	

svn, apache, excel, matlab, visual-studio, cocoa, osx, bash, spring, hibernate, scala, sharepoint, ajax, qt, drupal, linq, haskell, and magento. The average text length is at 8.31 words, and the maximum length is 34 words.

A pre-trained model called Dynamic Convolutional Neural Network (DCNN) [123] is used. The activations of the last layer of the network are used as our features. The feature dimension is 4096. DCNN forms local features for each word and combines them to produce a global feature vector for the whole text. The features generated by DCNN are based on a hierarchy of local features, so the context can be reflected. We use the word embeddings as the input to the selected pre-trained DCNN. Word2vec dictionary [124] is applied to train the word embeddings on the whole corpus of the Stack Overflow dataset. We also extract the feature of bag-of-words (BoW) with a dimension of 22956.

During the intra-class and extra-class discriminative analysis, the dimensions of deep features and BoW are reduced to 2000 and 10000 by W_x and W_y , respectively. In the subsequent intra-class correlation analysis, Q_x and Q_y are 2000×1500 and 10000×1500 , respectively. It results in a fused feature set with a dimension of 3000. By incorporating kernelized intra-class correlation analysis into mRMR, we further eliminate the irrelevant redundancy resulting in a dimension of 2060. For CCA, PCA is firstly applied to reduce the dimensions of deep features and BoW to 1900 and 10200, respectively. During CCA, W_x and W_y are 1900×1450 and 10200×1450 , respectively. This results in a fused feature set with dimension of 2900. mRMR is then used to select 2060 features for a fair comparison. MSCCA and MDCA result in a fused feature set with dimension of 2940 and 3020, respectively. Similarly, mRMR is then used to select 2060 features for a fair comparison. The nearest neighbor parameter k is set to 10. Ten-fold cross-validation was carried out for assessing the performance of our proposed approach. The proposed feature fusion algorithm is compared with other feature fusion methods as in the previous experiments.

Table 4.8 compares the classification accuracy obtained from each type of feature, as well as the fused feature by different methods. Comparing with the best result of the individual feature set obtained by DCNN (82.87%), we notice an improvement of about 5% achieved by our proposed fusion approach (87.63%). For individual feature, the performance of the feature extracted by DCNN is better than the feature of BoW. For the fused feature of BoW and DCNN, serial fusion is better

TABLE 4.8: Classification accuracies (%) in Stack Overflow dataset.

<i>Method \ Feature</i>	<i>BoW</i>	<i>DCNN</i>	<i>BoW + DCNN</i>
SVM	77.43	82.52	-
DCNN	-	82.87	-
Serial + mRMR + SVM	-	-	81.48
CCA/MCCA + mRMR + SVM	-	-	80.24
MSCCA + mRMR + SVM	-	-	84.15
MDCA + mRMR + SVM	-	-	83.29
IEDCA-IRE + SVM	-	-	87.63

than parallel fusion. Our proposed feature fusion algorithm outperforms all the comparing methods.

Figure 4.6 (a) to (d) summarize the performance attained by the methods considered including serial fusion, MCCA, MSCCA, MDCA and IEDCA-IRE. Serial fusion, MCCA, MSCCA and MDCA are used together with mRMR. Figure 4.6 (e) to (h) depict the redundancy rate [125] attained by various methods. Compared with these methods, IEDCA-IRE attains a relatively high redundancy rate in most cases, while exhibiting high improvements over the other methods in classification accuracy of D-case, SDUMLA-HMT, Stack Overflow and mAP of PASCAL VOC 2007 at the same time. This improvement can be attributed to the benefits brought by the relevant redundancy considered in IEDCA-IRE, as well as the irrelevant redundancy elimination. Since IEDCA-IRE retains the relevant redundancy, it has a relatively high redundancy rate in most cases.

4.3.5 Effect of Nearest Neighbor Parameter k

We empirically set k in the experiments. The parameter k is selected from $\{1, 2, \dots, l - 1\}$, where l represents the number of training samples for each class. To demonstrate the effect of k , the AR and PASCAL VOC-2007 databases are used. For AR database, seven images from each subject are randomly selected for training, while the remaining seven images are for testing. For PASCAL VOC-2007, the experimental setting is the same as in Section 4.4. Fig.4.7 shows the average classification accuracy over 10 runs versus the feature dimension for different k . As shown in Fig.4.7, the performance is not sensitive to the variation of the parameter k in the small range.

(a) D-case

(b) SDUMLA-HMT

(c) VOC 2007

(d) Stack Overflow

(e) D-case

(f) SDUMLA-HMT

FIGURE 4.6: Performance on D-case dataset, SDUMLA-HMT database, PASCAL VOC 2007 and Stack Overflow. The figures (a), (b), (c), (d) provide the classification accuracy of D-case dataset, SDUMLA-HMT database and Stack Overflow, as well as the mean average precision of PASCAL VOC 2007. The figures (e), (f), (g), (h) provide the redundancy rate.

(a)

(b)

FIGURE 4.7: The recognition performance versus feature dimension under a small range of k (a) AR database (b) PASCAL VOC-2007 database.

The experiments demonstrate IEDCA-IRE is stable for a small range of k . In the following, we will see whether the performance varies when k becomes large. PASCAL VOC-2007 is used. The experimental setting remains the same. Fig.4.8 shows the ten-run average precision results versus k . From Fig.4.8, we can find that, as k increases in a large range, the performance degrades. This can be explained when the neighborhood becomes large as k increases, the sample points may not correlate within local neighborhoods. It is suggested k should not be set too large.

FIGURE 4.8: The recognition performance versus different k on PASCAL VOC-2007 database.

4.4 Conclusions

In this chapter, we propose an enhanced feature fusion through redundancy irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis (IEDCA-IRE). It works by incorporating kernelized IEDCA into mRMR considering only relevant redundancy. As IEDCA-IRE falls in the field of CCA-based feature fusion, we have compared IEDCA-IRE with representative CCA-based feature fusion schemes including CCA, MCCA, MSCCA and MDCA. CCA and MCCA are unsupervised correlation analysis. MSCCA and MDCA are supervised correlation analysis. However, the fused feature set generated by the aforementioned correlation analysis all contains irrelevant redundancy as proved in Section 4.2.1. IEDCA-IRE is supervised and the irrelevant redundancy is eliminated in the fused feature set. The proposed IEDCA-IRE inherits the discriminating structure in the learned correlation features by incorporating the class information and thoroughly explores the intra-class correlation including both one-to-one pair-wise correlation and the correlation between different features within the same class. Moreover, more information is retained in the discriminative analysis to enhance the performance of learned correlation features. Extensive experiments on five different databases demonstrate the effectiveness of our proposed approach in the fusion of different features of single modality, multimodal feature fusion, fusion

of deep feature extracted from different deep neural network models and fusion of deep features and handcrafted features.

Chapter 5

Deep Supervised Multi-View Canonical Correlation Analysis

In previous Chapter 4, we kernelized intra-class correlation analysis, which is a CCA-based multi-view representation learning method. Such kernel extension allows the learning of nonlinear representations. However, the learned representation by the kernel CCA-based approach is limited by the fixed kernel. Moreover, the training time of kernel CCA scales poorly with the size of training data as it is a nonparametric method. To address this, we propose deep supervised multi-view canonical correlation analysis (DMCCA) in this chapter. DMCCA tries to learn a shared complex nonlinear representation from multiple views of data and simultaneously learn complex nonlinear transformations from each view to this shared space in which the resulting transformations are maximally informative to each other. It combines the canonical correlation analysis from multiple views of data with the flexibility of deep representation learning.

This chapter is organized as follows. In Section 5.1, we illustrate the motivation and the related multi-view canonical correlation analysis. The proposed deep supervised multi-view canonical correlation analysis (DMCCA) is elaborated in Section 5.2. In Section 5.3 and Section 5.4, we demonstrate the experiments and provide conclusions.

5.1 Introduction

5.1.1 Motivation

Canonical correlation analysis (CCA) [23] is a representative method in the studies of multi-view representation learning. CCA finds linear projections of two views of the data that are maximally correlated. Its kernel extension allows the learning of nonlinear representations. However, the learned representation by kernel CCA is limited by the fixed kernel. Moreover, the training time of kernel CCA scales poorly with the size of training data as it is a nonparametric method. Although CCA and the kernel extension can effectively model the relationship between two sets of variables, they are limited in dealing with high-level associations among multi-view data. With the development of deep neural networks, we propose deep supervised multi-view canonical correlation analysis (DMCCA) to solve this problem, with a strategy of learning a joint representation which is associated among multiple views of data at a higher level after learning several layers of view-specific representations from lower layers. DMCCA tries to learn a shared complex nonlinear representation from multiple views of data and simultaneously learn complex nonlinear transformations from each view to this shared space in which the resulting transformations are maximally informative to each other. It combines canonical correlation analysis from multiple views of data with the flexibility of deep representation learning. The label information is leveraged for learning better representations and embedded as a view into the whole networks at the training stage. During training, the label information and the network structure of other views of data are jointly trained to update the parameters of the whole networks and maximize the total correlation simultaneously.

5.1.2 Multi-View Canonical Correlation Analysis

Rather than define the canonical correlation problem in terms of finding those linear composites of two views of variables that are maximally correlated, in MCCA, it is defined in terms of finding a single canonical variate, together with a different linear combination of the variables comprising each of K views, with the objective of maximizing a mean square correlation of these K linear composites.

In matrix notation, assume K observation matrices $X_1, X_2, \dots, X_K$, where X_i is $m_i \times N$ (representing N observations on m_i variables). Let T denote the unknown group configuration matrix, where T is $q \times N$ (q is selected empirically). We seek a shared representation T and K transformation vectors A_i (where A_i has m_i components) such that:

$$R^2 = \sum_{i=1}^K [r(T, A_i^T X_i)]^2 \quad (5.1)$$

is a maximum, where $r(T, A_i^T X_i)$ denotes the formal product moment correlation between T and the linear composite $A_i^T X_i$. The maximum correlation over all A_i of $[r(T, A_i^T X_i)]^2$ is given by:

$$\max_{A_i} [r(T, A_i^T X_i)]^2 = \frac{TX_i^T (X_i X_i^T)^{-1} X_i T}{TT^T} \quad (5.2)$$

so that R^2 is equivalent to:

$$R^2 = \frac{1}{TT^T} \sum_{i=1}^K TX_i^T (X_i X_i^T)^{-1} X_i T^T \quad (5.3)$$

It has been proved that the maximization of R^2 is equivalent to the minimization of $\sum_{i=1}^K \text{tr} [(A_i^T X_i - T)(A_i^T X_i - T)^T]$ [46]. The group configuration matrix T can be obtained using the eigenequation:

$$\left(\sum_{i=1}^K X_i^T (X_i X_i^T)^{-1} X_i \right) T = T \Lambda \quad (5.4)$$

where Λ is a diagonal matrix with q largest eigenvalues of $\sum_{i=1}^K X_i^T (X_i X_i^T)^{-1} X_i$. The matrices A_i can be calculated as:

$$A_i = (X_i X_i^T)^{-1} X_i T^T \quad (5.5)$$

FIGURE 5.1: A schematic of DMCCA, consisting of deep networks for K views. The networks are jointly trained with label view so that the output layers of the networks are maximally correlated.

5.2 Deep Supervised Multi-View Canonical Correlation Analysis

In this section, we propose deep supervised MCCA (DMCCA): a multi-view representation learning approach that benefits from the hierarchy abstract power of deep representation learning. DMCCA learns a nonlinear mapping for each view and maximizes the correlation among the learned representations among different views. During training, DMCCA passes the input vectors from each view through multiple layers of nonlinear transformation and backpropagates the gradient of the MCCA objective with respect to parameters of the network, as shown in Figure 5.1.

