Audiovisual Emotion Recognition Using Entropy-estimation-based Multimodal Information Fusion

by

Zhibing Xie

Master of Science, The Hong Kong Polytechnic University, 2009

Bachelor of Engineering, Shanghai University, 2004

A Dissertation

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, 2015

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

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.

Audiovisual Emotion Recognition Using Entropy-estimation-based Multimodal Information Fusion

Doctor of Philosophy 2015

Zhibing Xie

Electrical and Computer Engineering

Ryerson University

Abstract

Understanding human emotional states is indispensable for our daily interaction, and we can enjoy more natural and friendly human computer interaction (HCI) experience by fully utilizing human's affective states. In the application of emotion recognition, multimodal information fusion is widely used to discover the relationships of multiple information sources and make joint use of a number of channels, such as speech, facial expression, gesture and physiological processes. This thesis proposes a new framework of emotion recognition using information fusion based on the estimation of information entropy. The novel techniques of information theoretic learning are applied to feature level fusion and score level fusion. The most critical issues for feature level fusion are feature transformation and dimensionality reduction. The existing methods depend on the second order statistics, which is only optimal for Gaussian-like distributions. By incorporating information theoretic tools, a new feature level fusion method based on kernel

entropy component analysis is proposed. For score level fusion, most previous methods focus on predefined rule based approaches, which are usually heuristic. In this thesis, a connection between information fusion and maximum correntropy criterion is established for effective score level fusion. Feature level fusion and score level fusion methods are then combined to introduce a two-stage fusion platform. The proposed methods are applied to audiovisual emotion recognition, and their effectiveness is evaluated by experiments on two publicly available audiovisual emotion databases. The experimental results demonstrate that the proposed algorithms achieve improved performance in comparison with the existing methods. The work of this thesis offers a promising direction to design more advanced emotion recognition systems based on multimodal information fusion and has great significance to the development of intelligent human computer interaction systems.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my Ph.D. supervisor, Dr. Ling Guan, for his continuous patience, guidance and motivation throughout my Ph.D. study and thesis writing. His support helped me in all the time of 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 mentor for my Ph.D. study.

Being a member of Ryerson Multimedia Research Laboratory (RML), I express my appreciation to all members of this laboratory, Dr. Matthew Kyan, Dr. Yifeng He, Dr. Yun Tie, Dr. Rui Zhang, Dr. Ning Zhang, Dr. Muhammad Talal Ibrahim, Dr. Adrian Bulzacky, Dr. Naimul Mefraz Khan, Xiaoming Nan, Ziyang Zhang, Dong Nan, Fei Guo, Kevin Tang and Lei Gao. I would also like to thank many people with whom I have collaborated in my research work during the past few years. In particular, I want to thank Dr. Yun Tie and Dr. Ning Zhang who have shared with me a great deal of knowledge and experience in the field of multimedia retrieval and analysis.

I would like to thank my thesis committee members for the insights they provided to this thesis. I am also grateful to the Department of Electrical and Computer Engineer-

ing at Ryerson University for providing helpful research environment which is inspiring and resourceful. The research experience significantly supported me in understanding and developing the links between my previous experience, theoretical knowledge, and industrial skills.

Contents

List of Tables

List of Figures

List of Abbreviations

AAM	Active Appearance Model
ANN	Artificial Neural Network
ANOVA	Analysis of Variations
AP	Acoustic Prosodic Information
BDPCA	Bi-Directional Principle Component Analysis
CCA	Canonical Correlation Analysis
CFA	Cross-modal Factor Analysis
CHMM	Coupled Hidden Markov Model
DBN	Dynamic Bayesian Network
DS	Dempster-Shafer
DWT	Discrete Wavelet Transform
EAR	Emotion Association Rules
EBS	Elastic Body Spline

List of Abbreviations

ECG	Electrocardiogram
EEG	Eelectroencephalogram
EMG	Electromyogram
EOG	Electrooculogram
GMM	Gaussian Mixture Model
GSR	Galvanic Skin Response
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HSV	Hue-Saturation-Value
ITL	Information Theoretic Learning
KCCA	Kernel Canonical Correlation Analysis
KECA	Kernel Entropy Component Analysis
KLDA	Kernel Linear Discriminant Analysis

List of Abbreviations

K-NN	K-Nearest Neighbors
KPCA	Kernel Principal Component Analysis
KSOM	Kohonen Self-Organizing Map
LDA	Linear Discriminant Analysis
LDP	Local Directional Pattern
LDPv	Local Directional Pattern Variance
LPC	Linear Predictor Coefficient
LPCC	Linear Predictive Cepstral Coefficient
MCC	Maximum Correntropy Criterion
MEE	Minimum Error Entropy
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
MMC	Meta-Multiclass

List of Abbreviations

MSE	Mean Square Error
NoSQL	Not Only SQL
PCA	Principle Component Analysis
PDE	Partial Differential Equation
PDF	Probability Density Function
RBF	Radial Basis Function
RML	Ryerson Multimedia Research Laboratory
SFFS	Sequential Forward Floating Search
SL	Semantic Label
SVM	Support Vector Machine
TRECVID	TREC Video Retrieval Evaluation

Chapter 1

Introduction

1.1 Background

Recognition of emotional states can help us estimate the desire and future behavior of a person. The users can express feeling and provide feedback through emotions. Different emotional states are usually associated with a broad range of behavioral cues and signals including auditory, visual and physiological presentation. These signals carry sufficient emotional information, making automatic detection of emotions possible. Emotion recognition finds its extensive applications in the area of human computer interaction (HCI), since the information about emotional states could be used to make communication with computers in a more human-like manner. Hence, the detection of user's emotional states

is a crucial element for developing more effective interfaces between humans and computers, especially in applications of, for example, affective computing, interactive video games, human robot interaction and many other emerging areas [1].

The natural emotional communication is usually performed through two methods including verbal way and non-verbal way. The verbal way involves speech while non-verbal way involves facial expression, body gesture, and sign language. Generally speaking, the data sources used for emotion recognition could come from more than one modality such as audio, video, electroencephalogram (EEG), electrocardiogram (ECG) and so on [2]. Moreover the human brains combine different information into a coherent one, integrate supplementary information, and derive a decision from various modalities of data. Due to complementarity and redundancy of the data coming from these channels, emotion recognition based on multiple modalities is expected to perform more robustly than single modal methods. Hence in affective applications, multimodal information fusion is used for the integration of related information from multiple modalities to enhance the performance of data classification or reduce the uncertainty and ambiguity in decision making. In this thesis, we use entropy-estimation-based strategies for integrating audio and video information which can lead to improved performance of affective behavior recognition.

Multimodal information fusion refers to a process which achieves more reliable and robust analysis performance by integrating a set of multiple data sources, extracted fea-

tures, and intermediate decisions [3]. The performance of unimodal based recognition system is usually far from satisfactory, because one modality only may not provide sufficient information and it is vulnerable to the drastic variation and noisy nature of the acquired signals. On the other hand, multiple types of sensors may carry redundant, complementary, or even contradictory information. The integration of multimodal data potentially provides a more discriminatory description of the intrinsic characteristics. Hence, utilizing useful data and eliminating conflict information based on effective fusion algorithms become an increasingly essential issue in numerous research areas.

Multimodal information fusion is also of great importance for human computer interaction application. Since the conventional human computer interfaces are considered too restrictive for natural interaction between human and computer, a great deal of effort has 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, for example audiovisual emotion recognition application. 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 multimodality. Therefore, the objective of this thesis is to propose a more effective framework for emotion recognition based on multimodal fusion which is able to utilize complementary information, eliminate redundant data and improve the accuracy of the overall recognition performance of interactive human computer communication.

1.2 Challenges

Due to the importance of emotion in human communication, many attempts have been made to make computers interact more naturally with human beings. But research on affective computing is a very challenging field due to a variety of reasons. Communication between humans is complex and we still don't have a complete understanding of how humans process emotions. The motivation for employment of information fusion techniques in bimodal emotion recognition is that we can have more accurate and reliable detection of audiovisual emotional states. However the benefits usually come with certain drawbacks, and in order to accomplish a task better, the challenges resulted from the analysis process cannot be ignored [4, 5]. The following issues are a number of important challenges to be addressed.

- The main reason is due to the dissimilar characteristics of the audiovisual modalities involved during fusion process. One issue of concern is that audiovisual modalities are usually captured at various rates in various formats. Furthermore, there is large uncertainty over emotion features, and the features are preferred to be independent of speaker, language, gender and culture. Hence, it is necessary to address several critical issues which include the identification of dissimilar characteristics between different modalities, the synchronization of audio and video modalities, and the

selection of optimal fusion algorithms.

- The second essential obstacle associated with audiovisual emotion recognition based on multimodal fusion techniques is the variation of input information. Different data sources often have different levels of confidence and uncertainty in various scenarios. The correlation of audiovisual modalities can be perceived at different levels, for example, at early level, intermediate level or late level. A problem to be solved is how to describe the temporal coupling relationship between audio and video streams and preserve their natural correlations over time. On the other hand, the independence among different modalities is also essential, since it provides additional information in accomplishing different tasks. Therefore, when processing multimodal information fusion for emotion recognition, both independence and correlation provide valuable insights under different scenarios.

- Another issue addressed differently by the research community is the architecture of information fusion for bimodal emotion recognition. The speech and facial features can be concatenated to construct joint feature vectors and then modeled by a single classifier at the early stage of fusion. But early fusion increases the dimensionality and may suffer from the problem of data sparseness. On the other hand, audiovisual signals can be modeled by the corresponding classifiers and then the recognition

results from each classifier are integrated at the late stage of fusion. Although late fusion enables us to interpret the role of multiple modalities, mutual correlation among multiple modalities is usually not taken into serious consideration. Hence optimal fusion architecture for bimodal emotion recognition is necessary.

To address these problems, the analysis and fusion process needs to treat the above issues properly. The optimal selection of the techniques and algorithms used for fusion is one of the primary elements to the performance and accuracy of emotion recognition systems. Therefore many research investigations have been carried out into the improvement of emotion recognition techniques, but we are still far from a satisfactory performance of affective computing systems.

1.3 Contribution of the Thesis

This thesis systematically studies audiovisual emotion recognition approaches using multimodal information fusion based on entropy estimation. A set of contributions which this thesis has made are summarized as follows.

- We present a novel framework of multimodal emotion recognition using information fusion approach based on entropy estimation. Audio and visual channels are utilized to classify and detect emotional states for intelligent human computer in-

terfaces. We propose a new dual-level framework of multimodal information fusion which consists of feature level fusion module based on kernel entropy component analysis and score level fusion module based on maximum correntropy criterion. Our extensive experimental study on eNTERFACE emotion database and RML emotion database demonstrates that the proposed methods are capable of providing improved performance. The comparison with other methods shows that the proposed two-stage fusion platform outperforms the traditional algorithms in terms of both accuracy and reliability. To our knowledge, the present work is the first investigation focusing on implementing entropy estimation based fusion approaches in bimodal emotion recognition system.

- We introduce kernel entropy component analysis based strategy for information fusion at feature level. Information fusion and dimensionality reduction of feature vectors are critical issues of feature level fusion. However, most previous methods depend on the analysis of the second order statistics which is only optimal for Gaussian-like distributions. In order to overcome this problem, our analysis provides a strategy for implementing the techniques of information theory into the application of feature level fusion. Moreover, the proposed feature level fusion framework is applied in the application of audio emotion recognition. Our objec-

tive is to boost the accuracy of speech emotional classification by fully utilizing the audio features. The proposed method is validated using data obtained through emotion databases. The developed application represents a feasible solution to speech emotion recognition which can easily be integrated into human computer interaction systems.

- Taking into consideration of the limitations of existing predefined rule fusion methods, we propose a novel approach based on maximum correntropy criterion for score level fusion. Since the distributions of matching scores at score level usually do not have clear boundaries and it is not prudent to make assumption of parametric probability density function (PDF) models, we utilize similarity metrics with information theoretic learning principles to integrate the matching scores. Our proposed score level fusion method is implemented in the application of audiovisual emotion recognition. Since humans express emotions through various channels, emotional states can be perceived by combining emotional cues derived from multiple modalities. It is believed that audio and visual channels are correlated and they may contain supplementary information. The novelty of the proposed method is an optimal fusion framework of audiovisual emotion recognition which integrates the advantages of information theoretic learning techniques and information fusion

strategies. In comparison with the results given by other methods, the experimental results clearly demonstrate that the proposed two-stage fusion method can boost the performance of multimodal emotion recognition system.

1.4 Organization of the Thesis

The organization of this thesis is as follows:

Chapter 1 presents the general background of emotion recognition and information fusion, the contribution of this research, and the organization of this thesis.

Chapter 2 provides a detailed review of the related works. It covers the introduction and related applications about emotion recognition and information fusion.

Chapter 3 presents the information-theoretically optimal tools in detail. It describes the connection between the techniques of information theory and the application of multimodal information fusion. A new application of speech emotion recognition based on feature level fusion using kernel entropy component analysis is described in this chapter. The extraction and fusion of audio features are discussed. The effectiveness of the proposed method is demonstrated through extensive experimentation.

Chapter 4 introduces a new score level fusion strategy based on entropy estimation. Different methods of score level fusion are discussed, and the tools of information theoretic

learning, such as correntropy and maximum correntropy criterion, are described. A score level fusion algorithm using maximum correntropy criterion is presented. A two-stage fusion framework of audiovisual emotion recognition combining feature level and score level is proposed. The design of system structure for bimodal emotion recognition is introduced. The experimental results on both feature level and score level are reported. Improved performances have been achieved in accuracy and reliability.

Chapter 5 summarizes the works presented in this thesis and proposes the possible directions for future research.

Chapter 2

Literature Review

2.1 Overview

Analysis and recognition of human emotional behavior have gained a lot of interest and emotion recognition is considered an essential step towards building efficient and practical intelligent human computer interfaces. Hence there are a number of studies in the literature which have been conducted to recognize emotions through various types of features, classifiers and fusion methods. In order to obtain more reliable estimation of human emotions, more information sources are taken into account. The multimodal fusion approaches increase the confidence of the results and decrease the level of ambiguity with respect to the emotions. The emotion recognition using multiple modalities is more

complicated due to the asynchronous characteristics of the emotion patterns and the correlation possibly occurring in different channels. Hence many fusion approaches have been exploited in recent years. This chapter provides a brief review of the related works in the fields of emotion recognition and information fusion.

2.2 Emotion Recognition

2.2.1 Introduction

Effective interpretation and analysis of human behavior characteristics are of fundamental significance in the design of intelligent human computer interaction systems [6]. 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. Among various interaction media, emotional sensitivity is believed to be a key element toward more human-like interaction. Obviously, analyzing emotion states in real time without human intervention could largely help understand the behavior of humans. Emotions carry information such as desire, intent and response to some events. It is believed that there exist a number of emotions which are basic and can be recognized

universally [7]. Emotional states are homogeneously expressed through bodily and physiological cues, such as speech, facial expression and other information sources. Among these modalities, audio and video streams are the most fundamental and natural communication means of human beings. Therefore, in order to translate users' emotions reliably, a number of user-independent emotion assessment techniques based on speech, facial expression and other data sources have been developed recently.

With the progress in machine learning and data mining, emotion recognition has been applied to several domains. However, due to the complex nature of human emotions, the task of emotion recognition based on multimodal information is challenging, and the applications are still limited to simple informational dialog systems. The difficulties can be summarized into three aspects [8]. First of all, since machine learning techniques are rarely independent of the application domain, it is essential to design optimal recognition algorithms which are the most suitable and efficient in characterizing different emotions without depending on speakers. The second obstacle is created by data variability, which is introduced by several facts, like speaking styles and speaking rates of different speakers, variations of head pose and lighting conditions, and subtle facial behaviors. The commonly used algorithms are vulnerable to these shortcomings. Last, the complicated correlation between different modalities results to another issue. Most existing works have not built up a close relationship between the features. Therefore, a proper selection

of fusion algorithms is believed to affect the classification performance significantly.

A concept of discrete emotion models has to be chosen to deal with multimodal emotion recognition, since the term emotion itself is an abstract concept which describes human feelings. Emotion models based on discrete categories have been proved useful in empirical studies. Hence we have to integrate emotions into quantifiable categories. It is widely accepted that several discrete categories of emotion have been shown to be universal across cultures and age groups [8]. A number of attempts have been made to define basic emotion categories. The investigation on emotion in psychology and neurophysiology reveals that the fundamental states of human emotion include happiness, sadness, anger, fear, surprise, and disgust [9]. Generally speaking, these six emotion states are not culturally determined, but universal to human nature biologically. In this thesis, these six principal emotion states are the focus of study. A wide investigation shows that some of the emotions are audio dominant, and the others are visual dominant. Hence it is important to study the joint characteristics of audio and visual channels. It is widely believed that when one modality is not good enough to determine a certain emotion, it is highly probable that the other one can help derive more complete, precise and discriminatory results. Therefore, the integration of audiovisual data will likely improve the performance of emotion recognition systems.

Many solutions have been proposed to process spoken utterance and facial expression

to identify the emotional information. However, this requires sufficient artificial intelligence. Although a few tentative systems have been developed for audiovisual emotion recognition, we are still far from having natural emotion interaction between man and machine due to the limits in efficiency, accuracy, and generality of the proposed systems. In [10, 11], we can find the review of cutting-edge works in multimodal emotion recognition. A number of multimodal systems have been proposed in the literature in a broad spectrum of scenarios, such as intelligent human computer interaction, security and surveillance, online entertainment and education, etc. Nowadays low cost devices can easily capture human emotion, which makes emotion recognition system more economically feasible for deployment. Hence automatic detection of user emotions has been applied to a variety of applications, including intelligent household robot for natural and friendly interaction with human beings [12] and fear type emotion recognition system dedicated to visual-audio surveillance [13].

