A DISCRIMINATIVE ANALYSIS FRAMEWORK FOR MULTI-MODAL INFORMATION FUSION

by

Lei Gao

Master of Information and Communication Systems, Zhengzhou University, 2011

Bachelor of Physical Engineering, Zhengzhou University, 2007

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2017

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

_____

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

_____

A DISCRIMINATIVE ANALYSIS FRAMEWORK FOR MULTI-MODAL INFORMATION FUSION

Doctor of Philosophy, 2017

Lei Gao

Electrical and Computer Engineering

Ryerson University

Abstract

Since multi-modal data contain rich information about the semantics presented in the sensory and media data, valid interpretation and integration of multi-modal information is recognized as a central issue for the successful utilization of multimedia in a wide range of applications. Thus, multi-modal information analysis is becoming an increasingly important research topic in the multimedia community. However, the effective integration of multi-modal information is a difficult problem, facing major challenges in the identification and extraction of complementary and discriminatory features, and the impactful fusion of information from multiple channels. In order to address the challenges, in this thesis, we propose a discriminative analysis framework (DAF) for high performance multi-modal information fusion.

The proposed framework has two realizations. We first introduce Discriminative Multiple Canonical Correlation Analysis (DMCCA) as the fusion component of the framework. DMCCA is capable of extracting more discriminative characteristics from multi-modal information. We demonstrate that optimal performance by DMCCA can be analytically and graphically verified, and Canonical Correlation Analysis (CCA), Multiple Canonical Correlation Analysis (MCCA) and Discriminative Canonical Correlation Analysis (DCCA) are special cases of DMCCA, thus establishing a unified framework for canonical correlation analysis.

To further enhance the performance of discriminative analysis in multi-modal information fusion, Kernel Entropy Component Analysis (KECA) is brought in to analyze the projected vectors in DMCCA space, and thus forming the second realization of the framework. By doing so, not only the discriminative relation is considered in DMCCA space, but also the inherent complementary representation of the input data is revealed by entropy estimation, leading to better utilization of the multi-modal information and better pattern recognition performance.

Finally, we implement a prototype of the proposed DAF to demonstrate its performance in handwritten digit recognition, face recognition and human emotion recognition. Extensive experiments show that the proposed framework outperforms the existing methods based on similar principles, clearly demonstrating the generic nature of the framework. Furthermore, this work offers a promising direction to design advanced multi-modal

information fusion systems with great potential to impact the development of intelligent human computer interaction systems.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisor, Prof. Ling Guan, for his continuous patience, guidance and motivation throughout my Ph.D. study and thesis writing. His support helped me through my research. He encouraged me to not only grow as a researcher but also as an independent thinker. I appreciate all his contributions of time and ideas to make my research experience both delightful and productive. I could not have imagined having a better advisor and supervisor for my Ph.D. study. His dedication, diligence, and enthusiasm will continue motivating me in my future career.

In addition, I am very grateful to Prof. Lin Qi and Prof. Enqing Chen for their insightful suggestions, valuable discussions, and great attention to details. I would also like to thank Dr. Baining Guo and Dr. Wenjun Zeng, for offering an internship at Microsoft Research Asia in 2016. I have benefited immensely from their mentorship.

I would like to thank the Electrical and Computer Engineering Department of Ryerson University for providing a very well equipped and technically supported Ryerson Multimedia Research Laboratory (RML). My thanks are due to the School of Graduate Studies of Ryerson University, NSERC and Canada Research Chair Program for providing financial support. And I am very thankful to my thesis committee members for their invaluable advice on the dissertation.

Being a member of Ryerson Multimedia Research Laboratory (RML), I express my appreciation to all members of this laboratory, Dr. Yifeng He, Dr. Yongjin Wang, Dr. Yun Tie, Dr. Rui Zhang, Dr. Ning Zhang, Dr. Xiaoming Nan, Dr. Naimul Mefraz Khan, Chengwu Liang, Ziyang Zhang, Dong Nan, Fei Guo, Li Fang, Kevin Tang and Jian Li. I would also like to thank the many people with whom I have collaborated in my research work during the past few years.

Finally, I would also like to thank my family, including my parents, my parents-in-law, for their consistent support and encouragement throughout these years. I especially would like to thank my wife Shujun Wei for her complete patience and love.

Table of Contents

List of Tables

List of Figures

List of Acronyms

HCI	Human Computer Interaction
HCC	Human Computer Communication
DAF	Discriminative Analysis Framework
DMCCA	Discriminative Multiple Canonical Correlation Analysis
KECA	Kernel Entropy Component Analysis
CCA	Canonical Correlation Analysis
DCCA	Discriminative Canonical Correlation Analysis
MCCA	Multiple Canonical Correlation Analysis
ITL	Information Theoretic Learning
LDPv	Local Directional Pattern Variance
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
AAM	Active Appearance Model
KCCA	Kernel canonical correlation analysis
GIS	Geographical Information Systems
SIMO	Single-input and Multiple-output
MSE	Mean Square Error
MEE	Minimum Error Entropy
KPCA	Kernel Principal Component Analysis
NLP	Natural Language Processing

GMA	Generalized Multi-view Analysis
MDA	Multi-view Discriminant Analysis
CFA	Cross-Modal Factor Analysis
FLDA	Fisher Linear Discriminant Analysis
KLDA	Kernel Linear Discriminant Analysis
PDF	Probability Density Function
MNIST	Mixed National Institute of Standards and Technology
RML	Ryerson Multimedia Lab
HOG	Histogram of Oriented Gradient
LBP	Local Binary Patterns
MFCC	Mel-frequency Cepstral Coefficient
FF	Formant Frequency

Chapter 1

Introduction

1.1 Motivation

Nowadays, with the development of science and technology, especially sensory technologies, human can obtain more and more data easily. How to analyze or process these massive data to gain useful information is becoming a challenging but necessary research topic. Moreover, in many fields of studies, information about a given phenomenon is obtained through different types of acquisition techniques and multiple sources, and the availability of such multi-modal data has been growing with extremely fast pace [1]. Therefore, effective utilization and integration of the contents across multiple distinct yet complementary sources of information for improving multimedia analysis and pattern recognition performance is becoming an increasingly important research topic in information science [2]. Due to the rich characteristics of natural processes and environments, and technological constraints, it is rare that a single modality provides complete understanding thereof. Thus, unimodal based pattern analysis and recognition systems usually affords low level of performance due to

the drastic variation and noisy nature of the acquired signals, which leads to insufficient and inaccurate pattern representation of the perception of interest [3].

The increasing availability of multiple data sets that contain information, obtained using different acquisition methods from the same system, introduces new degrees of freedom that raise questions beyond those related to analyzing each data set separately. Joint analysis of multiple data sets has since been the topic of extensive research, and leaped significantly forward in the late 1960s/early 1970s with the formulation of concepts and techniques of data fusion [4, 5]. However, until rather recently, these data fusion methodologies were largely confined within the limits of psychometrics and chemometrics, the communities in which they evolved. With recent technological advancement, the availability of data sets that correspond to the same phenomenon has increased, leading to the development of multi-modal information fusion. Multi-modal data are associated with high-impact commercial, social, biomedical, environmental, military applications. Thus, the drive to develop new and efficient analytical methodologies is high and reaches far beyond pure academic interest.

In general, natural integration of multiple media, their associated features, or the intermediate decisions to perform an analysis task is referred to as multi-modal fusion [6]. Multi-modal data contain a combination of information content from different sources in various presentation formats. The combination of multi-modal data may potentially provide a more complete and discriminatory description of the intrinsic characteristics of the patterns, and produce improved system performance than using a sin-

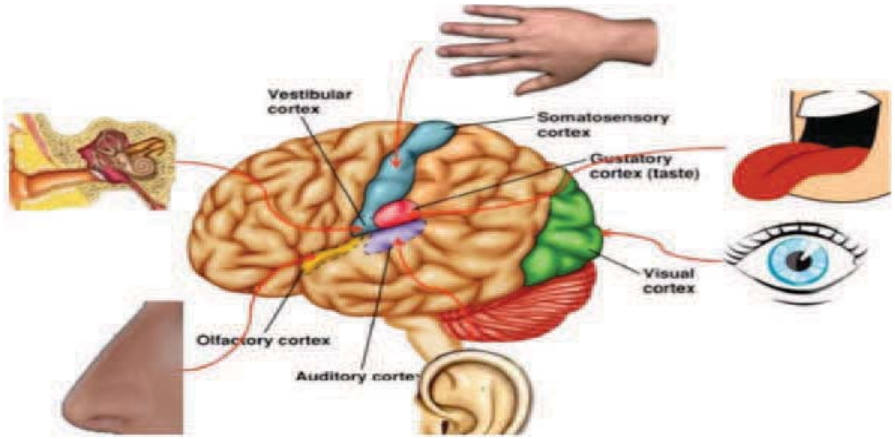

Figure 1.1: The natural multi-modal fusion system-Human Brain

gle modality only [7]. As we know, the human brain is arguably the best natural fusion system which collects information from different sensory modalities, such as sight, sound, touch, smell, self-motion, and taste, etc. to gain meaningful perceptual experiences shown in Figure 1.1 [3].

Generally speaking, an information fusion analysis task involves processing of multi-modal data to obtain valuable insights about the data, a situation, or a high level activity. Examples of information fusion analysis tasks include semantic concept detection, face recognition, audiovisual emotion recognition, human tracking, event detection, etc. Multimedia data used for these tasks could be sensory (such as audio, video, RFID) or non-sensory (such as WWW resources, database). The fusion of multiple modalities can provide complementary information and increase the accuracy of the overall decision making process. For instance, fusion of audio-visual features along with other textual information have become more effective in detecting events from a team sports video [8], which would otherwise not be possible by using a single information source.

Multi-modal information fusion is also of great importance for human computer interaction (HCI), human-computer communication (HCC), security/surveillance and many other areas. Specially, we take human computer interaction as an example. Since the conventional human computer interfaces [9, 10, 11] are considered too restrictive for natural interaction between human and computer, a great deal of efforts have been spent on numerous non-intrusive sensors so that users can conduct their activities in a more natural way without feeling the presence of these sensors. The intention of the users can be inferred from many data sources including voice, facial expression, gesture, and so on. This necessitates the employment of multi-modality data [12].

Motivations for multi-modal information fusion are many. They include obtaining a more unified picture and global view of the system at hand; improving decision making; exploratory research; answering specific questions about the system, such as identifying common versus distinctive elements across modalities or time; and in general, extracting knowledge from data for recognition. However, although massive work has already been done in the related field (see, for example, [13, 14, 15, 16, 17, 18] and references therein), the knowledge of how to actually exploit the additional diversity that multiple data sets offer is still at its very preliminary stage. For example, although multi-modal data can potentially provide a more complete and discriminatory description of the intrinsic characteristics of the pattern, multiple types of data may carry redundant, or even contradictory information. Hence, utilizing useful data and eliminating conflict information based on effective fusion algorithms become another increasingly essential research topic in information fusion.

In addition, multi-modal information fusion is at its preliminary stage for several reasons [19]. First, the data are generated by very complex systems: biological, environmental, sociological, and psychological, to name a few, driven by numerous underlying processes that depend on a large number of variables to which we have no access. Second, due to the augmented diversity, the number, type, and scope of new research questions that can be posed is potentially very large. Third, working with heterogeneous data sets such that the respective advantages of each data set are maximally exploited, and drawbacks suppressed, is not a task clearly defined.

Considering the above mentioned issues, an effective multi-modal information fusion scheme is urgently needed. Therefore, in this thesis, we propose a novel discriminative analysis framework (DAF) which is able to more effectively utilize complementary and discriminative information, eliminate redundancy and improve the overall recognition performance.

1.2 Objective

Research in multi-modal information fusion has achieved substantial advances, especially in recent years. Nevertheless, perfectly emulating the information fusion capacity of the human brain is still far from accomplished. Ideally, the fusion method should be capable of taking full advantage of information collected from multiple sources and bearing a better description of the intended perception.

This thesis proposes a novel DAF for multi-modal information fusion. The framework has two realizations. First, discriminative multiple canonical correlation analysis (DMCCA) is introduced as the fusion function of

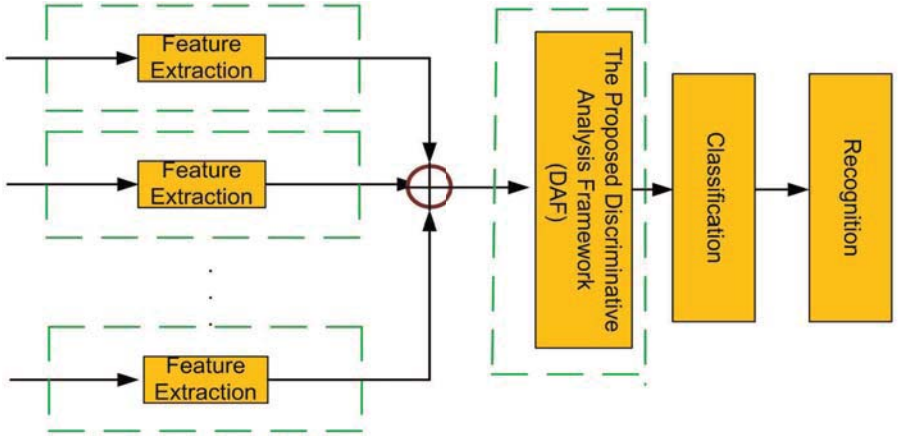

Figure 1.2: The general block diagram of the proposed DAF for multi-modal information fusion.

the framework to extract the discriminative representations from multiple data/information sources for multi-modal analysis and fusion. Then kernel entropy component analysis (KECA) is brought in to improve the performance of the DMCCA-based fusion function. A general block diagram of the proposed DAF is depicted in Figure 1.2. The circled areas in Figure 1.2 indicate fusing different features together. To achieve this objective, we need to address a number of challenges, which will be presented in the following section.

1.3 Challenges

The motivation for studying better information fusion techniques is to obtain a more reliable analysis and accurate recognition performance. However, the benefits usually come with a certain price, and to accomplish the task better, the challenges resulted from the analysis process have to be

properly addressed [17, 20]. Although many attempts have been made to improve information fusion techniques, it is still a very challenging field due to several reasons. The majority of these reasons arise from the data to be fused, imperfection and diversity of the sensory technologies, and the nature of the application environment which are summarized below:

* **Data imperfection** : data provided by sensors is always affected by some level of impreciseness as well as uncertainty in the measurements.

* **Outliers and spurious data** : the uncertainties in sensors may also come from the ambiguities and inconsistencies present in the environment [21].

* **Conflicting data** : fusion of such data can be problematic especially when the fusion system is based on evidential belief reasoning and Dempster's rule of combination [22].

* **Data modality** : sensor networks may collect the qualitatively similar (homogeneous) or different (heterogeneous) data such as auditory, visual, and tactile measurements of a phenomenon.

* **Data correlation** : this issue is particularly important and common in distributed fusion settings, e.g., wireless sensor networks where some sensor nodes are likely to be exposed to the same external noise biasing their measurements.

* **Data alignment/registration** : sensor data must be transformed from each sensor's local frame into a common frame before fusion occurs. Such an alignment problem is often referred to as data registration. Data registration is of critical importance to the successful deployment of fusion systems in practice.

* **Data dimensionality** : the measurement data could be pre-processed,

either locally at each of the sensor nodes or globally at the fusion center to be compressed into lower dimensional data, assuming a certain level of compression loss is allowed.

While many of the aforementioned problems have been identified and actively investigated, no existing information fusion algorithm is capable of addressing all these problems. In this thesis, the following challenges will be addressed.

1. First, although intuition indicates that fusion of multi-modal data should help in many information processing tasks, it is not necessarily always true. The major difficulties lie in the identification of the inherent relationship between different modalities, and the design of a fusion strategy that can effectively utilize the complementary information presented in different channels.

2. It is important for a fusion method to be able to identify the discriminatory representation amongst different modalities. Most existing methods only reveal the relation among different modalities while ignoring the discriminative relation among different classes [23]. In addition, multi-modal data may carry redundant or even contradictory information. It is necessary to extract more discriminatory description of the intrinsic characteristics from the multi-modal data. DMCCA is selected to fulfill this purpose in this work.

3. Finally, although there are numerous methods proposed for information fusion, the theoretical foundation of these methods largely depend on the second order statistics, such as variance, correlation, mean square error and so on. Since the second order statistics are only optimal for Gaussian-like distribution [24] and sensitive to the choice of input parameters [25],

a poor estimator is likely obtained if the underlining distribution greatly differs from Gaussian and possesses a large number of parameters. KECA was proposed to solve this problem [26]. It utilizes descriptor of entropy estimation to extract a more complementary representation of input data other than the second order statistics, leading to improved performance. However, unsupervised in nature, it only puts the information or data from different channels together without considering the discriminatory representation of the input information or data sources.

Although many investigations have been carried out to address these challenges, the performance of information fusion systems are still far from satisfactory. To obtain good fusion performance, this thesis focuses on developing methodology to properly treat the three issues.

1.4 Main Contributions

In this thesis, we propose a DAF which attempts to utilize the discriminant analysis and entropy-estimation of multi-modal data to enhance multi-modal information fusion. It is driven by several machine learning fundamentals, both supervised and unsupervised. For the supervised, benefiting from the discriminative power of DMCCA, we learn and select the discriminative representation among original multi-modal data. On the other hand, the unsupervised is used to extract the complementary information based on entropy-estimation from the extracted discriminative representation. Therefore, the proposed DAF is able to achieve the tasks of the identification and extraction of complementary and discriminatory representation simultaneously. The contributions are summarized below:

1. We present the DMCCA as the first realization of the DAF for

multi-modal information fusion, extracting the discriminative representation from original multi-modal data effectively. Furthermore, we mathematically verify that the best performance by DMCCA achieves when the number of projected dimensions is smaller than or equals to c, the number of the classes being studied. Based on this property, we may just start at dimension c and then perform a localized search around c to find the best performance, leading to significant reduction in computational cost. This is a particularly attractive feature when dealing with large scale problems.

2. We further verify that canonical correlation analysis (CCA), multiple canonical correlation analysis (MCCA) and discriminative canonical correlation analysis (DCCA) are special cases of DMCCA, thus establishing a unified framework for canonical correlation analysis.

3. We then propose KECA plus DMCCA (KECA+DMCCA) as another realization of the DAF for multi-modal information fusion. By combining the entropy-estimation property of KECA and the discriminative power of DMCCA, KECA+DMCCA transforms the original multiple information into the discriminative multiple canonical correlation analysis space to reveal discriminative representation and eliminate redundant information among different multiple variables. Then KECA is applied to the projected vectors in the DMCCA space. By doing so, not only the discriminative representations are considered, but also the complementary relationship of the input data is revealed by KECA, improving the recognition/classification accuracy. Moreover, we mathematically verify that the optimal performance by KECA+DMCCA achieves with c (the number of the classes being studied) independently projected vectors. It is a particularly attractive property when solving large scale problems.

4. Taking into consideration of the limitations of existing fusion methods, we propose a novel graphic approach for selecting optimal projection in multi-modal information fusion as an extension of contribution **1**. By graphically examining the transformation matrix, the proposed approach identifies the optimal projection and, in turn, the optimal feature sets in the transformed domain for final recognition.

1.5 Thesis Organization

The rest of this thesis is organized as follows.

Chapter 2 starts with an introduction of multi-modal information fusion, which is followed by a review on the recent advances on the three levels of information fusion: feature/data level, score level and decision level. An introduction to Information Theoretic Learning (ITL) methods to information fusion is then presented. Finally, we briefly discuss some representative applications in multi-modal information fusion fields.

Chapter 3 first briefly presents the fundamentals of CCA, DCCA and MCCA, and then formulates DMCCA. In the process, we analytically demonstrate that the optimally projected dimension by DMCCA can be quite accurately predicted, leading to both superior performance and substantial reduction in computational cost. After that, the relation between CCA, DCCA, MCCA and DMCCA is analyzed. Finally, a novel graphic approach for selecting optimal projection in multi-modal information fusion is presented.

Chapter 4 presents the entropy-estimation based discriminative analysis method integrating KECA and DMCCA to extract discriminatory representations and identify the inherent complementary relationship among

different modalities beyond the second order statistics, further improving the recognition accuracy.

Chapter 5 examines the performance of the proposed DAF in several applications, ranging from multi-feature fusion to multi-modal fusion.

Chapter 6 summarizes the works presented in this thesis and outlines possible directions for future research.

Chapter 2

Background

This chapter begins with a general overview of information fusion, including the definition of multi-modal information fusion and three fusion levels. Following that, we review the recent advances in the areas of information fusion, intelligent feature level fusion, Information Theoretic Learning (ITL) methods in information fusion, and some representative applications.

2.1 An Overview on Information Fusion

The proliferation of multimedia content and the advances in sensing technology have enabled and encouraged the design and development of computationally efficient and economically feasible multi-modal systems for a broad spectrum of application scenarios. Multimedia, by name and definition, contains a combination of information contents from different media sources in various content forms. Examples include audio, video, image and text, each of which can be deemed as a modality in a multi-modal multimedia representation. The integration of multi-modality data con-

tains more information about the semantics presented in the medium, and provides a more comprehensive description of the patterns or perceptions of interest [27].

Multi-modal information fusion refers to a process which achieves more reliable and robust analysis performance by integrating a set of multiple data sources, extracted features, and intermediate decisions [28]. It has drawn increasingly extensive interest in both research and industrial sectors, in a plethora of applications such as security and surveillance, video conferencing, video streaming, education and training, healthcare, database management, and human computer interaction (HCI). It is worth pointing out that multi-feature fusion is a special case of multimodal fusion. In multi-feature fusion, different sets of features are extracted from the same modality data but with different extraction methods, and thus highly likely carry richer information. Therefore, the fusion of the multi-feature sets could lead to better recognition results.

Regarding the existing approaches, there are three levels of information fusion: feature/data level, score level and decision level [29]. Data/feature level fusion combines the original data or extracted features through certain strategies before classification [30]. For instance, by concatenating the feature vectors of different modalities, a new vector is formed to represent the multi-modal information. Since feature level contains richer information about the raw data, the fusion at feature level is expected to perform better in some scenarios in comparison with fusion at score level and decision level. Moreover, fusion at the feature/data level has the advantages to provide the classifiers with better discriminatory representations by exploiting the co-variation and correlation between different modalities.

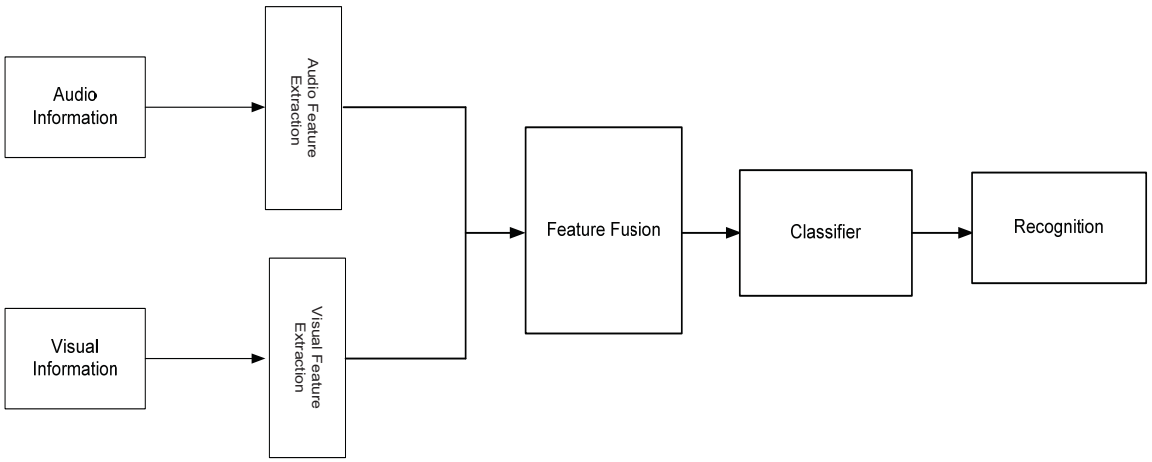

Figure 2.1: Feature level fusion for audiovisual information

For example, concatenating the feature vectors which have been extracted from two modalities, like audio and video signals, is a typical application of multi-modal information fusion. Figure 2.1 shows a schematic representation of audiovisual fusion at feature level. In Figure 2.1, features are extracted from different data channels, such as audio and video streams. The extracted features are first merged by feature fusion unit, and then the combined feature vector is input into classifiers for further analysis.

Fusion at the score level combines the scores generated from different modalities through a rule based scheme, or in a pattern classification sense in which the scores are taken as new input features of a classification algorithm [31]. At score level, it is possible to combine scores obtained from the same modalities or different ones. Its advantages include simple implementation and scalability. This level of fusion can be divided into two categories, combination and classification. Regarding combination, the input matching scores are combined by normalizing them into the same

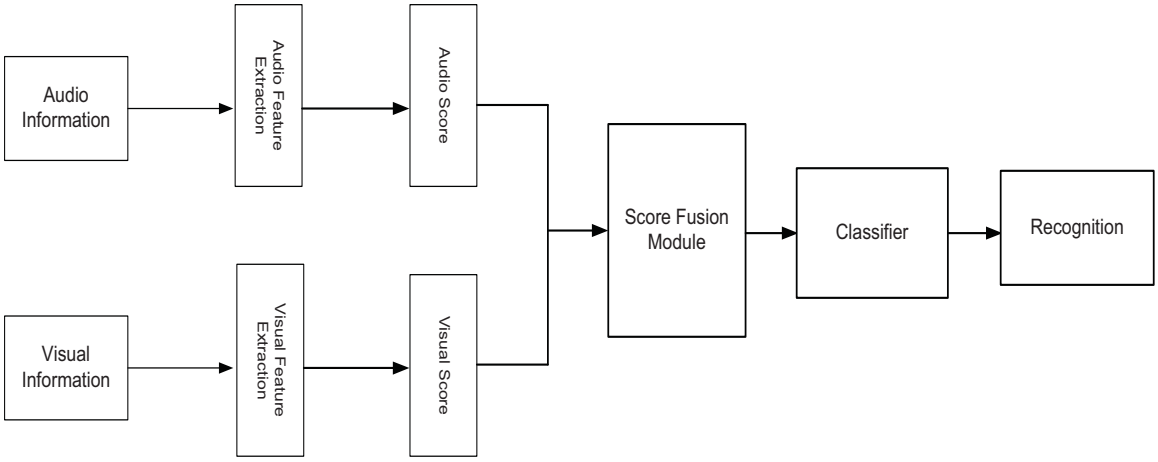

Figure 2.2: Score level fusion for audiovisual information

range. In terms of classification, the matching scores are viewed as input features for a second level classification. However, the fusion at score level has disadvantages, such as inability to utilize correlation at feature level and tedious learning process. Figure 2.2 shows a schematic representation of audiovisual fusion at score level. The data from different streams are extracted into feature vectors. The feature vectors are transformed into matching scores in score fusion module. Score fusion module integrates the scores and obtains the final result.