Assume K views in our data and let $X_i \in \mathbb{R}^{m_i \times N}$ represent the i^{th} input matrix.

The network for the i^{th} view consists of D_i layers, and each layer in the i^{th} view network has c_d^i units. The output of the first layer for the i^{th} view is $h_1^i = s(W_1^i x_i + b_1^i)$, where x_i is an instance for the i^{th} view and $s : \mathbb{R} \rightarrow \mathbb{R}$ is a nonlinear activation function. The output of the d^{th} layer for the i^{th} view is $h_d^i = s(W_d^i h_{d-1}^i + b_d^i)$, where $W_d^i \in \mathbb{R}^{c_d^i \times c_{d-1}^i}$ is the weight matrix for d^{th} layer of the i^{th} view. The output of last layer for the i^{th} view is represented by $f_i(X_i)$.

In DMCCA, the objective is to jointly learn network parameters W_d^i and b_d^i for different views such that $R^2 = \sum_{i=1}^K [r(T, A_i^T f_i(X_i))]^2$, the correlation among the learned representations across different views is as high as possible. W_d^i and b_d^i are trained to maximize the correlation using gradient-based optimization. To compute the gradient of $r(T, A_i^T f_i(X_i))$ with respect to W_d^i and b_d^i , we can compute its gradient with respect to $f_i(X_i)$ and then use backpropagation. As mentioned in Section 5.1, the maximization of R^2 is equivalent to the minimization of $\sum_{i=1}^K \text{tr} \left[(A_i^T f_i(X_i) - T)(A_i^T f_i(X_i) - T)^T \right]$. Let $Q_i = (f_i(X_i) f_i(X_i)^T)^{-1}$. Since $A_i = Q_i f_i(X_i) T^T$, this objective function can be rewritten as:

$$\begin{aligned} \min_{T \in \mathbb{R}^{q \times N}, A_i \in \mathbb{R}^{q_i \times q}} & \sum_{i=1}^K \text{tr} (A_i^T f_i(X_i) - T)(A_i^T f_i(X_i) - T)^T \\ & = \sum_{i=1}^K \text{tr} \left[T \left(f_i(X_i)^T Q_i f_i(X_i) - I \right) T^T \right] \end{aligned} \quad (5.6)$$

Note that $f_i(X_i)^T Q_i f_i(X_i)$ is symmetric and idempotent. As T are the eigenvectors for the q largest eigenvalues of $\sum_{i=1}^K f_i(X_i)^T Q_i f_i(X_i)$, we have the following:

$$\begin{aligned} & \sum_{i=1}^K \text{tr} \left[T (f_i(X_i)^T Q_i f_i(X_i) - I) T^T \right] \\ & = \text{tr} \left(T \left[\sum_{i=1}^K f_i(X_i)^T Q_i f_i(X_i) \right] T^T \right) - \sum_{i=1}^K \text{tr} (I_q) \\ & = \text{tr} \left(T \left[\sum_{i=1}^K f_i(X_i)^T Q_i f_i(X_i) \right] T^T \right) - Kq \end{aligned} \quad (5.7)$$

To compute the gradient and take derivative with respect to each output layer $f_i(X_i)$:

$$\frac{\partial O}{\partial f_i(X_i)} = A_i (T - A_i^T f_i(X_i)) \quad (5.8)$$

The details of the derivation are provided in Section 5.2.1. During training, the label information and the network structure of other views of data are jointly trained to update the parameters of the whole networks and maximize the total correlation simultaneously. The whole algorithm is summarized below.

Algorithm 5.1 Deep Supervised Multi-View CCA Algorithm

Input: multi-view training data matrix, $X_1, X_2, \dots, X_K$; learning rate η ; number of iterations S .

Output: $Y_1, Y_2, \dots, Y_K$.

Training Stage:

for iteration $t = 1, 2, \dots, S$ **do**

for each view $i = 1, 2, \dots, K$ **do**

$f_i(X_i) \leftarrow$ forward pass of X_i with weight W_i

end for

$A_1, A_2, \dots, A_K, T$ from MCCA of $f_1(X_1), f_2(X_2), \dots, f_K(X_K)$

for each view $i = 1, 2, \dots, K$ **do**

$\partial O / \partial f_i(X_i) \leftarrow A_i (T - A_i^T f_i(X_i))$

$\nabla W_i \leftarrow \text{backprop}(\partial O / \partial f_i(X_i), W_i)$

$W_i \leftarrow W_i - \eta \nabla W_i$

end for

end for

Testing Stage:

for each view $i = 1, 2, \dots, K$ **do**

$f_i(X_i) \leftarrow$ forward pass of X_i with weight W_i

end for

$A_1, A_2, \dots, A_K, T$ from MCCA of $f_1(X_1), f_2(X_2), \dots, f_K(X_K)$

for each view $i = 1, 2, \dots, K$ **do**

$Y_i \leftarrow A_i^T f_i(X_i)$

end for

5.2.1 Derivation of DMCCA Gradient

To perform backpropagation, we need to compute the gradient of the correlation objective with respect to the output layers for different views. Let O denote the objective function. The objective function O is:

$$\min_{T \in \mathbb{R}^{q \times N}, A_i \in \mathbb{R}^{o_i \times q}} \sum_{i=1}^K \text{tr}(A_i^T f_i(X_i) - T)(A_i^T f_i(X_i) - T)^T \quad (5.9)$$

As mentioned in Section 5.1.2, the solution of A_i is calculated through an eigenvalue problem. Let F_i represents $f_i(X_i)$. To compute the gradient with respect to the output layers, let's rewrite and further simplify this objective function as:

$$\sum_{i=1}^K \text{tr}(A_i^T F_i - T)(A_i^T F_i - T)^T = \sum_{i=1}^K \|A_i^T F_i - T\|_F^2 \quad (5.10)$$

As $A_i = (F_i F_i^T)^{-1} F_i T^T$, the above equation can be further rewritten as:

$$\begin{aligned} & \sum_{i=1}^K \text{tr}(A_i^T F_i - T)(A_i^T F_i - T)^T \\ &= \sum_{i=1}^K \|T F_i^T (F_i F_i^T)^{-1} F_i - T\|_F^2 \\ &= \sum_{i=1}^K \text{tr} \left[T (F_i^T (F_i F_i^T)^{-1} F_i - I) T^T \right] \end{aligned} \quad (5.11)$$

Note that $F_i^T (F_i F_i^T)^{-1} F_i$ is symmetric and idempotent. As T are the eigenvectors for the q largest eigenvalues of $\sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i$, we have the following:

$$\begin{aligned} & \sum_{i=1}^K \text{tr} \left[T (F_i^T (F_i F_i^T)^{-1} F_i - I) T^T \right] \\ &= \text{tr} \left(T \left[\sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i \right] T^T \right) - \sum_{i=1}^K \text{tr}(I_q) \\ &= \text{tr} \left(T \left[\sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i \right] T^T \right) - Kq \end{aligned} \quad (5.12)$$

Again, as T are the eigenvectors for the q largest eigenvalues of $\sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i$, the above equation can be further simplified as:

$$\text{tr}(T \left[\sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i \right] T^T) - Kq = \sum_{j=1}^q \lambda_j \left[\sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i \right] - Kq \quad (5.13)$$

To compute the gradient with respect to the output layers (i.e., $\partial O / \partial F_i$). Let $H = \sum_{i=1}^K F_i^T (F_i F_i^T)^{-1} F_i$ and $Q_i = (F_i F_i^T)^{-1}$. Using the chain rule:

$$\frac{\partial O}{\partial (F_i)_{ab}} = \sum_{cd} \frac{\partial O}{\partial H_{cd}} \cdot \frac{\partial H_{cd}}{\partial (F_i)_{ab}} \quad (5.14)$$

By Eq. 101 in [126] for the derivative of trace, $\partial O / \partial H_{cd} = (T^T T)_{cd}$. Then we can have:

$$\frac{\partial O}{\partial (F_i)_{ab}} = \sum_{cd} (T^T T)_{cd} \cdot \frac{\partial H_{cd}}{\partial (F_i)_{ab}} \quad (5.15)$$

Let $J_i = F_i^T Q_i F_i$, we can easily have $\partial H_{cd} / \partial (F_i)_{ab} = \partial (J_i)_{cd} / \partial (F_i)_{ab}$, where $(J_i)_{cd} = \sum_{ef} (F_i)_{ec} (Q_i)_{ef} (F_i)_{fd}$. Using the product rule:

$$\begin{aligned} \frac{\partial H_{cd}}{\partial (F_i)_{ab}} &= \frac{\partial (J_i)_{cd}}{\partial (F_i)_{ab}} \\ &= \delta_{cb} \sum_f (F_i)_{fd} (Q_i)_{af} + \delta_{db} \sum_f (F_i)_{ec} (Q_i)_{ea} + \sum_{ef} (F_i)_{ec} (F_i)_{fd} \frac{\partial (Q_i)_{ef}}{\partial (F_i)_{ab}} \\ &= \delta_{cb} (Q_i F_i)_{ad} + \delta_{db} (Q_i F_i)_{ac} + \sum_{ef} (F_i)_{ec} (F_i)_{fd} \frac{\partial (Q_i)_{ef}}{\partial (F_i)_{ab}} \end{aligned} \quad (5.16)$$

To compute $\partial (Q_i)_{ef} / \partial (F_i)_{ab}$ by the chain rule:

$$\begin{aligned}
\frac{\partial(Q_i)_{ef}}{\partial(F_i)_{ab}} &= \sum_{hl} \frac{\partial(Q_i)_{ef}}{\partial(Q_i^{-1})_{hl}} \cdot \frac{\partial(Q_i^{-1})_{hl}}{\partial(F_i)_{ab}} \\
&= - \sum_{hl} \left\{ (Q_i)_{eh} (Q_i)_{lf} [\delta_{ah} (F_i)_{lb} + \delta_{al} (F_i)_{hb}] \right\} \\
&= - \sum_l (Q_i)_{ea} (Q_i)_{lf} (F_i)_{lb} - \sum_m (Q_i)_{eh} (Q_i)_{af} (F_i)_{hb} \\
&= - (Q_i)_{ea} (Q_i F_i)_{fb} - (Q_i)_{af} (Q_i F_i)_{eb}
\end{aligned} \tag{5.17}$$

Substitute above equation into $\partial H_{cd}/\partial(F_i)_{ab}$, we get:

$$\begin{aligned}
\frac{\partial H_{cd}}{\partial(F_i)_{ab}} &= \frac{\partial(J_i)_{cd}}{\partial(F_i)_{ab}} \\
&= \delta_{cb} (Q_i F_i)_{ad} + \delta_{db} (Q_i F_i)_{ac} + \sum_{ef} (F_i)_{ec} (F_i)_{fd} \frac{\partial(Q_i)_{ef}}{\partial(F_i)_{ab}} \\
&= \delta_{cb} (Q_i F_i)_{ad} + \delta_{db} (Q_i F_i)_{ac} - (Q_i F_i)_{ac} (F_i^T Q_i F_i)_{bd} - (Q_i F_i)_{ad} (F_i^T Q_i F_i)_{bc} \\
&= (I - J_i)_{cb} (Q_i F_i)_{ad} + (I - J_i)_{db} (Q_i F_i)_{ac}
\end{aligned} \tag{5.18}$$