2.2.2 Related Works

Since emotion recognition starts to attract the attention of research community, research works have been conducted from many perspectives. Combining multiple modalities, features and classifiers becomes the subject of main research stream. In the previous studies, one of the aspects is the investigation on the features of emotion representation.

Another aspect of the related research is based on the classification techniques. A number of efforts have been reported that different types of classifiers are integrated in the systems. Moreover, a few advanced fusion frameworks have been developed to utilize complementary modalities, features and classifiers. The interdependency and correlation of the affective features are of importance for multiple-level fusion. It is believed that hybrid fusion which aims at combining the benefits of both low level fusion and semantic level fusion may be a good choice for fusion problem [14]. Hence information fusion has huge potential for the realization of efficient interactive emotion recognition systems. Here, certain main research works are described in brief.

Emotions are usually expressed through speech, facial expression, posture and as well as through physiological signals, for instance brain activity, heart rate, muscle activity, blood pressure and so on. Many emotion theories reveal that audio, video and physiological signals are useful for emotion recognition. There are a number of advantages of using multiple cues for emotion recognition, since one channel may not involve all emotions. This has motivated intensive research of emotion recognition in discovering the significant manner of the multiple features on specific emotions. Certain techniques for emotion recognition based on various modalities can be found in the literature. In the following, we present a number of typical applications in recent works in the field of multimodal emotion recognition.

One of the typical examples is a new multi-resolution approach to recognize and predict emotion from the measured physiological signals presented by Verma [15]. The multimodal physiological signals are electroencephalogram (EEG) and peripheral signals including blood volume pressure, respiration pattern, skin temperature, galvanic skin response (GSR), electromyogram (EMG) and electrooculogram (EOG) provided in the DEAP database. Discrete wavelet transform, a classical transform for multi-resolution signal analysis has been used. The features from each channel are considered with equal importance. Early fusion and late fusion based on support vector machine (SVM), multilayer perceptron (MLP), k-nearest neighbor (K-NN) and meta-multiclass (MMC) classifiers are compared respectively. The experimental results clearly prove the highest accuracy of the method based on SVM. Moreover, Maaoui et al. presented two methods based on feature level and decision level to integrate facial and physiological modalities to improve the accuracy and robustness of the emotion recognition system [2]. At feature level fusion, the mutual information approach is tested for selecting the most relevant features and principal component analysis is used to reduce the dimensionality. For decision level fusion, two methods including voting process and dynamic Bayesian networks are implemented. The system is validated using the data obtained through an emotion elicitation experiment based on international affective picture system. The experimental results show that feature level fusion is better than decision level fusion.

In contrast to emotion recognition through speech and facial expression, Houjeij et al. designed a system for emotional classification from human dialogue based on text and speech context [1]. The work focuses on music mood classification based on the combination of lyrics and audio features. The proposed system concatenates text and speech features and feeds them as an input to the classifier. A decision level fusion technique is used to obtain a weighed sum of classifier scores from the probability estimators of k-NN classifiers. The comparison of the experimental results obtained in each case shows that the hybrid text-speech approach achieves better accuracy than speech or text modality alone.

In addition to exploration of different modalities, research is also focused on finding reliable informative features and combining efficient classifiers in order to improve the performance of emotion recognition in practical applications. A variety of pattern recognition methods are utilized to construct a classifier and the widely used classification methods for emotion recognition include support vector machines (SVM), hidden Markov model (HMM), Gaussian mixture model (GMM), Bayesian network, neural network, decision tree and so on. The classification methods used in emotion recognition can be typically divided into linear and nonlinear methods. Linear method is based on linear weighted combination of object characteristics. On the other hand, non-linear method performs the classification by making a decision through non-linear integration of input

features. Many studies show that non-linear method is more appropriate for combining the multiple sources of information. And it is more widely used and more effective in classifying the overlapped emotional states. A variety of systems have been developed for emotion related applications, and some typical applications are described in the following paragraph.

One recent application presented by Milton et al. is a class-specific scheme to recognize emotions from speech signals [16]. The scheme is designed by multiple parallel classifiers, including k-nearest neighbor (k-NN), Gaussian mixture model (GMM), back propagation artificial neural network (ANN) and support vector machine (SVM) classifiers, each of which is optimized to a class. Each classifier for an emotional class is built by a feature identified from a pool of features and a classifier identified from a pool of classifiers, which can optimize the recognition of particular emotion. The classification is done in two levels. In the first level, features and classifiers for each emotion are employed, and in the second level, the predictions of the first level classifiers are combined using multiplication of probability, average of probability and un-weighted voting to take a final decision on the detected emotion. The experimental results show that the proposed scheme improves the emotion recognition accuracy. Furthermore, Vayrynen et al. presented a decision level fusion architecture based on multiple k-NN classifiers for multimodal speech prosody and vocal source [17]. The sum fusion rule and the sequential

forward floating search (SFFS) algorithm are used to produce expert classifiers. Automatic classification tests in five emotional classes demonstrate that higher emotional classification performance is achievable using both prosodic and vocal source features. In addition, Wu et al. presented an approach to emotion recognition of affective speech based on multiple classifiers using acoustic prosodic information (AP) and semantic label (SL) [18]. Three types of models, Gaussian mixture model (GMM), support vector machine (SVM) and multilayer perceptron (MLP) are adopted as the base-level classifiers. A meta-decision tree is employed for classifier fusion to obtain the AP-based emotion recognition confidence. For SL-based recognition, semantic labels derived from an existing knowledge database are used to automatically extract emotion association rules (EAR) from the recognized word sequence of affective speech. The maximum entropy model is utilized to characterize the relationship between emotional states and EAR. Finally, a weighted product fusion method is used to integrate the AP-based and SL-based recognition results for the final emotion decision.

Most of the previous works focus on combining different modalities and classifiers directly for automatic emotion recognition. However, base level integration may not perform well on all emotional states. Developing optimal design strategies for emotion recognition is always an active research field. Some studies have proven that more complicated hybrid fusion approaches can achieve higher recognition performance than

individual classifiers. Hence recently many researchers in the field of multimodal emotion recognition have exploited a synergistic combination of different modalities, features and classifiers. The multimodal integration of emotion recognition can be done by taking into account features and classifiers at different levels of analysis. We usually tackle the integration in the form of fusion at low, intermediate or high levels. Low level is also called early fusion, while high level fusion is also called late fusion. A few new approaches of multimodal analysis for the improved emotion recognition results have been proposed. The following applications are the typical examples in this aspect.

For example, Ooi et al. presented a new architecture of intelligent audio emotion recognition based on prosodic and spectral features [19]. It has two main paths in parallel and can recognize six emotions. The extraction of audio features is followed by bi-directional principle component analysis (BDPCA), linear discriminant analysis (LDA) and radial basis function (RBF) neural classification. Feature level and decision level fusion modules have also been used at the final stage to assist weight assignment and decision making. Simulation results and comparison have revealed good performance of the proposed recognizer. Moreover, Bejani et al. simulated human perception of emotion through combining emotion-related information using ANOVA feature selection method and multi-classifier neural networks [20]. Speech emotion recognition is based on prosody features and mel-frequency cepstral coefficients (MFCC), and facial expres-

sion recognition is based on integrated time motion image and quantized image matrix. A feature selection method based on the analysis of variations (ANOVA) is used. The experimental results show that using hybrid features and decision level fusion improves the outcome of unimodal systems. Another recent application described by Sayedelahl is a bimodal feature-decision fusion approach to enhance the performance of estimating emotions from spontaneous speech conversations [8]. The feature vectors consisting of audio information are extracted from the whole speech sentence, and they are combined with the video features of the individual key frames representing that sentence. The final estimation is calculated by a decision level fusion of predictions from all corresponding frames. The performance is compared with two fusion approaches, the decision level fusion using weighted linear combination and a simple feature level fusion. The experimental results show improvement in correlation between emotion estimation and audio references. In addition, Xu et al. described a multimodal emotion recognition fusion framework based on HMM and ANN [21]. This framework is designed to extract and integrate features from both video sequences and speech signals. It is constructed from two hidden Markov models (HMMs) representing video and audio streams respectively. Artificial Neural Network (ANN) is applied as the whole fusion mechanism. Two important phases for HMMs are facial animation parameter extraction from video sequences based on Active Appearance Model (AAM), and pitch/energy feature extraction from

speech signals. The experiments indicate that this approach has better performance and robustness than methods using video or audio separately.

All the studies reviewed above have demonstrated that human emotional states can be detected through vocal prosody, facial expressions, gestures and others physiological signals. And it is believed that the performance of emotion recognition systems can be improved by employing multimodal fusion. There are a number of research works in the field of emotion recognition which highlight the benefits of fusion mechanism, but we are still far from a satisfactory multimodal fusion framework for emotion recognition.

2.2.3 Publicly Available Databases

The first requirement to develop emotion recognition system is the acquisition and validation of emotion data. The performance and robustness of the recognition system are easily affected, if it is not well-trained with sufficient and suitable data in the databases. Therefore, we need publicly available databases to evaluate the performance of emotion recognition. Recent advances have motivated many researchers to create emotion databases. These databases contain audio, visual or audiovisual and physiological emotion data. The popular databases include eNTERFACE database, RML database, TRECVID database, BANCA database, M2VTS database, BIOMET database, MUCT database and DEAP database. There are a number of available datasets which can be used for various research

tasks, however, there lacks any standardization effort for a common database.

The eNTERFACE audiovisual emotion database is a reference database for testing and evaluating video, audio or joint audiovisual emotion recognition algorithms [22]. The evaluation of algorithms can be performed on this database for multimodal signal processing tasks, such as multimodal person identification or audiovisual speech recognition.

Ryerson Multimedia Research Laboratory (RML) also makes ongoing efforts to build multimodal databases related to emotion recognition. The RML emotion database is language and cultural background independent audiovisual emotion database [23]. The video samples were collected from eight human subjects, speaking six different languages and six basic human emotions are expressed. It contains 720 audiovisual emotional expression samples.

The TREC Video Retrieval Evaluation (TRECVID) is a TREC-style database for video analysis and retrieval evaluation [24]. TRECVID has test data from a small number of known professional sources, such as broadcast news organizations, TV program producers, and surveillance systems. Features from visual, audio and caption tracks in TRACVID datasets are extracted and utilized in various multimedia fusion tasks, for example video shot retrieval based on fusion methods, semantic video analysis using multimodal fusion, news video story segmentation using multi-level fusion, and semantic concept detection based on discriminative model fusion.

The BANCA database is a multimodal database intended for training and testing multimodal verification systems [25]. BANCA was captured in four European languages in two modalities (face and voice). The subjects were recorded in three different scenarios, controlled, degraded and adverse over 12 different sessions.

The XM2VTS database is a large multimodal database for face analysis [26]. It contains four recordings of 295 subjects captured onto high quality digital video. Each recording contains a speaking head shot and a rotating head shot. This database can provide color images, sound files, video sequences and 3D Model.

The BIOMET multimodal database for person authentication contains five different modalities which include audio, face images (3 cameras), hand image, fingerprint and on-line signature [27]. For the face images, a camera prototype was designed to suppress the influence of ambient light. Moreover a 3D acquisition system prototype and a standard digital camera were used.

The MUCT face database is a freely available database of face images [28]. A wide range of subjects were photographed. It consists of 3755 faces with 76 manual landmarks. Compared to existing publicly available 2D manually landmarked databases, the MUCT database provides more diversity of lighting, age, and ethnicity. The MUCT data is suitable for training and evaluating a wide assortment of models.

The DEAP database is a multimodal database for emotion analysis using physiological

signals [29]. It contains electroencephalogram (EEG) and peripheral signals from 32 participants. The EEG signals were recorded from 32 active electrodes, whereas peripheral physiological signals include GSR, skin temperature, blood volume pressure, respiration rate, electromyogram (EMG) and electrooculography (EOG). In addition, frontal face video was also recorded. The participants were recorded as each watched excerpts of music videos and rated each video in terms of the levels of arousal, valence, like, dislike, dominance and familiarity.

2.3 Multimodal Information Fusion

2.3.1 Levels of Fusion

Regarding the existing approaches, the schemes of information fusion are roughly classified into early, intermediate and late stages. In addition, many fusion schemes have been explored in the task-specific context. For a specific application, the existing fusion approaches can be divided into five modules including sensor level fusion, feature level fusion, score level fusion, decision level fusion and hybrid fusion [30].

Sensor Level Fusion

Sensor level fusion, which is also known as data level fusion, refers to the combination

of data from multiple sensors to perform inferences which may not be possible from a single sensor or source alone [31]. One of the typical examples of sensor level fusion is sensing a speech signal simultaneously with two different microphones. The data from different types of sources are integrated using techniques drawn from several disciplines, such as statistical estimation, signal processing, pattern recognition, and artificial intelligence. Sensor level fusion is believed to have the potential for enhancing the performance of analysis and recognition. However, due to a number of challenges, current techniques have failed to achieve its full potential. The reason sensor level fusion is not widely utilized is because it usually cannot be implemented for multimodal information fusion due to the incompatibility of data from heterogeneous data sources.

One of typical applications of sensor level fusion is spatially optimized data/pixel-level fusion of 3D shape and texture for face recognition described by Faisal [32]. Fusion functions are objectively optimized to model expression and illumination variations in linear subspaces for invariant face recognition. In addition to spatial optimization, multiple nonlinear fusion models are combined to enhance the learning capabilities. Moreover, Gilula et al. presented a framework of multi-level categorical data fusion using partially fused data [33]. This approach extends previous methodologies and applies to categorical variables with any number of levels. Using information from partially fused data, the method smoothly accommodates a Bayesian approach based on mixtures of

joint distributions constructed using evident dependence. In addition, Noore et al. described a novel wavelet based data level fusion algorithm which combines information from fingerprint, face, iris, and signature images of an individual [34]. This computationally efficient biometric fusion algorithm integrates information from four biometric images into a single composite image using multi-level discrete wavelet transformation. This approach reduces the memory size, and increases the recognition accuracy using multi-modal biometric features.

Feature Level Fusion

Feature level fusion is also called early fusion, referring to the combination of different feature vectors, obtained either with different modalities or by applying different feature extraction algorithms to the same modality [35]. This level of fusion requires a discriminative representation of the original data, so feature level fusion usually contains feature extraction process to achieve a compact representation of the original feature vectors. 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. Other advantages of fusion at feature level are simple implementation and requirement of only one learning phase. Moreover, fusion at the early stage has the advantages that it is able to provide the classifiers with better discriminatory ability

by exploiting the co-variation and correlation between different modalities. For example, concatenating the feature vectors which have been extracted from two modalities, like audio and video signals, is a typical application of multimodal information fusion.

However, sometimes the performance of feature level fusion is not satisfactory. The main reason is that the obtained features may not be compatible because of difference in the nature of modalities, since it is hard to achieve time synchronization and same format, and difficult to learn intrinsic correlation among heterogeneous features. Furthermore, the high dimensionality of the concatenated feature vector presents challenges for the design of classifiers, which largely increases the computational load. This problem is known as "curse of dimensionality". In order to overcome this disadvantage, some standard dimensionality reduction techniques, such as linear discriminant analysis (LDA) and principal component analysis (PCA), have been applied. Therefore, more sophisticated design is needed to process the concatenated data at feature level. Figure 2.1 shows a schematic representation of bimodal audiovisual fusion at feature level. In this figure, the data from different channels, such as audio and video streams, are extracted into feature vectors. The extracted features are first merged by feature fusion unit, and then the combined feature vector is input into classifiers for further analysis.

A number of systems based on feature level fusion have been developed. For example, Yang et al. described a feature level fusion framework of fingerprint and finger-vein for

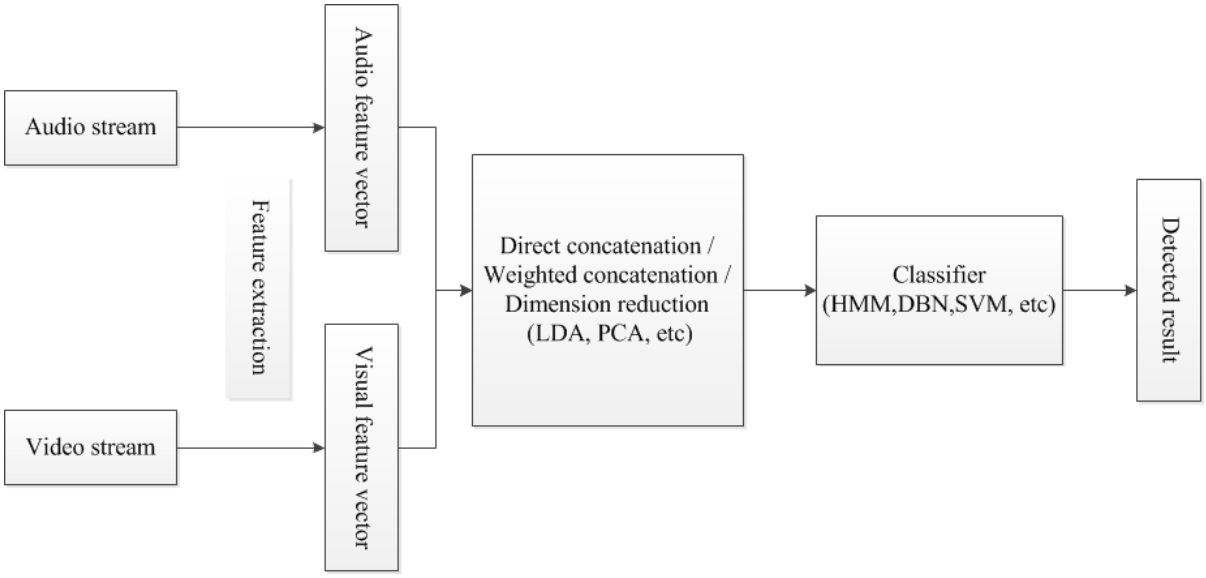

Figure 2.1: Feature level of bimodal audiovisual fusion.

personal identification [36]. 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 personal identification. This approach has a high capability in fingerprint-vein based personal recognition as well as multimodal feature level fusion. Furthermore, Ross et al. presented several feature level fusion strategies using hand and face biometrics, such as fusion of PCA and 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 [35]. 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 [37]. 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 M^2SA algorithm is adopted to reconstruct the missing entries of the incomplete training tensor. Furthermore, this framework is applied to the problem of face image analysis using Active Appearance Model (AAM) to validate its performance. Evaluations of AAM using the proposed framework are conducted with promising results.