There are a number of typical applications of score level fusion. For example, Karthik et al. presented quality-based score level fusion in multi-biometric system [32]. The quality of biometric samples has a significant impact on the accuracy of a matcher. Therefore, dynamically assigning weights to individual matchers based on the quality of samples can improve the overall recognition performance of a multi-biometric system. The likelihood ratio-based fusion scheme takes into account the quality of

the biometric samples while combining the match scores provided by the matchers. Another recent application is a score level fusion framework of multi-modal biometrics using triangular norms presented by Hanmandlu [33]. The scores from multiple biometrics are combined using triangular norms (T-norms). T-norms achieve better performance over the traditional methods like SVM and linear regression. In addition, Dass et al. described an optimal framework for combining the matching scores from multiple modalities using the likelihood ratio statistics of the generalized densities estimated from the genuine and impostor matching scores [34]. The fusion approaches for combining the generalized densities include copula models which consider the dependence between the matching scores, and the product rule which assumes independence between the individual modalities.

The decision level fusion usually generates the final results based on the decisions made from individual classifiers or modalities using rule based methods such as AND, OR, and majority voting [35, 36]. Generally speaking, decision level fusion needs employment of independent classifiers for every modality and integration of the likelihood scores based on the strategies of reliability estimation. The organization of the correspondence between the channels is made during the integration step only. However, the fusion at decision level might lose too much useful information.

Some researchers have successfully adopted decision level fusion strategy. For example, Zhou et al. presented a facial expression recognition method based on global and local features with decision level fusion [37]. Local directional pattern (LDP) global features of the whole face are extracted, which can guarantee basic expression difference and decrease the

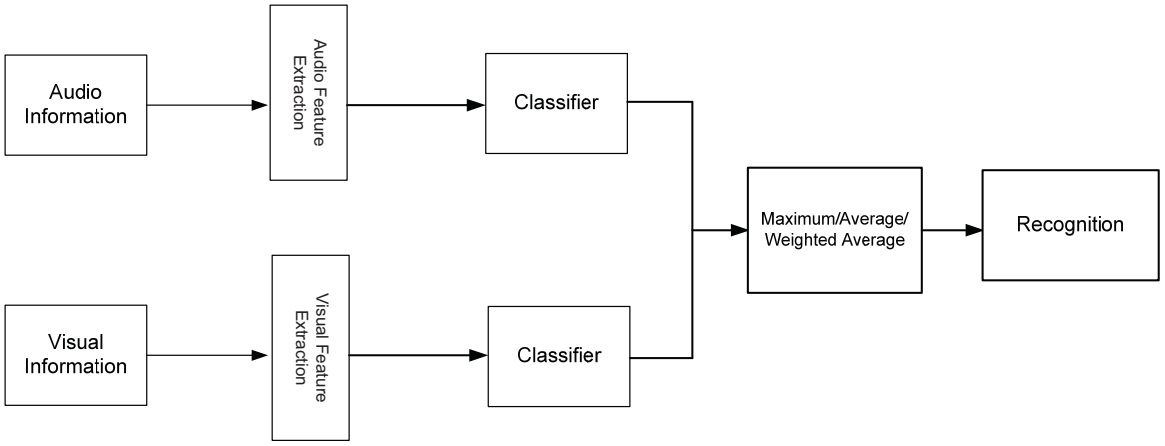

Figure 2.3: Decision level fusion for audiovisual information

influence of non-facial region. Local directional pattern variance (LDPv) descriptor is used to extract local features of regions of eyes and mouth, and extrude their contribution on expression changes. After feature extraction, instead of simple feature concatenation, a decision level fusion for global LDP feature and local LDPv feature is selected. Another interesting study is on decision level integration system for multi-modal emotional expression analysis presented by Metallinou [38]. Face, voice and head movement cues for emotion recognition are estimated and the classifiers are integrated using a Bayesian framework. The facial classifier has the best performance followed by the voice and head classifiers, and the multiple modalities seem to carry complementary information, especially for happiness. Decision fusion increases the average accuracy from 55% to about 62%. Wang et al. [7] proposed a Kernel Cross-Modal Factor Analysis method for audiovisual emotion recognition. It achieves 85.00% and 78.00% performance on RML database and eNTERFACE database,

respectively.

In general, the selection of a fusion level is dependent on the characteristics of the data and the requirements of the application problem on hand. The three fusion levels, the feature/data level fusion, decision level fusion and score level fusion, delegated by multi-classifier combination, have been researched extensively in pattern recognition, information fusion and human-computer interaction (HCI), and have been applied successfully to handwritten character recognition, face recognition and emotion recognition. [39, 40, 41].

2.2 Intelligent Feature Level Fusion

Although research in information fusion has advanced substantially in recent years, realistically emulating the information fusion capacity of the human brain is still far from accomplished. Major issues arise from the data to be fused, imperfection and diversity of the sensor technologies, and the nature of the application environment [13]. Therefore, intelligent feature level fusion has drawn significant attention from the research communities of multimedia and biometrics due to its capacity of information preservation and impressive progress has been made [42, 43].

The advantage of the feature level fusion is as follows. As different feature vectors extracted from the same pattern tend to reflect different characteristics of the pattern, optimally combining these features not only keeps the effective discriminant information, but also eliminates the redundant information to certain degree. This property is especially important to classification and recognition of large scale database in high dimensional feature space. Among them, serial feature fusion [44] was

the early winner. However, after serial feature extraction, how to select low-dimensional discriminative feature vectors for effective recognition remains an open challenge.

In addition, a number of systems based on feature level fusion have been developed. For example, Yang et al. described a feature level fusion framework using fingerprint and finger-vein for person identification [45]. The fingerprint and finger-vein features are first extracted using a unified Gabor filter framework. Then a supervised local-preserving canonical correlation analysis method is employed to generate fingerprint-vein feature vectors in feature level fusion. The nearest neighborhood classifier is used for person identification. This approach has a high capability in fingerprint-vein based person recognition as well as multi-modal feature level fusion. Ross et al. presented several feature level fusion strategies using hand and face biometrics, such as fusion of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) coefficients of face, fusion of face and hand modalities, and fusion of LDA coefficients corresponding to the R,G,B channels of a face image [46]. It is shown that the feature selection scheme ensures that redundant feature values are detected and removed before invoking the matcher. Recently, Feng et al. presented a common theoretical framework for multiple model fusion at feature level using multi-linear subspace analysis [47]. One disadvantage of multi-linear approach is that it is hard to obtain enough training observations for tensor decomposition algorithms. To overcome this difficulty, the M2SA algorithm [48] is adopted to reconstruct the missing entries of the incomplete training tensor. This framework is applied to the problem of face image analysis using Active Appearance Model (AAM) [49] to vali-

date its performance. Evaluations of AAM using the proposed framework are conducted with promising results.

Recently, there has been extensive interest in the analysis of correlation based approaches for multi-modal information fusion. The objective of correlation analysis is to identify and measure the intrinsic association between different modalities, by which the discriminant information carried by all modalities pertaining to certain semantic is determined. Hu et al. [50] proposed a large margin multi-metric learning (LM3L) method for face and kinship verification in the wild. It jointly learns multiple distance metrics under which the correlations of different feature representations of each sample are maximized, and the distance of each positive pair is less than a low threshold and that of each negative pair is greater than a high threshold, simultaneously. However, LM3L is only used to address the face and kinship verification problem at present.

Canonical correlation analysis (CCA) is a statistical method dealing with the mutual relationship between two random vectors, and a valuable multi-data processing method [51, 52, 53]. Sun et al. [54] proposed to use CCA to identify the correlation information of multiple feature streams of an image signal, and demonstrated the effectiveness of the method in handwritten character recognition and face recognition. CCA has also been applied to audiovisual based talking-face biometric verification [55], medical image analysis [56], and audio-visual synchronization [57]. However, in many practical problems dependencies between two signals cannot be described by simple linear correlation. If there is nonlinear correlation between the two variables, CCA may not correctly correlate this relationship. Kernel canonical correlation analysis (KCCA) [58, 59], a nonlinear

extension of CCA via the kernel trick to overcome this drawback, has been developed for the fusion of global and local features for target recognition [60], fusion of ear and profile face for multi-modal biometric recognition [61], fusion of text and image for spectral clustering [62], and fusion of labelled graph vector and the semantic expression vector for facial expression recognition [63], to name a few.

However, by CCA and KCCA, only the correlation between the pairwise samples is revealed. This correlation neither well represents the similarity between the samples in the same class, nor does evaluate the dissimilarity between the samples in different classes. To tackle the problem, a supervised learning method, namely discriminative CCA (DCCA) is proposed [64, 65, 23]. It simultaneously maximizes the within-class correlation and minimize the between-class correlation, thus potentially more suitable for recognition tasks than CCA.

Nevertheless, the CCA, KCCA, and DCCA merely deal with the mutual relationships between two random vectors, limiting the application of these techniques if there are multiple random vectors. Multi-set canonical correlation analysis (MCCA) is a natural extension of two-set canonical correlation analysis. It is generalized from CCA to deal with multi-modal features. The idea is to optimize characteristics of the dispersion matrix of the transformed variables to obtain high correlations between all new variables simultaneously. The method is not confined and the optimization takes place subject to different chosen constraints and orthogonality criteria. MCCA has been applied for inclusion in geographical information systems (GIS) [66], joint blind source separation [67] and blind single-input and multiple-output (SIMO) channels equalization [68]. However, MCCA

does not explore the discriminatory representation and is not capable of providing satisfactory recognition performance.

2.3 Information Theoretic Learning

In information fusion, there is another, probably more pressing issue, to be addressed. Although there are numerous methods proposed for information fusion, the theoretical foundation of these methods largely depend on the second order statistics, such as variance, correlation, mean square error and so on. Since the second order statistics are only optimal for Gaussian-like distribution [24] and sensitive to the choice of input parameters [25], a poor estimator is likely obtained if the underlining distribution greatly differs from Gaussian, failing to reveal the nature of input data. To overcome this problem, one of the new solutions is information theoretic learning (ITL), a terminology perhaps first used by Watanabe [69]. Using the ITL solutions, we can employ the mathematical theory of information initially developed by Claude Shannon [70]and Alfred Renyi [71] to quantify global scalar descriptors of the underlying probability density function.

Information theory was first conceptualized by Claude Shannon to deal with the problem of optimally transmitting messages over noisy channels [70]. The strategy proposed by Shannon was quickly accepted by the science and engineering communities and had an immediate impact on the design of communication systems. After the pioneering work of Shannon, information theory became a field of scientific studies and new discoveries have been brought to light based upon Shannon's fundamental concepts. Moreover, information theory has also been utilized in the areas of physic-

s, statistics, and biology as well as in field of engineering, for example machine learning and signal processing [72, 73, 74].

One of the most important ITL descriptors is entropy [71]. Hence there rises wide interest in better understanding the properties and applications of entropy. It is believed that entropy can quantify the data's statistical structure more precisely in comparison with the second order statistics which is still the mainstream of statistical signal processing [75]. In ITL, the second order moments are substituted by a geometric interpretation of data in functional space. In this functional space, variance is replaced by entropy, correlation is replaced by correntopy, and mean square error (MSE) is replaced by minimum error entropy (MEE) [25].

As a novel entropy-estimation-based information fusion method, kernel entropy component analysis (KECA) was proposed [26] and achieved 85.00% and 86.00% performance on RML dataset and eNTERFACE dataset, respectively. Unlike the existing methods which depend on the second order statistics, KECA is based on the information theory and preserves the maximum Renyi entropy of the input data with the smallest number of extracted features. It utilizes descriptor of information entropy to achieve improved performance [26]. This is the most significant property of KECA. Furthermore, KECA is a feature transformation technique projecting original space onto a feature subspace spanned by the kernel principal axes corresponding to the largest contribution of Renyi entropy [76]. Its mapping result is greatly different from the existing methods, such as kernel principal component analysis (KPCA), kernel canonical correlation analysis (KCCA), etc. By sorting the associated eigenvalues from the highest to the lowest, KECA selects the information with high significance and

ignores that with less significance based on the entropy estimation [77]. From the information-theoretic point of view, KECA is able to identify the optimal transformation which preserves as much information entropy as possible between input space and kernel feature subspace with the smallest number of features. Therefore, the information contents are maximally similar between two different feature spaces [78]. Moreover, from the viewpoint of information fusion, KECA helps derive a unsupervised fusion method which can realize a more complete and precise representation of multiple information sources [79].

However, straightforward utilization of KECA simply puts the information or features from different channels together without considering the intrinsic structure and relationship among them, likely resulting in unsatisfied performance.

2.4 Applications

With the rapid development of advanced multi-disciplinary technologies for acquiring, storing and transmitting massive big data, multi-modal information fusion has attracted growing attention recently, in both academia and industry. It has been applied to diverse domains, such as Internet of things, Robotics, Manufacturing, Engineering, Natural Language Processing (NLP) and medical informatics [80].

In practice, humans make extensive use of real-time big data simultaneously sourced from multiple cognitive sensors such as sight, sound, touch, smell, self-motion and taste, for both perceiving and interacting with the world. Therefore, effective interpretation and analysis of human behavior characteristics are of fundamental significance in the design of

intelligent human computer interaction systems. But the traditional human computer interfaces are not ideal for natural communication between humans and computers. Hence the need for more friendly and natural communication interface between humans and machines has arisen, and extensive efforts have been committed to improve non-intrusive sensors which could help users communicate freely [81]. Among them, voice and face information are two of the most natural, passive, and noninvasive types of traits [82, 83, 84, 85]. They can be easily captured by low-cost sensing devices, making them more economically feasible for potential deployment in a wide range of applications. In addition, many methods [86, 83, 87] have been proposed for information fusion based recognition tasks such as handwritten digit recognition, face recognition and so forth [88, 89, 90, 91, 92, 93, 94].

2.5 Summary

In this chapter, we first reviewed multi-modal information fusion from three different performance levels. After that, we presented the recent advances in intelligent feature level fusion. Then, we discussed the utilization of information theoretic learning (ITL), kernel entropy component analysis (KECA) in particular, in solving the second order statistic problems and related applications. Finally, some representative applications are presented.

Existing research work is the foundation of our study. By reviewing the related literature, we identified critical challenges, such as complimentary representations, discriminative representations, and second-order statistic-

s, which have not been properly addressed yet. These challenges inspire the research carried out in this thesis.

Chapter 3

Discriminative Multiple Canonical Correlation Analysis for Multi-modal Information Fusion

In this chapter, we study the first realization of the DAF using Discriminative Multiple Canonical Correlation Analysis (DMCCA) as the fusion function for multi-modal analysis as shown in Figure 3.1. The circled areas in Figure 3.1 indicate fusing different features together. We will analytically verify the following characteristics of DMCCA:

1. Benefiting from the discriminative characteristic of DMCCA, we can identify and extract the discriminatory representation among different modalities.

2. An important property of DMCCA is analytically verified. It shows that the number of projected dimensions corresponding to the optimal recognition accuracy is smaller than or equals to the number of classes being classified.

3. Canonical correlation analysis (CCA), multiple canonical correlation analysis (MCCA) and discriminative canonical correlation analysis (DC-

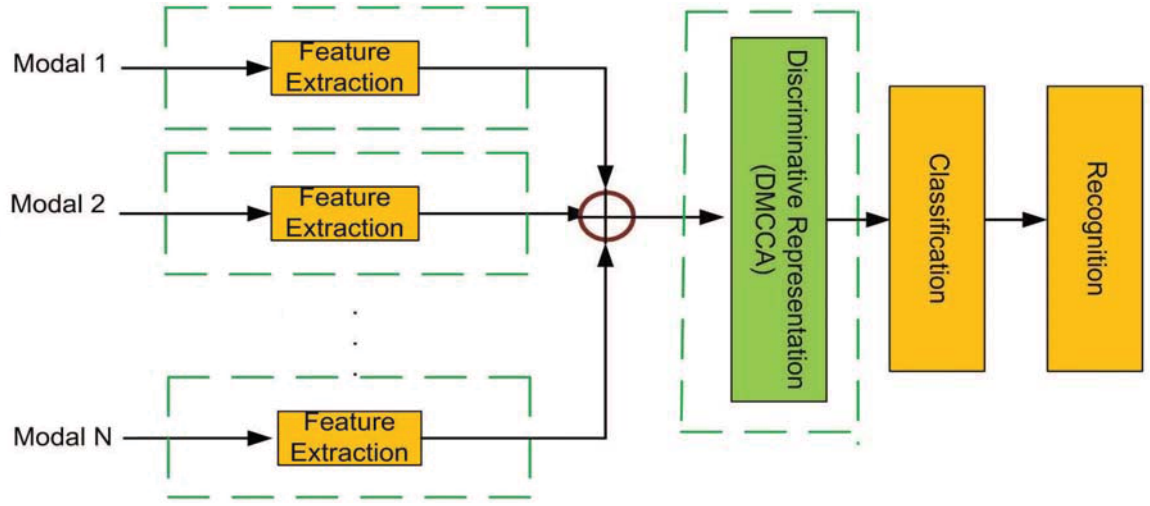

Figure 3.1: The proposed DAF using DMCCA as the fusion function for multi-modal information fusion.

CA) are special cases of DMCCA, thus establishing a unified framework for canonical correlation analysis for information fusion in the transformed domain.

4. We propose a novel graph representation approach for selecting optimal projection in multi-modal information fusion which substantially minimizes the effort of finding the optimal or near-optimal dimension of the features in the projected space.

3.1 Canonical Correlation Analysis

The aim of CCA is to find basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximized. Simultaneously, it needs to satisfy the canonical property that the first projection is uncorrelated with the

second projection, etc. To do so, all useful information, be it common to the two sets or specific to one of them, is maximumly preserved through the projections.

Let $x \in R^m, y \in R^p$ be two sets variables of the entries, and m, p being the dimensions in x and y, respectively. The CCA finds a pair of directions ω_1 and ω_2 to maximize the correlation between the projections of the two canonical vectors: $X = \omega_1^T x$, $Y = \omega_2^T y$, which can be written as follows:

$$\arg\max_{\omega_1,\omega_2} {\omega_1}^T R_{xy} \omega_2, \tag{3.1}$$

where $R_{xy} = xy^T$ is the cross-correlation matrix of the vectors x and y.

Simultaneously, x and y should satisfy the following condition to guarantee the first projection is uncorrelated with the second projection (canonical property):

$${\omega_1}^T R_{xx} \omega_1 = {\omega_2}^T R_{yy} \omega_2 = 1. \tag{3.2}$$

By solving the above optimization problem using the algorithm of Lagrange multipliers, we obtain the following relationship [57]:

$$\begin{bmatrix} 0 & \mathrm{R}_{xy} \\ R_{yx} & 0 \end{bmatrix} \omega = \mu \begin{bmatrix} \mathrm{R}_{xx} & 0 \\ 0 & R_{yy} \end{bmatrix} \omega, \tag{3.3}$$

where μ is the canonical correlation value and $\omega = [\omega_1^T, \omega_2^T]^T$ is the projected vector. Then equation (3.3) can be solved using the generalized eigenvalue (GEV) method.

3.2 Discriminative Canonical Correlation Analysis

The purpose of DCCA is to maximize the similarities of any pairs of sets of within-class while minimizing the similarities of pairwise sets of between-class, as mathematically expressed in [64, 65]:

$$T = \arg\max_{T} tr(T^T S_w T)/tr(T^T S_b T), \tag{3.4}$$

where T is the discriminant function and S_w, S_b relate to the within-class scatter matrix and between-class scatter matrix respectively.

The solution to equation (3.4) is obtained by solving the following GEV problem:

$$S_b T = \lambda S_w T. \tag{3.5}$$

For detailed information, please refer to [64].

3.3 Multiple Canonical Correlation Analysis

MCCA can be viewed as a natural extension of the two-set canonical correlation analysis [95]. Given M sets of random variables $x_1, x_2, \cdots x_M$ with the dimensions of $m_1, m_2, \cdots m_M$. The objective of MCCA is to find $\omega = [{\omega_1}^T, {\omega_2}^T \cdots {\omega_M}^T]^T$ which satisfies similar requirement as CCA and described as:

$$\arg\max_{\omega_1, \omega_2 \cdots \omega_M} \frac{1}{M(M-1)} \sum_{\substack{k,l=1 \\ k \neq l}}^{M} {\omega_k}^T C_{x_k x_l} \omega_l (k \neq l) \tag{3.6}$$

subject to

$$\sum_{k=1}^{M} {\omega_k}^T C_{x_k x_k} \omega_k = M, \tag{3.7}$$

where $C_{x_k x_l} = x_k {x_l}^T$. Solving (3.6) by the method of Lagrange multipliers yields

$$\frac{1}{M-1}(C - D)\omega = \beta D\omega, \tag{3.8}$$

where

$$C = \begin{bmatrix} x_1 {x_1}^T & \dots & x_1 {x_M}^T \\ \vdots & \ddots & \vdots \\ x_M {x_1}^T & \cdots & x_M {x_M}^T \end{bmatrix} \tag{3.9}$$

$$D = \begin{bmatrix} x_1 {x_1}^T & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_M {x_M}^T \end{bmatrix}. \tag{3.10}$$

and β is the generalized canonical correlation.

3.4 Concept of Discriminative Multiple Canonical Correlation Analysis

One of the major challenges in information fusion is to identify the discriminatory representation amongst different modalities. In this section, we introduce discriminative multiple canonical correlation analysis (DMCCA) to address this problem. Although Generalized multi-view analysis (GMA) [96] and Multi-view Discriminant Analysis (MDA) [97] are also proposed to solve the multi-view (multimodal) problem, there exist obvious differences among them. To be specific, the differences between DMCCA and GMA are:

1. In GMA, the discriminability is obtained within each feature, while in DMCCA it is achieved by using all features.

2. In GMA, the cross-view correlation is obtained only from observations corresponding to the same underlying sample, while in DMCCA it is obtained from all observations from different feature sets.

3. GMA has to deal with has a great number of parameters especially when the number of features is large. However, DMCCA works with a small number of parameters.

Different from the purpose of MDA to maximize the between-class variations and minimize the within-class variations from both intra-view and inter-view in the common space, the purpose of DMCCA is to simultaneously maximize the within-class correlation and minimize the between-class correlation, helping reveal the intrinsic structure and discriminatory representation from different sources/modalities, and improve the recognition accuracy.

The advantages of DMCCA for multi-modal information fusion rest on the following facts:

1. DMCCA involves modalities/features having a mixture of correlated (modality-common information) components and achieving the maximum of the correlation [57]. Therefore, DMCCA possesses the maximal commonality of multiple modalities/features.

2. The within-class and the between-class correlations of all modalities/features are considered jointly to extract more discriminative information, leading to a more discriminant common space and better generalization ability for classification from multiple modalities/features.

3.4.1 Derivation of the DMCCA

Let P sets of zero-mean and unit variance random features be $x_1 \in R^{m_1}, x_2 \in R^{m_2}, \cdots x_P \in R^{m_P}$ for c classes and $Q = m_1 + m_2 + \cdots m_P$. Concretely, DMCCA aims to seek the projection vectors $\omega = [{\omega_1}^T, {\omega_2}^T, \cdots {\omega_P}^T]^T$ ($\omega_1 \in R^{m_1 \times Q}, \omega_2 \in R^{m_2 \times Q}, \cdots \omega_P \in R^{m_P \times Q}$) for information fusion so that the within-class correlation is maximized and the between-class correlation is minimized. Based on the definition of CCA and MCCA, DMCCA is formulated as the following optimization problem:

$$\arg\max_{\omega_1, \omega_2 \cdots \omega_P} \rho = \frac{1}{P(P-1)} \sum_{\substack{k,m=1 \\ k \neq m}}^{P} {\omega_k}^T \tilde{C}_{x_k x_m} \omega_m \tag{3.11}$$

subject to

$$\sum_{k=1}^{P} {\omega_k}^T C_{x_k x_k} \omega_k = P, \tag{3.12}$$

where $\tilde{C_{x_k x_m}} = C_{w_{x_k x_m}} - \delta C_{b_{x_k x_m}} (\delta > 0), C_{x_k x_k} = x_k {x_k}^T$. $C_{w_{x_k x_m}}$ and $C_{b_{x_k x_m}}$ denote the within-class and between-class correlation matrixes, respectively.