Then we substitute $\partial H_{cd}/\partial(F_i)_{ab}$ into $\partial O/\partial(F_i)_{ab}$:

$$\begin{aligned}
\frac{\partial O}{\partial(F_i)_{ab}} &= \sum_{cd} (T^T T)_{cd} \cdot \frac{\partial H_{cd}}{\partial(F_i)_{ab}} \\
&= \sum_{cd} (T^T T)_{cd} (I - J_i)_{cb} (Q_i F_i)_{ad} + \sum_{cd} (T^T T)_{cd} (I - J_i)_{db} (Q_i F_i)_{ac} \\
&= 2 [Q_i F_i T^T T (I - J_i)]_{ab}
\end{aligned} \tag{5.19}$$

Eventually we obtain:

$$\frac{\partial O}{\partial f_i(X_i)} = \frac{\partial O}{\partial F_i} = 2 [Q_i f_i(X_i) T^T T (I - J_i)] \tag{5.20}$$

Since $A_i = (F_i F_i^T)^{-1} F_i T^T = Q_i F_i T^T$ and $J_i = F_i^T Q_i F_i$, $\partial O/\partial F_i$ can be further simplified as:

$$\begin{aligned}
\frac{\partial O}{\partial f_i(X_i)} &= 2 [Q_i F_i T^T T (I - J_i)] \\
&= 2A_i [T - A_i^T f_i(X_i)]
\end{aligned} \tag{5.21}$$

5.3 Experiments and Discussions

To test the effectiveness of our proposed DMCCA, experiments are performed on XRMB database [127], SDUMLA-HMT database [112] and AR database [113] respectively. Experiments are designed to fuse a variety of features from single modality as well as multiple modalities. Section 5.3.1 demonstrates the experiment of multimodal feature fusion, while Section 5.3.2 shows the experiment of different fingerprint fusion. Section 5.3.3 demonstrates the experiment of unimodal feature fusion.

5.3.1 Multimodal feature fusion: Phonetic Recognition on XRMB database

In this experiment, we test DMCCA for phonetic recognition on the University of Wisconsin X-ray Microbeam Database (XRMB) [127]. XRMB contains acoustic and articulatory recordings as well as phonetic labels. We present phoneme classification results on the acoustic vectors projected using CCA, DCCA and DGCCA. The acoustic features are mean- and variance-normalized 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives computed every 10ms over a 25ms window. The articulatory data consist of horizontal and vertical displacements of 8 pellets on the speaker’s lips, tongue and jaw, yielding a 16-dimensional vector at each sample. The articulatory measurements are downsampled to match the MFCC frame rate.

We compare phonetic classification results on the vectors projected using CCA, DCCA, and DMCCA. We consider articulatory and acoustic data as the two views. SVM is used as the standard classifier to facilitate fair comparisons. We use the same train/tune/test split of the data as in [128]. We run a set of cross-speaker

TABLE 5.1: Classification accuracies (%) in XRMB database.

<i>Method</i>	<i>Cross-speaker</i>	<i>Speaker-dependent</i>
Serial + PCA + SVM	54.92	73.84
Serial + LDA + SVM	56.75	74.53
Acoustic Feature + NN	55.63	73.24
Articulatory Feature + NN	50.85	68.17
Acoustic Feature + SVM	57.14	74.08
Acoustic Feature + Softmax	56.69	73.81
Articulatory Feature + SVM	52.28	69.66
Articulatory Feature + Softmax	52.04	69.21
PCA + CCA + SVM	55.26	75.19
LDA + CCA + SVM	55.83	76.42
DCCA + SVM	59.34	77.82
DMCCA + SVM	62.01	80.23

experiments following the setting in [128]. Following the pre-processing in [128], we get 273 and 112-dimensional feature vectors for the first and second view respectively. Each speaker has roughly 50,000 frames. For the label view in DMCCA, we use 39-dimensional one-hot vectors corresponding to the labels for each frame, following [128].

For the network size of three views, each containing three hidden layers with ReLu activation functions. Hidden layers for the acoustic view are all with width of 1024, and layers in the articulatory view all have width of 512. For the label view network, it has a width of 256. L2 penalty constants of 0.0001 and 0.01 are used to train the acoustic and articulatory view networks, respectively. The label view neural network had an L2 penalty of 0.0005. The output layer dimension of each network is set to 30 for DCCA and DMCCA. Covariance matrix regularization is set to 0.01.

- Speaker-dependent phonetic recognition - We first explore the performance of phonetic recognition using transformed CCA/DCCA/DMCCA-based features in a speaker-dependent setting. In particular, we train and test CCA/DCCA/DMCCA transformations on data from each of XRMB speakers.
- Speaker-independent phonetic recognition - To test the degree of speaker-independence of CCA/DCCA/DMCCA projections, we repeat the above experiments, but learn the projections on three of the XRMB speakers (“source” speakers) and test on the fourth (“target” speaker). The results are shown in Table 5.1.

FIGURE 5.2: Confusion matrix for speaker-dependent DCCA.

As shown in Table 5.1, DMCCA improves on both the linear and the nonlinear two-view DCCA for both cross-speaker and speaker-dependent cross-validated tasks. In all cases, DMCCA improves over CCA by around 5%, and DMCCA improves over DCCA by 2-3%.

To highlight the improvements of DMCCA over DCCA, Figure 5.2 and Figure 5.3 presents the confusion matrices on speaker-dependent test data. We observe large improvements in the classification of D, V and AW. The confusion matrices also highlight the common misclassifications that DMCCA improves upon. For example, DMCCA rectifies the frequent misclassification of D as T.

5.3.2 Different Fingerprint Fusion

In this experiment, SDUMLA-HMT database is used. This database contains 106 individuals. Experimental settings follow those in [51]. Fingerprints from thumb

FIGURE 5.3: Confusion matrix for speaker-dependent DMCCA.

and index of both hands are used. The fingerprints were enhanced following the method in [115]. After enhancement, the core point was detected, and convolution was performed on the enhanced fingerprints using Gabor filters which have eight orientations. Around the detected core point, circular tessellations consisting of 15 concentric bands were extracted. For each fingerprint, the feature dimension is 640.

Following [129], we use 8480 fingerprint images for training and another 8480 for testing. For the network size of four fingerprint views, each containing four hidden layers with ReLu activation functions. All the hidden layers for fingerprint view are with width of 1024. For the label view network, it has four hidden layers with a width of 512. The output layer dimension of each network is set to 100 for both DCCA and DMCCA.

Our proposed DMCCA is compared with other feature fusion algorithms as in the previous experiment. Besides, we also compare with Joint Sparse Representation

TABLE 5.2: Individual classification accuracies (%) for fingerprint fusion on SDUMLA-HMT database.

<i>Modality</i>	<i>Classification Accuracy</i>
Left thumb fingerprint + NN	76.40
Left index fingerprint + NN	87.49
Right thumb fingerprint + NN	80.87
Right index fingerprint + NN	88.15
Left thumb fingerprint + SVM	76.84
Left index fingerprint + SVM	88.53
Right thumb fingerprint + SVM	81.05
Right index fingerprint + SVM	89.58

TABLE 5.3: Classification accuracies (%) for fingerprint fusion on SDUMLA-HMT database.

<i>Method</i>	<i>Four Fingerprints</i>
Serial + PCA + SVM	89.94
Serial + LDA + SVM	90.21
PCA + MCCA + SVM	92.25
LDA + MCCA + SVM	92.53
JSR	93.69
SMDL	93.71
DMCCA + SVM	96.23
DMCCA + Softmax	95.86

(JSR) and SMDL [50], which are used in combination with a sparse representation classifier. Table 5.2 shows the classification accuracy for each modality, while Table 5.3 provides the classification accuracy for fingerprint fusion. Experimental results demonstrate our DMCCA outperforms other comparing methods.

5.3.3 Unimodal Feature Fusion: AR Database

In this experiment, AR database [113] is used. This database contains images from 100 subjects under various illuminations and facial expressions. The subset of AR database without occlusion is chosen. In the experiment, seven images from each subject are randomly chosen for training, while the remaining are used for testing. Ten independent recognition tests are conducted and the results are averaged. Two types of feature extraction methods were employed. 2D Wavelet transform (WT) is applied to extract the first type of feature. Pyramid decomposition scheme and the sub-band coefficients LL of the level of decomposition 3 were used. The coefficients generated by the mother function of Daubechies 4 (D4) were computed resulting in the features with dimension of 1444. Histogram of Oriented Gradients

TABLE 5.4: Classification accuracies (%) in AR database.

<i>Method</i>	<i>Classification Accuracy</i>
Serial + PCA + SVM	81.87
Serial + LDA + SVM	83.24
Wavelet + NN	81.73
HOG + NN	76.34
Wavelet + SVM	82.19
Wavelet + Softmax	81.08
HOG + SVM	77.92
HOG + Softmax	75.58
PCA + CCA + SVM	90.26
LDA + CCA + SVM	91.47
DCCA + SVM	96.06
DMCCA + SVM	98.58

(HOG) is the second type of feature extracted with nine orientations in 5×5 cells. UOCTTI variant [115] is adopted in calculating directed and undirected gradients of HOG. The dimension of HOG is 1984. Table 5.4 shows the average classification accuracies, obtained by the individual feature and the fused feature. For the network size of three views, each containing four hidden layers with ReLu activation functions. Hidden layers for Wavelet and HOG are all with width of 1024. For the label view network, it has a width of 512.

5.4 Conclusions

In this chapter, we have proposed deep supervised multi-view canonical correlation analysis (DMCCA). DMCCA tries to learn a shared complex nonlinear representation from multiple views of data and simultaneously learn complex nonlinear transformations from each view to this shared space in which the resulting transformations are maximally informative of each other. It combines canonical correlation analysis from multiple views of data with the flexibility of deep representation learning. The label information is leveraged for learning better representations and embedded as a view into the whole networks at training stage. During training, the label information and the network structure of other views of data are jointly trained to update the parameters of the whole networks and maximize the total correlation simultaneously. Various experiments demonstrate the effectiveness of our proposed approach.

Chapter 6

Feature Regrouping for CCA-Based Feature Fusion and Extraction Through Normalized Cut

In previous correlation-based multi-view representation learning approaches we proposed, the natural groupings of features are directly used. It is still unclear whether the natural groupings of features are optimal for CCA-based fusion. In this chapter, we propose a feature regrouping algorithm for CCA-based feature fusion and extraction through normalized cut (FR-NC). Feature correlation analysis is incorporated into normalized cut, in which the intra-group correlation is maximized, and the extra-group correlation is minimized simultaneously. CCA-based feature fusion is performed on the regrouped features. The proposed feature regrouping scheme can help to enhance the predictive power of the fused feature by discovering the feature groups across multiple feature sets through feature correlation analysis.

6.1 Introduction

In this section, we demonstrate the motivation behind our work first. Then, the related work in normalized cut is reviewed briefly.

6.1.1 Motivation

CCA-based feature fusion is performed directly on the feature sets. However, whether the natural groupings of features produce the best result is still a question mark. Despite all the advantages reported in the existing CCA-based feature fusion, they still encounter certain loss of information during the transformation in different feature spaces. Moreover, the complementary and mutual information existed in multiple feature sets cannot be fully utilized. We have done some tentative experiments on UCI dataset [130], D-case dataset [111] and SDUMLA-HMT database [112] by concatenating all the sets of features together and randomly split into different number of groups. CCA-based feature fusion is performed on the regrouped features. We notice the performance of classification is improved. This motivates us to find an effective way to divide the high-dimensional data into different groups for feature fusion. The correlation and the dependency between different feature sets need further exploration since the implication among different features can enhance the performance of pattern classification.