Score Level Fusion

Score level fusion, which is known as intermediate level fusion, refers to the combination of matching scores provided by different modalities [38]. Although feature sets have rich sources of information, the features from different modalities may not be compatible. Moreover large dimensionality of a feature space might lead to irrelevant and redundant information. On the other hand, fusion at decision level is considered to be rigid due to the lack of information content. Hence score level fusion is fairly popular due to easy availability of the scores and sufficient information to discriminate between genuine and imposter scores. At score level, it is possible to combine scores obtained from the same modalities or different ones. Its advantages include simple implementa-

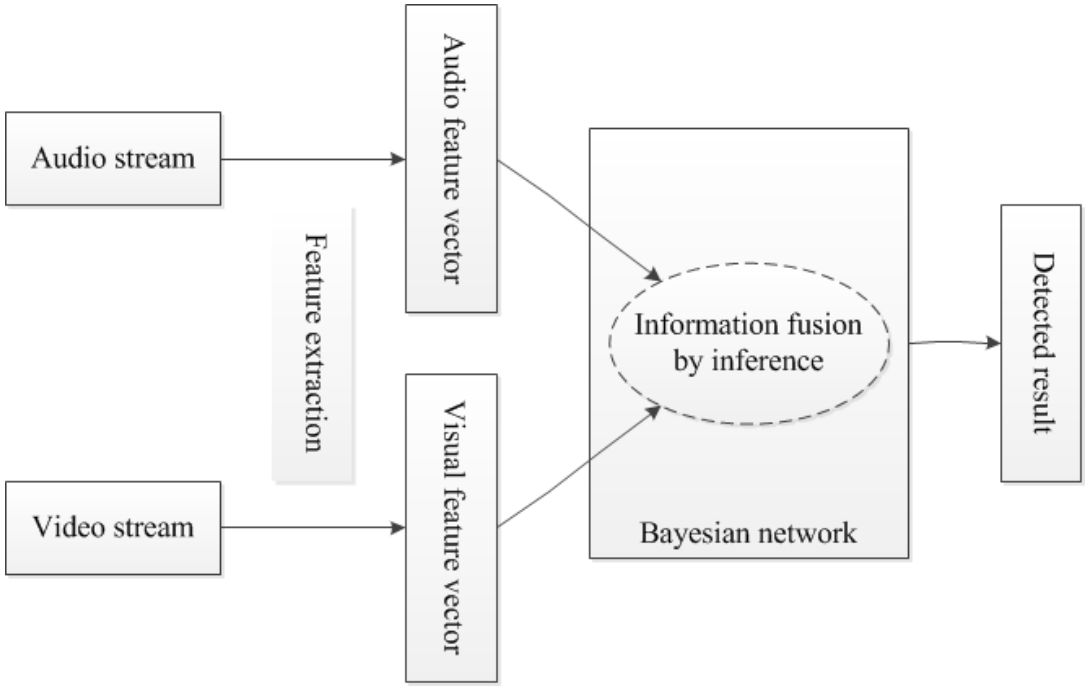

Figure 2.2: Score level of bimodal audiovisual fusion.

tion and scalability. This level of fusion can be divided into two categories, combination and classification [39]. Regarding combination, the score is combined by normalizing the input matching scores into the same 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, including failure to utilize correlation at feature level and tedious learning process. Figure 2.2 shows a schematic representation of bimodal 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 unit. Score fusion unit integrates the scores and obtains the final result.

There are a number of typical applications developed by research community. For example, Karthik et al. presented quality-based score level fusion in multi-biometric system [40]. 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 multimodal biometrics using triangular norms presented by Hanmandlu [41]. The scores from multiple biometrics are combined at score level 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 [42]. 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.

Decision Level Fusion

Decision level fusion is also called late fusion which refers to the combination of decisions from separate classifiers [43]. It involves the combination of likelihood values or probability scores obtained from separate single modality to come up with a combined decision. 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. Late fusion has some obvious advantages. The fusion at decision level usually processes information with the same representation. Moreover, scalability is also one of its merits. The fusion at this level allows extensive flexibility in the choice of individual classifiers. It also simplifies the algorithm development process. However, the fusion at decision level might lose too much useful information. Figure 2.3 shows the scheme of bimodal audiovisual fusion at decision level. In this figure, the decisions made by the classifiers are combined in decision fusion unit for further analysis. The final results are usually achieved by linear weighted sum or linear weighted product.

Some researchers have successfully adopted decision level fusion strategy. For example, Zhou et al. presented a facial expression recognition method based on global and

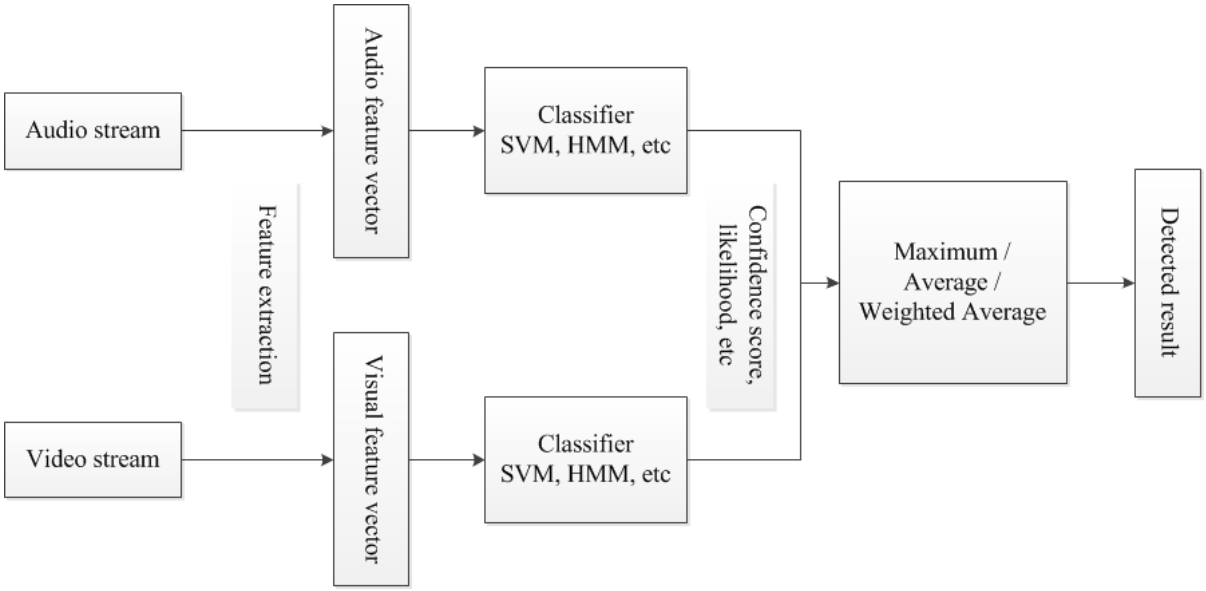

Figure 2.3: Decision level of bimodal audiovisual fusion.

local features with decision level fusion [44]. Local directional pattern (LDP) global features of the whole face are extracted, which can guarantee basic expression difference and decrease the 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. Moreover, another typical application is decision level integration system for multimodal emotional expression analysis presented by Metallinou [45]. 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 significantly increases the average accuracy from 55% to about 62%.

Hybrid Fusion

In hybrid fusion strategy, the correlation among different modalities and levels becomes more critical. The correlation is comprehended at various levels, and there are numerous forms of correlation utilized in multimodal fusion process. For instance, in the scenario of audiovisual speech recognition task, a hybrid fusion can be realized by a combination of feature level fusion with decision level fusion [46]. Recently various kinds of analysis problems have been solved by hybrid fusion. Singh et al. described a two-level hierarchical fusion of face images captured under visible and infrared light spectrum to improve the performance of face recognition [47]. Information fusion is performed at both image level and feature level to generate a fused feature vector. At image level fusion, two face images from different spectrums are fused using Discrete Wavelet Transform (DWT) based fusion algorithm. At feature level fusion, the amplitude and phase features are extracted from the fused image using 2D log polar Gabor wavelet. An adaptive SVM learning algorithm intelligently selects either the amplitude or phase features to generate a fused feature set for improved face recognition. Another work presented by Hussain et

al. is a hybrid fusion approach for detecting physiological features from multiple channels using machine learning techniques [48]. First the classification decision from individual channels is obtained and all selected features are merged to achieve the feature fusion. The decision fusion is performed based on weighted majority voting. The experimental results show that the best performance is achieved by using the hybrid framework.

2.3.2 Fusion Algorithms

The major fusion methods reported in the literature can be divided into three categories: rule based methods, classification based methods, and estimation based methods [30]. The following section describes a number of major strategies and their applications, including fixed rules, custom defined rules, support vector machine, probabilistic inference, Dempster-Shafer theory, and dynamic Bayesian network.

Fusion Based on Fixed Rules

The fusion based on fixed rules is the most popular method in this category including AND rule, OR rule, majority voting rule, maximum rule, minimum rule, sum rule, product rule, mean rule and so on [49]. In AND fusion, the outputs of different classifiers are compared to a preset threshold. An acceptance decision is reached only when all the classifiers agree. While in OR fusion, a positive decision is made as soon as one of the

classifiers makes an acceptance decision. In majority voting rule, the outputs of different classifiers are tested by a threshold and a decision is reached based on the majority of classifiers declaring the same decision. Maximum rule and minimum rule select the decision having the largest or least value amongst the modalities involved. In sum rule or product rule, the decision is computed by adding or multiplying the results for all modalities. A brief description of typical applications is as follow. Using linear weighted rule, Yang et al. assigned equal weights to the different modalities in a multimodal fusion system for people detection and tracking [50]. Moslem et al. presented a framework of multichannel uterine electromyogram (EMG) signals by using a weighted majority voting decision fusion rule [51]. In addition, a typical application for sum rule is a new classification framework based on sum-rule fusion of fuzzy k-NN classifiers presented by Chua [52].

Fusion Based on Custom Defined Rules

Since fixed rule fusion method is straightforward as well as computationally less expensive, it has been widely used. This method performs well if the weights of different modalities are appropriately determined. However, since the optimality of most fusion rules relies on the knowledge of probability distributions for all sensors, the overall performance is often worse than the expected result due to instabilities of the sensor's prob-

ability density functions. Unlike the above fixed rule approaches which use standard statistical rules, custom defined rule based fusion has the flexibility of adding rules based on the requirements of specific tasks. However, it is domain specific and requires proper knowledge. This fusion method is widely used in the areas of multimodal dialog systems, sports video analysis and so on. For example, Pfleger described a flexible and generic approach to multimodal fusion which is called context based multimodal integration [53]. This rule based integration approach is able to meet all demands which a fusion component for a multimodal dialogue system has to deal with. Another work is a new decision fusion rule presented for target detection in distributed sensor detection system [54]. The fusion method derives the overall decision based on multiple decisions from each individual sensor assuming that the probability distributions are not known.

Fusion Based on Support Vector Machine

Support vector machine (SVM) is one of the most effective classification techniques used in two-class problems [55]. Especially in the domain of multimedia, SVM has become increasingly popular, and it has been used for different tasks including concept classification, feature categorization, modality fusion, text categorization, face detection and so on. SVM is formalized as an optimization problem which finds the best hyperplane vectors by maximizing the margin between different sets. The transformation is

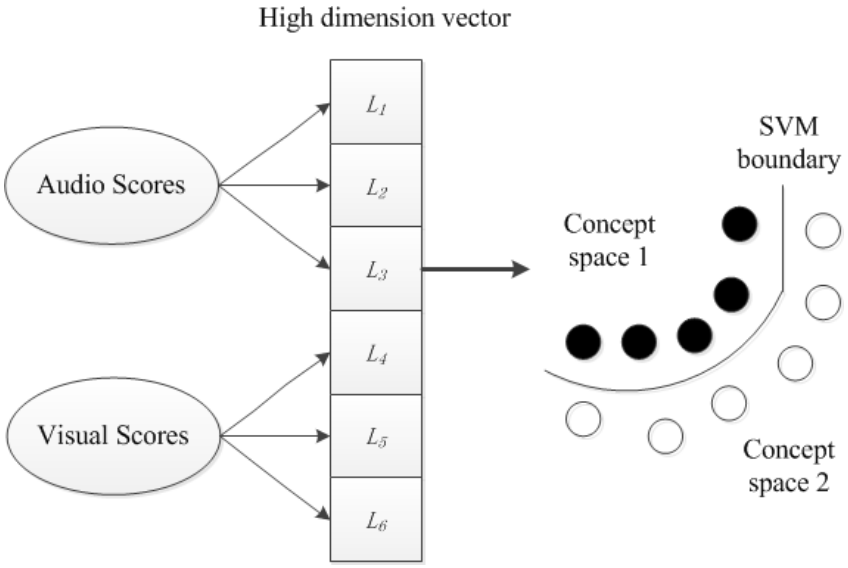

Figure 2.4: SVM based score classification of combined information.

non-linear and the transformed space is of higher dimensionality than the original one. Hence, generally speaking, SVM is considered as a supervised learning method and used as an optimal classifier. The following Figure 2.4 shows SVM based score classification of combined information from multiple intermediate data. There are many existing works which have used a fusion scheme based on SVM. For instance, a hybrid fusion approach has been realized as normalized early fusion and contextual late fusion for semantic indexing of multimedia resources based on visual and text cues [56]. In addition, Adams et al. presented a learning based approach to the semantic indexing of multimedia content using cues derived from audio, visual, and text features [57]. Furthermore, a facial expression recognition based on local directional pattern using SVM decision level fusion is one of the most recent works [44].

Fusion Based on Probabilistic Inference

Probabilistic inference has been widely used, and it is often referred to as classical sensor fusion method. Probabilistic inference can be applied at both feature level and decision level. In this method, multimodal information is combined according to probability theory, like Bayesian inference [58]. The observations obtained from multiple modalities are integrated, and an inference of the joint probability of an observation is derived. There are various advantages of probabilistic inference. For example, it allows for any prior knowledge about the likelihood of the hypothesis. Based on the new observations, it can update a priori probability in order to compute the posterior probability of the hypothesis. However, probabilistic method has some limitations. It requires priori and conditional probabilities. If the knowledge of priors has mutually exclusive hypotheses and uncertainty, the method would usually provide ambiguous results. The method of probabilistic inference, especially Bayesian inference, has been successfully used in multimodal information fusion. The research work for audiovisual speech recognition in [59] is one of the typical examples of Bayesian inference fusion at feature level for event detection in multimedia surveillance domain. Moreover, Xu et al. described a framework which utilizes both internal audiovisual features and various types of external information sources for event detection in team sports video [60]. Another work presented by

Pradeep et al. is to adopt a Bayesian inference fusion approach at hybrid levels, such as feature level and decision level [61].

Fusion Based on Dempster-Shafer Theory

Dempster-Shafer (DS) theory is an efficient method of combining accumulative evidences or for changing priors in the presence of new evidence [62]. Since DS evidence uses belief and plausibility values to represent the evidence and its corresponding uncertainty, it has the preference of many researchers. What is more, DS method can relax the restriction of Bayesian inference method on mutually exclusive hypotheses; therefore it can be regarded as an extension of the classical Bayesian theory. Some representative research works utilize DS fusion method for various multimedia analysis tasks, such as segmentation of satellite images [63], video classification [64], and finger print classification [65]. For example, Singh et al. presented a fingerprint classifier fusion algorithm using DS theory with update rule [65]. Another application is visual tracking system with spatio-temporal DS information fusion [66].

Fusion Based on Dynamic Bayesian Network

Bayesian inference can be extended into a network in which edges denote probabilistic dependencies and nodes represent observations or states of different modalities, like audio and video [67, 68]. Dynamic Bayesian network (DBN) is suitable for various multimedia

analysis tasks which require decisions to be performed using time-series data. DBN has advantages over the other methods in two aspects [69]. The first one is that DBN is able to model the multiple dependencies among the nodes. The second one is that the temporal dynamics of multimodal data can easily be integrated. The most popular form of DBN is hidden Markov model (HMM). Since the fusion of input data usually focuses on time-dependent patterns, HMM has been seen as one of the best choices when considering the different alternatives among statistical models. Single HMM has been widely used to process the joint audiovisual features. For example, Dumas et al. presented a multimodal fusion algorithm based on HMM for the development of adaptive interactive systems [70]. Furthermore, Nefian et al. described the use of statistical model coupled HMM (CHMM) for audiovisual integration in speaker dependent recognition [71]. Another recent application presented by Pinquier et al. is multiple feature fusion with hierarchical HMM for activity recognition based on wearable audiovisual sensors [72].