Let $x_i = [{x_{i1}}^{(1)}, {x_{i2}}^{(1)} \cdots {x_{in_1}}^{(1)}, \cdots {x_{i1}}^{(c)}, {x_{i2}}^{(c)} \cdots {x_{in_c}}^{(c)}] \in R^{m_i \times n}$, then

$$e_{n_{il}} = [\underbrace{0, 0, \cdots 0}_{\sum_{u=1}^{l-1} n_{iu}}, \underbrace{1, 1, \cdots 1}_{n_{il}} \underbrace{0, 0, \cdots 0}_{n - \sum_{u=1}^{l} n_{iu}}]^T \in \mathbf{R^n} \tag{3.13}$$

$$\mathbf{1} = [1, 1, \cdots 1]^{\mathbf{T}} \in \mathbf{R^n}, \tag{3.14}$$

where i is the number sequence of the random features, n is the total number of training samples, ${x_{ij}}^{(d)}$ denotes the jth sample in the dth class, respectively, and n_{il} is the number of samples in the lth class of x_i set.

$$\sum_{l=1}^{c} n_{il} = n, \tag{3.15}$$

where c is the total number of classes. Note that, as the random features satisfies the property of zero-mean, it can be shown that:

$$x_i \cdot \mathbf{1} = \mathbf{0}. \tag{3.16}$$

Then, the within-class correlation matrix between sets x_k and x_m, $C_{w_{x_k x_m}}$ can be written as:

$$\begin{aligned} C_{w_{x_k x_m}} &= \sum_{l=1}^{c} \sum_{h=1}^{n_{kl}} \sum_{g=1}^{n_{ml}} {x_{kh}}^{(l)} {x_{mg}}^{(l)T} \\ &= \sum_{l=1}^{c} (x_k e_{n_{kl}})(x_m e_{n_{ml}})^T \\ &= x_k A {x_m}^T, \end{aligned} \tag{3.17}$$

where

$$A = \left[\begin{pmatrix} H_{n_{i1}\times n_{i1}} & \cdots & 0 \\ \vdots & H_{n_{il}\times n_{il}} & \vdots \\ 0 & \cdots & H_{n_{ic}\times n_{ic}} \end{pmatrix} \right] \in R^{n\times n} \tag{3.18}$$

with $H_{n_{i1}\times n_{i1}}$ in the form of $n_{i1} \times n_{i1}$ and all the elements in $H_{n_{i1}\times n_{i1}}$ being unit values. Similarly, the between-class correlation matrix $C_{b_{x_k x_m}}$ is in the form of:

$$\begin{aligned} C_{b_{x_k x_m}} &= \sum_{l=1}^{c} \sum_{\substack{q=1 \\ l\neq q}}^{c} \sum_{h=1}^{n_{kl}} \sum_{g=1}^{n_{mq}} {x_{kh}}^{(l)} {x_{mg}}^{(q)T} \\ &= \sum_{l=1}^{c} \sum_{q=1}^{c} \sum_{h=1}^{n_{kl}} \sum_{g=1}^{n_{mq}} {x_{kh}}^{(l)} {x_{mg}}^{(q)T} - \sum_{l=1}^{c} \sum_{h=1}^{n_{kl}} \sum_{g=1}^{n_{ml}} {x_{kh}}^{(l)} {x_{mg}}^{(l)T} \\ &= (x_k \mathbf{1})(x_m \mathbf{1})^T - x_k A {x_m}^T \\ &= -x_k A {x_m}^T. \end{aligned} \tag{3.19}$$

Substituting equations (3.17) and (3.19) into (3.11) yields:

$$\begin{aligned} \arg\max_{\omega_1,\omega_2\cdots\omega_P} \rho &= \tfrac{1}{P(P-1)} \sum_{k,m=1}^{P} {\omega_k}^T \tilde{C}_{x_k x_m} \omega_m \\ &= \tfrac{1+\delta}{P(P-1)} \sum_{k,m=1}^{P} {\omega_k}^T x_k A {x_m}^T \omega_m, \end{aligned} \tag{3.20}$$

subject to

$$\sum_{k=1}^{P} {\omega_k}^T C_{x_k x_k} \omega_k = P. \tag{3.21}$$

Using Lagrangian multiplier criterion to solve (3.20) results in the following expression

$$\begin{aligned}
&\tfrac{1+\delta}{P-1}\left[\left[\begin{pmatrix} 0 & x_1Ax_2{}^T & x_1Ax_3{}^T\ldots & x_1Ax_P{}^T \\ x_2Ax_1{}^T & 0 & x_2Ax_3{}^T\ldots & x_2Ax_P{}^T \\ x_3Ax_1{}^T & x_3Ax_2{}^T & 0\ldots & x_3Ax_p{}^T \\ & \vdots & & \\ x_PAx_1{}^T & x_PAx_2{}^T & x_PAx_3{}^T\ldots & 0 \end{pmatrix}\right]\right]\omega \\
&= \rho\begin{pmatrix} x_1x_1{}^T & 0 & 0\ldots & 0 \\ 0 & x_2x_2{}^T & 0\ldots & 0 \\ 0 & 0 & x_3x_3{}^T\ldots & 0 \\ & \vdots & & \\ 0 & 0 & 0\ldots & x_Px_P{}^T \end{pmatrix}\omega
\end{aligned} \tag{3.22}$$

It is further rewritten in a compact form:

$$\frac{1+\delta}{P-1}(C-D)\omega = \rho D\omega, \tag{3.23}$$

where

$$C = \left[\left[\begin{pmatrix} x_1x_1{}^T & x_1Ax_2{}^T & x_1Ax_3{}^T\ldots & x_1Ax_P{}^T \\ x_2Ax_1{}^T & x_2x_1{}^T & x_2Ax_3{}^T\ldots & x_2Ax_P{}^T \\ x_3Ax_1{}^T & x_3Ax_2{}^T & x_3x_3{}^T\ldots & x_3Ax_p{}^T \\ & \vdots & & \\ x_PAx_1{}^T & x_PAx_2{}^T & x_PAx_3{}^T\ldots & x_Px_P{}^T \end{pmatrix}\right]\right] \tag{3.24}$$

$$D = \begin{bmatrix} \begin{pmatrix} x_1 {x_1}^T & 0 & 0 \dots & 0 \\ 0 & x_2 {x_2}^T & 0 \dots & 0 \\ 0 & 0 & x_3 {x_3}^T \dots & 0 \\ & \vdots & & \\ 0 & 0 & 0 \dots & x_P {x_P}^T \end{pmatrix} \end{bmatrix} \tag{3.25}$$

$$C - D = \begin{bmatrix} \begin{pmatrix} 0 & x_1 A {x_2}^T & x_1 A {x_3}^T \dots & x_1 A {x_P}^T \\ x_2 A {x_1}^T & 0 & x_2 A {x_3}^T \dots & x_2 A {x_P}^T \\ x_3 A {x_1}^T & x_3 A {x_2}^T & 0 \dots & x_3 A {x_P}^T \\ & \vdots & & \\ x_P A {x_1}^T & x_P A {x_2}^T & x_P A {x_3}^T \dots & 0 \end{pmatrix} \end{bmatrix} \tag{3.26}$$

$$\omega = [{\omega^T}_1, {\omega^T}_2, \cdots {\omega^T}_P]^T. \tag{3.27}$$

Since $(1 + \delta/P - 1)$ is a constant, it has no influence to the projection matrix ω, and thus will be ignored in the following analysis. Equation (3.23) is further written in the form of:

$$\begin{gathered} x_1 A {x_2}^T \omega_2 + x_1 A {x_3}^T \omega_3 + \cdots + x_1 A {x_P}^T \omega_P = \rho x_1 {x_1}^T \omega_1 \\ x_2 A {x_1}^T \omega_1 + x_2 A {x_3}^T \omega_3 + \cdots + x_2 A {x_P}^T \omega_P = \rho x_2 {x_2}^T \omega_2 \\ \vdots \\ x_P A {x_1}^T \omega_1 + x_P A {x_2}^T \omega_2 + \cdots + x_P A {x_{P-1}}^T \omega_{P-1} = \rho x_P {x_P}^T \omega_P \end{gathered} \tag{3.28}$$

Based on the definition of $\tilde{C}_{x_k x_m}$ and equation (3.23), the value of ρ plays a critical role in evaluating the relationship between within-class and between-class correlation matrixes. When the value of ρ is greater

than zero, the corresponding projected vector ω contributes positively to the discriminative power in classification while the projected vector ω corresponding to the non-positively values of ρ would result in reducing the discriminative power in classification. Clearly, the solution obtained is the eigenvectors associated to the positive eigenvalues in equation (3.23).

Commonly, it is known that the time taken greatly depends on the computation of the projective vectors to extract discriminative features. When the rank of eigen-matrix is very high, the computation of eigenvalues and eigenvectors will be time-consuming. To address this problem effectively, an important property of DMCCA is proved here. That is the number of projected dimension d corresponding to the optimal recognition accuracy is smaller than or equals to the number of classes, c, or mathematically:

$$d \leq c \tag{3.29}$$

Now we will show that d does satisfy inequality (3.29). From equation (3.18), the rank of matrix A satisfies

$$rank(A) \leq c \tag{3.30}$$

Then, equation (3.30) leads to:

$$rank(x_i A {x_j}^T) \leq \min(r_i, r_A, r_j), \tag{3.31}$$

where r_i, r_A, r_j are the ranks of matrices x_i, A, x_j ($i, j \in [1, 2, 3, ..., P]$), respectively.

Due to the fact that $rank(A) \leq c$, equation (3.31) satisfies

$$rank(x_i A {x_j}^T) \leq \min(r_i, c, r_j), \tag{3.32}$$

when c is less than r_i and r_j, equation (3.32) is written as

$$rank(x_i A {x_j}^T) \leq c \tag{3.33}$$

Otherwise, equation (3.32) satisfies

$$rank(x_i A {x_j}^T) \leq min(r_i, r_j) < \ c \tag{3.34}$$

On the other hand equation (3.28) can be written as follows:

$$\begin{cases} x_1 A({x_2}^T\omega_2 + {x_3}^T\omega_3 + \cdots + {x_P}^T\omega_P) = \rho x_1 {x_1}^T\omega_1 \\ x_2 A({x_1}^T\omega_1 + {x_3}^T\omega_3 + \cdots + {x_P}^T\omega_P) = \rho x_2 {x_2}^T\omega_2 \\ \vdots \\ x_P A({x_1}^T\omega_1 + {x_2}^T\omega_2 + \cdots + {x_{P-1}}^T\omega_{P-1}) = \rho x_P {x_P}^T\omega_P \end{cases} \tag{3.35}$$

Equation (3.35) is further expressed as:

$$\begin{cases} (\zeta_1)^{-1} x_1 A({x_2}^T\omega_2 + {x_3}^T\omega_3 + \cdots + {x_P}^T\omega_P) = \omega_1 \\ (\zeta_2)^{-1} x_2 A({x_1}^T\omega_1 + {x_3}^T\omega_3 + \cdots + {x_P}^T\omega_P) = \omega_2 \\ \vdots \\ (\zeta_P)^{-1} x_P A({x_1}^T\omega_1 + {x_2}^T\omega_2 + \cdots + {x_{P-1}}^T\omega_{P-1}) = \omega_P \end{cases} \tag{3.36}$$

where

$$\begin{cases} \zeta_1 = \rho x_1 {x_1}^T \\ \zeta_2 = \rho x_2 {x_2}^T \\ \quad \vdots \\ \zeta_P = \rho x_P {x_P}^T \end{cases} \quad (when\ x_i {x_i}^T \text{ is non} - \text{singular}) \tag{3.37}$$

or

$$\begin{cases} \zeta_1 = \rho x_1 {x_1}^T + \sigma_1 I_1 \\ \zeta_2 = \rho x_2 {x_2}^T + \sigma_2 I_2 \\ \quad\vdots \\ \zeta_P = \rho x_P {x_P}^T + \sigma_p I_p \end{cases} (when\ x_i {x_i}^T\ is \text{ singular}) \tag{3.38}$$

where $I_i \in R^{m_i \times m_i}$ and $\sigma_1, \sigma_2, ...\sigma_p$ are constants.

Since $rank(A) \leq c$ and $\omega_i \in R^{m_i \times Q} (i = 1, 2, ...P)$, based on equation (3.36), the rank of ω_i satisfies

$$rank(\omega_i) \leq c (i = 1, 2, ..., P) \tag{3.39}$$

Then the fused feature of $Y_i (i = 1, 2, ...P)$ can be written as follows:

$$\begin{cases} Y_1 = {\omega_1}^T x_1 \\ Y_2 = {\omega_2}^T x_2 \\ \quad\vdots \\ Y_p = {\omega_P}^T x_P \end{cases} \tag{3.40}$$

Let $(\omega)_d$ be the projected matrix with DMCCA achieving optimal performance and $(\omega)_d \in R^{Q \times d}$. $(\omega)_d$ is written as follows:

$$(\omega)_d = \begin{bmatrix} \omega_{11}, \omega_{12}, ..., \omega_{1d} \\ \omega_{21}, \omega_{22}, ..., \omega_{2d} \\ \vdots \\ \omega_{P1}, \omega_{P2}, ..., \omega_{Pd} \end{bmatrix} = \begin{bmatrix} (\omega_1)_d \\ (\omega_2)_d \\ \vdots \\ (\omega_p)_d \end{bmatrix} \tag{3.41}$$

Then ω, $(\omega)_d$ and $(\omega_i)_d(i = 1, 2, ...p)$ satisfy the relationship:

$$\begin{aligned}
\omega &= \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_P \end{bmatrix} = \begin{bmatrix} \omega_{11} & \omega_{12} & ... & \omega_{1d} & \omega_{1(d+1)} & ... & \omega_{1Q} \\ \omega_{21} & \omega_{22} & ... & \omega_{2d} & \omega_{2(d+1)} & ... & \omega_{2Q} \\ & & & \vdots & & ... & \\ \omega_{P1} & \omega_{P2} & ... & \omega_{Pd} & \omega_{P(d+1)} & ... & \omega_{PQ} \end{bmatrix} \\
&= \begin{bmatrix} (\omega_1)_d, \omega_{1(d+1)}, ..., \omega_{1Q} \\ (\omega_2)_d, \omega_{2(d+1)}, ..., \omega_{2Q} \\ \vdots \\ (\omega_P)_d, \omega_{P(d+1)}, ..., \omega_{PQ} \end{bmatrix} \\
&= [(\omega)_d, \underbrace{0, ..., 0}_{Q-d}] + [\omega - [(\omega)_d, \underbrace{0, ..., 0}_{Q-d}]](0 \in R^Q)
\end{aligned} \tag{3.42}$$

inserting (3.42) into (3.40) yields:

$$\begin{cases}
Y_1 = {\omega_1}^T x_1 = [(\omega_1)_d, \underbrace{0_1, ..., 0_1}_{Q-d}]^T x_1 + [\omega_1 - [(\omega_1)_d, \underbrace{0_1, ..., 0_1}_{Q-d}]]^T x_1 \\
Y_2 = {\omega_2}^T x_2 = [(\omega_2)_d, \underbrace{0_2, ..., 0_2}_{Q-d}]^T x_2 + [\omega_2 - [(\omega_2)_d, \underbrace{0_2, ..., 0_2}_{Q-d}]]^T x_2 \\
\vdots \\
Y_p = {\omega_p}^T x_p = [(\omega_p)_d, \underbrace{0_p, ..., 0_p}_{Q-d}]^T x_p + [\omega_p - [(\omega_p)_d, \underbrace{0_p, ..., 0_p}_{Q-d}]]^T x_p
\end{cases} \tag{3.43}$$

where $0_i(i = 1, 2, ..., p)$ is in the form R^{m_i}.

Based on equation (3.40), d should satisfy inequality (3.44)

$$d \leq \min(m_1, m_2, ..., m_P) \tag{3.44}$$

Simultaneously, since $rank(\omega_i) \leq c$ and Y_i possesses the same number of rows, d satisfies the relation (3.45)

$$d \leq min[rank(\omega_1), rank(\omega_2), ..., rank(\omega_p)] \leq c \tag{3.45}$$

considering (3.44) and (3.45) together, d should satisfy (3.46)

$$\begin{cases} d \leq \min(m_1, m_2, ..., m_P) \\ d \leq c \end{cases} \tag{3.46}$$

Now, we analyze the following two cases.

1. when c satisfies (3.47)

$$c \leq \min(m_1, m_2, ..., m_P) \tag{3.47}$$

combining (3.44) and (3.45) leads to

$$d \leq c \tag{3.48}$$

2. when c satisfies (3.49)

$$c > \min(m_1, m_2, ..., m_P) \tag{3.49}$$

combining (3.44) and (3.45) leads to

$$d \leq \min(m_1, m_2, ..., m_P) \leq c \tag{3.50}$$

In summary,

$$d \leq c \tag{3.51}$$

To achieve the optimal recognition accuracy, we select the c projected vectors from the eigenvectors associated with the c different largest eigenvalues in equation (3.23).

Since $[\omega_i - [(\omega_i)_d, \underbrace{0_i, ..., 0_i}_{Q-d}]]^T x_i$ and $[\underbrace{0_i, ..., 0_i}_{Q-d}]^T x_i$ have no contribution to the optimal fused result of Y_i, the optimal performance reached by DMCCA when, d, the projected dimension is less than or equals to c, as expressed in (3.52)

$$\begin{cases} Y_{1,optimal} = [(\omega_1)_d, \underbrace{0_1, ..., 0_1}_{Q-d}]^T x_1 = (\omega_1)_d{}^T x_1 \in R^d (d <= c) \\ Y_{2,optimal} = [(\omega_2)_d, \underbrace{0_2, ..., 0_2}_{Q-d}]^T x_2 = (\omega_2)_d{}^T x_2 \in R^d (d <= c) \\ \vdots \\ Y_{p,optimal} = [(\omega_p)_d, \underbrace{0_p, ..., 0_p}_{Q-d}]^T x_p = (\omega_p)_d{}^T x_p \in R^d (d <= c) \end{cases} \tag{3.52}$$

Thus, expressions in (3.52) lead to the proof of (3.29).

Specifically, if the feature space dimension equals Q, the computational complexity of DMCCA is on the order of *O(Q*c)*, instead of *O(Q*Q)*, as other transformation-based methods require (such as MCCA). Thus, this property is not only analytically elegant, but practically significant when c is small compared with the feature space dimension (which includes emotion recognition, digit recognition, English character recognition, and many others), where c ranges from a handful to a couple of dozens, but the feature space dimension could be hundreds or even thousands.

In summary of the discussion so far, the information fusion algorithm based on DMCCA is given below:

Step 1. Extract information from multi-modal sources to form the training sample spaces.

Step 2. Convert the extracted information into the normalized form and compute the matrices C and D.

Step 3. Compute the eigenvalues and eigenvectors of equation (3.23).

Step 4. Obtain the fused information expression from equation (3.52), which is used for classification.

3.4.2 Relation Between CCA, DCCA, MCCA, and DMCCA

In this subsection, we will demonstrate that CCA, MCCA and DCCA are special cases of DMCCA.

1) Relation with DCCA [23]: when P=2, equation (3.23) turns into the following form:

$$(C - D)\omega = \rho D\omega \tag{3.53}$$

where

$$C = \begin{bmatrix} x_1 {x_1}^T & x_1 A {x_2}^T \\ x_2 A {x_1}^T & x_2 {x_2}^T \end{bmatrix} \tag{3.54}$$

$$D = \begin{bmatrix} x_1 {x_1}^T & 0 \\ 0 & x_2 {x_2}^T \end{bmatrix} \tag{3.55}$$

Thus, it transforms into the method of DCCA, only dealing with the mutual relationships between two random vectors. Since DCCA is a special case of DMCCA, it also possesses this discriminative property. Note, the authors of [23] were the first to perform dimensionality reduction using this property with DCCA, but did not provide a concrete proof.

2) Relation with MCCA [66]: when A is an identity matrix, the matrices

C and D of the DMCCA can be written as:

$$C = \left[\begin{pmatrix} x_1 {x_1}^T & \dots & x_1 {x_P}^T \\ \vdots & \ddots & \vdots \\ x_P {x_1}^T & \cdots & x_P {x_P}^T \end{pmatrix} \right] \tag{3.56}$$

$$D = \left[\begin{pmatrix} x_1 {x_1}^T & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_P {x_P}^T \end{pmatrix} \right] \tag{3.57}$$

It transforms into the method of MCCA.

3) Relation with CCA [54]: when P=2 and A is an identity matrix, the matrixes C and D of the DMCCA can be described as:

$$C = \begin{bmatrix} x_1 {x_1}^T & x_1 {x_2}^T \\ x_2 {x_1}^T & x_2 {x_2}^T \end{bmatrix} \tag{3.58}$$

$$D = \begin{bmatrix} x_1 {x_1}^T & 0 \\ 0 & x_2 {x_2}^T \end{bmatrix} \tag{3.59}$$

It transforms into the method of CCA. Since there are no within-class and between-class correlation considered, CCA and MCCA do not possess the discriminative power as DMCCA. In addition, the best performance of CCA (and MCCA) is not predictable.

3.4.3 A Novel Graph Representation Approach for Selecting Optimal Projection

As aforementioned, information fusion is becoming a key research area with applications to various multimedia analysis tasks. Feature level fu-

sion has been considered as a most promising fusion method due to the rich information presented at this level. A critical operation of feature level fusion is the projection of the features onto a space which best presents the information for recognition. However, the identification of the optimal projection of the multi-modal features onto the projected space remains a difficult task.

In this subsection, we present a novel graph representation approach for selecting optimal projection in information fusion which substantially minimizes the effort of finding the optimal or near-optimal dimension of the features in the projected space.

In general, the solutions to a large number of multi-modal information fusion methods are obtained by utilizing the algorithm of matrix transformation. Some unsupervised and supervised examples are PCA, CCA, Cross-Modal Factor Analysis (CFA) [98], Entropy Component Analysis (ECA), Fisher Linear Discriminant Analysis (FLDA) and their kernel versions [99, 100, 7, 76, 101]. The solution to matrix transformation is usually the eigenvectors associated with the eigenvalues in a form of equation (3.23):

$$\frac{1}{P-1} inv(D) * (C - D)\omega = \rho\omega \tag{3.60}$$

where *inv()* refers to the inverse transform of a matrix. However, unless the covariance matrices D have full rank, the block matrix in equation (3.25) is singular. An approach [102] to dealing with singular covariance matrices and to controlling complexity is to add a multiple of the identity matrix $\lambda \mathrm{I}(\lambda > 0)$ to D. Thus, the generalized form of equation (3.60) can be written as:

$$\frac{1}{P-1} inv(D^{+}) * (C - D)\omega = \eta\omega \tag{3.61}$$

where

$$D^{+} = \begin{cases} D & \text{when } D \text{ is a full rank matrix} \\ D + \lambda \mathrm{I} & \text{when } D \text{ is a singular matrix} \end{cases}$$
$$\eta = \begin{pmatrix} \eta_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \eta_Q \end{pmatrix} \tag{3.62}$$

In equation (3.62), η_i is the criterion to seek the projection vectors for feature extraction. Hence, the value of η_i is the key parameter to the effect of selecting features. A larger η_i corresponds to the more discriminative features, while a smaller η_i corresponds to the less discriminative features.

Thus, it is reasonable to evaluate the final multi-modal information fusion results by criterion $J(\eta)$

$$J(\eta) = \sum_{i=1}^{Q} \eta_i \tag{3.63}$$

where η_i is the ith eigenvalue of equation (3.62).

Then we plot the graph of the proposed criterion $J(\eta)$. Close examination of the graph reveals that the proposed approach can accurately estimate the dimension of the features in the projected space for optimal recognition without actually carrying out the complete experiment. Therefore, it is not only analytically elegant, but practically significant especially for fusion in feature space of very high dimensions. Examples to demonstrate the effectiveness of this graph representation approach will be presented in Chapter 5.

3.5 Summary

In this chapter, we introduce a discriminative multiple canonical correlation analysis approach for multi-modal analysis and fusion. At first, it finds projection directions to maximize the within-class correlation and minimize the between-class correlation among multiple information/data to identify the discriminative representation between different modalities effectively. Second, based on the proposed DMCCA, we verify that the best performance by discriminative representation achieves when only a small fraction of the data needs to be analyzed. After that, we establish a unified framework for canonical correlation analysis for information fusion in the transformed domain. Finally, we present a graph representation method on selecting optimal projection in multi-modal information fusion. By examining the transformation matrix, the proposed approach identifies the optimal projection and, in turn, the optimal feature sets in the transformed domain for final recognition.

Chapter 4

KECA plus DMCCA for Multi-modal Information Fusion

The effective interpretation and integration of multiple information content are important for the efficacious utilisation of multimedia in a wide variety of application context. There are two major challenges for information fusion: 1) *how to identify the complementary representation from multiple information/data*; and 2) *how to extract discriminatory representation from individual channels or data sources.*

In chapter 3, DMCCA provides a way to find projection directions to maximize the within-class correlation and minimize the between-class correlation among multiple information/data sources in order to identify the discriminatory representation between different modalities. In this chapter, benefiting from the discriminatory representation extracted by DMCCA, we propose the second realization of the DAF, a novel method integrating KECA and DMCCA as the fusion component of the framework to address the two challenges simultaneously. Profiting from the proposed method, we can identify and extract the complementary and discriminative

representation synchronously, achieving improved recognition accuracy.

4.1 Introduction

It is a well known fact that the second order statistics such as variance and correlation is the theoretical foundation of the majority of existing information fusion methods [26]. For these methods, feature transformation is usually based on top eigenvalues and the corresponding eigenvectors of certain matrices. Since the second order statistics is only optimal for Gaussian-like distribution, it could be a poor estimator, if the distribution from multiple modalities differs greatly from Gaussian. This issue motivates researchers to apply Information Theoretic Learning (ITL) as an alternative to solve fusion problems [103].