Let us illustrate the shortcomings of existing CCA-based feature fusion schemes using a simple example. Figure 6.1 illustrates the feature correlation in SDUMLA-HMT database, before and after performing our proposed feature regrouping method. The details are provided in Section 6.3. In the figure, the light-coloured areas represent high correlation among features, while the dark-coloured areas represent low correlation. When the features only correlate with themselves, only the diagonal elements are light-coloured. Our findings on the databases in Section 6.3 indicate that features within the same feature set are relatively highly correlated, and features from different feature sets are relatively low correlated. Figure 6.1(a) demonstrates the correlation among the natural groupings of feature. The light-coloured off-diagonal parts in Figure 6.1(a) shows the feature within the same feature set exhibits high correlation, while the off-diagonal dark-coloured areas in Figure 6.1(a) indicate the features from different feature sets exhibits low correlation. However, in the light-coloured off-diagonal parts in Figure 6.1(a), there exists some dark-coloured area. This indicates that some features within the same feature set are sparsely lower correlated. It may increase the variances of the fused feature and lead to impaired performance. This problem is also mentioned in [131].

(a)

(b)

FIGURE 6.1: Visualization of correlation matrices in SDUMLA-HMT database before and after performing feature regrouping for CCA-based feature fusion and extraction through normalized cut (FR-NC), elements with zero values are represented in dark colour, elements with higher values are represented in light colour. (a) the correlation matrix of the natural feature groups. (b) the correlation matrix of the regrouped features using FR-NC.

6.1.2 Normalized Cut

Graph cut-based approach is widely used in image segmentation. In graph cut-based methods, the main idea is to partition a graph $G = (V, E)$ into connected components A_i such that $\cup A_i = V$ and $A_i \cap A_j = \emptyset$ through omitting the edges which link these components. In $G = (V, E)$, pixels are represented as graph nodes V , graph edges E connect pixels within the distance of graph radius. In graph cut-based approach, global objective functions, such as the normalized cut, are optimized to segment images. Wu and Leahy [132] defined a cut between two connected components A and B as:

$$cut(A, B) = \sum_{u \in A, v \in B} W(u, v) \quad (6.1)$$

where W measures the similarity between different pixels. A higher W indicates higher similarity between the pixels. Minimum cut is proposed [132] for image segmentation in which the smallest $(K-1)$ cuts from all possible cuts are chosen and K -subgraph partitions are built by deleting the corresponding edges. In minimum cut, the criteria favours the small segment formation which contains a small number of isolated nodes. This will result in over-segmentation. To address this problem, a new disassociation measure between two components is proposed by Shi and Malik [133]. Instead of considering the weights of total edges that connect the partitions, they use the fraction of total edge connections to all graph nodes for computation of cut cost. This new measurement is called normalized cut (Ncut).

Ncut [133] is a classical graph cut-based algorithm for image segmentation through using all the edge connections towards the nodes in a graph. In Ncut, the pixel graph is partitioned using the eigenvectors of the normalized Laplacian graph matrix, which globally minimizes the segmentation cost. Given a graph $G = (V, E)$, it tries to find a cut that can partition V into two parts A and B satisfying the following minimization criterion:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(B, V)}{assoc(B, V)} \quad (6.2)$$

where $assoc(A, V) = \sum_{u \in A, t \in V} W(u, t)$ is the total sum of connections between A and all nodes, $cut(A, B) = \sum_{u \in A, v \in B} W(u, v)$ is the total sum of edges connecting

A and B . It has been proved that the minimization of $Ncut(A, B)$ is equivalent to the minimization of energy E :

$$E = \min_y \frac{y^T (D - W) y}{y^T D y} \quad (6.3)$$

with the constraint of $y_i \in \{1, -1\}$ and $y^T D \mathbf{1} = 0$ ($\mathbf{1}$ represents vector of all ones). y is an indicator vector (details can be found in [133]). The solution to the above problem is the second smallest eigenvector of the generalized eigenvalue problem:

$$(D - W) y = \lambda D y \quad (6.4)$$

D is the diagonal degree matrix where $D_{ii} = \sum_j W_{ij}$ and W is the graph affinity matrix. Since Ncut can partition a graph into two balanced parts by minimizing the corresponding cut cost, it has also been well studied for pre-segmentation in saliency detection. Ncut has been extensively exploited for image segmentation [134]. In our proposed method, Ncut is explored in feature regrouping for CCA-based feature fusion and extraction.

6.2 Feature Regrouping for CCA-based Feature Fusion and Extraction through Normalized Cut

To solve this problem, we propose a feature regrouping algorithm for CCA-based feature fusion and extraction through normalized cut (FR-NC). The motivation of FR-NC is based on the observation that relevant features are highly correlated in the high dimensional data. The fused features generated by correlated feature groups through CCA are resistant to sample size variation. Such feature groups can generate discriminative CCA-based fused features and provides additional informative group structure. Specifically, we concatenate all the extracted features together. The correlations among each feature pair of multiple feature sets are explored. Ncut is used for feature regrouping, in which the intra-group correlation is maximized, and the extra-group correlation is minimized. Highly correlated

features can stay together and the variances in the fused feature set can be reduced to improve the performance of the fused feature. As demonstrated by the light-coloured diagonal areas in Figure 6.1(b), in the regrouped features, they are highly correlated within each partitioned group. The dark-coloured diagonal areas in Figure 6.1(b) demonstrate that features from the different partitioned groups are lower correlated with each other. Compared the correlation relationship among the natural groupings in Figure 6.1(a) and our regrouped features in Figure 6.1(b), the dark-coloured area in the light-coloured off-diagonal parts in Figure 6.1(a) is eliminated by our regrouping algorithm. Moreover, the correlation within each partitioned group is enhanced. This will improve the performance of CCA-based feature fusion. The proposed feature regrouping scheme can help to reduce the variances in the estimation and improve the predictive power of the fused feature by discovering the feature groups across multiple feature sets through feature correlation analysis. The feature regrouping framework for CCA-based feature fusion and extraction is illustrated in Figure 6.2.

6.2.1 Feature Regrouping Through Normalized Cut

In feature regrouping, we try to partition the concatenated feature set into disjoint groups $F_1, F_2, \dots, F_G$ by maximizing intra-group correlation and minimizing extra-group correlation. Ncut is incorporated in our feature regrouping algorithm. Let the concatenated feature set from different feature sets of n objects be denoted by $X = [x_1, x_2, \dots, x_n] \in R^{p \times n}$, where x_i can be treated as a sample, p is the concatenated feature dimension. The covariance matrix of the full feature set is calculated as:

$$C = \frac{1}{n} \sum_{k=1}^n (x_k - u)(x_k - u)^T \quad (6.5)$$

where $u = \frac{1}{n} \sum_{k=1}^n x_k$. The normalized correlation between feature i and j is defined as:

$$R_{ij} = \frac{|C_{ij}|}{\sqrt{C_{ii}C_{jj}}} \quad (6.6)$$

FIGURE 6.2: The flowchart of feature regrouping for CCA-based feature fusion and extraction through normalized cut.

where C_{ij} , C_{ii} and C_{jj} are elements in C . A feature set V can be partitioned into G disjoint groups, $\{F_1, F_2, \dots, F_G\}$, $\bigcup_{i=1}^G F_i = V$, $\bigcap_{i=1}^G F_i = \emptyset$ by normalized cut as:

$$Ncut(F_1, F_2, \dots, F_G) = \frac{cut(F_1, V - F_1)}{assoc(F_1, V)} + \frac{cut(F_2, V - F_2)}{assoc(F_2, V)} + \dots + \frac{cut(F_G, V - F_G)}{assoc(F_G, V)} \quad (6.7)$$

Inspired by the idea of *cut* in graph theory which represents the dissimilarity degree between two disjoint sets through computation of the weight sum of the removed edges connecting two parts, we redefine the concept of *cut* by introducing the idea of extra-group correlation as:

$$cut(F_k, V - F_k) = \sum_{i \in F_k, j \in V - F_k} R_{ij} \quad (6.8)$$

$assoc(F_k, V) = \sum_{i \in F_k, j \in V} R_{ij}$ is the total correlation from features in F_k to all features in the feature set. Through computing the cut cost as a fraction of the total correlation to all features in the feature set, a balanced partitioning can be generated. Similarly, we can define the normalized association sum within groups for the regrouping as:

$$Nassoc(F_1, F_2, \dots, F_G) = \frac{assoc(F_1, F_1)}{assoc(F_1, V)} + \frac{assoc(F_2, F_2)}{assoc(F_2, V)} + \dots + \frac{assoc(F_G, F_G)}{assoc(F_G, V)} \quad (6.9)$$

where $assoc(F_1, F_1)$, $assoc(F_2, F_2)$, $assoc(F_G, F_G)$ are total intra-correlations of features within each partitioned group. The relationship between $Ncut$, which can be regarded as the extra-group correlation of the feature regrouping, and the normalized association, which can be regarded as the intra-group correlation of the feature regrouping, can be derived as follows:

$$\begin{aligned} Ncut(F_1, F_2, \dots, F_G) &= \frac{cut(F_1, V - F_1)}{assoc(F_1, V)} + \frac{cut(F_2, V - F_2)}{assoc(F_2, V)} + \dots + \frac{cut(F_G, V - F_G)}{assoc(F_G, V)} \\ &= \frac{assoc(F_1, V) - assoc(F_1, F_1)}{assoc(F_1, V)} + \frac{assoc(F_2, V) - assoc(F_2, F_2)}{assoc(F_2, V)} + \dots \\ &\quad + \frac{assoc(F_G, V) - assoc(F_G, F_G)}{assoc(F_G, V)} \\ &= G - \left(\frac{assoc(F_1, F_1)}{assoc(F_1, V)} + \frac{assoc(F_2, F_2)}{assoc(F_2, V)} + \dots + \frac{assoc(F_G, F_G)}{assoc(F_G, V)} \right) \\ &= G - Nassoc(F_1, F_2, \dots, F_G) \end{aligned} \quad (6.10)$$

The optimal feature regrouping for CCA-based feature fusion is to minimize normalized cut, which is the extra-group correlation among regrouped features. This

is equivalent to maximizing the association within each regrouped features, i.e. the intra-group correlation of the regrouped features.

6.2.2 Optimization of Feature Regrouping Via Normalized Cut

Given a feature regrouping of a feature set into G sets $F_1, F_2, \dots, F_G$, let $d(i) = \sum_{j=1}^p R_{ij}$ be the total correlation from feature i to all other features, D be a $p \times p$ diagonal matrix in which d is on the diagonal, ($D_{ii} = \sum_{j=1}^p R_{ij}$), R be a $p \times p$ symmetrical matrix. Eq. (6.10) can be expressed in Rayleigh quotient and minimization can be achieved by solving a generalized eigenvalue problem:

$$(D - R)x = \lambda Dx \quad (6.11)$$

Eq. (6.11) can be rewritten as:

$$D^{-\frac{1}{2}}(D - R)D^{-\frac{1}{2}}y = \lambda y \quad (6.12)$$

where $y = D^{\frac{1}{2}}x$. It is easily verified that $y_0 = D^{\frac{1}{2}}\mathbf{1}$ is an eigenvector of Eq. (6.12) with eigenvalue of 0, and $\mathbf{1}$ is a $p \times 1$ vector of all ones. As $(D - R)$ is a *Laplacian* matrix which is positive semidefinite, $D^{-\frac{1}{2}}(D - R)D^{-\frac{1}{2}}$ is symmetric positive semidefinite. y_0 is the smallest eigenvector in Eq. (6.12). All the eigenvectors are perpendicular to each other. Let y_1 be the second smallest eigenvector, looking back to the generalized eigenvalue problem Eq. (6.11), $y_1^T y_0 = x_1^T D \mathbf{1} = \mathbf{0}$, x_1 is the second smallest eigenvector in Eq. (6.11). $x_0 = \mathbf{1}$ is the smallest eigenvector of Eq. (6.11).