2.4 Summary

This review chapter describes the background of emotion recognition and presents the related works from the aspects of multiple modalities, classification solution and fusion approaches. Publicly available databases for performance evaluation are covered in this

chapter. It also discusses several issues of multimodal information fusion, from fusion levels to fusion algorithms and their merits and drawbacks. In addition, this chapter presents the popular fusion methods and their typical applications, including fixed rules, custom defined rules, support vector machine, probabilistic inference, Dempster-Shafer theory, and dynamic Bayesian network.

Emotion recognition applications based on the fusion methods mentioned above have been developed by many researchers. Despite the fact that a great number of analysis tasks have been performed, the various applications of human computer interface have usually presented challenges to the systematic understanding of the models and techniques of emotion recognition. Hence there still exist some issues which have not yet been explored sufficiently. In the next chapters, novel emotion recognition frameworks focusing on the multimodal fusion at feature level and score level have been proposed.

Chapter 3

Feature Level Fusion

3.1 Overview

With the development of information technology, the research topic of identifying the emotional states from audio signals is attracting much attention, since speech conveys the abundant emotional information. In this chapter, the tools of information theoretic learning are described. A new audio emotion recognition application based on entropy-estimation-based feature level fusion is proposed. The system architecture and algorithms of speech emotion recognition application based on kernel entropy component analysis are also presented. In this design, both prosodic and spectral features are fully utilized. The extracted features are followed by feature level fusion module to select the most

significant features. At feature level, the audio features are combined to construct a joint feature vector, but high dimensional feature set may easily suffer from the problem of data sparseness, and stress the computational resources. To solve this disadvantage, a feature level fusion method based on kernel entropy component analysis is explored for audio emotion recognition. The performance of the proposed architecture is evaluated on eNTERFACE and RML emotion databases. The universal six emotions such as happiness, angry, sadness, disgust, surprise and fear are considered. The results and comparison from the experiments have demonstrated an improved performance of the proposed scheme.

3.2 Introduction of Feature Level Fusion

Regarding the fusion at feature level, the features from multiple modalities are integrated early to select discriminatory features, and the features are utilized as the streams in a multi-stream classification technique. Feature level fusion plays a significant role in the improvement of recognition accuracy. However, what should not be neglected is that the extracted features are incomplete and imprecise due to heterogeneous measurement of different modalities. Therefore, before classification is implemented, several critical steps have to be executed, for example, integrating complementary data, eliminating redundant

information and processing feature vectors. In order to create a subset of new features by a combination of the original features, some linear strategies are typically used to discard redundant components and reduce high dimensionality of the data. These linear methods consist of principal component analysis (PCA) [73], linear discriminant analysis (LDA) [74], canonical correlation analysis (CCA) [75], cross-modal factor analysis (CFA) [76], etc.

One of the widely used approaches is canonical correlation analysis (CCA). CCA is a statistical approach which combines linear dimensionality reduction and information fusion by computing maximally correlated linear projections. For example, a combination of early and late fusion strategies is applied with CCA to a problem of open-set speaker identification [75]. Unlike canonical correlation analysis, cross-modal factor analysis (CFA) is a novel method to represent the coupled patterns between two different subsets of features through cross-modal association. CFA provides a feature selection capability in addition to feature dimension reduction and noise removal. These advantages make CFA a promising tool for many multimedia analysis tasks. One of its applications is audiovisual based multimodal emotion recognition which identifies the optimal transformation capable of representing coupled patterns between audio and visual channels through cross-modal association [76]. Moreover, most of the previous methods assume that there exists linear relationship among the original data. But there exists non-

linear feature extraction in many situations. Therefore, kernel method is proposed to achieve non-linear transformation, which leads to kernel PCA, kernel CCA and kernel CFA [77–79].

In feature level fusion, high dimensional features may easily suffer from the problem of data sparseness and stress the computational complexity. However, the majority of previous methods transform high-dimensional data to low-dimensional features largely depending on the second order statistics. In these methods, feature transformation is usually based on top eigenvalues and the corresponding eigenvectors of certain matrices. Hence, the theoretical foundation of these existing methods mostly relies on the second order statistics, such as variance, correlation, mean square error and so on, which is only optimal for Gaussian-like distribution. For example, principal component analysis (PCA) uses the variance as the metric. Therefore the second order statistics is a poor estimator, if the distribution from multiple modalities differs greatly from Gaussian.

Since the existing strategies do not have the capability of revealing the nature of input information, this issue motivates us to apply kernel entropy component analysis (KECA) as an alternative to information fusion at feature level. Compared with the previous methods, kernel entropy component analysis is an information theoretical method which preserves the maximum Renyi entropy of the input data with the smallest number of extracted features [80]. It does not necessarily correspond to the top eigenvalues and

eigenvectors of the kernel matrix, but depends on the largest contribution of entropy estimation. Kernel entropy component analysis captures the nonlinear higher order statistics of the data and achieves feature transformation and fusion based on the estimation of entropy value. This thesis demonstrates the effectiveness of KECA in feature transformation and fusion on emotion recognition system, and compares the performance of the proposed solution with kernel principal component analysis (KPCA) and kernel canonical correlation analysis (KCCA) based methods.

3.3 Information Theoretic Learning

One of the common problems of data processing is how to extract the information contained in data. In our daily lives, we are always bombarded by huge amount of data; however most of them are often not our primary interest. The important clues of information processing questions usually hide either in time structure or in spatial redundancy. Hence new strategies have come up because of the pressure of distilling useful information from data effectively, and the old ways of dealing with this problem are forced to evolve and adapt to the new reality. One of the new frameworks is information theoretic learning (ITL), a terminology perhaps first used by Watanabe [81]. The purpose of information theoretic learning is to provide more practical machine learning applications. Using the

innovative frameworks, we can employ the mathematical theory of information initially developed by Claude Shannon and Alfred Renyi to quantify global scalar descriptors of the underlying probability density function.

As we all know, information theory was first conceptualized by Claude Shannon to deal with the problem of optimally transmitting messages over noisy channels [82]. 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. It provided a mathematical framework to formulate and quantify interaction beyond physical laws. After the pioneering work of Shannon, information theory became a scientific field and many research works have been expanded upon Shannon's fundamental concepts. Moreover, information theory has been also utilized in the areas of physics, statistics, and biology as well as in field of engineering, for example machine learning and signal processing [83–85].

Information theory could play a role in the topics of machine learning and data mining. For supervised machine learning strategies, the objective is to estimate the joint probability density function (PDF) between the inputs and the desired responses, which means that a priori knowledge is required in order to achieve the optimal design. However, machine learning applications and other advanced data mining applications cannot be dependent on parametric PDF models. Because in the real world, the data

sets are conducive to non-Gaussian distributions which even change in time. Hence the employment of information theory could provide the trade-off between model complexity and generalization accuracy. It could develop new cost functions for optimization and learning algorithms and quantify data better with information theoretic descriptors.

In the framework of information theoretic learning (ITL), one of the most important descriptors is entropy [86]. 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. In information theoretic learning, 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). Since information theoretic learning can use the traditional learning methods of adaptive filters, neural networks, and kernel learning without major modifications, we are interested in a general ITL methodology to implement adaptive algorithms with information theoretic criteria. The fundamental issue in ITL is how to estimate entropy directly from samples and how to use the descriptors of the data for new learning principles based on information theoretic concepts. The following content covers the concepts of information theoretic learning which are useful in characterizing optimal solutions for practical

machine learning application.

3.4 Feature Fusion Based on Entropy Estimation

3.4.1 Shannon Entropy

The concept of Shannon entropy was introduced as a measure of statistical uncertainty [87]. 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 could provide a mathematical framework to quantify and formulate the nature of information. Shannon entropy plays a central role in information theoretic studies. It is believed to be able to measure the amount of the information contained in a series of events, which can be expressed as follows.

$$H_s(X) = -\sum_k p(x_k) log p(x_k) \tag{3.1}$$

or

$$H_s(X) = -\int f_x(x) log f_x(x) \tag{3.2}$$

where $p(x_k)$ and $f_x(x)$ are the discrete and continuous probability density function of data set respectively, and k is the total number of data set 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 its single scalar which measures the uncertainty in a form of probability density. 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 engineering [88].

3.4.2 Renyi Entropy

In practice applications, Renyi entropy is one of the widely used generalizations of information entropy [89]. Renyi wanted to find the most general class of information measure which preserved the additivity of statistically independent systems. In econometrics, Renyi quadratic entropy has been used to quantify diversity. Renyi entropy of order α

of a random variable X is expressed as

$$H_\alpha(X) = \frac{1}{1-\alpha} log(\sum_{1}^{N} p_k^\alpha) \tag{3.3}$$

or

$$H_\alpha(X) = \frac{1}{1-\alpha} log(\int f_x^\alpha(x)dx) \tag{3.4}$$

where $\alpha \geq 1$. At a deeper level, Renyi entropy measure is much more flexible than Shannon entropy due to the parameter α [90]. An interesting observation is that Shannon entropy can be considered as a special case of Renyi entropy when α converges to one. We usually choose α=2 as the fundamental descriptor, because it gives us a computationally efficient entropy estimator. Here, continuous Renyi quadratic entropy is given by

$$H(X) = -log(\int p^2(x)dx) \tag{3.5}$$

where $p(x)$ is probability density function (PDF) generated by the data set $\boldsymbol{D} = \boldsymbol{x}_1, ..., \boldsymbol{x}_N$. 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 [91].

3.4.3 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 [92]. The bottleneck of non-linear problem is the large parameter 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 completed efficiently in the space provided by the algorithms expressed in inner products. 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.

If the input space consists of $x_i \in \mathbf{R}_d$ in the set X, the non-linear mapping is expressed as follows.

$$\begin{aligned} \phi : \mathbf{R}_d &\rightarrow F \\ x &\rightarrow \phi(x) \end{aligned} \tag{3.6}$$

where $F \in \mathbf{R}_l, l \geq d$. A kernel function is a function k that satisfies

$$k(x_i, x_j) =< \phi(x_i), \phi(x_j) > \quad x_i, x_j \in X \tag{3.7}$$

This is what is known as kernel trick. An explicit expression of non-linear mapping ϕ

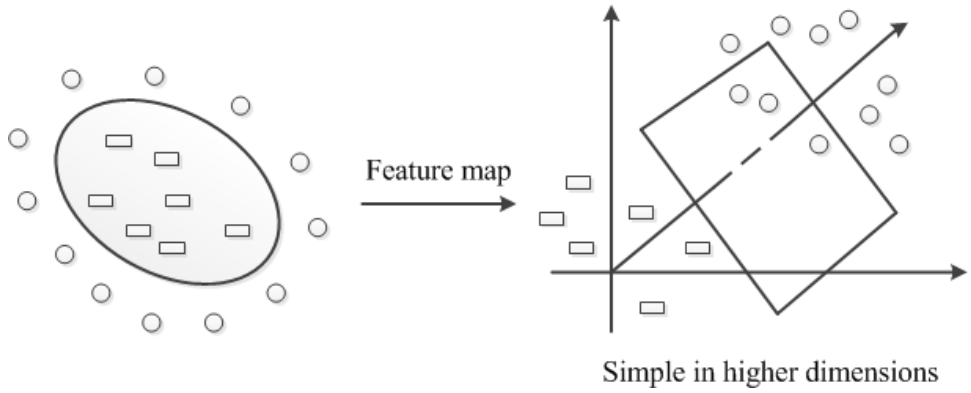

Figure 3.1: Nonlinear mapping of kernel method.

is difficult to determinate. However kernel trick 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 Mercer's condition, i.e., positive semi definite. Figure 3.1 illustrates the nonlinear mapping of kernel method. Some widely used kernel functions include linear kernel $k(x_i, x_j) =< \boldsymbol{x}_i, \boldsymbol{x}_j >$, polynomial kernel $k(x_i, x_j) = (< \boldsymbol{x}_i, \boldsymbol{x}_j > +1)^d$, exponential kernel $k(x_i, x_j) = exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{2\sigma^2})$ and Gaussian kernel. The Gaussian kernel is defined as follows:

$$k_\sigma(\boldsymbol{x}, \boldsymbol{y}) = exp(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}) \tag{3.8}$$

where σ is bandwidth working as a scale parameter which controls the width of Gaussian

kernel.

The kernel matrix $\boldsymbol{K}$ contains all the evaluation of kernel function k. From the kernel trick, 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

$$\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) =< \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) > \tag{3.9}$$

or in matrix form

$$\begin{aligned} &= \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix} \\ &= \begin{pmatrix} < \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_1) > & \cdots & < \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_n) > \\ \vdots & \ddots & \vdots \\ < \phi(\boldsymbol{x}_n), \phi(\boldsymbol{x}_1) > & \cdots & < \phi(\boldsymbol{x}_n), \phi(\boldsymbol{x}_n) > \end{pmatrix} \end{aligned} \tag{3.10}$$

where n is the number of data sets 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), kernel linear discriminant analysis (KLDA), and kernel k-means are

typical extensions of the corresponding linear algorithms by applying the kernel method on every inner product evaluation [93, 94].

3.4.4 Parzen Window Density Estimator

The strategies of density estimator can be divided into parametric method and nonparametric method. 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. A number of nonparametric density estimation methods are available, but we focus on Parzen windowing which is also known as kernel density estimation [95]. 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.

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. Hence we have to resort to a nonparametric estimation strategy. 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. Suppose there are N independent and identically distributed ($i.i.d.$) samples $\{x_1, ..., x_N\}$ from a random variable. The expression of Parzen window density estimator is given by

$$\tilde{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h}) \tag{3.11}$$

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.

3.4.5 The Proposed Feature Level Fusion Framework

The continuous Renyi quadratic entropy is given by

$$H(p) = -log(\int p^2(x)dx) = -logV(p) \tag{3.12}$$

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. In order to estimate $V(p)$, Parzen window density estimator is applied [95]. We rewrite Parzen window density estimator using the kernel notation as follows.

$$\tilde{p}(x) = \frac{1}{N\sigma} \sum_{x_i \in D} K(\frac{x - x_i}{\sigma}) = \frac{1}{N} \sum_{x_i \in D} k_\sigma(x, x_i) \tag{3.13}$$

where $k_\sigma(x, x_i)$ is the kernel centered at x_i and σ is kernel size. We assume a positive semi-finite Parzen window with Gaussian kernel $k_\sigma(\boldsymbol{x}, \boldsymbol{y}) = exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2})$. 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 $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. 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 [96].

$$\begin{aligned}\tilde{V}(p) &= \frac{1}{N}\sum_{i=1}^{N}\tilde{p}(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k_\sigma(x_i,x_j) \\ &= \frac{1}{N^2}[k_\sigma(x_1,x_1)+k_\sigma(x_1,x_2)+...+k_\sigma(x_1,x_N)] \\ &+ ...+[k_\sigma(x_N,x_1)+k_\sigma(x_N,x_2)+...+k_\sigma(x_N,x_N)] \\ &= \frac{1}{N^2}\mathbf{1}^T\mathbf{K}\mathbf{1}\end{aligned} \tag{3.14}$$

where element (i,j) of the $N \times N$ kernel matrix $\boldsymbol{K}$ is equal to $k(x_i,x_j)$, and $\mathbf{1}$ 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 in the form of algebra.

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

$$\mathbf{K} = \mathbf{EDE}^T \tag{3.15}$$

where $\boldsymbol{D}$ is a diagonal matrix storing the eigenvalues $\lambda_1, ..., \lambda_N$ and $\boldsymbol{E}$ is a matrix with the corresponding eigenvectors $\boldsymbol{\alpha_i}$,...,$\boldsymbol{\alpha_N}$ as columns. Hence the empirical Renyi entropy estimator equals to the elements of the corresponding kernel matrix. We rewrite the above expression to yield the following result [97].

$$\begin{aligned}
\tilde{V}(p) &= \frac{1}{N^2}\mathbf{1}^T\mathbf{K}\mathbf{1} \\
&= \frac{1}{N^2}\mathbf{1}^T\mathbf{E}\mathbf{D}\mathbf{E}^T\mathbf{1} \\
&= \frac{1}{N^2}\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} \alpha_1 & \cdots & \alpha_N \end{pmatrix}\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{pmatrix}\begin{pmatrix} \alpha_1^T \\ \vdots \\ \alpha_N^T \end{pmatrix}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= \frac{1}{N^2}\begin{pmatrix} \lambda_1 1^T\alpha_1 & \cdots & \lambda_N 1^T\alpha_N \end{pmatrix}\begin{pmatrix} \alpha_1^T 1 \\ \vdots \\ \alpha_N^T 1 \end{pmatrix} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}(\sqrt{\lambda_i}\alpha_i^T\mathbf{1})^2
\end{aligned} \tag{3.16}$$

where λ_i and $\boldsymbol{\alpha_i}$ are the i-th eigenvalue and eigenvector of kernel matrix $\boldsymbol{K}$, and $\mathbf{1}$ is a $N \times 1$ vector of ones.

The above expression is known as entropy-value in kernel entropy component analysis (KECA) [91]. Since the total entropy value is estimated by each term $\sqrt{\lambda_i}\alpha_i^T$, certain eigenvalues and the corresponding eigenvectors make more contribution than others. It is noted that both eigenvalues and eigenvectors make contributions to the entropy estimator. Instead of selecting the largest eigenvalues, kernel entropy component analysis selects eigenvalues and eigenvectors based on the largest entropy estimation. This is the most significant property of kernel entropy component analysis. Furthermore, kernel entropy component analysis 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 eigenvalues 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 to derive a semi-supervised fusion method which can realize a more complete, precise and discriminant representation of multiple information sources. Information fusion based on KECA can reduce the dimensionality of input feature vector, while it retains 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 data which are always complementary and redundant.