In ITL, one of the most important descriptors is entropy. Therefore, there rises wide interest in better understanding the properties and applications of entropy. It is known that entropy can quantify the statistical structure of a dataset more precisely in comparison with the second order statistics which is still the mainstream of statistical signal processing [104]. As a recently proposed entropy measurement method, KECA has been studied in order to obtain more appropriate representations for information fusion than the second order statistics. However, as an unsupervised method, KECA only preserves the maximum Renyi entropy of the input data with the smallest number of extracted features without considering the similarity between the samples in the same class, or the dissimilarity between the samples in different classes, resulting in losing discriminatory representation of the multi-modal information [105]. Therefore, in this chapter, we propose using KECA plus DMCCA (KECA+DMCCA) as the

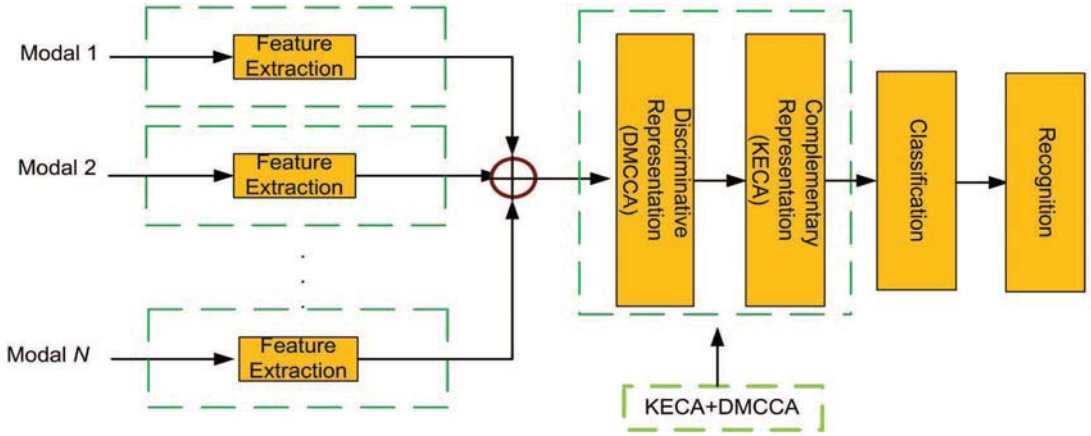

Figure 4.1: KECA+DMCCA as the fusion component in the proposed DAF

second realization of the fusion component of the proposed DAF to solve this problem. The schematic representation of the proposed method is shown in Figure 4.1.

The remainder of this chapter is organized as follows. Section 4.2 briefly describes entropy estimation. Based on entropy estimation, we review the method of KECA in multi-modal information fusion in Section 4.3. Section 4.4 presents the proposed realization KECA+DMCCA of the DAF. Finally, we summarize the chapter in Section 4.5.

4.2 Entropy Estimation

4.2.1 Shannon Entropy

The concept of Shannon entropy was introduced as a measure of statistical uncertainty. In the field of thermodynamics, Shannon entropy is a physical concept which correlates with the quantity of kinematic randomness, while in the area of information theory, entropy is no longer a physical concept

and it stands for a concept which provides a mathematical tool to quantify and formulate the nature of information [106]. Shannon entropy plays a central role in information theory. This entropy is believed to be able to measure the amount of the information contained in a series of events, which can be expressed as follows.

$$H(X) = -\int f(X) \log f(X), \tag{4.1}$$

or in the discrete form

$$H(X) = -\sum_{m} p(X_m) \log p(X_m), \tag{4.2}$$

where $f(X)$ and $p(X_m)$ are the continuous and discrete probability density function of data sets respectively, and m is the total number of data sets in the discrete case.

The concept of information is so rich that perhaps there is no single definition which is able to quantify information properly. Entropy can be interpreted as a means of quantifying information content. A fundamental property of entropy is that with a single scalar, it measures the uncertainty in a form of probability density [107]. It can also be extended to measure dissimilarity between data. Furthermore, the entropy measure has been showed to be an appropriate descriptor of the hyper-volume spanned by a high dimensional probability density. Therefore, Shannon theory is used to derive a set of estimators to apply entropy as cost functions in machine learning. It has been applied in a variety of fields from basic sciences such as biology and physics to different engineering discipline.

4.2.2 Renyi Entropy

In 1960, Alfrd Rnyi introduced a parameterized family of uncertainty measures $H_\alpha(X)$, now known as the Renyi entropy [108]. In information theory, the Renyi entropy generalizes the Hartley entropy, the Shannon entropy, the collision entropy and the min entropy. In practical applications, Renyi entropy is one of the widely used generalizations of information entropy [108]. Renyi wanted to find the most general class of information measure which preserved the additivity of statistically independent systems. Renyi entropy of order α of a random variable X is written as

$$H_\alpha(X) = \frac{1}{1-\alpha}\log(\int f^\alpha(X)dX), \tag{4.3}$$

or

$$H_\alpha(X) = \frac{1}{1-\alpha}\log(\sum_{m=1}^{N} {p_{X_m}}^\alpha), \tag{4.4}$$

where $\alpha >= 1$. At a deeper level, Renyi entropy measure is much more flexible than Shannon entropy due to the parameter α. An interesting observation is that Shannon entropy can be considered as a special case of Renyi entropy when α converges to one [109]. We usually choose $\alpha = 2$ as the fundamental descriptor, because it gives us a computationally efficient entropy estimator. Here, continuous Renyi quadratic entropy is given by

$$H(X) = -\log(\int p^2(X)dx), \tag{4.5}$$

where $p(X)$ is probability density function (PDF) generated by the data sets $x_1, x_2, ...x_N$. The main reason why Renyi quadratic entropy is employed is that the entropy value can be elegantly estimated by PDF $p(X)$.

Then the entropy can be estimated by replacing probability density function with non-parametric density estimator.

4.2.3 Kernel Method

Kernel method is widely used in nonlinear problem of data analysis, and one of the most well-known applications is support vector machine [110]. The bottleneck of nonlinear problem is a large number of high-dimensional classifiers; hence the computation would become expensive. Kernel method provides a way to simplify the computation, and the calculation can be executed efficiently in the space provided by the algorithms expressed in inner products [111]. Therefore, the fundamental principle of kernel method is mapping the original data onto a feature space by a non-linear transformation and employing linear algorithms in the new space.

Given a set X including samples $x_k \in R^n$. Each vector x_k is projected from the input space, R^n, to a high dimensional feature space, R^f, by a nonlinear mapping function $\varphi : R^n \to R^f$, $f > n$. Then a kernel function is a function k that satisfies

$$k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle \,. \tag{4.6}$$

It is known that an explicit expression of non-linear mapping φ is difficult to determine. However, kernel lets us calculate inner products in a feature space of possibly infinite dimensionality directly without having to deal with the explicit mapping φ. This means that any linear machine learning algorithm expressed via inner products can solve nonlinear problems by operating in a high-dimensional feature space. However, the kernel function must satisfy the Mercers condition, i.e., positive

semi definite. Some widely used kernel functions include linear kernel $k(x_i, x_j) = \langle x_i, x_j \rangle$, polynomial kernel $k(x_i, x_j) = (< x_i, x_j > +1)^o$ and Gaussian kernel $k(x_i, x_j) = exp((-\|x_i - x_j\|^2)/2\sigma^2)$.

The kernel matrix K contains all the evaluation of kernel function k. From the kernel, we know that this matrix also contains all evaluation of inner products between the data points in the feature space. The expression is given by $K_{i,j} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$, or in matrix form

$$\begin{aligned} K_{n,n} &= \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ & \vdots & \\ k(x_n, x_1) & & k(x_n, x_n) \end{bmatrix} \\ &= \begin{bmatrix} \langle \varphi(x_1), \varphi(x_1) \rangle & \cdots & \langle \varphi(x_1), \varphi(x_n) \rangle \\ & \vdots & \\ \langle \varphi(x_n), \varphi(x_1) \rangle & & \langle \varphi(x_n), \varphi(x_n) \rangle \end{bmatrix}, \end{aligned} \tag{4.7}$$

where n is the number of samples in the original space.

Hence, kernel method is used to develop nonlinear generalization of any algorithm which can be cast in terms of inner products. For instance, kernel principal component analysis (KPCA) and kernel linear discriminant analysis (KLDA) are typical extensions of the corresponding linear algorithms by applying the kernel method on every inner product evaluation.

4.3 Kernel Entropy Component Analysis

4.3.1 Parzen Window Density Estimator

The strategies of density estimator can be divided into parametric method and nonparametric method [112]. Parametric models are restricted in their

representation capability, but we have to make assumptions of signal models and have knowledge of the signals which we are dealing with. On the other hand, nonparametric density estimation technique provides the freedom of representing signal distributions based on the observed samples. Nonparametric estimators yield well-behaved gradient algorithms which can optimize adaptive system parameters [113]. A number of nonparametric density estimation methods are available, but we focus on Parzen windowing which is also known as kernel density estimation. Parzen windowing is a computationally simple approach which can yield both continuous and smooth estimation of information-theoretic quantities for adaptive signal processing and learning algorithms [114].

As already stated, we need to deal with the issue of estimating entropy directly from samples in a nonparametric way, since it is not prudent to make an assumption of a parametric probability density function (PDF) model. It is essential to develop cost measures derived directly from data without further assumptions to capture as much data structure as possible. We use the direct approach of estimating the scalar value of Renyi quadratic entropy from samples by using Parzen window density estimator. It could estimate the probability distribution without any assumptions of parameters or shapes. Parzen windowing can be viewed as natural implementation of kernel function and creates a close connection between information theory and kernel method. Given N independent and identically distributed samples $\{x_1, x_2, ...x_N\}$ from a random variable. The expression of Parzen window density estimator is given by

$$\bar{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h}), \tag{4.8}$$

where $K()$ is the kernel and h is a smoothing parameter called width. In the general framework of Parzen windowing, the rectangular kernels can be replaced by smoother kernel functions, for example Gaussian distribution function. Parzen windowing provides density estimation of information theoretic quantities, and a non-parametric density estimator is obtained by replacing the actual PDF by its Parzen window density estimator. Therefore, by utilizing Parzen windowing method, the non-parametric estimator for entropy does not require an explicit estimation of probability density function.

4.3.2 KECA with Application to Information Fusion

The continuous Renyi quadratic entropy is given by

$$H(p) = -\log(\int p^2(x)dx) = -\log V(p), \tag{4.9}$$

where $V(p) = \int p^2(x)dx = E\{p(x)\}$. $V(p)$ is considered as expectation w.r.t. the density $p(x)$. In order to estimate the value of entropy, we only need to consider the quantity $V(p) = \int p^2(x)dx$, since the logarithm is a monotonic function. To estimate $V(p)$, Parzen window density estimator is applied. Parzen window density estimator using the kernel notation on N samples is written as follows

$$\bar{p}(x) = \frac{1}{N\sigma}\sum_{x_j \in D} K(\frac{x - x_j}{\sigma}) = \frac{1}{N}\sum_{x_j \in D} k_\sigma(x, x_j), \tag{4.10}$$

where $k_\sigma(x, x_j)$ is the kernel centered at x_j, and σ is kernel size. We assume a positive semi-finite Parzen window with Gaussian kernel. There is no single best method to choose the kernel size, so we need to be careful

and establish best procedures to select σ of the kernel. The convolution theorem for Gaussian function states that the convolution of two Gaussian functions is another Gaussian function with $\sigma = \sqrt{{\sigma_1}^2 + {\sigma_1}^2}$. In other words, the integral of the product of two Gaussians is exactly evaluated as the value of the Gaussian computed at the difference of the arguments and whose variance is the sum of the variances of the two original Gaussian functions. Hence we rearrange terms of Parzen window density estimator and obtain the following nonparametric estimator for Renyi entropy.

$$\bar{V}(p) = \frac{1}{N}\sum_{i=1}^{N}\bar{p}(x_i) = \frac{1}{N}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}k_\sigma(x_i, x_j) = \frac{1}{N^2}1^T K 1, \qquad (4.11)$$

where element (i, j) of the $N \times N$ kernel matrix K is equal to $k_\sigma(x_i, x_j)$, and 1 is a $N \times 1$ vector containing all ones. Therefore Renyi quadratic entropy is compactly expressed in terms of the kernel matrix. This result is obtained by noticing that the Gaussian maintains the functional form under convolution. However, other kernel functions cannot result in such convenient evaluation of the integral. It is shown that entropy value is a scalar, but one of the intermediate steps is to estimate the PDF, which is much harder in high-dimensional spaces. By employing continuous Renyi quadratic entropy, we can bypass the explicit need to estimate the PDF and obtain the entropy evaluation of the data using algebraic operations.

Furthermore, Renyi entropy estimator can be expressed in terms of eigenvalues and eigenvectors of the kernel matrix through eigen-decomposition. The eigen-decomposition of K is shown below

$$K = EDE^T, \qquad (4.12)$$

where D is a diagonal matrix storing the eigenvalues $\lambda_1, \lambda_2, ...\lambda_N$ and E is a matrix with the corresponding eigenvectors $\alpha_1, \alpha_2, ...\alpha_N$ as columns. Hence the empirical Renyi entropy estimator equals to the elements of the corresponding kernel matrix. Substituting (4.12) into (4.11) yields the following expression

$$\bar{V}(p) = \frac{1}{N^2} 1^T K 1 = \frac{1}{N^2} 1^T E D E^T 1 = \frac{1}{N^2} \sum_{i=1}^{N} \left(\sqrt{\lambda_i} {\alpha_i}^T 1\right)^2, \tag{4.13}$$

where λ_i and α_i are the i-th eigenvalue and eigenvector of kernel matrix K.

The above expression is known as entropy-value in KECA. The total entropy value is estimated by the joint contribution from all the $\sqrt{\lambda_i}{\alpha_i}^T$. Apparently, since both eigenvalues and eigenvectors make contributions to the entropy estimator, instead of selecting the largest eigenvalues, KECA selects eigenvalues and eigenvectors based on the largest entropy estimation. This is the most significant property of KECA. Furthermore, KECA is a feature transformation technique projecting original space onto a feature subspace spanned by the kernel principal axes corresponding to the largest contribution of Renyi entropy. Its mapping result is greatly different from the existing methods, such as kernel principal component analysis (KPCA), kernel canonical correlation analysis (KCCA), etc.

By sorting the associated entropy from the highest to the lowest, KECA selects the information with high significance and ignores the data with less significance based on the entropy estimation. From the information-theoretic point of view, KECA is able to identify the optimal transformation which preserves as much as information entropy between input space

and kernel feature subspace with the smallest number of features. Therefore, the information contents are maximally similar between two different feature spaces. Moreover, from the viewpoint of information fusion, KECA helps derive a fusion method which can realize a more complete and precise representation of multiple information sources. Information fusion based on KECA can reduce the dimensionality of input feature vector, while retaining most of the useful information content of the original data. The motivation of information fusion based on KECA is rooted in the fact that the data carried by different modalities usually have intrinsic association. It is essential to take full advantage of the correlation between them and extract the most discriminant and representative patterns from the input data.

To exploit the complementary representation of multi-modal data, an optimal mathematical framework for feature level information fusion based on KECA is presented [24]. The following steps summarize the procedure of feature transformation and fusion based on KECA.

(1) The feature vector $X = [x_1, x_2, ...x_N]$ is the input data which requires feature transformation and fusion.

(2) Gaussian function is chosen as kernel function and the kernel matrix K with elements $K_{i,j} = k(x_i, x_j)$ is obtained.

(3) Conduct the eigen-decomposition of K and calculate $K = EDE^T$.

(4) Choose the first n largest entropy estimation corresponding to $\sqrt{\lambda_i}{\alpha_i}^T$.

(5) From ${\phi^T}_{keca}\phi_{keca} = (D^{\frac{1}{2}}E^T)^T(D^{\frac{1}{2}}E^T) = EDE^T = K$ to calculate the kernel feature space data set $\phi_{keca} = (D^{\frac{1}{2}}E^T)$.

(6) Complete the feature transformation by ϕ_{keca}.

Since KECA does not suffer from the limitation of Gaussianity which

is inherent in cost functions based on the second order moments, better information fusion performance is achieved by information theoretic descriptors of entropy combined with nonparametric PDF estimators. The proposed method reduces the dimensionality of the features by eliminating data redundancy and utilizes data complementarity in the form of entropy measures. KECA brings about robustness and generality, and improves performance in many realistic scenarios. Nevertheless, as an unsupervised method, KECA merely puts the information or features from different channels together simply without considering the intrinsic structure and relationship, likely resulting in unsatisfied recognition performance. To solve these issues, the second realization of DAF, KECA plus DMCCA (KECA+DMCCA) is proposed in section 4.4.

4.4 The Proposed KECA+DMCCA

In this section, we present using KECA plus DMCCA (KECA+DMCCA) as the fusion component of the proposed framework for information fusion. With this method, we are more likely to identify and extract the complementary and discriminative representation among different modalities simultaneously, thus potentially improving recognition accuracy.

Given P sets of zero-mean random features $x_1 \in R^{m_1}, x_2 \in R^{m_2}, \cdots x_P \in R^{m_P}$ for c classes and $Q = m_1 + m_2 + \cdots m_P$. Let X_1, X_2, ..., $\cdots X_P$ denote the projections of the P discriminative vectors in discriminative multiple canonical correlation analysis space and n is the total number of training samples, i.e.

$$X_1 = {\omega_1}^T x_1;\ X_2 = {\omega_2}^T x_2; ...\ X_P = {\omega_P}^T x_P, \tag{4.14}$$

where $\omega_1, \omega_2, ...\omega_P$ are the projected vectors in DMCCA space.

Combining equations in (4.14) into matrix vector form leads to

$$\begin{cases} X_1 = {\omega_1}^T x_1 \\ X_2 = {\omega_2}^T x_2 \\ \vdots \\ X_P = {\omega_P}^T x_P \end{cases}$$

$$= \begin{bmatrix} {\omega_1}^T & 0 & \cdots & 0 \\ 0 & {\omega_2}^T & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & {\omega_P}^T \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_P \end{bmatrix} \tag{4.15}$$

$$= \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_P \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_P \end{bmatrix}$$

Then equation (4.15) is further written as

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{bmatrix} = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_P \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_P \end{bmatrix} \tag{4.16}$$

Since the number of training samples is n, X can also be expressed as

$$X = \left[\widetilde{X}_1, \widetilde{X}_2, ... \widetilde{X}_n\right] \tag{4.17}$$

Based on the definition of KECA, KECA+DMCCA is formulated as

$$V(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n} k_\sigma(\widetilde{X}_i, \widetilde{X}_j) = \frac{1}{n^2} 1^T \widetilde{K}\, 1. \tag{4.18}$$

The projection of KECA+DMCCA onto the ith principal axis in the kernel feature space is defined as

$$\Phi_{KECA+DMCCA} = (D'_i)^{\frac{1}{2}} E'^T_i, \tag{4.19}$$

where D'_i is the ith eigenvalue and E'_i is the corresponding ith eigenvectors of $\widetilde{K}$.

Since the rank of ω_i in DMCCA satisfies relation (4.20)

$$rank(\omega_i) \leq c(i = 1, 2, ..., P) \tag{4.20}$$

and

$$\omega = [\omega^T{}_1, \omega^T{}_2, \cdots \omega^T{}_P]^T, \tag{4.21}$$

the number of maximum linearly independent group of ω in DMCCA is less than or equals to the number of the classes, which can be written as follows

$$\omega_j^c = [a'_1, a'_2, ...a'_c] \begin{bmatrix} \omega'_1 \\ \omega'_2 \\ \vdots \\ \omega'_c \end{bmatrix}, \tag{4.22}$$

where ${\omega^c}_j$ is the jth column vector of the matrix ω and $a_1^{'}, a_2^{'}...a_c^{'}$ are the weights of the maximum linearly independent group $\omega_1^{'}, \omega_2^{'}...\omega_c^{'}$.

Since there are Q projected vectors in DMCCA space, the total entropy of the DMCCA is written as follows:

$$H_{total} = -P_{{\omega^c}_1} log P_{{\omega^c}_1} - P_{{\omega^c}_2} log P_{{\omega^c}_2} - ... - P_{{\omega^c}_Q} log P_{{\omega^c}_Q}, \tag{4.23}$$

where $P_{{\omega^c}_j}$ is the probability of ${\omega^c}_j$ (j=1,2...Q).

Equation (4.23) is further rewritten as:

$$\begin{aligned} H_{total} = & -P_{{\omega^c}_1 = a_1\omega_1^{'} + a_2\omega_2^{'} + ... + a_c\omega_c^{'}} log P_{{\omega^c}_1 = a_1\omega_1^{'} + a_2\omega_2^{'} + ... + a_c\omega_c^{'}} \\ & -P_{{\omega^c}_2 = b_1\omega_1^{'} + b_2\omega_2^{'} + ... + b_c\omega_c^{'}} log P_{{\omega^c}_2 = b_1\omega_1^{'} + b_2\omega_2^{'} + ... + b_c\omega_c^{'}} - ... \\ & -P_{{\omega^c}_Q = q_1\omega_1^{'} + q_2\omega_2^{'} + ... + q_c\omega_c^{'}} log P_{{\omega^c}_Q = q_1\omega_1^{'} + q_2\omega_2^{'} + ... + q_c\omega_c^{'}} \end{aligned}, \tag{4.24}$$

where $a_1, a_2...a_c$, $b_1, b_2...b_c$,...,$q_1, q_2...q_c$ are the weights of ω_1,ω_2,...,ω_Q. As $\omega_1^{'}, \omega_2^{'}...\omega_c^{'}$ are the vectors of the maximum linearly independent group, they are independent variables. $P_{{\omega^c}_j}$ (j=1,2...Q) is further written as:

$$P_{{\omega^c}_j = j_1\omega_1^{'} + j_2\omega_2^{'} + ... + j_c\omega_c^{'}} = j_1 P_{\omega_1^{'}} + j_2 P_{\omega_2^{'}} + ... + j_c P_{\omega_c^{'}}, \tag{4.25}$$

where $j_1, j_2...j_c$ are the weights of ${\omega^c}_j$. Therefore, equation (4.24) is rewritten as follows:

$$H_{total} = F(\omega_1^{'}, \omega_2^{'}, ..., \omega_c^{'}), \tag{4.26}$$

where F is a function only containing variables $\omega_1^{'}$, $\omega_2^{'}$,...$\omega_c^{'}$. We can obtain the total information entropy of DMCCA with the first c projected vectors corresponding to largest eigenvalues to form the maximum linearly independent group. Moreover, it is known that the entropy will not be improved during the transform of KECA. Hence, we can achieve the opti-

mal results of information fusion with KECA+DMCCA by the c linearly independent projected vectors $\omega_1^{'}, \omega_2^{'}...\omega_c^{'}$.

Specifically, benefiting from this property, if the feature space dimension equals to Q and the number of the training samples is n, the computational complexity of KECA+DMCCA is on the order of $O(n^2 * c)$, instead of $O(n^2 * Q)$. Thus, this property is not only analytically elegant, but practically significant when c is small compared with the feature space dimension. Examples include emotion recognition, digit recognition, and many others, where c ranges from a handful to a couple of dozens, but the feature space dimension could be hundreds or even thousands.

In summary, the information fusion algorithm based on KECA+DMCCA is given below:

Step 1. Extract information from multi-modal sources to form the training sample spaces.

Step 2. Convert the extracted information into the normalized form and compute the matrices C and D in equations (3.20) and (3.21).

Step 3. Compute the eigenvalues in the matrix ρ and eigenvectors in the matrix ω of equation (3.19).

Step 4. Select the c projected vectors from the eigenvectors collection associated with the c different largest eigenvalues in equation (3.19).

Step 5. Find the discriminative projections of the original multi-modal feature/data in DMCCA space.

Step 6. KECA is applied to the discriminative projections achieving information fusion of KECA+DMCCA.

In essence, KECA+DMCCA transforms the original multiple input information/data sources into the DMCCA space at first. Since DMCCA

can be seen as a way of guiding discriminative feature selection toward the underlying semantics to find basis vectors for different sets of variables, it reveals discriminative representation among different multiple variables. In addition, based on the definition of canonical correlation, the transformed sets of linear combinations are those with the largest correlation subject to the condition that they are orthogonal to the former canonical variables. Therefore, it also eliminates redundant information effectively before KECA is implemented. After that, KECA is applied to the discriminative vectors in the DMCCA space. Thus, the discriminatory and complementary representations of input data beyond the second order statistics are revealed together, improving the recognition accuracy.

4.5 Summary

In this chapter, firstly, we study the entropy estimation and KECA, which are expected to reveal more complementary representation than the second order statistics from the multiple input sources. After that, we investigate the second realization of the proposed DAF for information fusion, which integrates KECA and DMCCA together. Based the proposed fusion component, not only the discriminative representation is considered, but also the complementary representation of input data is revealed in the space of KECA, instead of that of the second order statistics. Moreover, we mathematically verify that the optimal performance by KECA+DMCCA achieves with c independent projected vectors. It is a particularly attractive property when solving large scale problems. Finally, since the kernel method is applied to the proposed discriminative method, it provides a more effective method to solve nonlinear problems in information fusion.

Chapter 5

Experimental Results and Analysis

A multimedia analysis task involves processing multi-modal data to obtain valuable insights about the data, a situation, or a higher level activity. In what follows, firstly, we present feature extraction and implementation of the proposed DAF in handwritten digit recognition, face recognition and human emotion recognition. After that, we conduct experiments on Mixed National Institute of Standards and Technology (MNIST) handwritten digit database, ORL face database, and Ryerson Multimedia Lab (RML) [91]and eNTERFACE [115] emotional database to demonstrate the generic nature and the effectiveness of the proposed DAF for multi-modal information fusion.

5.1 Introduction

For handwritten digit recognition, its main application areas fall in postal mail sorting, bank check processing and form data entry. Face recognition is a vitally important research area spanning multiple fields and disciplines. It is essential for effective communications and interactions among

people with applications to bankcard identification, mugshot searching, surveillance systems, etc. Emotion recognition has been playing an important role in our daily social interactions and activities. Automated machine recognition of human emotion has been recognized as a key technology for building a more natural and friendly communication interface between humans and computers. The emotional state of an individual can be inferred from different sources such as voice, facial expressions, body language, ECG, and EEG. Among them, voice and face are two of the most natural, passive, and noninvasive types of traits, the study of which is the focus of this thesis.

5.1.1 Handwritten Digit Recognition

Handwritten digit recognition plays a significant role in several applications such as cheque processing and the automatic sorting of postal mail [116]. Recognition of handwritten digits is a difficult task due to the wide variety of styles, sizes and orientations of digit samples for the same writer and between different writers. In addition, there are two challenges in handwritten digit recognition due to the nature of the handwriting style [117].

a. Different writing styles and pens lead to strongly varying appearances.

b. The inherent variation in writing styles at different instances.

Recently, numerous works have been developed for the evaluation of handwritten digit recognition algorithms. They differ in the feature extraction and classification stages employed. Nishida [118] proposes a grammar-like model for applying deformations to structures composed of primitive strokes. Lam and Suen [119] use a two-stage method for recognition, in

which samples are first classified by their structure using a tree classifier. Cheung et al. [120] model characters with a spline, and assume that the spline parameters have a multivariate Gaussian distribution. A Bayesian approach is then used to determine the character class, with the model parameters as prior and the image data parameters as likelihood. Revow et al. [121] model digits as ink-generating Gaussian "beads" strung along a spline outline. Characters are matched through deformation of the spline and adjustment of the bead parameters.