In Rayleigh quotient, if A is real symmetric with the constraint that x is orthogonal to the $j - 1$ smallest eigenvectors $x_1, \dots, x_{j-1}$, then the quotient of $\frac{x^T A x}{x^T x}$ can be minimized using the next smallest eigenvector x_j and the eigenvalue λ_j is the corresponding minimum value. From Rayleigh quotient, the second smallest eigenvector in Eq. (6.11) can be obtained, which is the solution to normalized cut:

$$y_1 = \arg \min_{y^T y_0 = 0} \frac{y^T D^{-\frac{1}{2}} (D - R) D^{-\frac{1}{2}} y}{y^T y} \quad (6.13)$$

$$x_1 = \arg \min_{x^T D \mathbf{1} = 0} \frac{x^T (D - R) x}{x^T D x} \quad (6.14)$$

Solving this generalized eigenvalue problem takes $O(p^3)$ operations, p is the number of features. This takes extremely long in real-time applications when the feature dimension is high. Since some features in a high dimensional feature set are highly correlated with each other, this makes the feature correlation matrix R sparse. Subsequently, $D^{-\frac{1}{2}} (D - R) D^{-\frac{1}{2}}$ which is the A in Rayleigh quotient, is also sparse. This leads us to use Lanczos algorithm to reduce the computational cost in solving the generalized eigenvalue problem.

Now the eigenvectors are computed, the correlated feature set can be partitioned into two groups by the second smallest eigenvector. After the correlated feature set is extracted into two groups, the algorithm can be recursively applied on the two extracted feature groups instead of using the subsequent smallest eigenvalues (e.g. the third smallest eigenvalue). This is because the approximation error accumulates with every eigenvector when we need to meet the condition that all the eigenvectors are perpendicular to each other. In this case, the subsequent smallest eigenvalues are unreliable. Therefore, certain stopping criterion is imposed in the feature regrouping:

- The regrouping stops when an isolated feature appears in the extracted group. The feature groups for fusion are the extracted groups before the isolated feature appears.
- The regrouping stops when the extraction is unstable due to the varying eigenvectors.

The whole process of feature regrouping for CCA-based feature fusion and extraction through normalized cut is summarized in Algorithm 6.1.

Algorithm 6.1 The Feature Regrouping Algorithm**Input:** Natural groupings of features.**Output:** Regrouped features.

1. Concatenate all the extracted features together. Partition the concatenated feature set into two groups according to a predefined correlation threshold value. This aims to separate correlated features from uncorrelated ones and lay a good foundation for subsequent partitioning of correlated features. The isolated uncorrelated features can be regarded as one group for fusion.
2. Construct a feature correlation matrix R based on the covariance matrix C of the whole feature set, $R_{ij} = \frac{|C_{ij}|}{\sqrt{C_{ii}C_{jj}}}$
3. Construct a $p \times p$ diagonal matrix D where the diagonal $d(i) = \sum_{j=1}^p R_{ij}$.
4. Compute the eigenvectors and the smallest eigenvalues of $(D - R)x = \lambda Dx$
5. Bipartition the correlated feature group through the second smallest eigenvector.
6. Check if the subsequent partition on the extracted groups is needed and recursively partition if necessary.

6.3 Experiments

In this section, we evaluate the effectiveness of our proposed feature regrouping scheme for CCA-based feature fusion and extraction through normalized cut (FR-NC). Experiments are conducted on D-case dataset [111], SDUMLA-HMT database [112] and PASCAL VOC 2007 [121]. The experiments are designed to concatenate a variety of features from single modality as well as multiple modalities. Feature regrouping is conducted on the concatenated features. CCA-based feature fusion is performed on the regrouped features. Moreover, we also evaluate our proposed model in the fusion of regrouped deep features.

6.3.1 Unimodal Feature Fusion

In this experiment, the effectiveness of FR-NC is demonstrated in the case of regrouped feature fusion from the concatenation of various unimodal features using D-case dataset [111]. The D-case dataset is provided by IEEE AASP challenge involving acoustic scene classification and event detection. Here, we use the scene classification dataset to evaluate our proposed model. The scene classification dataset consists of ten different scenes: tube, subway station, supermarket, restaurant, quiet street, park, open-air market, office, busy street and bus. For each

scene, ten audio clips are provided and each of them is 30 seconds long. All the audio clips are sampled at 44.1 kHz. In our experiment, three types of feature extraction methods were employed. For each audio clips, the first type of features, Mel-frequency cepstral coefficients (MFCC) were extracted. The frame window is set to 25 ms, and the step between successive windows is 10 ms. The frequency range is limited to 0-900 Hz. The periodogram is calculated for each frame. Mel filter bank is applied, followed by taking the discrete cosine transform (DCT) of the calculated energies. Recurrence Quantification Analysis (RQA) is the second type of feature extracted. The basic idea of RQA is to quantify the patterns that emerge in the recurrence plots by certain metrics. The details can be found in [114]. Texture windows of 40 MFCC frames representing 400ms of audio were used in computing the recurrence plots. Cosine similarity of 0.03 between MFCC frames is adopted. Discrete Wavelet Transform (DWT) is the third type of features employed in this experiment. DWT provides frequency and time resolution which closely matches the human ear. A six-level wavelet decomposition using the Daubechies4 (D4) wavelet was used. For each level, both the mean and the variance of the wavelet coefficients over each scene segment were computed and used as features.

We concatenate all the feature sets and apply feature regrouping based on normalized cut (FR-NC) on the concatenated feature set. Different numbers of feature groups are extracted and fused. Five-fold cross-validation for ten runs is performed and we average the recognition results. The performance of the fused feature set from the original feature sets is compared with the fused feature from regrouped features. Several state-of-the-art feature fusion algorithms mentioned in Chapter 2 are used such as serial fusion, conventional CCA, MSCCA [108], MDCA [109] and IEDCA [31]. SVM is used as the classifier.

Table 6.2 shows the classification accuracy of the individual type of features, while Table 6.2 compares the classification accuracy of the fused features from the original six feature sets and regrouped features. As demonstrated in the results, serial fusion does not always work. In the case of MFCC and DWT, the serial-fused feature is less discriminative than the individual feature. CCA-based feature fusion performs well since they maximize the correlation between different features. Supervised CCA-based feature fusion techniques perform even better due to the incorporation of class information into correlation analysis. By using feature regrouping based

TABLE 6.1: Classification accuracies (%) by SVM for different features of D-case database.

<i>Feature Type</i>	<i>Classification Accuracy</i>
MFCC (Channel 1)	65.02
MFCC (Channel 2)	64.66
RQA (Channel 1)	55.38
RQA (Channel 2)	54.97
DWT (Channel 1)	64.84
DWT (Channel 2)	56.23

on normalized cut (FR-NC), the maximization of intra-group correlation and the minimization of extra-group correlation are achieved simultaneously. FR-NC provides enhanced feature sets for fusion and help generate a more discriminative fused feature. As shown in Table 6.2, by partitioning the combined MFCC, RQA and DWT features into three groups, classification accuracy can improve from 76.72% to 78.21% through fusion of IEDCA. Compared with the fusion of the original feature sets directly, the performance has been enhanced by fusing the regrouped features from the concatenated feature set. MSCCA improved from 70.76% to 71.34%. The classification accuracy of MSCCA can be further improved to 71.91% when we extracted four groups from the concatenated feature set. MDCA improved from 74.95% to 75.04%. MCCA has also improved from 65.28% to 66.10% with PCA, and from 69.36% to 70.25% with LDA. The best classification accuracy is achieved when the concatenated features are partitioned into three or four groups. The improvement of the fused feature set by IEDCA from the regrouped feature is larger than the fused feature set through other fusion methods. This is because both IEDCA and FG-NC use the intra and extra correlation analysis, the fused feature vector is more discriminative than other fused features.

6.3.2 Multimodal Feature Fusion: SDUMLA-HMT Database

In this experiment, SDUMLA-HMT database is used. This database is a comprehensive collection of multiple modalities: finger vein, face, iris, gait and fingerprint from 106 individuals. Experimental settings follow [51]. Irises of both eyes and fingerprints from the thumb and index of both hands are used. Figure 6.3 graphically demonstrates the whole procedure of the FR-NC application in multimodal feature fusion. Segmentation was performed on iris images. The valid part of the iris is illustrated in Figure 6.3. The segmented area is normalized and a binary iris

TABLE 6.2: Classification accuracies (%) for feature fusion of regrouped features in D-case database.

<i>Method</i>	<i>NO. of Groups</i>		<i>Concatenated Feature Set from All Six Feature Sets</i>				<i>Original Six Feature Sets</i>
	2	3	4	5	6		
Serial + PCA	-	-	-	-	-	-	61.94
Serial + LDA	-	-	-	-	-	-	63.51
PCA + CCA/MCCA	63.96	66.10	65.79	64.23	63.05		65.28
LDA + CCA/MCCA	68.32	69.48	70.25	68.59	67.62		69.36
MSCCA	68.95	71.34	71.91	69.35	68.51		70.76
MDCA	72.38	75.04	74.67	72.13	71.93		74.95
IEDCA	73.64	78.21	77.83	76.32	74.48		76.72

FIGURE 6.3: The whole process of the proposed feature regrouping for multi-modal CCA-based feature fusion in SDUMLA-HMT database.

template is generated from the segmented iris as shown in Figure 6.3. Log-Gabor features are extracted from each binary iris template. For fingerprints, they were enhanced following the method in [115]. After enhancement, the core point was detected and convolution was performed on the enhanced fingerprints using Gabor filters with eight orientations. Around the detected core point, circular tessellations consisting of 15 concentric bands were extracted. The entire preprocessing is shown in Figure 6.3.

We concatenate all the six feature sets and apply FR-NC on the concatenated feature set. Different numbers of feature groups are extracted and fused. Five-fold cross-validation for ten runs is performed and we average the recognition results. The performance of the fused feature set from the original six feature sets is compared with the performance of the fused feature from our regrouped features. Different feature fusion algorithms are used as in the previous experiment. SVM is used as the classifier. Table 6.3 shows the classification accuracy for each modality, while Table 6.4 compares the classification accuracy of the fused features from the original six feature sets and the regrouped features. Experimental results demonstrate our feature regrouping approach enhances the performance of feature

TABLE 6.3: Classification accuracies (%) by SVM for individual modalities of SDUMLA-HMT database.

<i>Modality</i>	<i>Classification Accuracy</i>
Left Iris	61.05
Right Iris	55.79
Left Thumb Fingerprint	76.84
Left Index Fingerprint	88.53
Right Thumb Fingerprint	81.05
Right Index Fingerprint	89.58

fusion. We found that fusion of regrouped features through IEDCA performed better than MSCCA and MDCA. The best classification accuracy is achieved when the concatenated features are partitioned into three groups.

6.3.3 Deep Feature Regrouping

In this experiment, we apply FR-NC on deep features generated by VGG-16 [10]. Feature fusion is conducted on the regrouped deep features. PASCAL VOC-2007 [121] dataset is used. PASCAL VOC-2007 contains about 10,000 images of twenty object classes which have been split into train, validation, and test sets. VGG is a well-known deep neural network which performs favourably to the state-of-the-art for classification in ImageNet Large-Scale Visual Recognition Challenge 2012 (ILSVRC). The activations of the last layer of the network are used as our features, resulting in 4096-dimensional feature vectors. For image preprocessing, since the raw images come in various sizes, we uniformed them by cropping them into 224×224 . The performance is evaluated by mean average precision (mAP) across all categories. FR-NC is applied to the extracted deep features. Different numbers of feature groups are extracted and fused. Various feature fusion algorithms are used as in the previous experiment. The fused features are fed into SoftMax for classification.