To exploit the complementary nature of multimodal data, we conclude an optimal mathematical framework for feature level information fusion based on KECA [98]. The following steps summarize the procedure of feature transformation and fusion based on KECA.

(1) The feature vector $\boldsymbol{X} = \{x_1, x_2, ...x_N\}$ is the input data which requires feature transformation and fusion.

(2) Gaussian function is chosen as kernel function and the kernel matrix $\boldsymbol{K}$ with elements $K_{ij} = k(x_i, x_j)$ can be obtained.

(3) As mentioned above, we conduct the eigen-decomposition of $\boldsymbol{K}$ and calculate $\mathbf{K} =$

$\mathbf{EDE}^T$, where $\boldsymbol{D}$ is a diagonal matrix storing the eigenvalues $\lambda_1, ..., \lambda_N$ and $\boldsymbol{E}$ is a matrix with the corresponding eigenvectors $\boldsymbol{\alpha_i}$,...,$\boldsymbol{\alpha_N}$ as columns.

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

(5) Then we can conclude that $\phi_{eca}^T\phi_{eca} = (\mathbf{D}^{\frac{1}{2}}\mathbf{E}^T)^T\mathbf{D}^{\frac{1}{2}}\mathbf{E}^T = \mathbf{EDE}^T = \mathbf{K}$ and calculate the kernel feature space data set $\phi_{eca} = (\mathbf{D}^{\frac{1}{2}}\mathbf{E}^T)$.

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

The proposed criterion does not suffer from the limitation of Gaussianity which is inherent in cost functions based on the second order moments. It 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. This brings robustness and generality, and improves performance in many realistic scenarios.

3.5 The Application to Audio Emotion Recognition

3.5.1 System Design

Speech is one of the most essential and natural verbal channels to transmit human affective states. Moreover, speech is easily accessible for emotion recognition. Speech emotion recognition is useful for many applications such as customer satisfaction assess-

ment, safety in automotive, medical diagnosis tool and so on. A detailed review of the cutting-edge works for audio emotion recognition can be found in [99]. The performance of speech emotion recognition based on multimodal information fusion has also been demonstrated by certain works; however, it is far from an ultimate solution due to unsatisfactory accuracy and efficiency of the proposed solutions. Moreover, the relationship between audio features needs more study. In this thesis, a new fusion solution for audio signal at feature level has been investigated.

The main areas of emotion recognition from speech signals include acquisition of emotional speech signals, feature extraction, feature selection and classification. The final recognition performance heavily depends on feature analysis strategy and classification method. Furthermore, interdependency and correlation of affective features need to be considered. Hence the primary objective of multimodal fusion is to improve the classification results by exploiting the complementary nature of different modalities. In this system, the prosodic features and MFCC (Mel-frequency cepstral coefficient) features are integrated at feature level to get a combined audio feature vector, since it is believed that combining the benefits of continuous features and spectral features is a good choice for audio emotion recognition. Prior to feeding the feature vectors to classification stage, dimensionality reduction and feature fusion should be performed, since the performance of classification critically depends on the discriminant ability of the features. In this thesis,

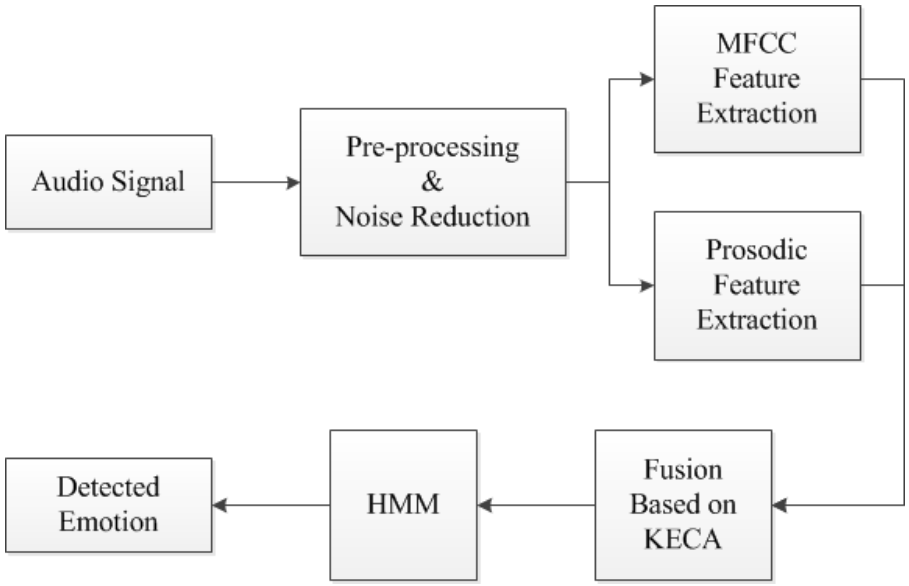

Figure 3.2: System block diagram of information fusion for audio emotion recognition.

a new feature level fusion method based on kernel entropy component analysis (KECA) is used to improve the emotion recognition performance [100]. The resulting feature vectors are input into classifiers to obtain the detected emotion states. Figure 3.2 describes a system block diagram of the proposed architecture of audio emotion recognition.

Audio feature integration is one of the typical examples of early fusion. The goal of feature level fusion is to get reduced set of features by applying transformation on the initial feature vectors. This leads to a change in the representation of the data which could make the data better visualized and understood. The feature fusion strategy achieves two kinds of improvement. The first one is to integrate all audio features by maximally preserving the content of information and select subset features which retain the original

feature characteristics. In most cases, not all features add useful information to the classification problem, since some of them may carry redundant information. The complex classifiers do not work well if the input features do not represent the underlining characteristics of data. Through feature level fusion, the generalization capability of the system can be enhanced and the interpretation ability of models can be improved. The second one is to transform the original space into a transformed space. The extracted features are merged into a single high-dimensional feature set. Since a large feature vector contains rich information about modalities, we usually expect an increase in classification accuracy theoretically. But in practice, the classifiers usually yield unreliable results. Simple concatenation could make classification results meaningless, if the available data are not ample. The growing feature vectors also result to stressful computational resources for classification model training. In order to alleviate these problems, dimensionality reduction is achieved by generating a new feature vector in transformed domain. This could alleviate the effect of "curse of dimensionality" and speed up learning process.

During the stage of classification, the resulting features are viewed as an input to a hidden Markov model (HMM) with Gaussian mixture observation probability density functions. HMM is based on probability algorithm to model sequential data. It is used as a classifier because of its outstanding performance of modeling the temporal characteristics of audio signals. HMM model is trained by the sequences of feature vectors

representative of the input signal corresponding to each emotion category. It calculates likelihood probability value for matching feature vector of each sample. The output of HMM is the optimal likelihood of the different classes which can be considered as the state of the detected emotion. In the next section, the discussion about audio feature extraction and fusion is presented.

3.5.2 Audio Feature Extraction and Fusion

Human speech contains not only linguistic content but also contains certain 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 feature and spectral feature [101, 102]. Continuous speech features (or prosodic speech features) such as pitch and energy are extracted from each frame. On the other hand, spectral speech features are calculated as statistics of all speech features.

Continuous Speech Features

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. By adding new mathematical derivatives of these features, we can get a large number of features. By using these local feature vectors, complex classifiers such as hidden Markov model (HMM) and support vector machine (SVM) can be trained reliably and hence the parameters can be accurately estimated [103]. For example, Suzuki et al. presented an emotion recognition method for synthesized speech based on normalization of prosodic features [104]. Furthermore, Wu et al. presented an approach to emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information [105].

Spectral Speech Features

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 less number of spectral features, the algorithms of feature selection based on spectral feature are executed faster, and the training of classifiers is more efficient. The widely used spectral features include MFCC (Mel-

frequency cepstral coefficient) and LPCC (Linear predictive cepstral coefficient) [106]. It has been shown that the features based on cepstral analysis such as MFCC and LPCC clearly outperform the performance of linear based features like linear predictor coefficient (LPC). In this thesis, we focus on MFCC which could provide the cepstral coefficients derived from mel-scale frequency filter-bank. In order to capture the phonetically important characteristics of speech, it is modeled using a filter bank with filters linearly spaced in lower frequencies and logarithmically in higher frequencies. MFCC can extract the significant emotion components from audio data and represent them according to a Mel-Frequency scale which is identical to the behavior of the human ear. Hence it is a popular analytical tool in the field of speech recognition. One typical example of spectral features is an application of speech emotion recognition using MFCC and wavelet features presented by Krishna [106]. Moreover, Nalini et al. presented a speech emotion recognition system using residual phase and MFCC features with auto-associative neural network [107].

Fusion of Audio Features

Since continuous and spectral features have been claimed to be efficient in distinguishing different emotional states, the integration of two types of features conveys more complementary relationship, which leads to higher classification accuracy. If one modal-

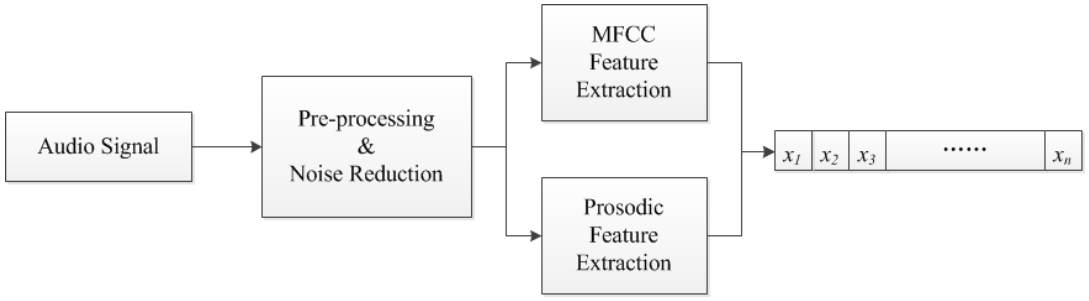

Figure 3.3: Block diagram of audio feature extraction.

ity fails to detect an emotion, the other modality can help improve the performance. In addition to the improved classification performance obtained by multimodal strategy, the system can still continue working as single modality emotion detection, if one of the modalities is absent due to temporary problems. Both continuous features and spectral features have their own advantages and limitations. The complementary relationship of these modalities leads to higher classification accuracy and better adaptability.

A procedure of feature selection and fusion should be conducted to extract the audio features and apply information fusion on the extracted features [108]. Figure 3.3 describes the procedure of the extraction for audio signal. At the stage of pre-processing, we need to reduce the effects of noise in the speech signal. The collected audio data usually contain noise from the background and the recording machine. The presence of noise makes the feature extraction less accurate. In our work, we perform noise reduction by transforming the wavelet coefficients based on thresholding algorithm [109]. First the coefficients of wavelet transform for audio signal is computed. After applied thresholding algorithm,

the inversion of the thresholded coefficients leads to the denoised signal. Compared with the traditional low pass filtering method, this method has the advantage of reducing the noise efficiently without corrupting the original signal. In order to exclude the silence periods, leading and trailing edges are then eliminated since they do not provide useful information. The audio signal should be windowed into a succession of sequences which are called frames. Within each frame, the signal is considered to be stationary and the analysis is reliable [109]. After windowed into frames, a short time analysis of audio signal can be performed within a short time interval of articulatory stability. A Hamming window of size 512 points with 50% overlap between adjacent windows is used to realize speech frames [79]. Since prosodic features are related to the rhythmic content of audio signal, they are normally represented by the statistics of intensity, pitch, fundamental frequencies, formant frequencies, etc. In this thesis, the statistics and variations of pitch, energy, pause length, speaking rate and formant frequencies are extracted as continuous features [79]. Regarding spectral features, the MFCC features are employed to extract spectral features in this thesis. MFCC has physical connection with human ears and is widely used to mimic human auditory system by ceptral analysis. The MFCC features are calculated based on speech frames. Since most of the information energy is stored in the first few coefficients of MFCC features, the first thirteen coefficients are used as useful features. The statistics of each coefficient, such as median, mean, standard deviation,

maximum value and minimum value, are integrated to form a feature vector [99]. The continuous and spectral feature vectors of successive frames are concatenated as the audio feature vectors. Then we apply kernel entropy component analysis to find transformed features in another space based on entropy estimation. This approach of information fusion leads to not only the improved classification performance, but also the adaptability and scalability of emotion recognition system. In the following section, the feasibility and superiority of the proposed solution are demonstrated though extensive experiments.

3.5.3 Experiments

Experimental Databases

In order to evaluate the performance of the proposed strategy and make a comparison with the existing methods, extensive experiments have been conducted on two publicly available databases, eNTERFACE emotion database [22] and RML emotion database [23]. The example images of the two databases are displayed in Figure 3.4. The RML emotion database contains 720 audiovisual emotional expression samples from eight human subjects speaking six languages. In RML database, six different accents include English, Mandarin, Urdu, Punjabi, Persian, and Italian. A set of six principal human emotions consists of anger, disgust, fear, sadness, surprise, and happiness. The

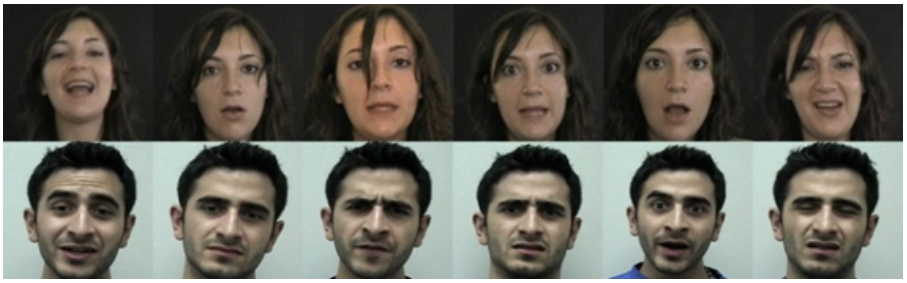

Figure 3.4: Example images of eNTERFACE database (top row) and RML database (bottom row).

experimental subjects were provided with a list of emotional sentences and were directed to express their emotions as naturally as possible by recalling the emotional happening, which they had experienced in their lives. A total number of ten different sentences were provided for each emotional class. The samples were recorded at a sampling rate of 22,050 Hz and a frame rate of 30 frames per second (fps) using a single channel 16 bit digitization. The eNTERFACE database contains 42 subjects coming from 14 different nationalities. The database also provides the six basic emotion states. All the experiments were driven in English. Each subject listened to six successive short stories which elicit a particular emotion, after that they had to react to each of the situations. The samples were recorded with a sampling rate of 48,000 Hz and a frame rate of 25 fps.

The performance of emotion recognition systems heavily relies on the qualities of the training and testing data and their similarity to real samples. A good database can achieve better research results. Selecting an emotion database is a task with several chal-

lenges, such as the quality of the samples, good emotional performances, management of huge amounts of information and so on. Some of the existing databases fall short in many necessary aspects. They usually have low levels of recording quality and suffer from copyright issues. In this thesis, we select eNTERFACE emotion database and RML emotion database which are designed to fulfill the need of a common database for multimodal emotion recognition. These two databases can be used as reference databases for testing and evaluating audio, video or joint audio-visual emotion recognition algorithms. These two databases have a good quality in representing characteristics of the emotional data. The data samples were recorded using digital video cameras and the recording of speech signal was realized through high-quality microphones. The background consists of a monochromatic panel that covered the entire area behind the subject, which makes face detection and tracking easier. Moreover, these two databases present rich emotional contents and target emotional reactions with specific tasks instead of sporadic periods of emotion. In my opinion, eNTERFACE database is the best database for our experiment purpose in comparison with other available databases. Hence eNTERFACE database is employed in this thesis to perform audiovisual emotion analysis. Since RML database is developed by our lab and it is used in our previous papers, it is also picked in this thesis for a comparison study. These two databases are freely distributed database for research purposes. Therefore we use eNTERFACE database and RML database as the

main benchmark databases for the performance evaluation of the emotion recognition system.

Experimental Setup

In the experiments, each sample is truncated to 2 second long clip which presents both audio and video data and each sample is divided into 10 segments. The dimensionalities of audio features are set to 240 [23]. The total numbers of samples from eNTERFACE and RML databases are 600 and 400 respectively. There is no overlap between the training and testing samples. The class distribution of the experimental data is balanced. The experimental evaluation procedure is based on leave-one-out cross-validation basis which is the most frequently used approach for testing classification performance. According to this approach, one sample is selected from the entire data set as the testing data for each time, while the rest of the data is used as the training data. This procedure continues until all the individual samples have been held out once. This procedure can make good use of the available data. The ratio of the number of correctly classified samples over the total number of samples is equal to the recognition accuracy.

Experimental Results

In our experiments, we compare our KECA based solution with two related traditional algorithms which are KPCA and KCCA based methods. Kernel principal compo-

nent analysis (KPCA) and kernel canonical correlation analysis (KCCA) are kernelized versions of PCA and CCA. Principle Components Analysis (PCA) is a traditional approach of reducing the dimensionality of a data set. It is an unsupervised method to find the lower-dimensional representation of the data which preserves most the data's variance. Canonical correlation analysis (CCA) seeks to maximize the correlation between principle directions in two or more separate data domains or modalities. It is a supervised technique which provides dimensionality reduction and obtains more significant amount of information. By generalizing the kernel method into PCA and CCA, we obtain the kernelization forms, KPCA and KCCA, which are widely used in nonlinear feature extraction. Hence we compare our proposed solution with KPCA and KCCA based methods in the aspects of dimensionality reduction and feature transformation.