In general, the performance of handwritten digit recognition depends on the feature extraction approaches. For feature extraction of digit recognition, various approaches using the global features and the local structural features, have been presented in [122].

Gabor transform has been widely applied to handwritten digit recognition, face recognition and emotion recognition, etc [123, 124, 125, 126]. An important property of Gabor transform is that it has optimal joint localization, or resolution in both the spatial and the spatial-frequency domains to extract global features. In addition, it has been shown to be a good fit to the receptive field profiles of simple cells in the striate cortex. The Gabor filter, based on a multi-channel filtering theory, is designed for information processing in the early stages of the human visual systems.

As a local feature extraction method, Zernike moments are widely used to handwritten digit recognition [116]. In general, moments are pure statistical measure of pixel distribution around the center of gravity of the image and allow capturing global shapes information [127]. They describe numerical quantities at some distance from a reference point or axis. Zernike moments are a class of orthogonal moments and have been shown

to be effective in terms of image representation. Advantages of Zernike moments can be summarized as follows:

a. The magnitude of Zernike moment has rotational invariant property.

b. They are robust to noise and shape variations to some extent.

c. Since the basis is orthogonal, they have minimum redundant information.

d. An image can better be described by a small set of its Zernike moments than any other types of moments such as geometric moments.

In this thesis, Gabor features and Zernike moments features are applied together to the information fusion on handwritten digit recognition problem.

5.1.2 Face Recognition

Automatic recognition of human faces has been an active research area in recent years. In addition to the importance of pure research, it has a number of commercial and law-enforcement applications such as surveillance, security, telecommunications and human-computer intelligent interaction, etc. Various approaches for face recognition have been proposed and they can be roughly classified into either analytic or holistic approaches.

Analytic approaches use things such as distances and angles between fiducial points on the face, shapes of facial features and local features. The main advantage of analytic approaches is to allow for a flexible deformation at the key feature points so that pose changes can be compensated for. While analytic approaches compare the salient facial features detected from the face, holistic approaches make use of the information derived from the whole face. More detailed literature on face recognition approaches

can be found in [128, 129, 130].

Despite remarkable progresses so far, the general task of face recognition remains a challenging problem. This is mainly due to the complex distortions that can be caused by variations in illumination, facial expressions and poses. It is widely believed that local features in face images are more robust against such distortions and a spatial frequency analysis is often desirable to extract such features. With good characteristics of spatial frequency localization, Gabor transform is a good candidate for this purpose [125].

In addition, in the evaluation the proposed method for face recognition, we extracted the histogram of oriented gradient (HOG) and local binary patterns (LBP) to represent the global features. Therefore, in this thesis, benefiting from fusing the HOG, LBP and Gabor wavelets features together, it is expected to achieve improved performance.

5.1.3 Emotion Recognition

5.1.3.a Audio Emotion Recognition

Speech is one of the most essential and natural verbal channels to transmit human affective states and it is easily accessible for emotion recognition. A detailed review of the cutting-edge works for audio emotion recognition can be found in [131]. The performance of speech emotion recognition based on information fusion has also been investigated by numerous works in the literature [24, 132].

Human speech contains not only linguistic content but also cues showing emotions of the speakers. Since audio features have been heavily used in emotion recognition, one of the important issues is the extraction of

speech features which characterize the emotional states efficiently without depending on lexical content or speakers. The widely used features are categorized into continuous features and spectral features [133, 134, 135].

Continuous speech features have been heavily used in emotion recognition, since they have been found to represent the most significant characteristics of emotional content in verbal communication. It is believed that continuous features such as pitch and energy convey much of the temporal information and always serve as the primary indicator of a speaker's emotion states. Continuous features are known as prosodic features. Because of temporal information present in speech signals, continuous speech features are superior in terms of classification time and accuracy.

In addition to time-dependent continuous features, spectral features are often selected as another representation for speech signals. Spectral features have different representations of the signal nature. Moreover, due to the fact that there are limited number of spectral features to study, the algorithms of feature selection based on spectral features are executed faster, and the training of classifiers is more efficient. The widely used spectral features include MFCC (Mel-frequency cepstral coefficient) and Formant Frequency (FF), which will be used in this thesis.

5.1.3.b Visual Emotion Recognition

Visual information is a most direct method of interaction between human and machine, and rich emotional information can be conveyed through the human face. In general, the face region is detected from the image first, and then facial expression information is extracted from the observed facial images [91].

Many solutions have been proposed to process facial expressions and identify the emotional information. Existing solutions for visual emotion recognition can be roughly categorized into two groups. One is to treat the human face as a whole unit [136], and the other is to represent the face by prominent components, such as the mouth, eyes, nose, eyebrow, and chin [137]. The analysis of facial components is critically dependent on the accurate localization of the local features. Further, focusing on only a few facial components, the representation of the discriminant characteristics of human emotions might be inadequate. In this thesis, we perform visual emotion recognition by treating the face as a holistic pattern and the visual information is represented by Gabor wavelet features.

5.2 Feature Extraction

In this subsection, we present the feature extraction on handwritten digit recognition, face recognition, audio and visual emotion recognition, which corresponds to the left most block in Figure 1.2 redrawn here as Figure 5.1.

5.2.1 Handwritten Digit Feature Extraction

Gabor filters, which operate directly on gray-level handwritten digit images, is chosen to extract features for handwritten digit recognition. Gabor filters have several advantages. First, Gabor features have been used for capturing local information in both spatial and frequency domains from images, as opposed to other global techniques such as Fourier Transforms.

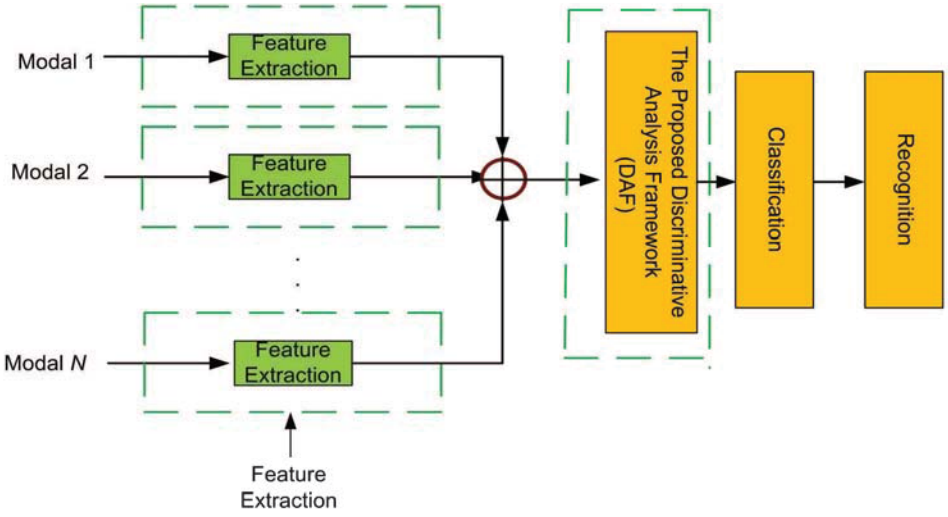

Figure 5.1: Feature extraction in the proposed DAF

Second, Gabor filters are orientation specific. This property allows us to analyze stroke directions in the handwriting. Third, the filtering output is robust to various noise components since Gabor filters use information from all pixels in the kernel.

Despite the advantages, Gabor filter based feature selection methods are normally computationally expensive due to high dimensional Gabor features. In this thesis, we use 24 Gabor filters; 4 for scaling and 6 for orientation. In addition, we consider the mean and standard deviation of the magnitude of the transform coefficients of each filter as the features.

The other feature, Zernike polynomials are orthogonal series of basis functions normalized over a unit circle. The complexity of these polynomials increases with increasing polynomial order. To calculate the Zernike moments, the image (or region of interest) is first mapped to the unit disc using polar coordinates, where the center of the image is the origin of the unit disc. The pixels falling outside the unit disc are not considered. The

coordinates are then decided by the length of the vector from the origin to the coordinate point. An important attribute of the geometric representations of Zernike polynomials is that lower order polynomials approximate the global features of the shape/surface, while the higher ordered polynomials capture local shape/surface features. Zernike moments are a class of orthogonal moments and have been shown to be effective in terms of image representation [138].

In this thesis, we extracted three features for handwritten digit recognition:

24-dimensional: the mean of the digit images transformed by the Gabor filters.

24-dimensional: the standard deviation of the digit images transformed by the Gabor filters.

36-dimensional: Zernike moment features.

5.2.2 Face Feature Extraction

Face recognition using Gabor features has attracted considerable attention in computer vision, image processing, pattern recognition, and so on. Gabor filters can exploit salient visual properties such as spatial localization, orientation selectivity, and spatial frequency characteristics [139]. In this thesis, we uses 4 filters for scaling and 6 filters for orientation. Moreover, we take the mean and standard deviation of the magnitude of the transform coefficients of each filter as the features.

Furthermore, in the evaluation the proposed method for face recognition, we extracted the histogram of oriented gradient (HOG) [140] and local binary patterns (LBP) [141] features to represent the global features.

Therefore, we extracted the following three kinds of features:
36-dimensional: HOG feature.
33-dimensional: LBP feature.
48-dimensional: Gabor transformation feature with the mean and standard deviation of the face images transformed by each filter.

5.2.3 Audio Feature Extraction

For emotional speech, a good reference model is the human hearing system. Previous works have explored several different types of features. In this thesis, three of the most popular audio features, Prosodic, MFCC and Formant Frequency (FF), are utilized to represent audio characteristics in emotion recognition.

The collected emotional data usually contain noise due to the background and "hiss" of the recording machine. Generally, the presence of noise will corrupt the signal, and make the feature extraction and classification less accurate. In this work, we perform noise reduction by thresholding the wavelet coefficients [91]. Leading and trailing edges are then eliminated since they do not provide useful information. To perform spectral analysis for feature extraction, the preprocessed speech signal is segmented into speech frames using a Hamming window of 512 points with 50% overlap.

Prosodic Feature

Prosody is mainly related to the rhythmic aspects of the speech, and is normally represented by the statistics and variations of fundamental frequency, intensity, speaking rate, etc. In this thesis, we extracted 25

Index	Feature Description
1	Pitch Mean,
2	Pitch Median
3	Pitch Standard Deviation
4	Pitch Max
5	Pitch Range
6	Pitch Variation Rate
7	Rising/Falling Ratio
8	Rising Pitch Slope Max
9	Falling Pitch Slope Max
10	Rising Pitch Slope Mean
11	Falling Pitch Slope Mean
12	Pitch Rising Range Max
13	Pitch Falling Range Max
14	Pitch Rising Range Mean
15	Pitch Falling Range Mean
16	Overall Pitch Slope Mean
17	Overall Pitch Slope Standard Deviation
18	Overall Pitch Slope Median
19	Energy Mean (dB)
20	Energy Median (dB)
21	Energy Standard Deviation (dB)
22	Energy Max (dB)
23	Energy Range (dB)
24	Average Pause Length
25	Speaking Rate

Figure 5.2: Extracted prosodic features

prosodic features as listed in Figure 5.2 [91]. The pitch is estimated based on the Fourier analysis of the logarithmic amplitude spectrum of the signal [142]. The energy features are extracted in time domain and represented in decibel (dB). Pitch variation rate R_{var} and pitch rising/falling ratio R_{rf} are calculated respectively as

$$R_{var} = \frac{N_{rise} + N_{fall}}{N_{frame}} \tag{5.1}$$

$$R_{rf} = \frac{N_{rise}}{N_{fall}} \tag{5.2}$$

where N_{frame} is the number of speech frames, N_{rise} and N_{fall} are the number of speech frames with continuous rising and falling pitch respectively.

Speaking rate is approximated by

$$R_{spk} = \frac{1}{mean\_segment\_length} = \frac{N}{\sum_{i=1}^{N} T_i} \tag{5.3}$$

where T_i is the length of voiced segment i and N is the number of voiced segments. The voiced segments are defined as the segments of speech signal between pauses.

Pitch slope of each rise and each fall is calculated as

$$S_{pitch} = \frac{|f_{\max} - f_{\min}|}{t_{end} - t_{start}} \tag{5.4}$$

where $f_{\max}$ and $f_{\min}$ denote the maximum and minimum pitch value on the rise (fall) respectively. t_{start} and t_{end} represent the starting and ending time of the rise (fall).

MFCC

Mel-frequency Cepstral Coefficient (MFCC) is a popular and powerful analytical tool in the field of speech recognition. The purpose of MFCC is to mimic the behavior of human ears by applying cepstral analysis. In this thesis, the implementation of MFCC feature extraction follows the same procedure as described in [143]. The MFCC is computed based on speech frames. However, the lengths of the utterances are different, and thus the total number of coefficients is different. In order to facilitate classification, the features of each utterance mapped to the feature space should have the same length. Furthermore, with a feature vector of high dimension, the computational cost is high. Usually, in speech recognition, the total

number of coefficients being used is between nine and thirteen. This is because most of the signal energy is compacted in the first few coefficients due to the properties of the cosine transform. In this thesis, we take the first 13 coefficients and then calculate the mean, median, standard deviation, max, and min of each coefficients as the extracted features to produce a total number of 65 MFCC features.

Formant Frequency

The formant frequency estimation is based on modeling the speech signal as if it were generated by a particular kind of source and filter [91]. To find the best matching system, we use the formant frequency features. In order to make the size of the formant frequency features uniform, and come up with a compromise between the imitation efficiency of the vocal tract system and dimensionality of the feature space, we take the mean, median, standard deviation, max and min of the first three formant frequencies as the extracted features. In this way, we extract a total number of 15 formant frequency features from each utterance.

In summary, three of the audio features are described as follows:

25-dimensional Prosodic features.

65-dimensional MFCC features (the mean, median, standard deviation, max, and min of the first 13 MFCC coefficients).

15-dimensional Formant Frequency features (the mean, median, standard deviation, max and min of the first three formant frequencies).

The procedure of audio feature extraction for emotion recognition is shown in Figure 5.3.

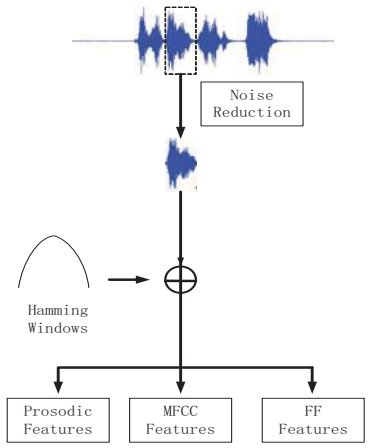

Figure 5.3: Extracted audio features

5.2.4 Visual Feature Extraction

In this thesis, we perform visual analysis by treating the face as a holistic pattern. A face detect scheme based on HSV color model is used to detect the face from the background. The visual information is represented by Gabor wavelet features.

Different approaches of face detection have been studied in the past. The face detection scheme that we used in this thesis is the Planar envelope approximation method [144] in HSV color space. After applying skin segmentation, some non-skin regions such as small isolated blobs and narrow belts are inevitably observed in the resultant image as their color falls into the range of skin color space. We apply morphological operations to implement the cleaning procedure. As shown in Figure 5.4 [91], the detected face region is mapped back to the original image, and cropped such that the major components of the face are included. The cropped face region is normalized to a gray-level image of size 128×128 as the input

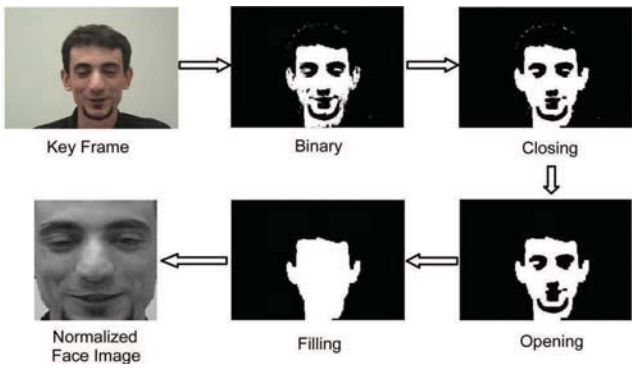

Figure 5.4: Procedure of the applied face detection scheme

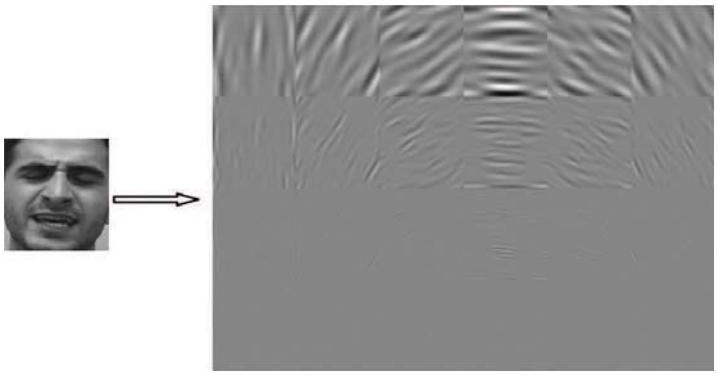

Figure 5.5: Example of Gabor wavelet transformed image

to the Gabor filter bank.

Using Gabor wavelet features to represent facial expressions have been explored and shown to be very effective in the literature [145]. It allows description of spatial frequency structure in the image while preserving information about spatial relations. In this thesis, the Gabor filter bank is designed using the algorithm proposed in [139], which consists of filters in 4 scales and 6 orientations. Figure 5.5 shows an example of Gabor wavelet transformed face image [91]. For an input image of size of 128×128, $128 \times 128 \times 4 \times 6 = 393216$ Gabor coefficients are generated. With a feature space of such high dimensionality, the computational cost is also

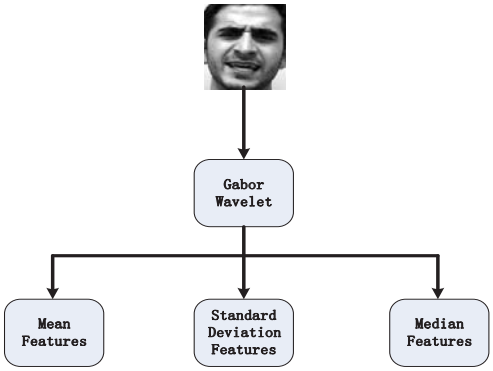

Figure 5.6: Extracted visual features

high, and thus this full feature space is unsuitable for practical applications. We therefore consider the mean, standard deviation and median of the magnitude of the transform coefficients of each filter as the features. This results in a feature vector of 72 dimensions.

In summary, three of the visual features are extracted as follows:

24-dimensional Gabor transformation features: the mean of the face images transformed by each filter.

24-dimensional Gabor transformation features: the standard deviation of the face images transformed by each filter.

24-dimensional Gabor transformation features: the median of the face images transformed by each filter.

The procedure of visual feature extraction for emotion recognition is shown in Figure 5.6.

5.3 Classification Method

For recognition, we use the algorithm proposed in [146]. The procedure is summarized below:

Given two sets of features represented by two feature matrices

$$X^1 = [x^1{}_1, x^1{}_2, x^1{}_3, ...x^1{}_d] \tag{5.5}$$

and

$$X^2 = [x^2{}_1, x^2{}_2, x^2{}_3, ...x^2{}_d]. \tag{5.6}$$

$dist[X^1X^2]$ is defined as

$$dist[X^1X^2] = \sum_{j=1}^{d} \left\| x^1{}_j - x^2{}_j \right\|_2, \tag{5.7}$$

where $\|a - b\|_2$ denotes the Euclidean distance between the two vectors a and b.

Let the feature matrices of the N training samples be $F_1, F_2, ...F_N$ and each sample belong to some class C_i $(i = 1, 2...c)$, then for a given test sample I, if

$$dist[I, F_l] = \min_j dist[I, F_j](j = 1, 2...N) \tag{5.8}$$

and

$$F_l = C_i, \tag{5.9}$$

the resulting decision is $I = C_i$.

5.4 Experimental Performance Evaluation and Analysis on DMCCA

In this section, we evaluate the effectiveness of the proposed DAF with DMCCA as the fusion function on MNIST handwritten digit database,

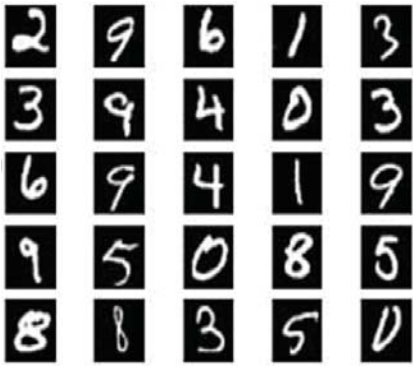

Figure 5.7: Example images from the MNIST database

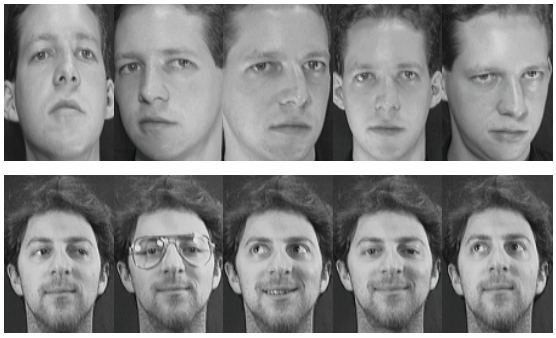

Figure 5.8: Images of two persons in the ORL database

ORL face database, and human emotion recognition on RML and eNTERFACE audiovisual databases, which corresponds implicitly to the DAF block in Figure 5.1 or explicitly in Figure 3.1.

The MNIST database, or modified NIST database, is constructed out of the original NIST database. All the digits are size normalized, and centered in a fixed size image where the center of gravity of the intensity lies at the center of the image with 28 * 28 pixels which take on binary values. Example images from MNIST database are shown in Figure 5.7. In the experiments, we select 1500 samples to form the training subset and 1500 samples as the testing subsets, respectively.

The ORL database (http://www.cam-orl.co.uk) contains images from 40 individuals, each providing 10 different images. Each image is normalized and centered in a gray-level image with size 64*64, or 4096 pixels in total. Ten sample images of two subjects from the ORL database are shown in Figure 5.8. In the experiment, the proposed algorithm is tested on the whole ORL database. The evaluation is based on cross-validation, where each time five images of each subject are randomly chosen for training, while the remaining five images are used for testing. Thus, the training sample set size is 200 and the testing sample set size is 200.

The RML database [91] consists of video samples of the six principal emotions (angry, disgust, fear, surprise, sadness and happiness), performed by eight subjects speaking six different languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian). The fame rate for the videos is 30 fps with audio recorded at a sampling rate of 22050 Hz. The spatial resolution of the image frames is 720*480 pixels, and the face region has an average size of 112*96 pixels. The eNTERFACE database [115] contains video samples from 43 subjects, also expressing the six basic emotions, with a sampling rate of 48000 Hz for English audio channel and a video frame rate of 25 fps. The image frames have a size of 720*576 pixels, with the average size of the face region around 260*300 pixels. Example facial expression images from RML and eNTERFACE are shown in Figure 5.9.

In the experiment, 288 samples of eight subjects from RML database and 456 samples of ten subjects from eNTERFACE database are selected, respectively. We divide the samples from RML database into training and testing subsets containing 192 and 96 samples, respectively. For eNTERFACE database, samples are divided into training and testing subsets

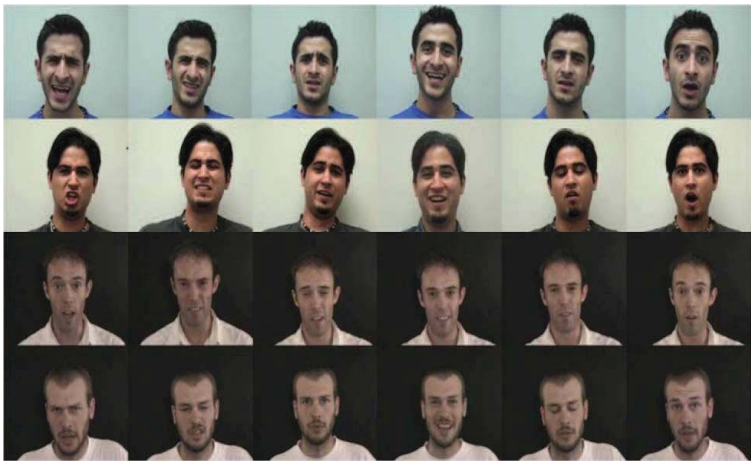

Figure 5.9: Example facial expression images from the RML (Top two rows) and eNTERFACE (Bottom two rows) Databases

including 360 and 96 samples, respectively.

5.4.1 Handwritten Digit Recognition

For handwritten digit recognition, mean, standard deviation and Zernike moments correspond to the feature extraction block in Figure 3.1. The performance of mean, standard deviation and Zernike moments is first evaluated shown in Table 5.1. The recognition accuracy is calculated as the ratio of the number of correctly classified samples over the total number of testing samples.

From Table 5.1, the standard deviation (52.60%) and Zernike moment (70.20%) features achieve better performance than the mean (49.13%), and therefore will be used in CCA and DCCA which only take two sets of features. In addition, the performance based on the method of serial fusion with standard deviation & Zernike moment and that with all the

Table 5.1: Results of handwritten digit recognition with a single feature

Single Feature	Recognition Accuracy
Mean	49.13%
Standard Deviation	52.60%
Zernike	70.20%

three features are implemented, respectively. The experimental results are shown in Table 5.2.

Table 5.2: Results of handwritten digit recognition by serial fusion

Serial Fusion	Recognition Accuracy
Standard Deviation & Zernike	70.20%
All of the three features	70.33%

Next, the comparison among DMCCA, serial fusion, CCA, MCCA, and DCCA are implemented. The overall recognition rates are given in Figure 5.10, with DMCCA providing the best performance, clearly showing the discriminative power of the DMCCA for information fusion in handwritten digit recognition. From the figure, it is clear that the application of DMCCA achieves the best performance when the projected dimension d equals to $9 < 10 = c$, the number of classes, confirming nicely with the mathematical analysis in Chapter 3. Moreover, the optimal recognition accuracies with different methods are presented in Table 5.3.