Table 6.5 displays the mAP of the fused feature through different fusion methods from different number of regrouped features, while Table 6.6 provides the per-class average precision (AP) as well as the overall mAP of VGG and different fusion methods on the VOC-2007 dataset for the fusion of regrouped features when the deep feature is partitioned into 30 groups. We can see that VGG achieved the mAP of 75.70% in Table 6.6. FR-NC enhances the performance of classification

TABLE 6.4: Classification accuracies (%) of multimodal fusion of regrouped features in SDUMLA-HMT database.

Method	NO. of Groups		Concatenated Feature Set from All Six Feature Sets						Original Six Feature Sets
	2	3	4	5	6				
Serial + PCA	-	-	-	-	-		90.53		
Serial + LDA	-	-	-	-	-		91.58		
PCA + CCA/MCCA	92.30	93.24	92.59	91.54	90.43		92.63		
LDA + CCA/MCCA	93.14	94.71	93.08	92.67	91.26		93.68		
MSCCA	94.93	96.52	95.84	94.04	93.21		95.79		
MDCA	95.11	98.28	96.43	94.87	93.46		97.89		
IEDCA	96.64	99.16	97.06	96.23	95.18		98.95		

TABLE 6.5: Mean average precision (%) for the fusion of regrouped deep features from VGG-16 in PASCAL VOC-2007 database.

<i>Method</i>	<i>NO. of Groups</i>	<i>10</i>	<i>20</i>	<i>30</i>	<i>40</i>	<i>50</i>
	<i>Deep Feature</i>					
PCA + CCA/MCCA		70.08	72.39	75.94	74.36	72.46
LDA + CCA/MCCA		73.13	75.21	77.08	76.31	74.87
MSCCA		74.64	76.52	78.64	78.95	76.98
MDCA		76.59	77.43	79.11	78.04	77.82
IEDCA		77.35	78.19	80.24	79.56	78.97

TABLE 6.6: Classification results (%) for regrouped deep feature fusion (NO. of groups =30) from VGG in VOC-2007 database.

<i>Class</i>	<i>Method</i>	<i>VGG</i>	<i>PCA+MCCA</i>	<i>LDA+MCCA</i>	<i>MSCCA</i>	<i>MDCA</i>	<i>IEDCA</i>
86.91	87.26	88.51	89.44	90.25	90.82
79.84	80.45	82.36	83.97	84.08	84.96
82.17	83.59	85.64	87.39	87.76	89.83
73.46	76.18	80.70	83.84	85.21	87.61
47.83	48.24	51.08	53.46	52.18	55.34
82.98	81.36	81.62	82.73	82.54	82.87
86.59	86.61	87.04	87.52	88.04	88.69
86.73	86.85	86.98	88.34	87.96	88.28
59.87	59.94	60.45	62.68	63.54	64.76
79.64	79.88	79.96	80.43	81.79	82.31
68.86	68.97	69.14	70.84	70.16	70.57
83.97	83.86	84.06	84.68	85.47	86.28
84.18	83.34	83.57	84.82	85.94	85.36
81.84	81.67	81.95	82.19	83.06	82.84
89.08	89.40	90.26	91.78	92.33	93.52
47.09	47.16	49.38	50.29	51.86	51.64
72.68	73.09	74.56	78.51	79.84	81.40
59.73	60.24	61.40	63.58	62.18	65.96
89.65	89.72	90.54	91.60	92.57	94.78
70.91	71.04	72.36	74.62	75.36	76.94
mAP		75.70	75.94	77.08	78.64	79.11	80.24

in terms of mAP. When the deep feature is partitioned into 30 feature groups for fusion, MCCA improved the mAP to 77.08% with LDA compared to using deep features directly with the mAP of 75.70%. The mAP can be further improved to 78.64%, 79.11% and 80.24% through MSCCA, MDCA and IEDCA respectively. For MSCCA, the mAP can reach 78.95% when we partition 40 groups of the deep feature for fusion. Our experiments show that the fusion of regrouped features can enhance the performance of the classification.

6.4 Conclusions

In this chapter, we propose a feature regrouping scheme for CCA-based feature fusion and extraction through normalized cut. It works by incorporating feature correlation into normalized cut. The proposed feature regrouping scheme for CCA-based feature fusion and extraction can maximize the intra-group correlation and minimize the extra-group correlation simultaneously. Moreover, the performance of the fused feature can be enhanced by fusing the regrouped features. Extensive experiments on different databases demonstrate the effectiveness of our proposed approach in the fusion of different features of single modality, multimodal feature fusion and fusion of regrouped deep features.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

In this thesis, we have proposed different correlation-based multi-view representation learning approaches to enhance learning performance. The challenge in this study is how to learn a good association among multiple views of data. Throughout Chapter 2, the related work for various correlation-based multi-view representation learning and deep representation learning models are reviewed. Our work presented in this thesis is inspired by these methods.

In Chapter 3, we explore correlation-based multi-view representation learning for pattern classification tasks. In contrast to traditional CCA-based feature representation learning approaches, which only consider one-to-one pairs-wise correlation, we consider both the pairs-wise correlation and the correlation across different features within the same class. In this way, class separability is enhanced, and the proposed correlation-based representation learning method can generate more discriminative representations. In correlation-based multi-view representation learning, the dimensionality of the representations from each view (each feature set) is usually first reduced using principal component analysis (PCA), linear discriminant analysis (LDA), etc. to ensure non-singularity and invertibility of covariance matrices. One issue with this standard CCA-based representation learning is that the reduced feature sets generated by PCA or LDA may neglect certain correlation information among different feature sets which is useful for correlation analysis, and this in turn may degrade the following classification performance. To address this

issue, our proposed correlation-based representation learning method can retain enough dimensions of feature for the correlation analysis and preserve the complete and effective discriminative structure underlying multi-view representations. In each feature set, class separability is enhanced by the discriminative analysis through the intra-class and extra-class nearest neighbour of each sample in the feature set. The effectiveness of our proposed method is demonstrated through both the pairs-wise correlation and the correlation across different features within the same class for improved pattern classification.

In Chapter 4, we further analyze the representations generated by correlation-based multi-view representation learning and find they may still contain redundancies due to the correlation criterion. These redundancies may be relevant or irrelevant to class labels. The irrelevant redundancies may degrade the pattern recognition performance, while the relevant redundancies can make the pattern recognition system more robust. To address this limitation, we propose an enhanced feature representation learning scheme through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis (IEDCA-IRE). IEDCA-IRE explores the intra-class correlation including both the pairs-wise correlation like CCA-based feature fusion approaches and the correlation across different features within the same class. By incorporating kernelized IEDCA into minimum Redundancy Maximum Relevance (mRMR) criterion, only the relevant redundancy is retained in the fused feature.

In Chapter 5, we review traditional CCA-based multi-view representation learning, which tries to find linear projections of two views of the data that are maximally correlated. Its kernel extension allows the learning of nonlinear representations. However, the learned representation by kernel CCA is limited by the fixed kernel. Moreover, the training time of kernel CCA scales poorly with the size of training data as it is a nonparametric method. Although CCA and the kernel extension can effectively model the relationship between two sets of variables, they are limited in dealing with high-level associations among multiple views of data. With the development of deep neural networks, we propose deep supervised multi-view canonical correlation analysis (DMCCA) to solve this problem, with a strategy of learning a joint representation which is associated among multiple views of data at a higher level after learning several layers of view-specific representations from lower layers. DMCCA tries to learn a shared complex nonlinear representation from multiple

TABLE 7.1: Classification accuracy (%) comparison of proposed methods over existing correlation-based multi-view representation methods on SDUMLA-HMT database.

<i>Method</i>	<i>Classification Accuracy</i>
Concatenation[135]	86.63
Concatenation + PCA[135]	89.94
Concatenation + LDA[135]	90.21
Parallel + PCA[136]	72.42
Parallel + LDA[136]	76.21
PCA + CCA[137]	92.25
LDA + CCA[137]	92.53
MCCA[25]	93.47
MSCCA[108]	94.16
MDCA[109]	94.42
DCCA[127]	92.53
MLP[138]	91.26
JSR[50]	93.69
SMDL [51]	93.71
IEDCA (proposed in Chapter 3)	95.58
IEDCA-IRE (proposed in Chapter 4)	96.74
DMCCA (proposed in Chapter 5)	96.23
FR-NC (proposed in Chapter 6)	96.64

views of data and simultaneously learn complex nonlinear transformations from each view to this shared space in which the resulting transformations are maximally informative of each other. It combines canonical correlation analysis from multiple views of data with the flexibility of deep representation learning. The label information is leveraged for learning better representations and embedded as a view into the whole networks at training stage. During training, the label information and the network structure of other views of data are jointly trained to update the parameters of the whole networks and maximize the total correlation simultaneously.

In Chapter 6, we propose a feature regrouping algorithm for CCA-based feature fusion and extraction through normalized cut (FR-NC). Feature correlation analysis is incorporated into normalized cut, in which the intra-group correlation is maximized, and the extra-group correlation is minimized simultaneously. CCA-based feature fusion is performed on the regrouped features. The proposed feature regrouping algorithm aims to provide enhanced fused features for pattern classification.

The main contribution of this thesis is to address the limitation of existing correlation-based multi-view representation learning methods. Three correlation-based multi-view representation learning models are proposed in Chapter 3, Chapter 4 and Chapter 5, respectively. Chapter 6 elaborates a feature regrouping algorithm for

correlation-based multi-view representation learning. The learning performance improved by these proposed methods over existing correlation-based multi-view representation methods is illustrated in Table 7.1. We summarize the main contributions of this thesis as follows:

- Traditional correlation-based multi-view representation learning approach may not preserve the complete and effective discriminative structure underlying multi-view representations. In Chapter 3, we propose intra-class and extra-class discriminative correlation analysis (IEDCA) to provide more discriminative and complete representations through correlation-based multi-view representation learning. The proposed IEDCA explores both the pairwise correlation like CCA-based feature fusion approaches and the correlation across different features within the same class.
- Most representation generated by correlation-based multi-view representation learning may still contain redundancies due to the correlation criterion. In Chapter 4, to explore the correlation underlying different views and jointly optimize the learned representation to remove the irrelevant redundancies brought by the correlation-based multi-view representation approach, we incorporate the proposed IEDCA into minimum Redundancy Maximum Relevance (mRMR) and further improve the learning performance.
- The nonlinear representation produced by kernelized correlation-based multi-view representation learning approach is limited by a fixed kernel. Moreover, the training time scales poorly with the size of the training data. In Chapter 5, we combine the flexibility of deep representation learning with the canonical correlation analysis from multiple views of data and propose deep supervised multi-view CCA (DMCCA). DMCCA can learn flexible nonlinear representations by passing multi-view data through multiple stacked layers of nonlinear transformation.
- In most correlation-based multi-view representation learning approach, the natural groupings of features are directly used. The complementary and mutual information existed in multi-view representations may not be fully utilized. In Chapter 6, we propose a feature regrouping approach for correlation-based multi-view representation learning to reduce the variances of learned representations and enhance the learning performance.