Figure 3.5, Figure 3.6, Figure 3.7 and Figure 3.8 describe the comparison of overall recognition accuracy between KECA, KPCA and KCCA on eNTERFACE database and RML database in terms of percentage. The parameter σ stands for kernel size in all figures and it is set from 0.2 to 0.8 with a step size of 0.2 for comparison. The recognition accuracy is estimated by the ratio of the correctly classified samples compared to the total samples. From these figures, it can be seen that KECA has better accuracy than KPCA and KCCA, and it outperforms KPCA and KCCA in dimensionality reduction and accuracy performance. The overall recognition accuracy of the proposed method

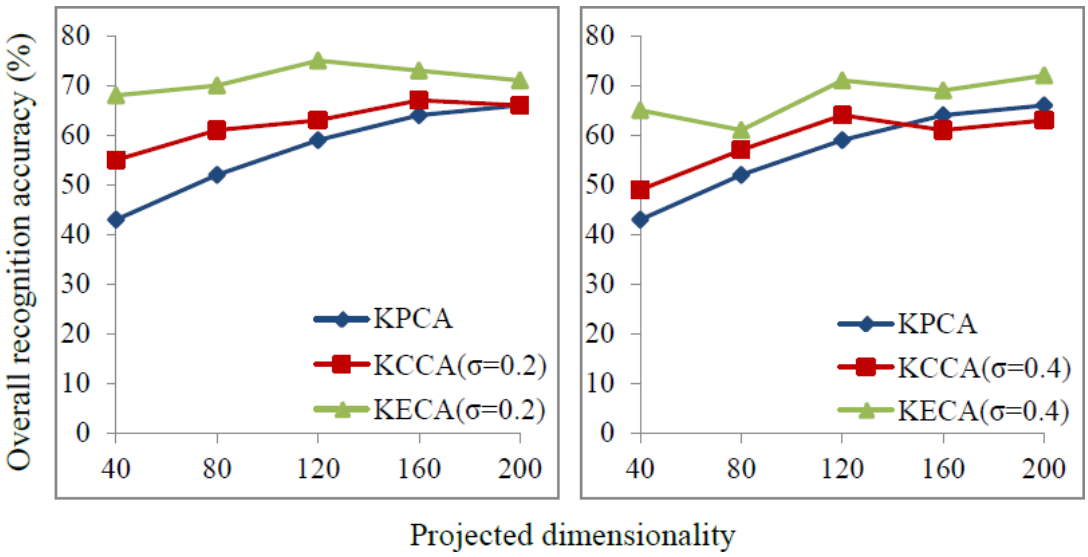

Figure 3.5: Experimental results of eNTERFACE database. The feature level fusion is based on KECA, KPCA and KCCA. Left: σ=0.2; Right: σ=0.4

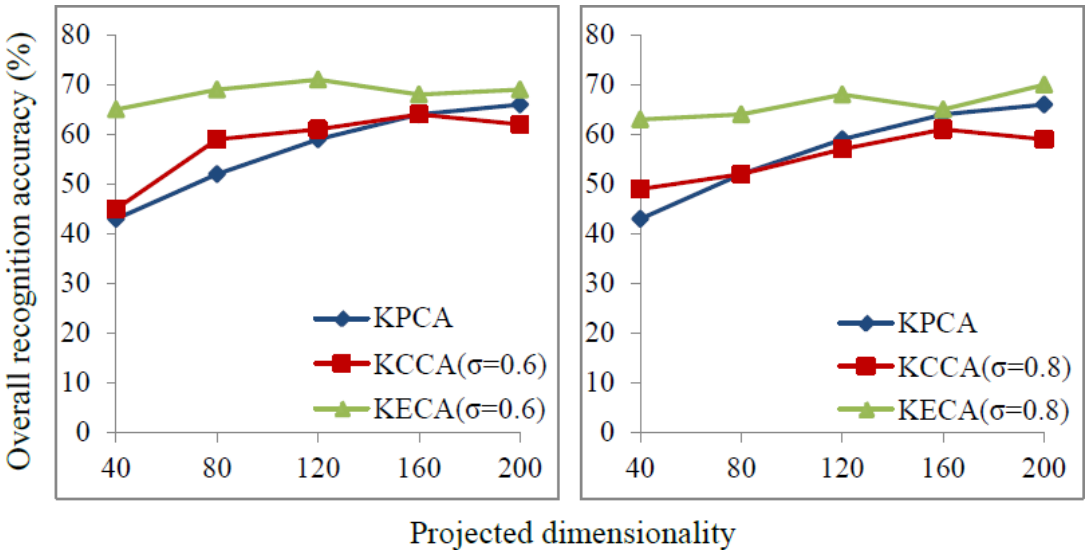

Figure 3.6: Experimental results of eNTERFACE database. The feature level fusion is based on KECA, KPCA and KCCA. Left: σ=0.6; Right: σ=0.8

based on KECA levels off even if the projected dimensionality is low. This shows that the extracted features at high dimension may contain redundant or noisy data. After

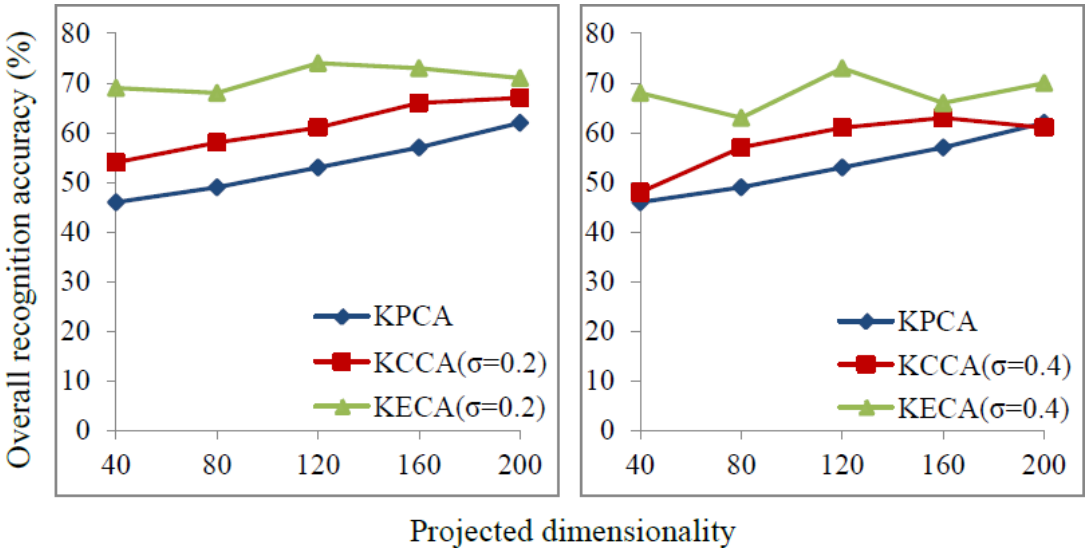

Figure 3.7: Experimental results of RML database. The feature level fusion is based on KECA, KPCA and KCCA. Left: σ=0.2; Right: σ=0.4

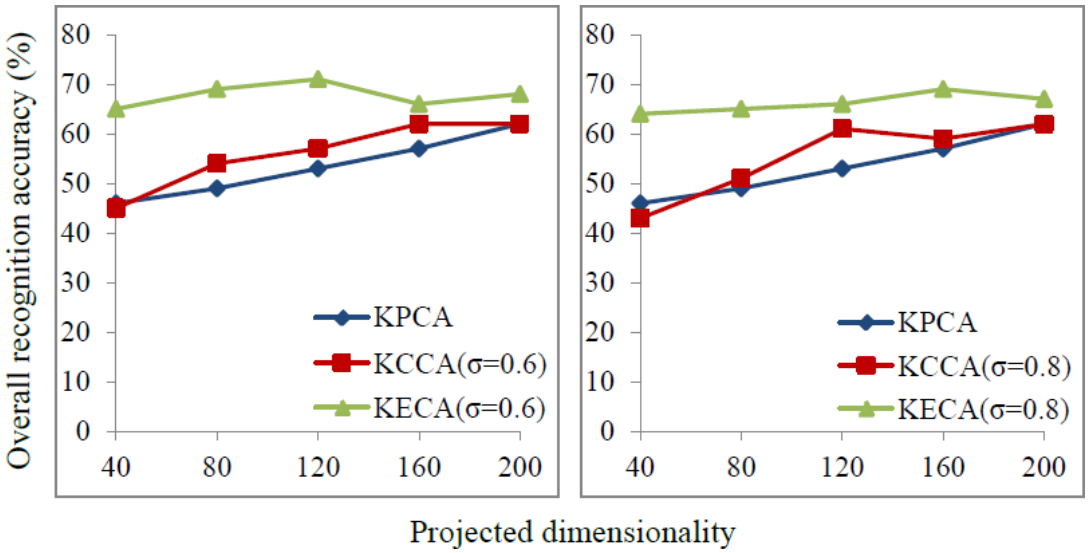

Figure 3.8: Experimental results of RML database. The feature level fusion is based on KECA, KPCA and KCCA. Left: σ=0.6; Right: σ=0.8

processed by fusion method based on KECA, most of useful information is preserved and stable accuracy is achieved. On the other hand, the performance of KCCA and KPCA is

largely degraded if the projected dimensionality decreases, which means that the fusion method is not properly selected and the degraded results are generated. These results demonstrate the improvement of dimensionality reduction and the ability of preserving useful information of the proposed solution.

Table 3.1, Table 3.2 and Table 3.3 display three 6 by 6 confusion matrices of average performance based on KECA, KPCA and KCCA on eNTERFACE database and RML database. The kernel size σ is set to 0.2. The projected dimension is set to 120. Confusion matrix is a widely used visualization tool for assessing classification performance. In these confusion matrices, the percentages of samples correctly and incorrectly classified for each class are indicated. In a confusion matrix, the rows represent the actual class of the samples while the columns represent the detected class of the samples. The element $M[i][j]$ at the i-th row and the j-th column indicates the classification percentage of samples belonging to class i which are recognized as class j. Hence the diagonal elements in confusion matrix show the percentage of correct classification for different classes, while the off-diagonal elements denote the percentage of misclassification. Confusion matrix is a simple and understandable way to display the results of recognition systems based on the percentages of correct and incorrect classification. From these tables, it is obviously noted that the results of KECA are more satisfactory. Compared with the methods based on KPCA and KCCA, the strategy based on KECA has better accuracy and stability.

Actual Emotion(%)	Detected Emotion(%)					
	Happiness	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Happiness	**76.54**	4.42	5.51	5.13	6.21	2.19
Disgust	6.34	**69.04**	6.46	6.65	5.19	6.32
Fear	5.41	3.53	**77.31**	3.09	3.31	7.35
Angry	4.12	2.81	4.86	**76.87**	6.45	4.89
Surprise	3.14	5.51	5.42	4.55	**77.34**	4.04
Sadness	7.52	5.02	6.56	6.89	4.5	**69.51**

Table 3.1: Confusion matrix of average performance on two databases based on KECA.

Actual Emotion(%)	Detected Emotion(%)					
	Happiness	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Happiness	**60.23**	10.23	4.25	8.54	12.23	4.52
Disgust	13.2	**58.23**	7.23	7.14	4.11	10.09
Fear	8.21	7.24	**55.21**	8.67	12.34	8.33
Angry	14.51	5.67	4.65	**54.89**	9.87	10.41
Surprise	8.41	8.34	10.23	6.55	**54.13**	12.34
Sadness	6.7	9.87	9.56	6.89	9.66	**57.32**

Table 3.2: Confusion matrix of average performance on two databases based on KPCA.

Actual Emotion(%)	Detected Emotion(%)					
	Happiness	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Happiness	**63.51**	8.24	9.65	4.67	7.48	6.45
Disgust	7.51	**62.24**	7.4	5.41	6.12	11.32
Fear	12.41	4.57	**59.78**	10.63	4.56	8.05
Angry	7.51	4.53	4.05	**66.1**	9.25	8.56
Surprise	8.61	7.44	10.54	6.53	**62.09**	4.79
Sadness	5.41	6.45	12.51	11.36	5.07	**59.2**

Table 3.3: Confusion matrix of average performance on two databases based on KCCA.

Figure 3.9 and Figure 3.10 show the comparison of average recognition accuracy between fusion and non-fusion methods on eNTERFACE database and RML database. The recognition accuracy is defined as the percentage of correctly classified samples. It is observed that the fusion result is superior to that obtained by non-fusion methods with

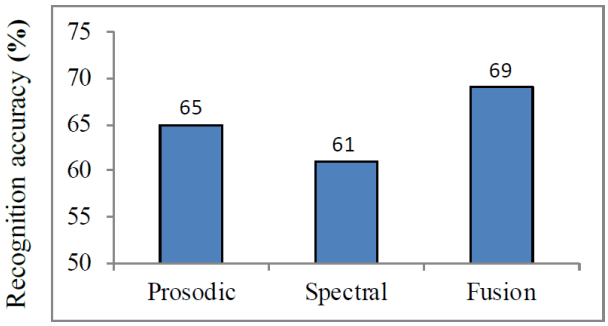

Figure 3.9: Comparison between fusion result and non-fusion result on eNTERFACE database.

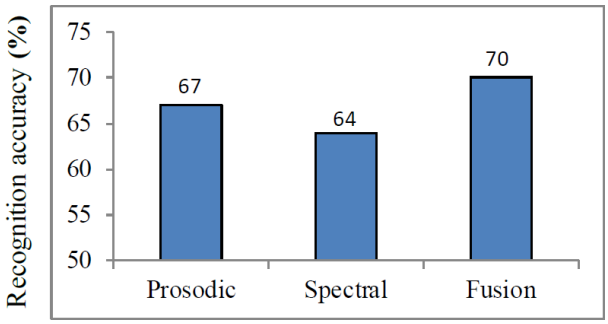

Figure 3.10: Comparison between fusion result and non-fusion result on RML database.

prosodic feature or MFCC feature involved. The integration of prosodic and spectral features enhances the performance of recognition system. This confirms the effectiveness of the fusion solution based on kernel entropy component analysis. Therefore, from the above experimental results, it is concluded that the overall performance of KECA outperforms the existing methods, like KPCA and KCCA. In addition, the experimental results and comparisons have revealed the good performance of the proposed audio emotion recognition system. It provides best recognition accuracy and more efficient dimensionality reduction, and more stable recognition results.

3.6 Summary

This chapter proposes a novel solution of feature level fusion using entropy estimation to audio emotion recognition. It has presented systematic analysis of information entropy based methods for addressing the challenging problems of feature vector fusion and dimensionality reduction in feature level fusion. Detailed mathematical analysis on information theoretical tools is presented. The connection between information theory and information fusion is discussed. To improve the performance of emotion recognition as well as achieve a more sophisticated fusion strategy for human computer interface, a new feature level fusion strategy based on kernel entropy component analysis is applied in the application of speech emotion recognition in this chapter. The experimental results are also presented.

Chapter 4

Dual-Level Fusion

4.1 Overview

Since humans rarely express their emotions exclusively, several channels such as speech and facial expression need to be considered. In this chapter, combined with the work presented in Chapter 3, we introduce a dual-level (feature level and score level) audiovisual fusion method for emotion recognition. Since speech and facial expression are generally complementary to each other in emotion recognition, correlation between audio and visual channels is fully utilized at score level, leading to enhance the performance of score level fusion, and, through the introduction of the dual-level fusion method, to improve audiovisual emotion recognition. In our system, the prosodic and MFCC features are

extracted from audio signals. Visual stream is analyzed and identified using Gabor filter and EBS (Elastic Body Spline) model. Since score level fusion can explore the contributions of various modalities without increasing the dimensionality, feature level fusion and score level fusion modules are used to jointly assist in decision making. In order to improve the recognition performance at score level, a new score level fusion method based on information theoretic tools, maximum correntropy criterion (MCC) in particular, is proposed. The performance of the proposed audiovisual emotion recognition system is evaluated on eNTERFACE and RML emotion databases. Comparison with existing bimodal emotion recognition methods has also included. On the basis of the experimental results, it is shown that the proposed solution provides better performance than single modality and other fusion strategies compared.

4.2 Introduction of Score Level Fusion

Like feature level fusion, score level fusion is also commonly practiced in multimodal information fusion. It usually combines the scores generated from multiple modalities through a rule based scheme which is realized through algebraic operations, such as weighted sum and multiplication, or through a pattern classification strategy in which the scores are treated as the input of classification algorithms [110]. The advantage of

score level fusion is that it contains rich information about quantitative similarity measurement and relatively easy to process. In certain cases, it is able to achieve theoretically optimal performance. Since score level fusion offers good tradeoff between performance and complexity, various fusion applications based on score level have been developed.

Generally speaking, the techniques of score level fusion can be divided into three categories including rule based, classifier based, and density based [111]. These three approaches have their own limitations, and none of them guarantees optimal performance. The merits and drawbacks of the existing methods, and their typical examples are described as follows.

4.2.1 Rule Based Fusion

In rule based fusion, all matching scores are first transformed into a comparable scale in a common domain and then processed through a certain algebraic combination rule to make final decision [112]. This method takes into consideration of two factors, one is normalization function, and the other is fusion rule. Given training samples of scores, score normalization is often incorporated to achieve better performance. The choice of normalization schemes and combination weights is data dependent and requires extensive empirical evaluation. The typical examples of the fusion rules are summation, average, product, minimum, maximum, median, etc. For instance, one simple fusion approach is

a sum rule where single modal scores are summed up to provide a final score, while a weighted sum based rule computes the combined score based on a weighted sum of the matching scores [113]. Hence this fusion method has the ability to adapt to variations in the input data and the significance of individual modality is leveraged in the final decision. Moreover, rule based fusion techniques require no training process and few consideration of matching score distribution, and it is easy to implement. However, this fusion is usually heuristic and does not guarantee optimality.