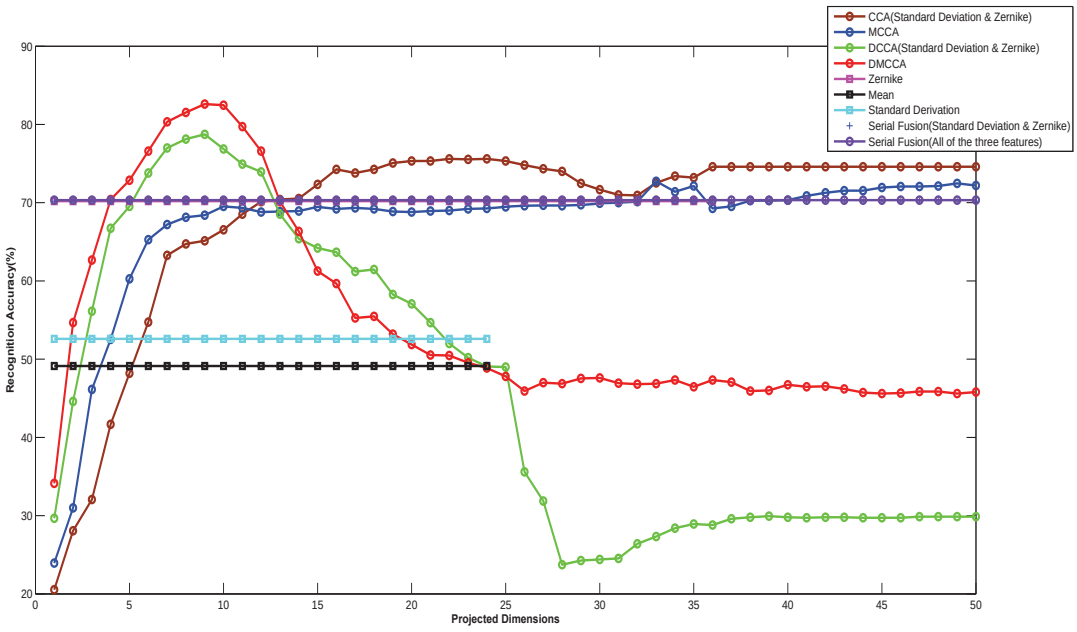

Figure 5.10: Handwritten digit recognition experimental results of different methods on MNIST Database

5.4.2 Face Recognition

For face recognition, HOG, LBP and Gabor features in a face image correspond to the feature extraction block in Figure 3.1.

The performance of using HOG, LBP and Gabor features is shown in Table 5.4. From Table 5.4, it suggests we use the HOG (90.50%) and Gabor (85.50%) which provide the best individual performance as the input to CCA and DCCA. We also experimented on the method of serial fusion with HOG & Gabor features, and all the three features, respectively, with the performance shown in Table 5.5.

Table 5.3: The optimal handwritten digit recognition accuracies with different methods

Methods	Optimal Recognition Accuracy
DMCCA	82.60%
MCCA	73.60%
DCCA	79.27%
CCA	75.60%

Table 5.4: Results of face recognition with a single feature

Single Feature	Recognition Accuracy
HOG(ORL)	90.50%
LBP(ORL)	77.50%
Gabor(ORL)	85.50%

The performance by the methods of CCA, MCCA, DCCA, and DMCCA is shown in Figure 5.11. From the experimental results, clearly, DMCCA provides more effective modeling to handle the face recognition problem. Moreover, DMCCA achieves the optimal performance when the projected dimension d is equal to 28, which is less than the number of classes (c=40). The optimal recognition accuracies with different methods are presented in Table 5.6.

5.4.3 Emotion Recognition

In this subsection, we will first evaluate feature fusion in emotion recognition using audio features and visual features, respectively. Then the

Table 5.5: Results of face recognition by serial fusion

Serial Fusion	Recognition Accuracy
HOG & Gabor(ORL)	77.50%
All of the three features(ORL)	77.50%

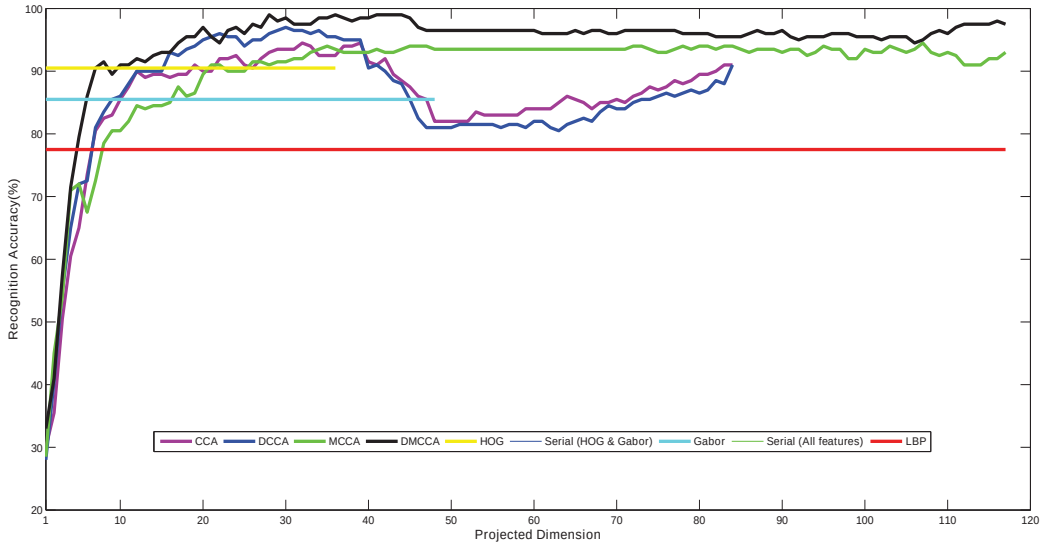

Figure 5.11: Face recognition experimental results of different methods on ORL Database

evaluation will move on to audiovisual bimodal emotion recognition.

5.4.3.a Audio Emotion Recognition

In the experiments of audio emotion recognition, Prosodic, MFCC and Formant Frequency features correspond to the feature extraction block in Figure 3.1. For bench mark purpose, the performance of using Prosodic, MFCC and Formant Frequency features in emotion recognition is first evaluated, and tabulated in Table 5.7.

Table 5.6: The optimal face recognition accuracies with different methods

Methods	Optimal Recognition Accuracy
DMCCA	98.00%
MCCA	94.50%
DCCA	97.00%
CCA	94.50%

Table 5.7: Results of emotion recognition with single audio feature

Single Feature	Recognition Accuracy
Prosodic(RML)	45.83%
MFCC(RML)	34.38%
Formant Frequency(RML)	22.92%
Prosodic(eNTERFACE)	55.21%
MFCC(eNTERFACE)	39.58%
Formant Frequency(eNTERFACE)	31.25%

Table 5.7 suggests we should use the Prosodic (45.83%, 55.21%) and MFCC (34.38%, 39.58%) features which perform better than Formant Frequency individually, in CCA and DCCA which only need to take two sets of features. We also experimented on the method of serial fusion on RML and eNTERFACE databases with Prosodic & MFCC features, and all of the three features, respectively. The results are shown in Table 5.8.

Then, we compare the performance of DMCCA with serial fusion, C-

Table 5.8: The experimental results of audio emotion recognition with serial fusion

Serial Fusion	Recognition Accuracy
Prosodic & MFCC(RML)	36.46%
All of the three features(RML)	29.17%
Prosodic & MFCC(eNTERFACE)	40.63%
All of the three features(eNTERFACE)	34.38%

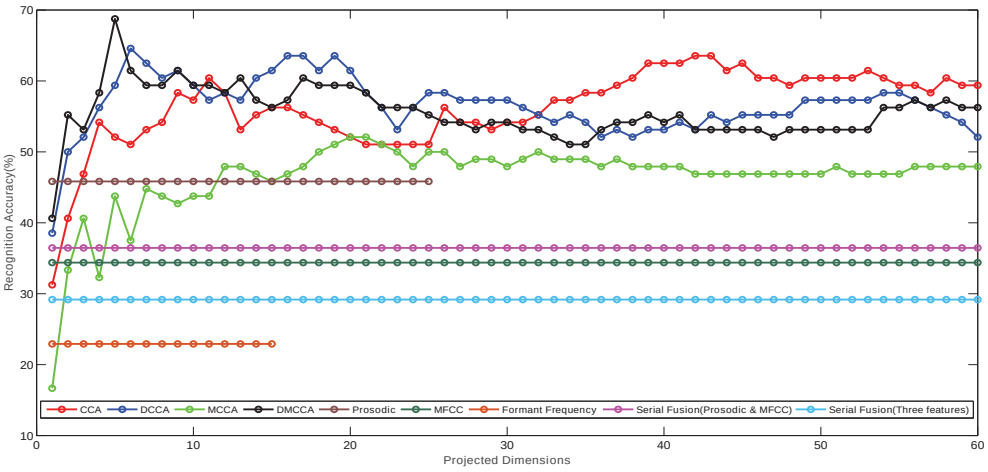

Figure 5.12: Audio emotion recognition experimental results of different methods on RML Database

CA, MCCA and DCCA. The overall recognition accuracies are shown in Figure 5.12 and Figure 5.13. Moreover, the optimal recognition accuracies with different methods are presented in Table 5.9. Clearly, the discrimination power of the DMCCA provides a more effective modelling of the relationship between different features and achieves better performance than the other methods.

Table 5.9: The optimal audio emotion recognition accuracies with different methods

Methods	Optimal Recognition Accuracy
DMCCA (RML)	68.75%
MCCA (RML)	54.17%
DCCA (RML)	64.58%
CCA (RML)	63.54%
DMCCA (eNTERFACE)	72.92%
MCCA (eNTERFACE)	63.54%
DCCA (eNTERFACE)	68.75%
CCA (eNTERFACE)	65.63%

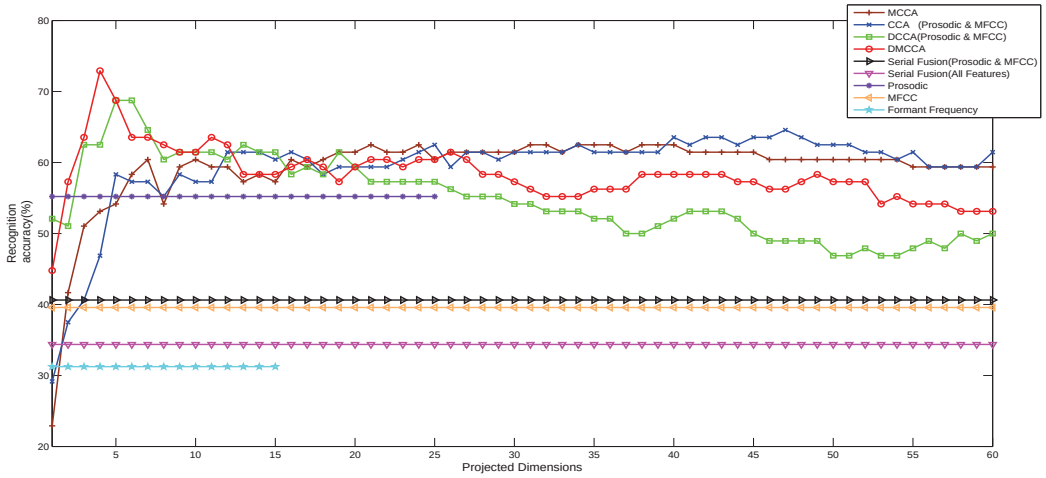

Figure 5.13: Audio emotion recognition experimental results of different methods on eNTERFACE Database

5.4.3.b Visual Emotion Recognition

In the experiments of visual emotion recognition, mean, standard deviation and median features correspond to the feature extraction block in Figure 3.1. We apply DMCCA to fuse the visual features extracted from RML and eNTERFACE databases, respectively. For benchmark purpose, the performance of using mean, standard deviation and median features is evaluated. The results are shown in Table 5.10.

Table 5.10: Results of visual emotion recognition with single Gabor feature

Single Feature	Recognition Accuracy
Mean(RML)	60.42%
Standard Deviation(RML)	65.63%
Median(RML)	56.25%
Mean(eNTERFACE)	75.00%
Standard Deviation(eNTERFACE)	80.21%
Median(eNTERFACE)	72.92%

From Table 5.10, it is observed that the features of mean (60.42%, 75.00%) and standard deviation (65.63%, 80.21%) achieve better performance in visual emotion recognition compared with the feature of median (56.25%, 72.92%). Thus, in the following experiments, we will use mean and standard deviation features in CCA and DCCA. The experiments using serial fusion with mean & standard deviation features, and all of the three features are also performed and summarized in Table 5.11.

Table 5.11: The experimental results of visual emotion recognition with serial fusion

Serial Fusion	Recognition Accuracy
Mean & Standard Deviation(RML)	71.83%
All of the three features(RML)	64.58%
Mean & Standard Deviation(eNTERFACE)	79.17%
All of the three features(eNTERFACE)	80.21%

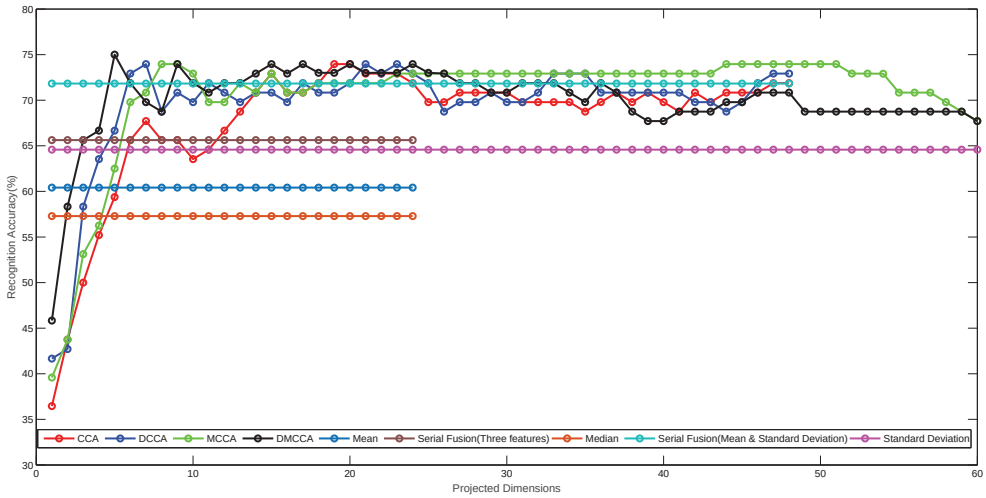

Figure 5.14: Visual emotion recognition experimental results of different methods on RML Database

The overall recognition results are illustrated in Figure 5.14 and Figure 5.15. In addition, the optimal recognition accuracies with different methods are presented in Table 5.12. Again, it shows that the proposed DMCCA outperforms the other methods.

Table 5.12: The optimal visual emotion recognition accuracies with different methods

Methods	Optimal Recognition Accuracy
DMCCA (RML)	76.04%
MCCA (RML)	72.92%
DCCA (RML)	72.92%
CCA (RML)	72.92%
DMCCA (eNTERFACE)	82.29%
MCCA (eNTERFACE)	77.08%
DCCA (eNTERFACE)	80.21%
CCA (eNTERFACE)	76.04%

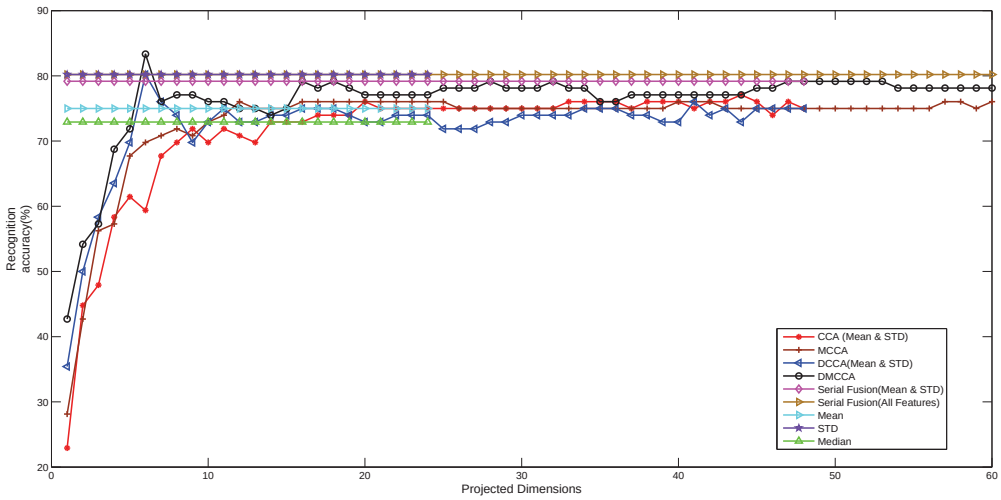

Figure 5.15: Visual emotion recognition experimental results of different methods on eNTERFACE Database

5.4.3.c Audiovisual Emotion Recognition

For audiovisual emotion recognition, audio features (Prosodic, MFCC & Formant Frequency) and visual features (mean, standard deviation & median) correspond to the left most feature extraction block in Figure 3.1. From the previous experiments, it is shown that the Prosodic features in audio and standard deviation of Gabor Transform coefficients in visual images are more likely to result in better performance in emotion recognition compared with other features. Therefore, in the following, we will use Prosodic and standard deviation for the methods of CCA and DCCA in audiovisual multimodal fusion. Besides, the results of serial fusion on all the six audiovisual features are also investigated, and the overall recognition accuracy is 30.28% for RML database and 35.42% for eNTERFACE database. The performance by the methods of serial fusion, CCA, MCCA, DCCA, audio multi-feature DMCCA, visual multi-feature DMCCA and audiovisual DMCCA for the two datasets are shown in Figure 5.16 and Figure 5.17, respectively. Moreover, the optimal recognition accuracies with different methods are presented in Table 5.13.

5.4.4 Computational Efficiency

From the experimental results, clearly, the discrimination power of the DMCCA provides a more effective modeling of the relationship among multiple information sources. Another advantage of DMCCA is the computational efficiency, especially when the number of classes being studied is small. Without loss of generality, we take the audiovisual emotion recognition as an example to demonstrate this advantage. Since Ekman's six

Table 5.13: The optimal audiovisual emotion recognition accuracies with different methods

Methods	Optimal Recognition Accuracy
DMCCA (RML)	82.29%
MCCA (RML)	77.08%
DCCA (RML)	68.75%
CCA (RML)	61.46%
DMCCA (eNTERFACE)	85.42%
MCCA (eNTERFACE)	80.17%
DCCA (eNTERFACE)	77.08%
CCA (eNTERFACE)	64.00%

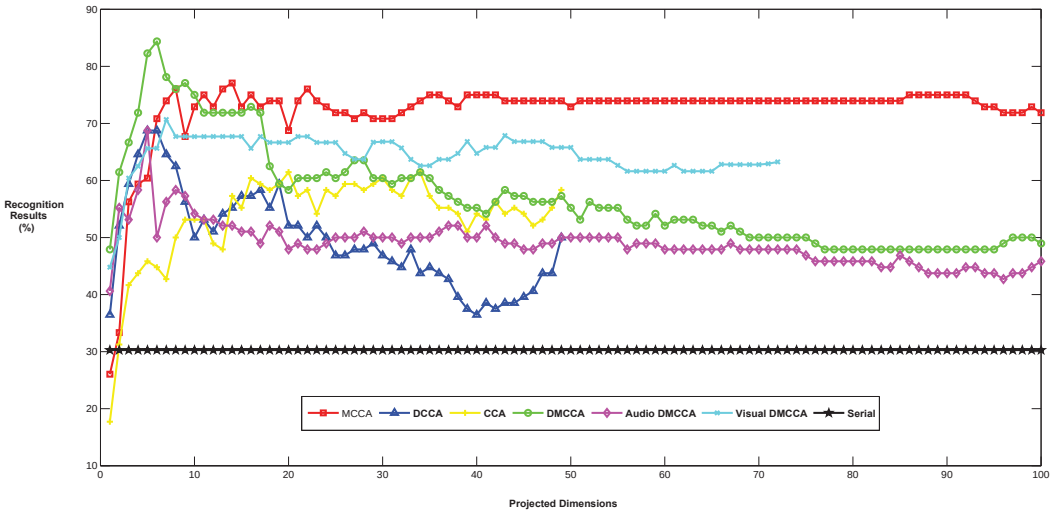

Figure 5.16: Audiovisual emotion recognition experimental results by different methods on RML Database

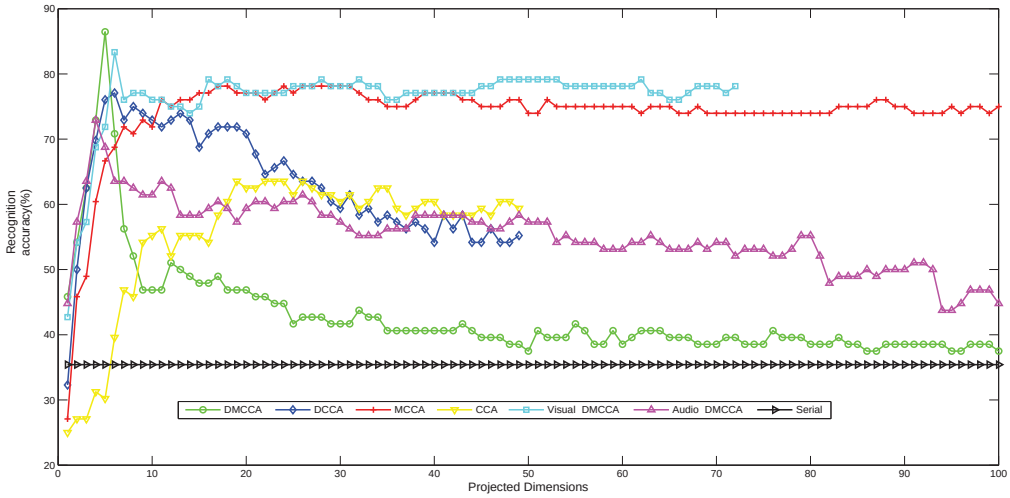

Figure 5.17: Audiovisual emotion recognition experimental results by different methods on eNTERFACE Database

basic emotional states are used in the work, c equals to six and the dimension of audiovisual features (M=177) is equal to dimension of audio features (105) plus dimension of visual features (72). Therefore, the ratio of $O(M^*c)$ to $O(M^*M)$ is about 1:30 and the level of efficiency by the proposed over MCCA is quite significant. To further show the efficiency of the proposed method, we investigate the actual running time of the proposed method and that of the MCCA in emotion recognition. All experiments are performed on a PC with Windows 7 operation system, Intel i7-3.07GHz CPU & 10 G RAM and the algorithms are coded in MATLAB 2013b. For the RML database, the running time of the proposed method is 129.43s while that of MCCA is 11043s. The ratio of computational times is 129.43 : 11043 = 1: 85.3. For the eNTERFACE Database, the running time of the proposed method is 224.3s while that of MCCA is 15048s. The

ratio is 224.3 : 15048 = 1: 67. In both cases, the proposed method shows remarkable advantage in terms of computational efficiency over MCCA.

5.4.5 Comparison with the Method of Embedding DCCA (EDCCA)

To further demonstrate the effectiveness of DMCCA on multimodal fusion, we applied the method of DCCA with embedding (named as EDCCA) to bimodal emotion recognition and compared with DMCCA. There are six sets of features for bimodal emotion recognition. Three are audio: Prosodic, MFCC and Formant Frequency; and three are visual: Mean, Standard Deviation and Median calculated from Gabor wavelet transform. Naturally, we embed the three audio features together and the three visual features together, and perform EDCCA on the two embedded features. The overall recognition accuracies by EDCCA on RML and eNTERFACE datasets are shown in Figure 5.18 and Figure 5.19.

From Figure 5.16 to Figure 5.19, we observe the following in terms of optimal performance:
1) RML: 82.29% (DMCCA, Green line in Figure 5.16) > 79% (EDCCA) > 69% (DCCA, Blue line in Figure 5.16)
2) eNTERFACE: 85.42% (DMCCA, Green line in Figure 5.17) > 78% (EDCCA) > 77% (DCCA, Blue line in Figure 5.17)

Therefore, properly embedding all features in two sets did improve the performance of DCCA in this example, but it is still inferior to the performance of DMCCA. Moreover, when the fusion involves features from three or more modalities, it is difficult, if not impossible, to design a reasonable embedding strategy. On the other hand, with a sound theoretical founda-

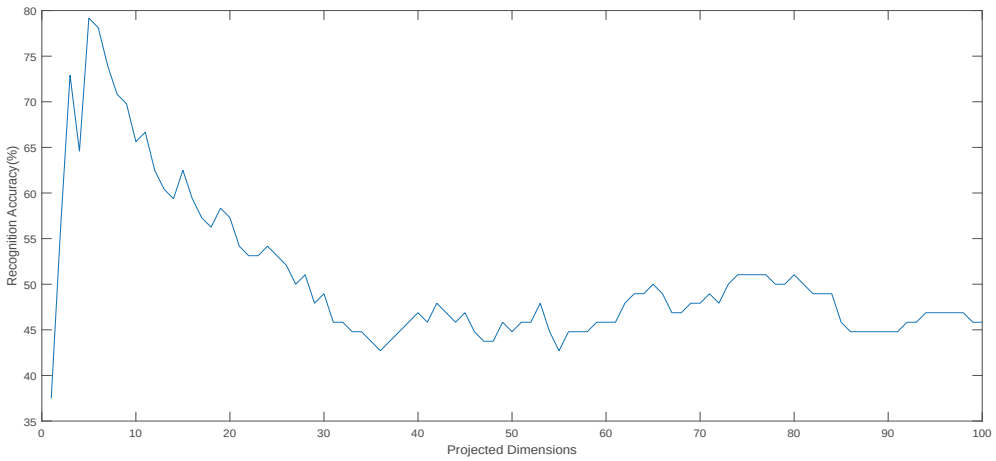

Figure 5.18: Performance on audiovisual fusion on emotion recognition with the method of EDCCA (RML Dataset)

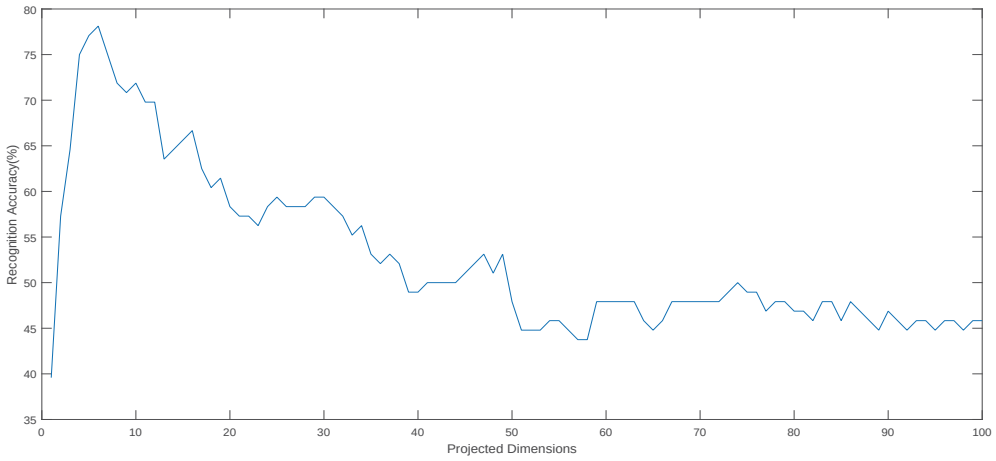

Figure 5.19: Performance on audiovisual fusion on emotion recognition with the method of EDCCA (eNTERFACE Dataset)

tion, DMCCA can handle fusion involving any number of modalities.