In summary, learning a good association through correlation-based multi-view representation learning approach can preserve the complete and effective discriminative structure underlying multi-view representations. Moreover, correlation-based multi-view representation (or feature) learning introduces one function to model a particular view and jointly optimizes all the functions to exploit the redundant views of the same input data to improve learning performance. This learning mechanism has a wide range of applications.

7.2 Future Works

The role of information fusion in pattern recognition is becoming more and more important. There still exists a lot of work to explore in multi-view representation learning. We plan to focus on the correlation-based multi-view representation learning methods. The future research work related to the work in this thesis and correlation-based multi-view representation learning are summarized in the following:

- 1 **What type of deep neural network model is suitable for correlation-based multi-view representation learning:** The deep architectures trained on large datasets can be used to a specific domain with multi-view measurements of an underlying data source, for example, cross-media retrieval. Due to the powerful representation ability with multiple levels of abstraction, deep learning-based multi-view representation learning has attracted much attention. From the perspective of neural network, multi-view representation learning learns the mid-level features from each view first, and integrates them into a single compact representation. Some of the representative examples are multi-view convolutional neural network [20], multi-modal recurrent neural network [21] and multi-modal autoencoder [22].
- 2 **What is the appropriate criterion for learning good correlation-based multi-view representation:** Canonical correlation analysis (CCA) is a representative technique in the studies of multi-view representation learning. It finds linear projections of two views of the data that are maximally

correlated. However, CCA-based multi-view representation learning encounters certain limitations. To learn a good association among multiple views of data, other correlation criteria may be further explored.

- 3 **How to deal with low-quality input data:** Correlation-based multi-view representation learning approaches may be affected by low-quality input data, for example, noisy and missing values. This challenge may further degrade the performance of multi-view representation learning methods. A good mechanism which can deal with low-quality input data can help correlation-based multi-view representation learning framework be more robust and immune to the interference of these noisy parts.

Appendix A

Analysis of IEDCA Fused Feature

Let $M_{xy} = C_{xx}^{-1}C_L C_{yy}^{-1}C_L^T$ and $M_{yx} = C_{yy}^{-1}C_L^T C_{xx}^{-1}C_L$, the optimization in Section 4.2.1 becomes:

$$\begin{cases} C_L C_{yy}^{-1} C_L^T W_x = \lambda^2 C_{xx} W_x \\ C_L^T C_{xx}^{-1} C_L W_y = \lambda^2 C_{yy} W_y \end{cases} \quad (\text{A.1})$$

M_{xy} and M_{yx} have the same nonzero eigenvalues, which satisfies $1 > \lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_s^2 > 0$, where $s = \text{rank}(C_L)$. Let:

$$\begin{cases} J_1 = C_{xx}^{-1/2} C_L C_{yy}^{-1} C_L^T C_{xx}^{-1/2} \\ J_2 = C_{yy}^{-1/2} C_L^T C_{xx}^{-1} C_L C_{yy}^{-1/2} \end{cases} \quad (\text{A.2})$$

Then we can obtain that M_{xy} and J_1 have the same nonzero eigenvalues. Similarly, M_{yx} and J_2 also have the same nonzero eigenvalues. Therefore, M_{xy} , M_{yx} , J_1 and J_2 share the nonzero eigenvalues $\lambda_1^2, \lambda_2^2, \dots, \lambda_s^2$. Let $F = C_{xx}^{-1/2} C_L C_{yy}^{-1/2}$, we can get:

$$\begin{cases} J_1 = F F^T \\ J_2 = F^T F \end{cases} \quad (\text{A.3})$$

By singular value decomposition (SVD), $F = \sum_{i=1}^s \lambda_i u_i v_i^T$, u_i and v_i ($i = 1, 2, \dots, s$) are the orthogonal eigenvectors of J_1 and J_2 corresponding to the nonzero eigenvalue λ_i^2 . Consider $W_x = [a_1, a_2, \dots, a_s]$, $W_y = [b_1, b_2, \dots, b_s]$, a_i and b_i indicate

the directions of the i -th pair of canonical projective vectors. Let $a_i = C_{xx}^{-1/2}u_i$, $b_i = C_{yy}^{-1/2}v_i$ ($i = 1, 2, \dots, s$). Since u_i is the eigenvector of J_1 corresponding to λ_i^2 , $J_1 u_i = \lambda_i^2 u_i$. It can be easily proved that a_i is the eigenvector of M_{xy} corresponding to λ_i^2 :

$$\left(C_{xx}^{-\frac{1}{2}}C_L C_{yy}^{-1}C_L^T C_{xx}^{-\frac{1}{2}}\right)C_{xx}^{\frac{1}{2}}a_i = \lambda_i^2 C_{xx}^{\frac{1}{2}}a_i \quad (\text{A.4})$$

$$\left(C_{xx}^{-1}C_L C_{yy}^{-1}C_L^T\right)a_i = \lambda_i^2 a_i \quad (\text{A.5})$$

$$M_{xy}a_i = \lambda_i^2 a_i \quad (\text{A.6})$$

Similarly, b_i is the eigenvector of M_{yx} corresponding to λ_i^2 . From SVD, u_i , u_j , v_i and v_j are orthogonal unit vectors,

$$u_i^T u_j = a_i^T C_{xx}^{1/2} C_{xx}^{1/2} a_j = a_i^T C_{xx} a_j = 0 \quad (\text{A.7})$$

$$v_i^T v_j = b_i^T C_{yy}^{1/2} C_{yy}^{1/2} b_j = b_i^T C_{yy} b_j = 0 \quad (\text{A.8})$$

Since u_i is the eigenvector of J_1 corresponding to λ_i^2 , $J_1 = FF^T$

$$FF^T u_i = \lambda_i^2 u_i \quad (\text{A.9})$$

$$(F^T F)(F^T u_i) = \lambda_i^2 (F^T u_i) \quad (\text{A.10})$$

This indicates $F^T u_i$ is the eigenvector of J_2 corresponding to λ_i^2 . Since $u_i^T u_i = v_i^T v_i$, $v_i = F^T u_i / \lambda_i$,

$$a_i^T C_L b_j = u_i^T C_{xx}^{-1/2} C_L C_{yy}^{-1/2} v_j = u_i^T F F^T u_j / \lambda_j = \lambda_j u_i^T u_j = 0 \quad (\text{A.11})$$

Similarly, $a_i^T C_L b_i = \lambda_i$. In summary:

$$a_i^T C_{xx} a_j = 0 \quad (\text{A.12})$$

$$b_i^T C_{yy} b_j = 0 \quad (\text{A.13})$$

$$a_i^T C_L b_j = 0 \quad (\text{A.14})$$

$$a_i^T C_L b_i = \lambda_i \quad (\text{A.15})$$

Since the correlation criterion is to maximize the correlation between each pair through the calculation of λ_i , the mutual information in each pair may contain both the relevance and non-relevance redundancy. Let $U = W_x^T X' H = (x_1^*, x_2^*, \dots, x_s^*)^T$, $V = W_y^T Y' H = (y_1^*, y_2^*, \dots, y_s^*)^T$, it can be proved that different pairs of fused features are uncorrelated:

$$\text{cov}(x_i^*, x_j^*) = a_i^T C_{xx} a_j = 0 \quad (\text{A.16})$$

$$\text{cov}(y_i^*, y_j^*) = b_i^T C_{yy} b_j = 0 \quad (\text{A.17})$$

$$\text{cov}(x_i^*, y_j^*) = a_i^T C_L b_j = 0 \quad (\text{A.18})$$

$$\text{cov}(x_i^* + y_i^*, x_j^* + y_j^*) = \text{cov}(x_i^*, x_j^*) + 2\text{cov}(x_i^*, y_j^*) + \text{cov}(y_i^*, y_j^*) = 0 \quad (\text{A.19})$$

Author's Publications

- **Zuobin Wu**, Kezhi Mao, and Gee-Wah Ng, “Deep Supervised Multi-View Canonical Correlation Analysis,” *in preparation*.
- **Zuobin Wu**, Kezhi Mao, and Gee-Wah Ng, “Enhanced feature fusion through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis,” *Neurocomputing*, vol. 335, pp. 105-118, 2019.
- **Zuobin Wu**, Kezhi Mao, and Gee-Wah Ng, “Feature regrouping for CCA-based feature fusion and extraction through normalized cut,” *2018 21st International Conference on Information Fusion (Fusion)*, pp. 2275-2282, IEEE, 2018.
- **Zuobin Wu**, Kezhi Mao, and Gee-Wah Ng, “Effective feature fusion for pattern classification based on intra-class and extra-class discriminative correlation analysis,” *2017 20th International Conference on Information Fusion (FUSION)*, pp. 1-8, IEEE, 2017.
- Nikhil Vij, **Zuobin Wu**, Malin Björnsdotter, Justin Dauwels, and François Benoît Vialatte, “A multimodal approach to analysis of steady state visually evoked potentials,” *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 183-188, IEEE, 2014.

Bibliography

- [1] Didier Dubois, Weiru Liu, Jianbing Ma, and Henri Prade. The basic principles of uncertain information fusion. an organised review of merging rules in different representation frameworks. *Information Fusion*, 32:12–39, 2016.
- [2] Gabriel Villarrubia, Juan F De Paz, Dechen Pelki, Fernando de la Prieta, and Sigeru Omatu. Virtual organization with fusion knowledge in odor classification. *Neurocomputing*, 231:3–10, 2017.
- [3] Jiao Du, Weisheng Li, Ke Lu, and Bin Xiao. An overview of multi-modal medical image fusion. *Neurocomputing*, 215:3–20, 2016.
- [4] Xiaochun Zhang and Chuancai Liu. Image annotation based on feature fusion and semantic similarity. *Neurocomputing*, 149:1658–1671, 2015.
- [5] Behzad Mirmahboub, Mohamed Lamine Mekhalfi, and Vittorio Murino. Person re-identification by order-induced metric fusion. *Neurocomputing*, 275: 667–676, 2018.
- [6] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.
- [7] Tara N Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel-rahman Mohamed, George Dahl, and Bhuvana Ramabhadran. Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 64:39–48, 2015.
- [8] Hong Liang, Xiao Sun, Yunlei Sun, and Yuan Gao. Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking*, 2017(1):1–12, 2017.

- [9] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2): 645–657, 2017.
- [10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep neural networks. In *NIPS*, pages 1097–1105, 2012.
- [13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [14] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, 17(11):2049–2058, 2015.
- [15] Xu Liu, Licheng Jiao, Jiaqi Zhao, Jin Zhao, Dan Zhang, Fang Liu, Shuyuan Yang, and Xu Tang. Deep multiple instance learning-based spatial-spectral classification for pan and ms imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):461–473, 2017.
- [16] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.
- [17] Gang Ma, Xi Yang, Bo Zhang, and Zhongzhi Shi. Multi-feature fusion deep networks. *Neurocomputing*, 218:164–171, 2016.
- [18] Pengjie Tang, Hanli Wang, and Sam Kwong. G-ms2f: Googlenet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing*, 225: 188–197, 2017.

- [19] Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing*, 281:78–85, 2018.
- [20] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941. IEEE, 2016.
- [21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137. IEEE, 2015.
- [22] Cesar Cadena, Anthony R Dick, and Ian D Reid. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and Systems*, volume 5, pages 1–9, 2016.
- [23] Guangfeng Lin, Hong Zhu, Xiaobing Kang, Caixia Fan, and Erhu Zhang. Feature structure fusion and its application. *Information Fusion*, 20:146–154, 2014.
- [24] Yichao Lu and Dean P Foster. Large scale canonical correlation analysis with iterative least squares. In *Advances in Neural Information Processing Systems*, pages 91–99, 2014.
- [25] Weifeng Liu, Xinghao Yang, Dapeng Tao, Jun Cheng, and Yuanyan Tang. Multiview dimension reduction via hessian multiset canonical correlations. *Information Fusion*, 41:119–128, 2018.
- [26] Siamak Mehrkanoon and Johan AK Suykens. Regularized semipaired kernel CCA for domain adaptation. *IEEE transactions on neural networks and learning systems*, 29(7):3199–3213, 2017.
- [27] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.
- [28] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.
- [29] Allan Aasbjerg Nielsen. Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data. *Science*, 21(4):555–567, 1994.