4.2.2 Classifier Based Fusion

Classifier based fusion concatenates the matching scores into vectors and treats the vector as feature vectors which are classified into one of the possible classes [114]. The widely used classifiers include support vector machine (SVM), neural network, linear discriminant analysis (LDA), k-NN and so on [115, 116]. A number of challenges for classifier based fusion include cost of misclassification and choice of classifier. Depending on different applications, the cost of accepting an impostor may be different from the cost of rejecting a genuine user. Moreover given a variety of classifiers, selecting and training a classifier which gives the optimal performance is not trivial. In this fusion, the classifiers need to learn a decision boundary between the classes based on the training set of matching scores. However, in the application of multimedia, especially emotion

recognition, the decision boundary between different states can be quite complex. The traditional classifiers are usually not capable of obtaining the boundaries. Moreover, the output scores from different modalities may be not homogeneous.

4.2.3 Density Based Fusion

Density based fusion is based on likelihood ratio and it requires explicit estimation of genuine and impostor matching score densities [111]. The matching scores are firstly transformed into posteriori probability and final decision is made according to pre-defined rules. The score distribution estimation of this approach usually relies on the methods such as the well-known naive Bayesian and Gaussian mixture model (GMM) [117]. However, in density based fusion, the densities are usually unknown and have to be estimated from a set of training scores based on parametric or non-parametric methods. The density based approach has the advantage that it directly achieves optimal performance at any desired operating point, provided the score densities are estimated accurately. On the other hand, the main challenges for this fusion are accurate estimation of the probability distributions of different modalities and huge number of training samples. For instance, genuine matching scores are limited, and it is difficult to estimate the density of matching scores which may not obey a certain distribution model [118]. Hence density based fusion is hard to carry out.

Most of the existing methods have not taken full advantage of intrinsic characteristics of the matching score from different modalities [119]. Their performances are usually degraded due to lack of sufficient training data and noisy training samples. In this thesis, we propose an optimal framework which explicitly takes into consideration the inherent relationship of multiple data sources at score level. Different modalities describe different semantic expressions, especially in multimedia related analysis tasks. It is highly possible that the presence of one modality presents an explicit or implicit impact over the other modalities. For instant, in the application of emotion recognition, speech and facial expression usually have an effect to another if the user expresses certain emotional behavior. Hence it is important to effectively identify and utilize the relationship between different modalities at score level.

Moreover, the experiments show that there are no sharp boundaries between different emotion states, which reveal the nature of human emotion. Since the number of emotion classes is large, the boundaries between different classes tend to be complex and hard to separate. So it is necessary to develop a method which is discriminant and representative enough to guarantee good generalization capabilities. In this thesis, a novel score level fusion method based on maximum correntropy criterion (MCC) is proposed. Correntropy is effective similarity measurement in information theory which has the stability to variation or noise [120]. Derived from information theoretic learning, maximum corren-

tropy criterion is a new evaluation function based on correntropy. It has the advantage that it provides a robust adaptation principle in presence of non-Gaussian signals. The traditional methods assume the signals to be Gaussian. But in practice, the input data are usually nonlinear and non-Gaussian with large noise and variation. Therefore, using maximum correntropy criterion as a cost function is well justified due to its ability to characterize the entire structure of the data.

4.3 Similarity Metric with ITL Principles

The structure of information is a vague and difficult concept to quantify, however it might comply with identifiable patterns which can be distinguished by the shape of probability density function. The major problem of data structure measures is how to evaluate the metrics without imposing unrealistic assumptions about the data distributions. Unlike the conventional methods, information theoretic strategies are particularly promising to capture data structure beyond the second order statistics [121]. This stems from the fact that entropy and other information theoretic measures are scalars which summarize the information contained in the data distributions in relevant ways. The core ideas of information theoretic learning algorithms evaluate pairwise data interactions and extract more information from data than the results based on single data samples such

as mean or variance. Moreover, because information theoretic learning provides efficient nonparametric estimators of entropy, it allows the description of data structure in a more meaningful way than what is commonly done in traditional methods.

In the cases of supervised learning, labels are available and they are usually of the same dimensionality of the system output, which is one of the significant features of supervised learning. Therefore a composite random variable can be defined, and the error containing information about the differences in the distribution of the desired output and the system output can simplify the adaptation. For the cases of unsupervised learning, we only have the input data. Finding the structure in the data requires a method which is able to quantify similarity or criteria elucidating relationships among the data samples in details. The evaluation of similarity or dissimilarity based on information theoretic descriptors is gradually applied to the area of unsupervised learning. Because entropy and the concept of similarity are effective descriptors of probability density function, they usually form the foundations for unsupervised learning.

From an analytical viewpoint, the best way for categorization is to use the information about the input data distribution to separate inputs into groups which share the same region in data space. However, this method works well only when the data are clearly separated and compact, but it usually fails if the data are close to each other, or overlapped with each other, and produce nonlinear boundaries. A better way of measuring

the distance between data is to weight the distance between samples nonlinearly. This motivates us to obtain a reasonable nonlinear weighting function based on a form of similarity or dissimilarity estimation. The fundamental issue is how to solve this problem by using local learning rules in practical situations. Moreover it is important to understand that the second order statistics cannot solve this problem well.

4.4 Correntropy and Maximum Correntropy Criterion

Similarity is one of key elements to quantify temporal signals or static measurements, but it is difficult to define mathematically. In traditional signal processing, the measurement of similarity is usually implemented based on the second order statistics usually in the forms of correlation and mean square error (MSE). The optimality of the second order statistics largely depends on the assumption of Gaussianity and linearity. Hence successful pattern recognition solutions from these methodologies rely heavily on the Gaussianity and linearity assumptions. There are many examples of how engrained the second-order moment descriptors of probability density function are in engineering thinking.

However, in practical applications, it is insufficient to assume that the data distribution is Gaussian. Recently the concept of information theoretic learning has been

extended to the processing of nonlinear and non-Gaussian signals with large variation and noise. ITL could preserve the nature of data since the evaluation function is directly estimated from the data. In this thesis, we utilize correntropy as a new localized similarity measurement based on the theoretical framework of ITL. Correntropy is a bivariate function which produces a scalar, but it contains the second and higher order PDF moments, which are expressed by the kernel used in its definition [122]. Correntropy essentially generalizes the conventional correlation function to high-dimensional spaces. It could help us deal with variability and uncertainty and it is intrinsically different from the conventional techniques. Moreover, correntropy induces a new information theoretic metric which behaves similarly to the 2-norm, 1-norm or zero-norm distance under different scenarios. This geometric interpretation elucidates the robustness of correntropy for noises and outliers.

Correntropy is a generalized similarity measure which exploits higher order moments of the probability density function (PDF). It is shown that correntropy is directly related to the probability of how similar the random variables are in a neighborhood of the joint space controlled by the kernel bandwidth. The bandwidth of kernel works as a spotlight, and controls the observation window. In the observation window, similarity is assessed. This adjustable window can provide an effective mechanism to eliminate the detrimental effect of noises and outliers, and it is intrinsically different from the use of a

threshold in conventional techniques. One merit of correntropy is simplicity of estimation directly from samples. The original definition can be properly called auto-correntropy which only applies to a single random process [123]. Auto-correntropy is enhancement of the auto-correlation function widely used in time series analysis. Auto-correntropy defines a generalized correlation function in terms of inner products of vectors in a kernel feature space. The feature of auto-correntropy is that it combines statistical and temporal information.

In comparison with auto-correntropy function, cross-correntropy measures the statistical dependency between random variables at two different times [123]. Cross-correntropy is simply called correntropy which extends the definition to the general case of two arbitrary random variables. A general form of correntropy for two random variables X and Y is formally defined as [124].

$$V_\sigma(X,Y) = E[k_\sigma(X,Y)] \tag{4.1}$$

where $E[.]$ denotes the mathematic expectation and $k_\sigma(.)$ is a kernel function satisfying Mercer's theory. It is observed that correntropy can also be used as a similarity measure in the joint space, but it differs from mean square error (MSE). If the joint distribution of random variables is unknown and only a finite number of samples (x_i, y_i) $i = 1, 2, ..., m$

are given, the sample estimator of correntropy can be calculated as follows.

$$\tilde{V_\sigma}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}[k_\sigma(x_i, y_i)] \tag{4.2}$$

where $k_\sigma(x_i, y_i) = exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2})$ which is the Gaussian kernel with a bandwidth σ.

Since correntropy is related to the similarity between two random variables, a large correntropy stands for a close relationship between these variables. Hence, the maximum value of correntropy is called maximum correntropy criterion (MCC) which is defined in [124] as

$$MCC(\sigma) \Longleftrightarrow max\frac{1}{N}\sum_{i=1}^{N}[k_\sigma(x_i, y_i)] \tag{4.3}$$

As we can see, correntropy can be viewed as a generalized correlation function between random variables. It has been shown that the capabilities of preserving nonlinear and non-Gaussian characteristics enable correntropy to be employed as a measure for determining nonlinear dynamics. Regarding the geometric interpretation, correntropy behaves as a metric equivalent to the 2-norm distance if the data points are close, while it provides the results similar to the 1-norm distance as the data points get further apart. Eventually it approaches the zero-norm as data points are far apart. This geometric interpretation elucidates the robustness of correntropy. Moreover, MCC derived from correntropy at-

tracts much attention in machine learning community, since it could serve as a more robust adaptation metric applied for parameter optimization. Hence, the mechanism of MCC offers a feasible and powerful alternative to the second order statistics. MCC is inherently insensitive to outliers and it is more robust to deal with imprecision and uncertainty of nonlinear and non-Gaussian signals without the knowledge of the underlying data density.

In the conventional methods, mean squared error (MSE) is contrast to correntropy. MSE is a quadratic function of similarity measure in the joint space. MSE is not always the best possible criterion to use in adaptive learning. In fact, the MSE function performs well when the statistics are zero mean and the distribution is Gaussian. In many practical cases, these conditions are violated by noises which can make the distribution non-Gaussian. This makes sense to study alternative functions and take a different approach using information theoretical concepts. There are advantages of using MCC algorithm over MSE when the signals of interest are non-Gaussian. MCC has a great advantage with respect to MSE in terms of computational complexity. Moreover, correntropy cost functions may provide faster convergence for the same adjustment.

4.5 The Proposed Score Level Fusion Framework

Extensive experiments demonstrate that there are no sharp boundaries between the emotional states [99]. Hence it is not easy to distinguish emotional states by maximizing the geometric distance for emotion application. Since the human perception of emotion is heavily based on the integration of different patterns from voice and facial expression, our strategy is to identify the intrinsic association between different modalities based on a measure of similarity. The proposed method based on maximum correntropy criterion optimization technique is developed to transform the score level data into more discriminant information which maximizes the inter-class dissimilarity and minimizes the intra-class similarity. In this thesis, the input samples are $X = \{\{x_i^l\}_{i=1}^{n_l}\}_{l=1}^{C}$, where n_l of samples belongs to class $\omega_l (l = 1, 2, ..., C)$. We define $S_{inter} = \sum_{l=1}^{C}(u_l - u)(u_l - u)^T$ as the inter-class scatter matrix and define $S_{intra} = \sum_{l=1}^{C}\sum_{i=1}^{n_l}(x_i^l - u_l)(x_i^l - u_l)^T$ as the intra-class scatter matrix, where u_l is the mean vector of class ω_l and u is the global mean of all the samples. The optimal transformation is realized by maximizing the ratio of $J_R = (R^T S_{inter} R)/(R^T S_{intra} R))$. If the denominator of the cost function is simplified to $R^T S_{intra} R = I$, the optimization problem can be converted to a problem of maximizing $J_R = R^T S_{inter} R$ [125].

The traditional cost functions provide global measurement and all data points in

the joint domain contribute equally. Unlike the existing methods, correntropy is a local similarity measurement whose value is mainly determined by the kernel function. Moreover noise and variation have less impact on the correntropy measurement. Due to the properties of correntropy, maximum correntropy criterion can be applied to achieve more robust and stable optimization performance and handle non-Gaussian distribution with large variation. Substituting $x_i = (u_l - u)$ and $y_i = RR^T(u_l - u) = RV_l$ into the expression of maximum correntropy criterion leads to correntropy based optimization problem as follows [125].

$$\begin{aligned} \max_R J_{MCC} &= \sum_{l=1}^{C} k_\sigma((u_l - u) - RV_l) \\ s.t. \quad & R^T S_{intra} R = I \end{aligned} \tag{4.4}$$

Since R is orthonormal, the expression can be rewritten to the following cost function.

$$\begin{aligned} \max_R J_{MCC} &= \sum_{l=1}^{C} k_\sigma(\sqrt{{U_l}^T U_l - U_l^T R R^T U_l}) \\ s.t. \quad & R^T S_{intra} R = I \end{aligned} \tag{4.5}$$

where $U_l = u_l - u$ and k_σ is Gaussian kernel with a bandwidth σ. The cost function can be optimized by many methods. In order to achieve fast convergence, a half quadratic based algorithm is applied to solve the nonlinear optimization problem of maximum

correntropy criterion based method [126]. The algorithm can be summarized as follows.

(1) The input vectors consist of the matching scores from the score level.

(2) Then initialize the basic vectors R where $R^T R = I$.

(3) Next we need to deal with iteration. In each iteration, we calculate the expression $p_l = -k_\sigma(\sqrt{{U_l}^T U_l - U_l^T R R^T U_l})$. We continue update R according to $S_{intra}^{-1} S_{inter} W R = \lambda R$ until the convergence.

(4) Then we obtain transform matrix R.

Score level fusion faces several challenges, especially for multimedia applications. The matching scores generated from different modalities, like audio and video channels, are diversified and heterogeneous. For example, an emotion recognition system usually deals with the vast variance and diversity of vocal data and facial expressions. The audiovisual signals are influenced by noise, occlusions, and subtle changes in face expression. Hence, it is necessary to convert these scores into a discriminant representation with the same nature. Moreover the uncertainty and potential imprecision about classifiers should be considered. However, the existing approaches are mostly rule-based fusion methods which largely depend on weighted level of each modality. They do not account for the dynamic properties of input signals. Their principles of classification are usually based on geometric distance, which is heuristic and does not reveal the nature of the signals. Another interesting observation is that there is not even a single feature which is signifi-

cant for all the classes. This actually reveals the nature of human emotion and there are no clear boundaries between emotions. One emotion might have similar patterns with some of the other emotions, while another emotion may have different patterns with the rest. Hence the human perception of emotion is based on the integration of different patterns. Many previous works have been restricted to pre-defined rules [111]. They usually do not have a satisfactory performance in realistic scenarios. Therefore, we need new fusion strategies which could result in small intra-class variation, large inter-class variation, and robustness to noise.

In this chapter, we employ maximum correntropy criterion (MCC) as similarity measurement to improve the accuracy, efficiency, robustness, and fault tolerance of score level fusion. In practical application, the accurate estimation of the joint densities for all matching scores is not always possible because of the limited availability of training data. The underlying reason to select the strategy based on MCC is that the estimation of entropy requires no prior, or relatively minimal, knowledge of the data structure. Experimental results demonstrate that emotional states do not have clear-cut boundaries; hence it is difficult to characterize the joint characteristics of different modalities. To address this problem, the proposed MCC based framework can transform the scores into a more discriminant domain before the matching stage and take full advantage of as much data as possible to distinguish two or more different emotions. Hence, the complementary

relationship of multiple modalities can be effectively utilized at score level. Compared with most traditional methods which are empirically determined and computationally expensive, our approach improves robustness and fault tolerance of score level fusion in emotion recognition application. Therefore, the MCC based score level fusion method provides robust classification ability and high flexibility in various scenarios.

4.6 The Application to Audiovisual Emotion Recognition

4.6.1 System Design

As pointed out previously, since human emotional behavior is a dynamic, complex and multimodal process, bimodal information from speech signal and facial expression will be considered in this work. Facial expression is acknowledged as one of the most direct channels to express human emotions on non-verbal communication. Recent trends in the research field of emotion recognition emphasize the combination of audio and video streams to achieve improvement in the recognition performance [127]. In audiovisual emotion recognition, the mutual correlation of the two modalities utilized by information fusion has flexibility in various scenarios. When one modality is corrupted, the other

modality plays a more important role and helps to improve the performance. Moreover matching scores at score level contain sufficient information to distinguish and they are relatively easy to obtain. Therefore, based on the proposed score level fusion method and the feature level fusion method presented in Chapter 3, this chapter presents a new dual-level fusion solution for audiovisual signals at feature level and score level [128]. In this system, information fusion at feature level is based on kernel entropy component analysis (KECA) described in Chapter 3. The obtained features are utilized as input for HMM classifiers, and the resulting scores from HMM are analyzed by the score level fusion method based on MCC to obtain the final decision. Figure 4.1 sketches an overview of the architecture of the proposed recognition system. It contains four main blocks: a) feature extraction, b) feature level fusion module, c) classification, and d) score level fusion module.