From the above experimental results, it can be seen that the recog-

nition accuracy of serial fusion is generally worse than CCA and related methods, and fusion does not help as shown in Table 5.2, Table 5.8 and Table 5.11, justifying that simply putting the features from different channels together without considering the intrinsic structure and relationship results in low recognition accuracy.

An important finding of the research is that, the exact location of optimal recognition performance occurs when the number of projected dimension d is smaller than or equals to the number of classes, confirming nicely with the mathematical analysis presented in Chapter 3. The significance here is that, we only need to calculate the first d (is smaller than or equals to the number of classes) projected dimensions of DMCCA to obtain the desired recognition performance, eliminating the need of computing the complete transformation processes associated with most of the other methods, and thus substantially reducing the computational complexity to obtain the optimal recognition accuracy.

5.4.6 Graphical Identification of The Optimal Performance by DMCCA

In this subsection, we present the calculation of $J(\eta)$ with DMCCA for selecting optimal projection with the results shown in Figure 5.20 to Figure 5.23, which graphically illustrate the relationship between optimal projected dimensions and the recognition performance using the proposed criterion $J(\eta)$ in equation (3.63). In Figure 5.20 and Figure 5.21, criterion $J(\eta)$ reaches the maximum when the projected dimension is nine for MNIST database and 28 for ORL database, respectively. In the Figures 5.22, criterion $J(\eta)$ reaches the maximum when the projected dimension

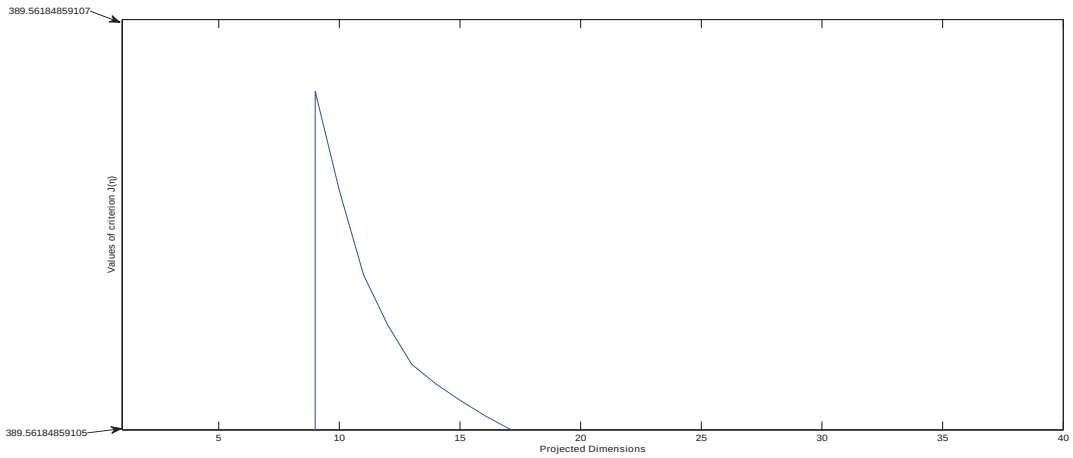

Figure 5.20: The calculation of $J(\eta)$ with the DMCCA for handwritten digit recognition on MNIST Database

is six for RML database which is equal to the number of classes (c=6). Similarly, the dimension of five is observed for the eNTERFACE database as shown in Figure 5.23. The graphical presentation again confirms nicely with the mathematical analysis presented in Chapter 3.

5.5 Performance Evaluation and Analysis with KECA plus DMCCA

We applied the DAF with KECA plus DMCCA (KECA+DMCCA) as the fusion function on handwritten digit recognition, face recognition, human emotion recognition, which corresponds implicitly to the DAF block in Figure 5.1 or explicitly in Figure 4.1. The same experiment setup as that used in testing DMCCA is employed for consistence.

To demonstrate the effectiveness of the entropy estimation in informa-

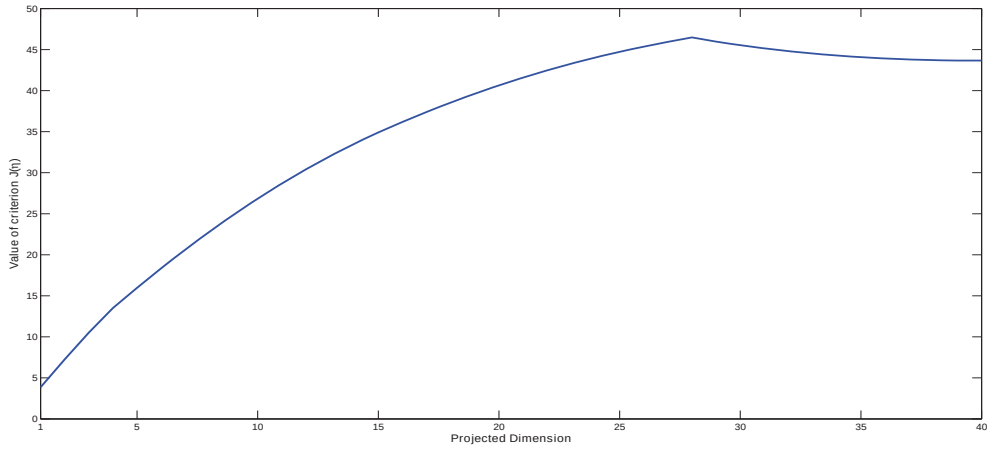

Figure 5.21: The calculation of $J(\eta)$ with the DMCCA for face recognition on ORL Database

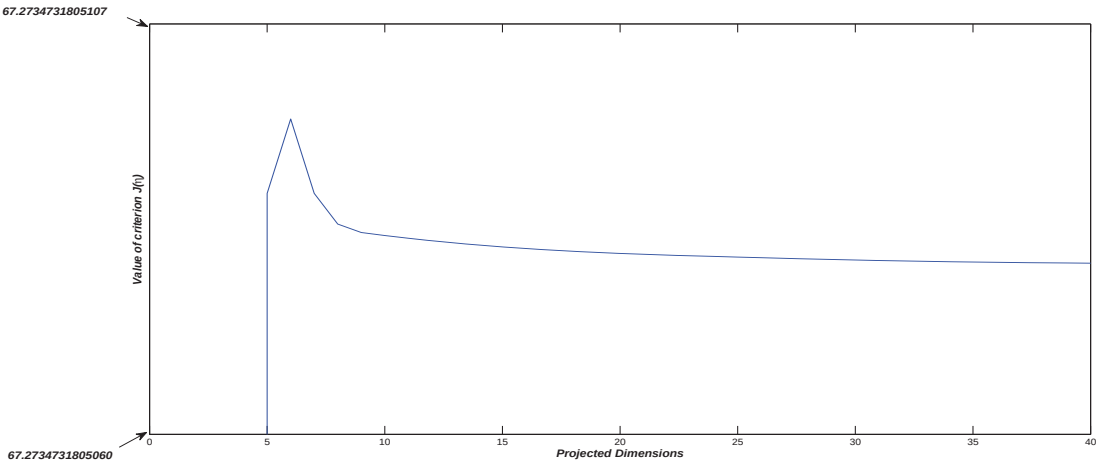

Figure 5.22: The calculation of $J(\eta)$ with the DMCCA for audiovisual emotion detection on RML Database

tion fusion, the method of KPCA is also implemented and the optimal accuracy is given in different experiments.

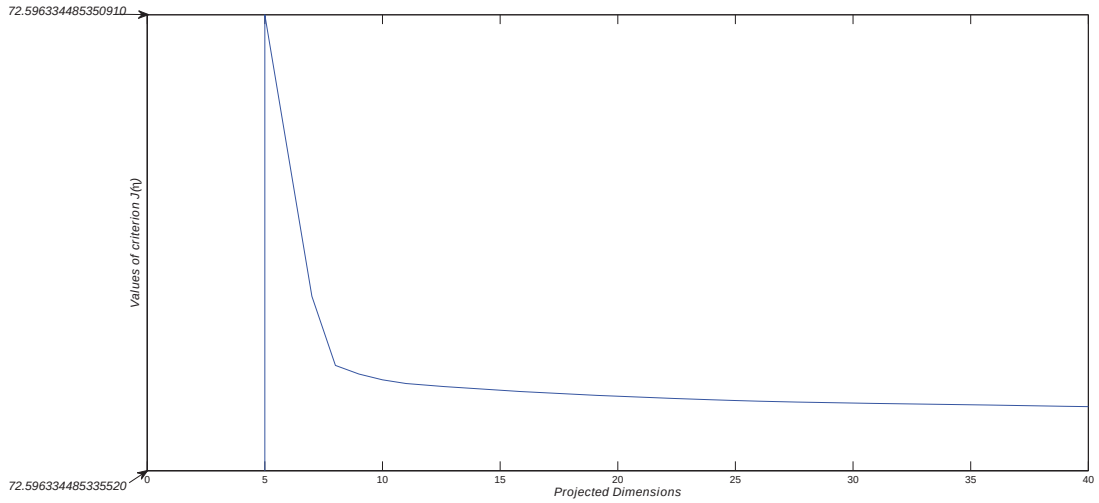

Figure 5.23: The calculation of $J(\eta)$ with the DMCCA for audiovisual emotion detection on eNTERFACE Database

5.5.1 Handwritten Digit Recognition

During the experiments, we also implemented the methods of DMCCA and KECA for the purpose of comparison. Since the kernel functions and the corresponding parameters affect the performance of kernel based algorithms significantly, we have conducted extensive experiments using Gaussian functions with σ = 1, 10, 100, 1000, 10000. The performance comparison is shown in Figure 5.24 and the optimal accuracies with different methods are given in Table 5.14. Obviously, the proposed KECA+DMCCA outperforms DMCCA, KECA and KPCA.

5.5.2 Face Recognition

In the experiment, similarly, the training sample set size is 200 and the testing sample set size is 200. The Gaussian functions with σ =1, 10,

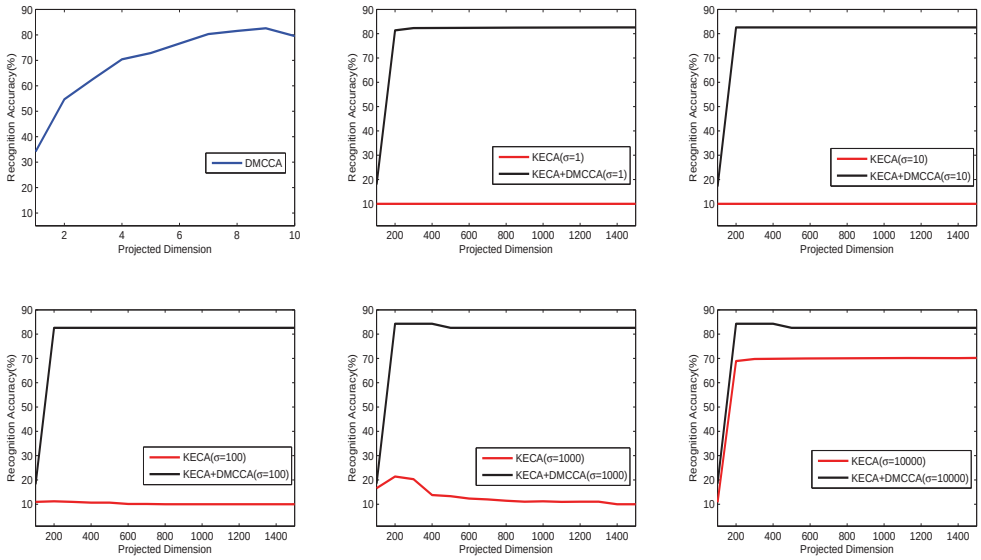

Figure 5.24: Experimental results of handwritten digit recognition with DMCCA, KECA and KECA+DMCCA on MNIST database (σ =1, 10, 100, 1000, 10000)

100, 1000, 10000 is implemented to demonstrate the effectiveness of the proposed framework. Then, the performance by the methods of DMCCA, KECA, and KECA+DMCCA is shown in Figure 5.25 and the optimal recognition accuracies are shown in Table 5.15. From the experimental results, clearly, KECA+DMCCA provides more effective modeling to handle the face recognition problem.

5.5.3 Emotion Recognition

5.5.3.a Audio Emotion Recognition

In the experiments, we used Gaussian functions as the kernel with σ =1, 10, 100, 1000, 10000. Then, we compare the performance of KE-

Table 5.14: The optimal handwritten digit recognition accuracies with different methods

Methods	Optimal Recognition Accuracy
KPCA	67.47%
KECA	70.27%
DMCCA	82.60%
KECA+DMCCA	84.27%

Table 5.15: The optimal face recognition accuracies with DMCCA, KECA, KPCA and KECA+DMCCA

Methods	Optimal Recognition Accuracy
KPCA	85.00 %
KECA	94.50%
DMCCA	98.00%
KECA+DMCCA	99.00%

CA+DMCCA with DMCCA and KECA. The overall recognition accuracies are shown in Figure 5.26 and Figure 5.27. Furthermore, the optimal recognition accuracies with different methods are summarized in Table 5.16. Clearly, KECA+DMCCA outperforms DMCCA, KECA and KPCA.

Note, KECA merely puts the information or features from different channels together without considering the intrinsic structure and relationship. Therefore, when it is used in information fusion and there are very different performances among original features such as the performances

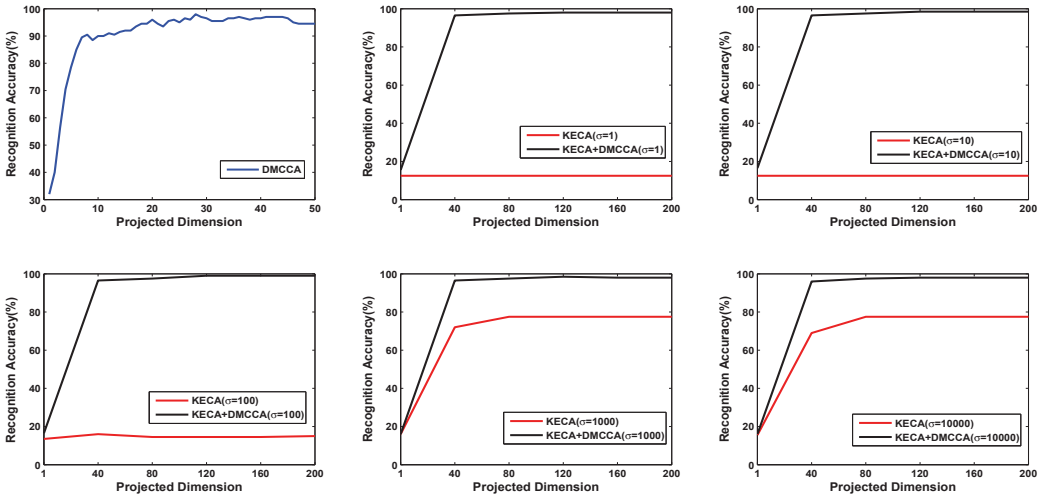

Figure 5.25: Experimental results of face recognition with DMCCA, KECA and KECA+DMCCA on ORL database (σ =1, 10, 100, 1000, 10000)

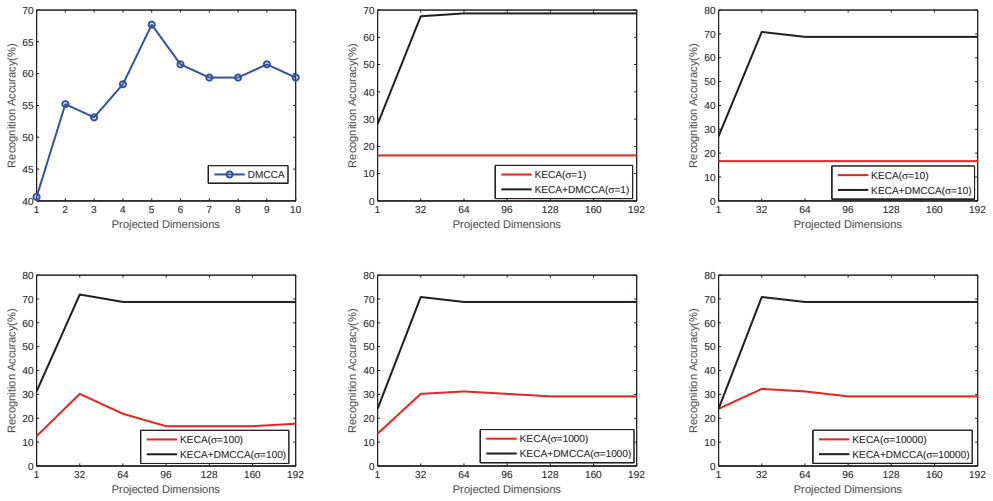

Figure 5.26: Experimental results of audio emotion recognition with DMCCA, KECA and KECA+DMCCA on RML database (σ =1, 10, 100, 1000, 10000)

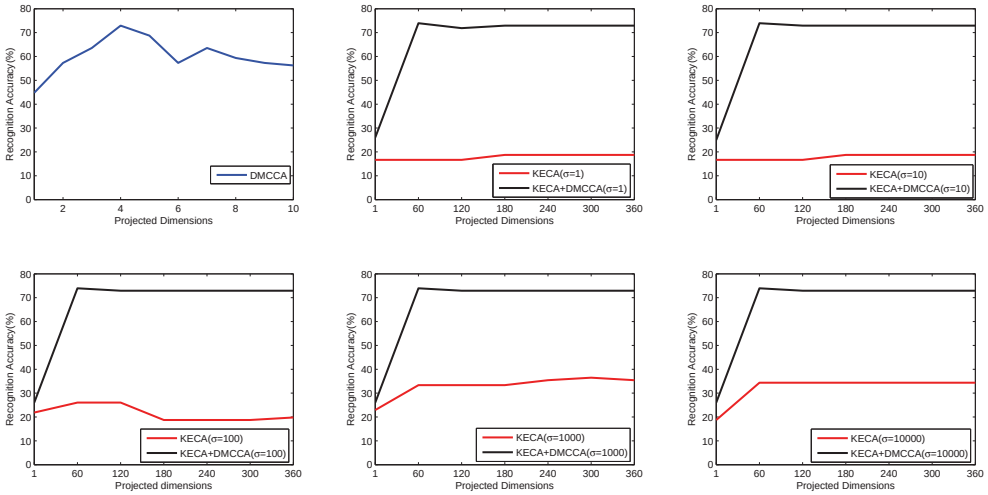

Figure 5.27: Experimental results of audio emotion recognition with DMCCA, KECA and KECA+DMCCA on eNTERFACE database (σ =1, 10, 100, 1000, 10000)

of Prosodic(45.83%, 55.21%), MFCC(34.38%, 39.38%) and Formant Frequency(22.92%, 31.25%) on RML and eNTERFACE database, there is no guarantee that KECA achieves higher recognition accuracy than the single feature. On the other hand, in KECA+DMCCA, not only the discriminative representations are considered by DMCCA, but also the complementary representations of the input data are revealed in the space of KECA, improving the recognition or accuracy than the original features.

5.5.3.b Visual Emotion Recognition

In this subsection, we conduct experiments using KECA+DMCCA on RML and eNTERFACE visual emotion database, respectively. The overall recognition accuracies are shown in Figure 5.28 and Figure 5.29. In addition, the optimal recognition accuracies with different methods are

Table 5.16: The optimal audio emotion recognition accuracies with DMCCA, KECA, KPCA and KECA+DMCCA

Methods	Optimal Recognition Accuracy
KPCA (RML)	27.08%
KECA (RML)	32.29%
DMCCA (RML)	68.75%
KECA+DMCCA (RML)	71.88%
KPCA (eNTERFACE)	25.00%
KECA (eNTERFACE)	35.41%
DMCCA (eNTERFACE)	72.92%
KECA+DMCCA (eNTERFACE)	75.00%

presented in Table 5.17. Again, the comparison shows that the proposed KECA+DMCCA outperforms the other methods.

5.5.3.c Audiovisual Emotion Recognition

In the following experiments, three audio features (Prosodic, MFCC, Formant Frequency) and three visual features (Mean, Standard Deviation, Median) are used. The performance of DMCCA, KECA and KECA+DMCCA is illustrated in Figure 5.30 and Figure 5.31. The optimal recognition accuracies with different methods are given in Table 5.18.

From the above experimental results, it can be seen that the recognition accuracies of KPCA and KECA are generally worse than DMCCA and KECA+DMCCA justifying that simply putting the features from d-

Table 5.17: The optimal visual emotion recognition accuracies with DMCCA, KECA, KPCA and KECA+DMCCA

Methods	Optimal Recognition Accuracy
KPCA (RML)	57.29%
KECA (RML)	67.71%
DMCCA (RML)	76.04%
KECA+DMCCA (RML)	78.13%
KPCA (eNTERFACE)	20.83%
KECA (eNTERFACE)	77.08%
DMCCA (eNTERFACE)	82.29%
KECA+DMCCA (eNTERFACE)	84.38%

ifferent channels together without considering the intrinsic relationship and discriminative representation results in low recognition accuracy. On the other hand, since the discriminative representations are considered by DMCCA and the complementary representation of the input data is revealed by KECA, the performance of KECA+DMCCA is better than the other methods compared in all cases.

5.6 Summary

In this chapter, the proposed DAF with DMCCA and KECA+DMCCA as the fusion functions is applied to handwritten digit recognition, face recognition and emotion recognition.

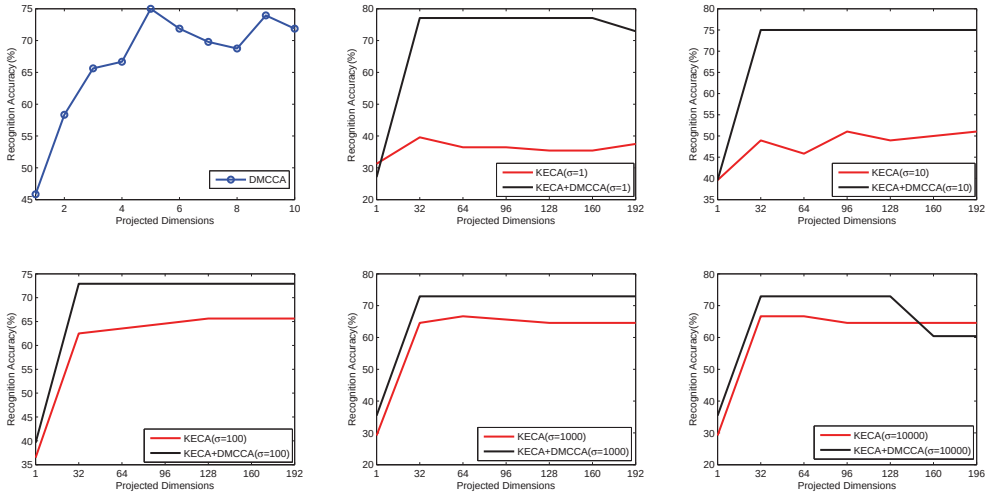

Figure 5.28: Experimental results of visual emotion recognition with DMCCA, KECA and KECA+DMCCA on RML database (σ =1, 10, 100, 1000, 10000)

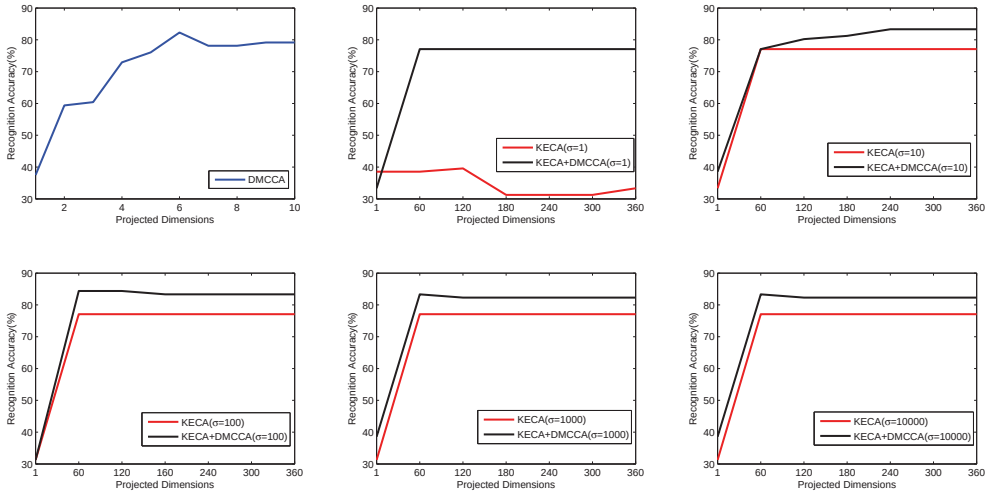

Figure 5.29: Experimental results of visual emotion recognition with DMCCA, KECA and KECA+DMCCA on eNTERFACE database (σ =1, 10, 100, 1000, 10000)

Table 5.18: The optimal audiovisual emotion recognition accuracies with DMCCA, KECA, KPCA and KECA+DMCCA

Methods	Optimal Recognition Accuracy
KPCA (RML)	27.08%
KECA (RML)	37.50%
DMCCA (RML)	82.29%
KECA+DMCCA (RML)	86.46%
KPCA (eNTERFACE)	26.04%
KECA (eNTERFACE)	40.61%
DMCCA (eNTERFACE)	85.42%
KECA+DMCCA (eNTERFACE)	88.54%

Since DMCCA can be seen as a way of guiding discriminative feature selection toward the underlying semantics to find basis vectors for different sets of variables, it reveals discriminative representations among different multiple variables. In addition, based on the definition of canonical correlation, the transformed sets of linear combinations are those with the largest correlation subject to the condition that they are orthogonal to the former canonical variables. Therefore, it also eliminates redundant information effectively. Hence DMCCA improves recognition performance with substantially reduced dimensionality of the feature space, leading to efficient practical pattern recognition.