- [30] Nicholas Asendorf and Raj Rao Nadakuditi. Improving multiset canonical correlation analysis in high dimensional sample deficient settings. In *Signals, Systems and Computers, 2015 49th Asilomar Conference on*, pages 112–116. IEEE, 2015.
- [31] Wu Zuobin, Mao Kezhi, and Gee-Wah Ng. Effective feature fusion for pattern classification based on intra-class and extra-class discriminative correlation analysis. In *2017 20th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2017.
- [32] Zuobin Wu, Kezhi Mao, and Gee-Wah Ng. Enhanced feature fusion through irrelevant redundancy elimination in intra-class and extra-class discriminative correlation analysis. *Neurocomputing*, 335:105–118, 2019.
- [33] Zuobin Wu, Kezhi Mao, and Gee-Wah Ng. Deep supervised multi-view canonical correlation analysis. *In Preparation*, 2020.
- [34] Wu Zuobin, Mao Kezhi, and Gee-Wah Ng. Feature regrouping for CCA-based feature fusion and extraction through normalized cut. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 2275–2282. IEEE, 2018.
- [35] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260, 2010.
- [36] Liang Sun, Shuiwang Ji, and Jieping Ye. A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1024–1031, 2008.
- [37] Dean P Foster, Sham M Kakade, and Tong Zhang. Multi-view dimensionality reduction via canonical correlation analysis. *Toyota Technical Institute-Chicago*, Retrieved from https://repository.upenn.edu/statistics_papers/150, 2008.
- [38] Liang Sun, Betul Ceran, and Jieping Ye. A scalable two-stage approach for a class of dimensionality reduction techniques. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 313–322, 2010.

- [39] Matthew B Blaschko and Christoph H Lampert. Correlational spectral clustering. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [40] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136, 2009.
- [41] Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.
- [42] Brian McWilliams, David Balduzzi, and Joachim M Buhmann. Correlated random features for fast semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 440–448, 2013.
- [43] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. Multi-view learning of word embeddings via CCA. In *Advances in neural information processing systems*, pages 199–207, 2011.
- [44] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.
- [45] Yi Zhang and Jeff Schneider. Multi-label output codes using canonical correlation analysis. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 873–882, 2011.
- [46] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
- [47] Gene H Golub and Hongyuan Zha. The canonical correlations of matrix pairs and their numerical computation. In *Linear algebra for signal processing*, pages 27–49. Springer, 1995.
- [48] Haim Avron, Christos Boutsidis, Sivan Toledo, and Anastasios Zouzias. Efficient dimensionality reduction for canonical correlation analysis. In *International Conference on Machine Learning*, pages 347–355, 2013.

- [49] Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(1):115–126, 2011.
- [50] Sumit Shekhar, Vishal M Patel, Nasser M Nasrabadi, and Rama Chellappa. Joint sparse representation for robust multimodal biometrics recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):113–126, 2014.
- [51] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and William Kenneth Jenkins. Multimodal task-driven dictionary learning for image classification. *IEEE Transactions on Image Processing*, 25(1):24–38, 2016.
- [52] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [53] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.
- [54] Alexandre d’Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48, 2005.
- [55] David A Torres, Douglas Turnbull, Bharath K Sriperumbudur, Luke Barrington, and Gert RG Lanckriet. Finding musically meaningful words by sparse CCA. In *Neural information processing systems (NIPS) workshop on music, the brain and cognition*, 2007.
- [56] Ami Wiesel, Mark Kliger, and Alfred O Hero III. A greedy approach to sparse canonical correlation analysis. *arXiv preprint arXiv:0801.2748*, 2008.
- [57] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.
- [58] Xi Chen, Liu Han, and Jaime Carbonell. Structured sparse canonical correlation analysis. In *Artificial Intelligence and Statistics*, pages 199–207, 2012.

- [59] Bernardo B Gatto and Eulanda M dos Santos. Discriminative canonical correlation analysis network for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4487–4491. IEEE, 2017.
- [60] Tingkai Sun, Songcan Chen, Jingyu Yang, Xuelei Hu, and Pengfei Shi. Discriminative canonical correlation analysis with missing samples. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 6, pages 95–99. IEEE, 2009.
- [61] Yong Shin and Cheong Park. Analysis of correlation based dimension reduction methods. *International Journal of Applied Mathematics and Computer Science*, 21(3):549–558, 2011.
- [62] Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23(7-8):2031–2038, 2013.
- [63] Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.
- [64] Raman Arora and Karen Livescu. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7135–7139. IEEE, 2013.
- [65] Anusua Trivedi, Piyush Rai, Scott L DuVall, and Hal Daumé III. Exploiting tag and word correlations for improved webpage clustering. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 3–12, 2010.
- [66] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, pages 966–973. IEEE, 2010.
- [67] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.
- [68] Malte Kuss and Thore Graepel. The geometry of kernel canonical correlation analysis (108). *Max Planck Institute for Biological Cybernetics*, 2003.

- [69] Kenji Fukumizu, Francis R Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383, 2007.
- [70] David R Hardoon and John Shawe-Taylor. Convergence analysis of kernel canonical correlation analysis: theory and practice. *Machine learning*, 74(1): 23–38, 2009.
- [71] Jia Cai and HongWei Sun. Convergence rate of kernel canonical correlation analysis. *Science China Mathematics*, 54(10):2161, 2011.
- [72] Raman Arora and Karen Livescu. Kernel CCA for multi-view learning of acoustic features using articulatory measurements. In *Symposium on Machine Learning in Speech and Language Processing*, pages 34–37, 2012.
- [73] Matthew Brand. Incremental singular value decomposition of uncertain data with missing values. In *European Conference on Computer Vision*, pages 707–720. Springer, 2002.
- [74] Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.
- [75] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in neural information processing systems*, pages 476–484, 2012.
- [76] Quoc Le, Tamás Sarlós, and Alexander Smola. Fastfood-computing hilbert space expansions in loglinear time. In *International Conference on Machine Learning*, pages 244–252, 2013.
- [77] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.
- [78] Suzanna Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in neural systems*, 7(1):7–31, 1996.
- [79] Pei Ling Lai and Colin Fyfe. Canonical correlation analysis using artificial neural networks. In *ESANN*, pages 363–368. Citeseer, 1998.

- [80] Pei Ling Lai and Colin Fyfe. A neural implementation of canonical correlation analysis. *Neural Networks*, 12(10):1391–1397, 1999.
- [81] William W Hsieh. Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, 13(10):1095–1105, 2000.
- [82] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.
- [83] Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, 2015.
- [84] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450. IEEE, 2015.
- [85] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.
- [86] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [87] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
- [88] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.
- [89] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.
- [90] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Multiview matrix completion for multilabel image classification. *IEEE Transactions on Image Processing*, 24(8):2355–2368, 2015.

- [91] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible CNN framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2016.
- [92] Yukhe Lavinia, Holly H Vo, and Abhishek Verma. Fusion based deep CNN for improved large-scale image action recognition. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 609–614. IEEE, 2016.
- [93] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788. IEEE, 2016.
- [94] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.
- [95] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.
- [96] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988. IEEE, 2017.
- [97] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.
- [98] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.
- [99] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*, pages 1405–1416. BMVA Press, 2014.
- [100] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9. IEEE, 2015.
- [101] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NIPS*, pages 3856–3866, 2017.

- [102] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708. IEEE, 2017.
- [103] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724. IEEE, 2014.
- [104] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014.
- [105] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.
- [106] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. DeViSE: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.
- [107] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005.
- [108] Yunhao Yuan, Peng Lu, Zhiyong Xiao, Jianjun Liu, and Xiaojun Wu. A novel supervised CCA algorithm for multiview data representation and recognition. In *Chinese Conference on Biometric Recognition*, pages 702–709. Springer, 2015.
- [109] Mohammad Haghighat, Mohamed Abdel-Mottaleb, and Wade Alhalabi. Discriminant correlation analysis: Real-time feature level fusion for multi-modal biometric recognition. *IEEE Transactions on Information Forensics and Security*, 11(9):1984–1996, 2016.
- [110] Kezhi Z Mao. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):629–634, 2004.

- [111] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.
- [112] Yilong Yin, Lili Liu, and Xiwei Sun. Sdumla-hmt: a multimodal biometric database. In *Chinese Conference on Biometric Recognition*, pages 260–268. Springer, 2011.
- [113] Liya Ding and Aleix M Martinez. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2022–2038, 2010.
- [114] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5-6):237–329, 2007.
- [115] Sharat Chikkerur, Chaohang Wu, and Venu Govindaraju. A systematic approach for feature extraction in fingerprint images. In *Biometric Authentication*, pages 344–350. Springer, 2004.
- [116] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- [117] Yoshihiro Kanno and Yakov Ben-Haim. Redundancy and robustness, or when is redundancy redundant? *Journal of Structural Engineering*, 137(9):935–945, 2011.
- [118] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.
- [119] Yunhui He. An efficient method to solve small sample size problem of non-linear discriminant vectors in feature space for face recognition. In *Computational and Information Sciences (ICCIS), 2011 International Conference on*, pages 117–120. IEEE, 2011.
- [120] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

- [121] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [122] Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31, 2017.
- [123] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, 2014.
- [124] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.
- [125] Zheng Zhao, Lei Wang, Huan Liu, and Jieping Ye. On similarity preserving feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):619–632, 2013.
- [126] Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook, nov 2012. URL <http://www2.imm.dtu.dk/pubdb/p.php>, 3274, 2012.
- [127] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594. IEEE, 2015.
- [128] Raman Arora and Karen Livescu. Multi-view learning with supervision for transformed bottleneck features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2499–2503. IEEE, 2014.
- [129] Bruno Matarazzo Dur, Jonas Mendon, Clodoaldo Aparecido de Moraes Lima, et al. Biometric recognition based on fingerprint: A comparative study. In *2017 Workshop of Computer Vision (WVC)*, pages 84–89. IEEE.

- [130] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- [131] Lei Han and Yu Zhang. Discriminative feature grouping. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2631–2637, 2015.
- [132] Zhenyu Wu and Richard Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1101–1113, 1993.
- [133] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.
- [134] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016.
- [135] S Khellat-Kihel, Reza Abrishambaf, João L Monteiro, and M Benyettou. Multimodal fusion of the finger vein, fingerprint and the finger-knuckle-print using kernel fisher analysis. *Applied Soft Computing*, 42:439–447, 2016.
- [136] Jian Yang, Jing-yu Yang, David Zhang, and Jian-feng Lu. Feature fusion: parallel strategy vs. serial strategy. *Pattern recognition*, 36(6):1369–1381, 2003.
- [137] Yun-Hao Yuan, Quan-Sen Sun, Qiang Zhou, and De-Shen Xia. A novel multisets integrated canonical correlation analysis framework and its application in feature fusion. *Pattern Recognition*, 44(5):1031–1040, 2011.
- [138] Appaji M Abhishek and HN Suma. Comparison of biosignal electrocardiography (ecg), fingerprint and iris as biometrics authentication tool. In *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES)*, pages 303–308. IEEE, 2014.