Different form the methods of DMCCA and KECA, KECA+DMCCA transforms the multiple input information/data into the discriminative multiple canonical correlation analysis space at first. After that, KECA

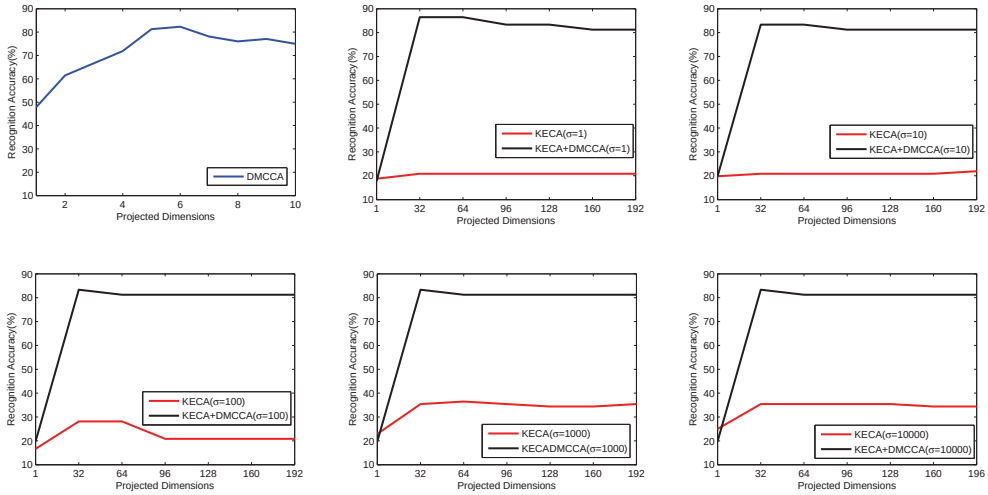

Figure 5.30: Experimental results of audiovisual emotion recognition with DMCCA, KECA and KECA+DMCCA on RML database (σ =1, 10, 100, 1000, 10000)

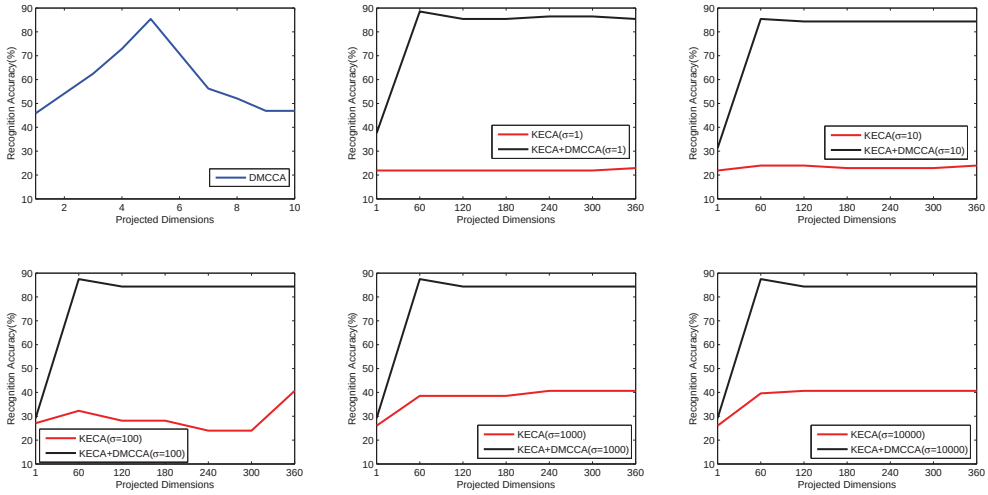

Figure 5.31: Experimental results of audiovisual emotion recognition with DMCCA, KECA and KECA+DMCCA on eNTERFACE database (σ =1, 10, 100, 1000, 10000)

is applied to the discriminative vectors in the DMCCA space. Therefore, not only the complementary representations of input data are revealed by entropy estimation, but also the discriminative representations are considered by DMCCA. After processed by the proposed fusion method, most of useful information is properly preserved and improved recognition accuracy is achieved. Experimental results show that the proposed DAF outperforms the existing methods based on similar principles.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

With the rapid development of advanced multi-disciplinary technologies for acquiring, storing and transmitting massive amount of data, multi-modal information processing has attracted rapidly growing attention recently, in both academia and industry. Multi-modal data/information research challenges, particularly related to fusion and perception, are ubiquitous in diverse domains, such as Internet of things, Robotics, Manufacturing, Engineering, Natural Language Processing (NLP) and Medical Informatics. Next-generation cognitive agents will require to be appropriately equipped with multi-modal information fusion and perception capabilities to carry out cognitive tasks such as perception, action, affective and cognitive learning, decision making and control, social cognition, language processing and communication, reasoning, problem solving, and consciousness.

Despite recent progress in the multi-disciplinary area of multi-modal fusion, there remain outstanding challenges for effectively exploiting multi-

modal information in practical environments, in particular for untapped real-world applications in diverse disciplines. In this thesis, a discriminative analysis framework (DAF) is introduced to handle multi-modal information fusion. First, discriminative multi-modal canonical correlation analysis (DMCCA) is proposed for multi-modal information fusion to extract the discriminative representation from the input multi-modal data. After that, DMCCA is integrated with kernel entropy component analysis (KECA) to further improve the performance of the DAF for multi-modal information fusion. Then the proposed DAF is applied to handwritten digit recognition, face recognition and emotion recognition to demonstrate the generic nature and effectiveness of the proposed framework.

After a comprehensive background study in Chapter 2, we introduce DMCCA for multi-modal analysis and fusion in Chapter 3. DMCCA finds projection directions to maximize the within-class correlation and minimize the between-class correlation among multiple information/data sources in order to identify the discriminative representation among different modalities effectively. In addition, we verify that the best performance by discriminative representation achieves when only a small fraction of the data needs to be analyzed. Furthermore, a unified framework for canonical correlation analysis is established for information fusion in the transformed domain. Finally, we present a method on graph representation for selecting optimal projection in multi-modal information fusion.

In Chapter 4, firstly, we study entropy estimation and KECA, which are expected to reveal more complementary representations than the second order statistics from the multiple input sources. After that, we investigated the proposed discriminative method KECA plus DMCCA (KE-

CA+DMCCA) for information fusion. Based the proposed method, not only the discriminative representations are considered by DMCCA, but also the complementary representations of input data are revealed in the space of KECA, resulting in further improved recognition performance.

In Chapter 5, we evaluate the effectiveness of the proposed framework on MNIST handwritten digit database, ORL face database, RML emotion database and eNTERFACE emotion database. Experimental results show that the proposed framework outperforms the methods based on similar principles.

6.2 Future Work

Based on the current work, we propose the following possible directions for future research. In this thesis, a connection between information theory and information fusion has been built, but the concepts of information theory studied in the thesis are only entropy. Actually, there are more sophisticated tools in information theory such as joint entropy, mutual entropy, mutual information, etc., which can potentially help explore more complementary representations among different modalities. The application of information theory enables us to consider the problem of information fusion from the viewpoint of the nature of information instead of statistics. We believe that the direction of integrating information theory and information fusion needs more work.

The second direction is from big data perspective. Since high volumes of multimedia, such as audio, video and images, are being generated daily, we should consider information fusion of multimedia data as a problem of big data. For the big data, it requires more sophisticated algorithms

for content analysis than those working on previous databases with limited data. Therefore, we should pay attention on how to effectively realize information fusion for massive multimedia data from widely distributed data sources.

Moreover, based on the above discussions, it is clear that research on information fusion algorithms and systems is becoming more and more common-place. There are a number of areas in the information fusion community that will most likely be highly active in the near future. With information fusion algorithms extending their applications from the statistics domain to many other fields such as robotics, sensor networks, and image processing, the need for standard fusion evaluation strategies applicable independent of the given application domain will grow more than ever. As a result, the fusion community will be driven towards development and wide spread adoption of such strategies, the investigation of which will be our future research focus. Undoubtedly, this trend will motivate more extensive research on topics related to the performance of information fusion systems. We believe multi-modal information fusion for multimedia analysis is a promising research area and will play increasing important roles in our future life.

References

[1] L. Guan, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki, and M. Ibrahim, "Multimodal information fusion for selected multimedia applications," *International Journal of Multimedia Intelligence and Security*, vol. 1, no. 1, pp. 5–32, 2010.

[2] J. A. Balazs and J. D. Velásquez, "Opinion mining and information fusion: a survey," *Information Fusion*, vol. 27, pp. 95–110, 2016.

[3] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[4] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[5] R. A. Harshman, "Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis," 1970.

[6] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115–2125, 2003.

[7] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[8] H. Xu and T.-S. Chua, "Fusion of audiovisual features and external information sources for event detection in team sports video," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 44–67, 2006.

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[10] K. Nordby, P. Helmersen, D. Gilmore, and S. Arnesen, *Human-computer Interaction: Interact'95*. Springer, 2016.

[11] H.-I. Suk and S.-W. Lee, "A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 286–299, 2013.

[12] Z. Xu and N. Zhao, "Information fusion for intuitionistic fuzzy decision making: an overview," *Information Fusion*, vol. 28, pp. 10–23, 2016.

[13] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, pp. 28–44, 2013.

[14] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[15] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol. 36, pp. 189–195, 2014.

[16] F. Bießmann, S. Plis, F. C. Meinecke, T. Eichele, and K.-R. Muller, "Analysis of multimodal neuroimaging data," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 26–58, 2011.

[17] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[18] L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 586–600, 2015.

[19] A. McIntosh, F. L. Bookstein, J. V. Haxby, and C. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *Neuroimage*, vol. 3, no. 3, pp. 143–157, 1996.

[20] Y. Levin-Schwartz, V. D. Calhoun, and T. Adali, "Quantifying the interaction and contribution of multiple datasets in fusion: Application to the detection of schizophrenia," *IEEE Transactions on Medical Imaging*, 2017 (To appear).

[21] M. Kumar, D. P. Garg, and R. A. Zachery, "A generalized approach for inconsistency detection in data fusion from multiple sensors," in *American Control Conference, 2006*, pp. 6–10, IEEE, 2006.

[22] P. Smets, "Analyzing the combination of conflicting belief functions," *Information Fusion*, vol. 8, pp. 387–412, 2007.

[23] T.-K. Sun, S.-C. Chen, Z. Jin, and J.-Y. Yang, "Kernelized discriminative canonical correlation analysis," in *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR'07. International Conference on*, vol. 3, pp. 1283–1287, IEEE, 2007.

[24] Z. Xie and L. Guan, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis," *International Journal of Semantic Computing*, vol. 7, no. 01, pp. 25–42, 2013.

[25] Z. Xie, Y. Tie, and L. Guan, "A new audiovisual emotion recognition system using entropy-estimation-based multimodal information fusion," in *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*, pp. 726–729, IEEE, 2015.

[26] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pp. 1–6, IEEE, 2013.

[27] G. O. Deák, M. S. Bartlett, and T. Jebara, "New trends in cognitive science: Integrative approaches to learning and development," *Neurocomputing*, vol. 70, no. 13, pp. 2139–2147, 2007.

[28] F. Castanedo, "A review of data fusion techniques," *The Scientific World Journal*, vol. 2013, pp. 1–19, 2013.

[29] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 116–134, 2007.

[30] M. J. Roemer, G. J. Kacprzynski, and R. F. Orsagh, "Assessment of data and knowledge fusion strategies for prognostics and health management," in *Aerospace Conference, 2001, IEEE Proceedings.*, vol. 6, pp. 2979–2988, IEEE, 2001.

[31] N. Poh and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, 2006.

[32] K. Nandakumar, Y. Chen, A. K. Jain, and S. C. Dass, "Quality-based score level fusion in multibiometric systems," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, pp. 473–476, IEEE, 2006.

[33] M. Hanmandlu, J. Grover, A. Gureja, and H. M. Gupta, "Score level fusion of multimodal biometrics using triangular norms," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1843–1850, 2011.

[34] K. N. S. Dass and A. Jain, "A principled approach to score level fusion in multimodal biometric systems," *Proceedings of Audio and Video-based Biometric Person Authentication*, vol. 3546, pp. 1049–1058, 2005.

[35] S. Prabhakar and A. K. Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognition*, vol. 35, no. 4, pp. 861–874, 2002.

[36] D. Zhang, F. Song, Y. Xu, and Z. Liang, "Decision level fusion," in *Advanced Pattern Recognition Technologies with Applications to Biometrics*, pp. 328–348, IGI Global, 2009.

[37] J. Zhou, T. Xu, and J. Gan, "Facial expression recognition based on local directional pattern using svm decision-level fusion," in *Proceedings of Tenth International Conference on Computability and Complexity in Analysis, Nancy, France*, vol. 810, pp. 126–132, 2013.

[38] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 2462–2465, IEEE, 2010.

[39] Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 90–94, 1995.

[40] A. Constantinidis, M. C. Fairhurst, and A. F. R. Rahman, "A new multi-expert decision combination algorithm and its application to the detection of circumscribed masses in digital mammograms," *Pattern Recognition*, vol. 34, no. 8, pp. 1527–1537, 2001.

[41] X.-Y. Jing, D. Zhang, and J.-Y. Yang, "Face recognition based on a group decision-making combination approach," *Pattern Recognition*, vol. 36, no. 7, pp. 1675–1678, 2003.

[42] J. Yang and J.-Y. Yang, "Generalized K–L transform based combined feature extraction," *Pattern Recognition*, vol. 35, no. 1, pp. 295–297, 2002.

[43] C. Liu and H. Wechsler, "A shape-and texture-based enhanced fisher classifier for face recognition," *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 598–608, 2001.

[44] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognition*, vol. 36, no. 6, pp. 1369–1381, 2003.

[45] J. Yang and X. Zhang, "Feature-level fusion of fingerprint and finger-vein for personal identification," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 623–628, 2012.

[46] A. A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Defense and Security*, pp. 196–204, 2005.

[47] Z.-H. Feng, J. Kittler, W. Christmas, and X.-J. Wu, "Feature level multiple model fusion using multilinear subspace analysis with incomplete training set and its application to face image analysis," in *International Workshop on Multiple Classifier Systems*, pp. 73–84, Springer, 2013.

[48] X. Geng, K. Smith-Miles, Z.-H. Zhou, and L. Wang, "Face image modeling by multilinear subspace analysis with missing values," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 881–892, 2011.

[49] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[50] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Asian Conference on Computer Vision*, pp. 252–267, Springer, 2014.

[51] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[52] W. K. Härdle and L. Simar, "Canonical correlation analysis," in *Applied Multivariate Statistical Analysis*, pp. 443–454, Springer, 2015.

[53] W. Härdle and L. Simar, "Canonical correlation analysis," *Applied Multivariate Statistical Analysis*, pp. 321–330, 2007.

[54] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.

[55] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2, pp. 229–233, IEEE, 2007.

[56] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 39–50, 2010.

[57] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[58] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[59] G. Lisanti, S. Karaman, I. Masi, and A. Del Bimbo, "Multi channel-kernel canonical correlation analysis for cross-view person re-identification," *arXiv preprint arXiv:1607.02204*, 2016.

[60] J. Zhao, Y. Fan, and W. Fan, "Fusion of global and local feature using KCCA for automatic target recognition," in *Image and Graphics, 2009. ICIG'09. Fifth International Conference on*, pp. 958–962, IEEE, 2009.

[61] X. Xu and Z. Mu, "Feature fusion method based on KCCA for ear and profile face based multimodal recognition," in *Automation and Logistics, 2007 IEEE International Conference on*, pp. 620–623, IEEE, 2007.

[62] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[63] M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu, "Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 202–207, IEEE, 2000.

[64] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005–1018, 2007.

[65] O. Arandjelović, "Discriminative extended canonical correlation analysis for pattern set matching," *Machine Learning*, vol. 94, no. 3, pp. 353–370, 2014.

[66] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

[67] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[68] J. Vía, I. Santamaría, and J. Pérez, "Canonical correlation analysis (CCA) algorithms for multiple data sets: Application to blind simo equalization," in *Signal Processing Conference, 2005 13th European*, pp. 1–4, IEEE, 2005.

[69] S. Watanabe, *Pattern recognition: Human and mechanical*. John Wiley & Sons, Inc., 1985.

[70] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[71] A. Renyi, *Foundations of probability*. Courier Corporation, 2007.

[72] J. Wu, J. Sun, L. Liang, and Y. Zha, "Determination of weights for ultimate cross efficiency using shannon entropy," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5162–5165, 2011.

[73] R. Arellano, J. Contreras, and M. Genton, "Shannon entropy and mutual information for multivariate skew-elliptical distributions," *Scandinavian Journal of Statistics*, vol. 40, no. 1, pp. 42–62, 2013.

[74] P. Li and C.-H. Zhang, "A new algorithm for compressed counting with applications in shannon entropy estimation in dynamic data.," in *COLT*, pp. 477–496, 2011.

[75] A. Greven, G. Keller, and G. Warnecke, *Entropy*. Princeton university press, 2014.

[76] R. Jenssen, "Kernel entropy component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 847–860, 2010.

[77] G. Montavon, M. L. Braun, T. Krueger, and K.-R. Muller, "Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 62–74, 2013.

[78] L. Gómez-Chova, R. Jenssen, and G. Camps-Valls, "Kernel entropy component analysis for remote sensing image clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 2, pp. 312–316, 2012.

[79] Z. Zhang and E. R. Hancock, "Kernel entropy-based unsupervised spectral feature selection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 05, pp. 1260–1277, 2012.

[80] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[81] P. H. Foo and G. W. Ng, "High-level information fusion: An overview.," *Journal of Advances in Information Fusion*, vol. 8, no. 1, pp. 33–72, 2013.

[82] V. Kılıç, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2015.

[83] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, 2000.

[84] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio–visual affective expression recognition through multistream fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[85] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *Computer Vision–ECCV 2016 Workshops*, pp. 337–348, Springer, 2016.

[86] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[87] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

[88] Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 424–428, 2007.

[89] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[90] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 1136–1139, IEEE, 2006.

[91] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[92] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[93] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.

[94] A. Sayedelahl, R. Araujo, and M. S. Kamel, "Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pp. 1–6, IEEE, 2013.

[95] J. Peng, Q. Li, A. A. A. El-Latif, and X. Niu, "Linear discriminant multi-set canonical correlations analysis (LDMCCA): an efficient approach for feature fusion of finger biometrics," *Multimedia Tools and Applications*, vol. 74, no. 13, pp. 4469–4486, 2015.

[96] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2160–2167, IEEE, 2012.

[97] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[98] J. Wang, Y. Zhou, K. Duan, J. J.-Y. Wang, and H. Bensmail, "Supervised cross-modal factor analysis for multiple modal data classification," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pp. 1882–1888, IEEE, 2015.

[99] C. Varon, C. Alzate, and J. A. Suykens, "Noise level estimation for model selection in kernel PCA denoising," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2650–2663, 2015.

[100] C. Alzate and J. A. Suykens, "A regularized kernel CCA contrast function for ICA," *Neural Networks*, vol. 21, no. 2, pp. 170–181, 2008.

[101] Q. Liu, H. Lu, and S. Ma, "Improving kernel fisher discriminant analysis for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42–49, 2004.

[102] T. Melzer, M. Reiter, and H. Bischof, "Appearance models based on kernel canonical correlation analysis," *Pattern Recognition*, vol. 36, no. 9, pp. 1961–1971, 2003.

[103] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.

[104] K. M. Wong and S. Chen, "The entropy of ordered sequences and order statistics," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 276–284, 1990.

[105] L. Gao, L. Qi, and L. Guan, "Information fusion based on kernel entropy component analysis in discriminative canonical correlation space with application to audio emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 2817–2821, IEEE, 2016.

[106] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[107] I. Dincer and Y. A. Cengel, "Energy, entropy and exergy concepts and their roles in thermal engineering," *Entropy*, vol. 3, no. 3, pp. 116–149, 2001.

[108] K. Życzkowski, "Rényi extrapolation of Shannon entropy," *Open Systems & Information Dynamics*, vol. 10, no. 03, pp. 297–310, 2003.

[109] P. Bromiley, N. Thacker, and E. Bouhova-Thacker, "Shannon entropy, Renyi entropy, and information," *Statistics and Information Series*, vol. 4, pp. 1–8, 2004.

[110] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[111] A. Elisseeff, J. Weston, *et al.*, "A kernel method for multi-labelled classification," in *NIPS*, vol. 14, pp. 681–687, 2001.

[112] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[113] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *European Conference on Computer Vision*, pp. 751–767, Springer, 2000.

[114] R. Jenssen, J. C. Principe, D. Erdogmus, and T. Eltoft, "The Cauchy–Schwarz divergence and parzen windowing: Connections to graph theory and Mercer kernels," *Journal of the Franklin Institute*, vol. 343, no. 6, pp. 614–629, 2006.

[115] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pp. 1–8, IEEE, 2006.

[116] C.-L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques," *Pattern Recognition*, vol. 36, no. 10, pp. 2271–2285, 2003.

[117] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[118] H. Nishida, "A structural model of shape deformation," *Pattern Recognition*, vol. 28, no. 10, pp. 1611–1620, 1995.

[119] L. Lam and C. Y. Suen, "Structural classification and relaxation matching of totally unconstrained handwritten zip-code numbers," *Pattern Recognition*, vol. 21, no. 1, pp. 19–31, 1988.

[120] K.-W. Cheung, D.-Y. Yeung, and R. T. Chin, "A Bayesian framework for deformable pattern recognition with application to handwritten character recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1382–1388, 1998.

[121] M. Revow, C. K. Williams, and G. E. Hinton, "Using generative models for handwritten digit recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 592–606, 1996.

[122] A. K. Jain and D. Zongker, "Representation and recognition of handwritten digits using deformable templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, pp. 1386–1390, 1997.

[123] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[124] T. P. Weldon, W. E. Higgins, and D. F. Dunn, "Efficient Gabor filter design for texture segmentation," *Pattern Recognition*, vol. 29, no. 12, pp. 2005–2015, 1996.

[125] Y. Hamamoto, S. Uchimura, M. Watanabe, T. Yasuda, Y. Mitani, and S. Tomita, "A Gabor filter-based method for recognizing handwritten numerals," *Pattern Recognition*, vol. 31, no. 4, pp. 395–400, 1998.

[126] C.-J. Lee and S.-D. Wang, "Fingerprint feature extraction using Gabor filters," *Electronics Letters*, vol. 35, no. 4, pp. 288–290, 1999.

[127] P. K. Singh, R. Sarkar, and M. Nasipuri, "A study of moment based features on handwritten digit recognition," *Applied Computational Intelligence and Soft Computing*, vol. 2016, pp. 1–17, 2016.

[128] Z. Chen, W. Huang, and Z. Lv, "Towards a face recognition method based on uncorrelated discriminant sparse preserving projection," *Multimedia Tools and Applications*, vol. 74, no. 1, pp. 1–15, 2015.

[129] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2368–2378, 2014.

[130] Z. Wang, Z. Miao, Q. J. Wu, Y. Wan, and Z. Tang, "Low-resolution face recognition: a review," *The Visual Computer*, vol. 30, no. 4, pp. 359–386, 2014.

[131] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235–238, 2012.

[132] M.-J. Han, K.-T. Song, F.-Y. Chang, *et al.*, "A new information fusion method for svm-based robotic audio-visual emotion recognition," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pp. 2656–2661, IEEE, 2007.

[133] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[134] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[135] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[136] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1588–1595, 2004.

[137] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1449–1461, 2004.

[138] S.-J. Ryu, M. Kirchner, M.-J. Lee, and H.-K. Lee, "Rotation invariant localization of duplicated image regions based on Zernike moments," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1355–1370, 2013.

[139] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[140] D. Huang, C. Zhu, Y. Wang, and L. Chen, "HSOG: a novel local image descriptor based on histograms of the second-order gradients," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4680–4695, 2014.

[141] A. Satpathy, X. Jiang, and H.-L. Eng, "LBP-based edge-texture features for object recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1953–1964, 2014.

[142] P. C. Wong, E. Skoe, N. M. Russo, T. Dees, and N. Kraus, "Musical experience shapes human brainstem encoding of linguistic pitch patterns," *Nature Neuroscience*, vol. 10, no. 4, pp. 420–422, 2007.

[143] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pp. 145–148, IEEE, 2006.

[144] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 264–277, 1999.

[145] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition," *Pattern Analysis and Applications*, vol. 9, no. 2-3, pp. 273–292, 2006.

[146] B. Shekar, M. S. Kumari, L. M. Mestetskiy, and N. F. Dyshkant, "Face recognition using kernel entropy component analysis," *Neurocomputing*, vol. 74, no. 6, pp. 1053–1057, 2011.