As shown in this figure, at first feature extraction from the audio and visual channels has been performed, separately. The bimodal features derived from the emotion samples are not suitable to be used directly in classification and more representative features transformed from original data are needed. We process the extracted audio and visual features by the proposed KECA based feature level fusion approach. As presented in Chapter 3, KECA provides audio and visual information with dimensional reduction and information fusion at the feature level, and the resulting features from successive

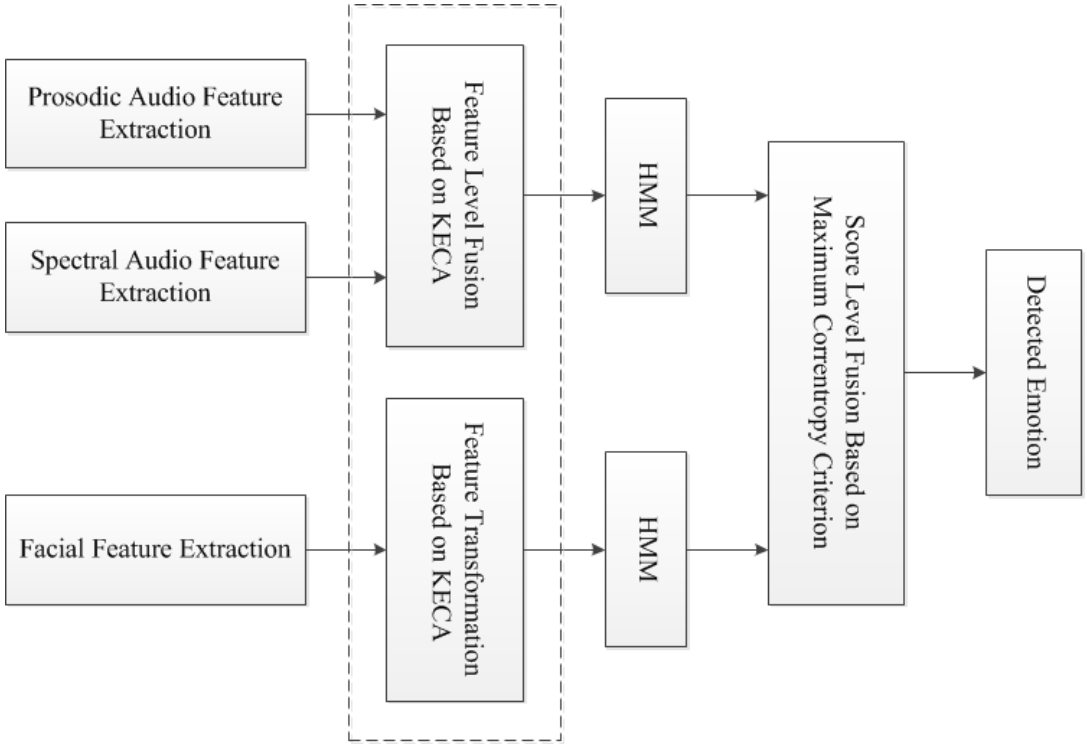

Figure 4.1: System block diagram of multimodal fusion solution for audiovisual emotion recognition.

segments are subsequently modeled by hidden Markov model (HMM). HMM identifies the inherent temporal structure of the features, and contains the likelihood of the samples with respect to different classes. HMMs can exploit the information of correlation among audiovisual modalities which are used to describe the temporal dynamics of the emotion cues between the audio and visual signal streams. Two HMMs are constructed for audio and visual channels respectively. The obtained features of audio and visual channels in the transformed domain are considered as the input to their respective HMM, and the outputs of each modality are treated as the matching scores. In score level fusion, the

intrinsic relationship of audio and visual channels is utilized through the introduction of MCC, and the final recognition results are obtained through score level fusion from the obtained audiovisual score values. The proposed score level fusion method is to further assist the decision making in the final fusion module. Therefore, the fusion of audiovisual information is performed at feature level and score level. The dual-level fusion strategy integrates multiple streams by considering the correlation and contribution of different streams to obtain improved recognition result.

Score level fusion is a process of combining intermediate results. The benefits of score level fusion used in emotion recognition system are twofold. The first is to solve the problem based on "divide and conquer". The underlying data set and the corresponding feature distribution of audio and visual channels could be too complex for a single classification module to learn. By using a divide and conquer approach, the bimodal feature spaces are divided into several distributions. Each part is handled by a classification module, which makes the problem easier. The second merit of score level fusion is the efficiency of training. The computational efficiency suffers from training and evaluation of a single classifier with large databases. In many audiovisual applications, partitioning of data into different subsets and combining the intermediate results to a final decision often prove to be more robust, time-saving, and competitive strategies.

The key issues in bimodal emotion recognition include the synchronization of fusion

process and the appropriate levels at which the information is to be fused. It is not always true that different modalities produce complimentary data in the fusion process. Asynchrony between audio and visual streams is a major problem for the early fusion. If we need to conduct the fusion process in the early fusion stage, we should fully understand the temporal structures of different modalities, for example the temporal correlation between speech and facial expression. However this issue is one of the unexplored areas of audiovisual emotion recognition. Therefore, we generally utilize late integration models, like score level fusion, to get the ultimate classification decision. Hence audio and visual streams are analyzed separately at early fusion and integrated dependently at late fusion, which allows them not occur simultaneously.

4.6.2 Visual Feature Extraction and Analysis

Visual information is the most direct method of interaction between human and machine, and rich emotional information is conveyed through the human face. Our goal is to investigate the role of static and dynamic information conveyed by facial expressions during emotional speech. We need to compute compact facial representations which preserve the useful information of the facial shapes and movements. Moreover, we need to model and recognize emotions by the knowledge of speech-related facial expressions occurring in parallel. The insights gained from emotional face analysis could be applied to

improve automatic emotion recognition and create more natural human computer interfaces. Many solutions have been proposed to process facial expressions and identify the emotional information. However, this requires efficient algorithms. One typical example is an emotion recognition model from geometric facial features using self-organizing map presented by Majumder [129]. A comprehensive data driven model using extended Kohonen self-organizing map (KSOM) has been developed whose input is a 26 dimensional facial geometric feature vector comprising eye, lip and eyebrow feature points. In addition, Metallinou et al. presented visual emotion recognition application using compact facial representations and viseme information [130]. The detailed motion-captured facial information of ten speakers of both genders is analyzed during emotional speech.

Visual Feature Extraction Based on Gabor Filter

Utilizing visual feature to understand emotional expressions has been demonstrated to be a robust approach for emotion recognition. The human facial expressions mostly originate from the movements of facial muscles. In existing emotion recognition schemes based on facial expression analysis, feature-based or region-based approaches are usually employed to determine the emotional states in the given image or video sequence [131]. In this thesis, the visual features are generated by the representation of specific regions in face images based on Gabor library [132]. Gabor transform based feature extraction has a

high degree of correlation with facial emotion expressions and minimum loss of primitive information. Gabor filters have received a special attention in research community due to their resemblance to the models of visual processing in primary visual cortex. Gabor wavelets capture the properties of orientation selectivity, spatial localization and spatial frequency in both space and frequency domains. Figure 4.2 illustrates Gabor filters with five spatial frequencies and eight orientations. It can be seen that Gabor filters exhibit strong characteristics of spatial locality and orientation selectivity. They have been extensively and successfully used in a number of object detection tasks. Numerous previous works have experimentally shown that Gabor based approaches can provide effective solution for facial expression recognition [133]. In this thesis, Gabor wavelets are applied with different spatial frequency properties to extract visual features from a sequence of facial region images.

In video channel, the visual features are extracted from the middle image of each shot instead of all images in order to reduce the computational complexity. After the facial region is detected from each image using Planar envelope approximation method in HSV color space, the detected region is segmented using skin color and processed by morphological operations to clear non-skin region. A set of multi-scale and multi-orientation Gabor wavelet coefficients are extracted by convolving face images normalized to a size of 64×64 with a bank of Gabor filters at 5 spatial frequencies and 8 orientations.

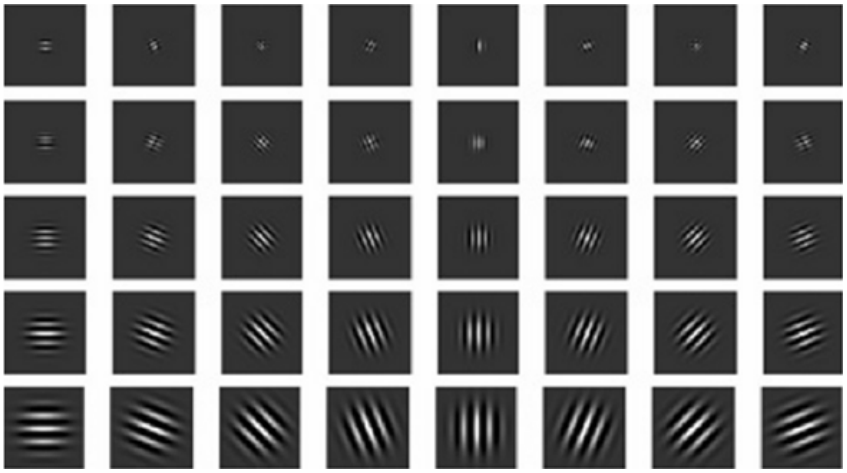

Figure 4.2: Representation of Gabor filters corresponding to 5 spatial frequencies and 8 orientations.

Based on the Gabor representations, visual features can be generated. Gabor filters are invariant against distortion, rotation and scale, and have the useful property of robustness against variations in illumination. They perform well for the task of facial expression analysis in image-based approaches. In order to reduce the computational complexity, the Gabor coefficients are down-sampled to a size of 32×32 in each sub-band and the resulting coefficients lead to the extracted feature vectors [23]. Although the visual feature vectors are largely down-sampled, the data points of the feature vectors are so sparse that the finite set of sampling data may not provide adequate classification capability. Working with well-selected feature sets can remove the irrelevant information in the original features and reduce computational load with a decreased dimensionality. In order to alleviate this problem, the extracted visual features are transformed using

KECA based method. The resulting features are modeled by HMM and the scores provided by HMM are used to implement information fusion with audio channel at the score level in the application of emotion recognition.

Visual Feature Extraction Based on EBS Model

Besides the Gabor filter based method of visual feature extraction, we extract visual features from a deformable Elastic Body Spline (EBS) model to make a comparison with other methods [134]. EBS algorithm mathematically describes the equilibrium displacement of the facial expressions subjected to muscular forces using a Navier partial differential equation (PDE). To find an appropriate physical property for an expression, muscular forces are assumed to distribute on the homogeneous isotropic elastic body of the facial region to obtain smooth deformation. Solving the PDEs, we can form the splines as linear combination of translated versions of the solution. The spline relaxes to an affine transformation as the distance from the point approaches infinity. In previous works, one constant EBS physical property is computed for a facial expression. However, different muscle fiber has a different way of deformation. The Poisson's ratio should be position dependent and the elastic property at different locations should not be the same. In our method, we use different Poisson's ratios to model the facial muscle fiber.

The algorithm for deformation feature extraction can refer to [134]. First initialize

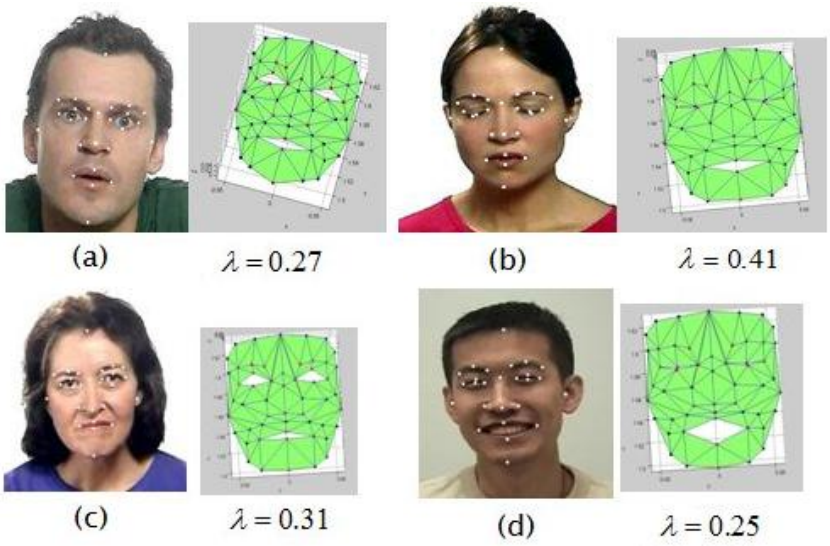

Figure 4.3: EBS facial model construction with different Poisson's ratio λ (a) male anger facial expression (b) female sadness facial expression (c) female anger facial expression (a) male happiness facial expression.

the control point positions X_i for the neutral face. Then set the Poisson's ratio λ for each facial region to 0.01. Next calculate the EBS between neutral and expressive faces using the method in [134]. We compute the distance d_t according to the t-th characteristic feature points at the boundary of two facial regions. Then we summarize the distance for the characteristic feature point at the mth boundary of the nth facial region as D_{min} so that we get $D_{min} = absolute(\sum d_t)_{m,n}$. The following step is iteration for all facial regions. The iteration procedure is to change λ from 0.02, 0.03, ... to 0.5 in one facial region and keep other regions unchanged. Then we find the new distance $D_{m,n}$, keep $\sum_{m,n}(D_{m,n}) \leq \sum_{m,n}(D_{min})$ and update D_{min}. After iteration, we find the minimum D_{min} to fix λ and the EBS coefficients for each facial region. Using this approach, we

can find a set of Poisson's ratios λ for all facial regions and construct more reasonable facial features for the final decision.

Figure 4.3 illustrates EBS facial model construction with different Poisson's ratios. The proposed method could extract the prominent characteristics of facial expressions and accurately describe the primary factors to discriminate between expressions. It is more robust under different illumination conditions and subtle facial expressions. The visual features transformed by KECA based feature level fusion method are viewed as an input to the classifier which can provide the classification scores for video channels. The resulting scores from HMM are further processed at the score level. At the score level, integrated with the matching score from audio channel using maximum correntropy criterion based score level fusion method, a more stable and accurate recognition decision of emotion states is achieved.

4.6.3 Experiments

The experimental databases and experimental setup used in this section are the same as those used in the application of audio emotion recognition. The dimensionalities of audio features and visual features are set to 240 and 175 respectively [79]. Apart from working with unimodal classifiers, we conduct experiments on both early fusion and late fusion of audio and visual features. The six emotion states are used in the experiments.

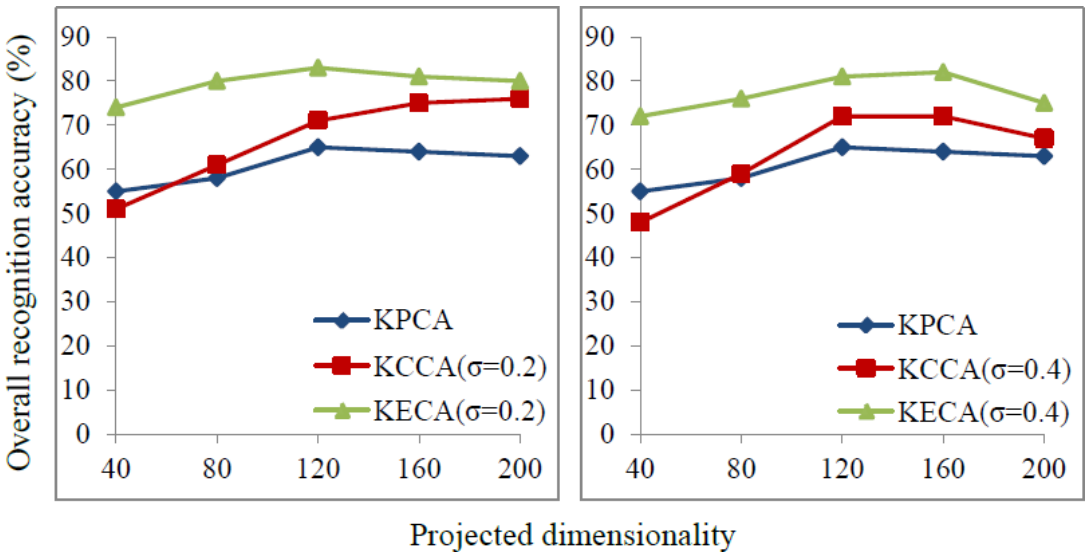

Figure 4.4: Experimental results of eNTERFACE database. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.2; Right: σ=0.4

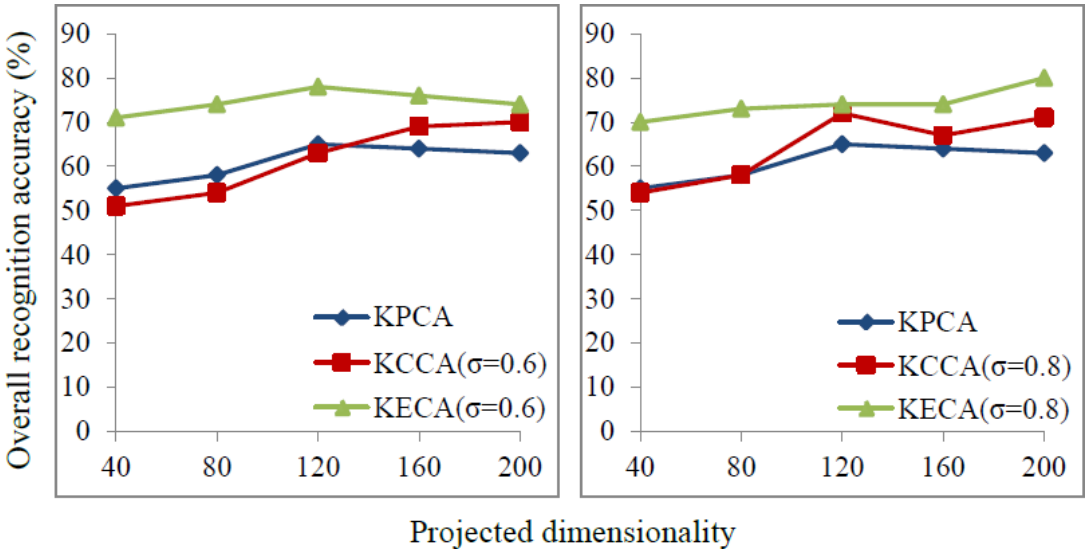

Figure 4.5: Experimental results of eNTERFACE database. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.6; Right: σ=0.8

Comparison of feature level fusion

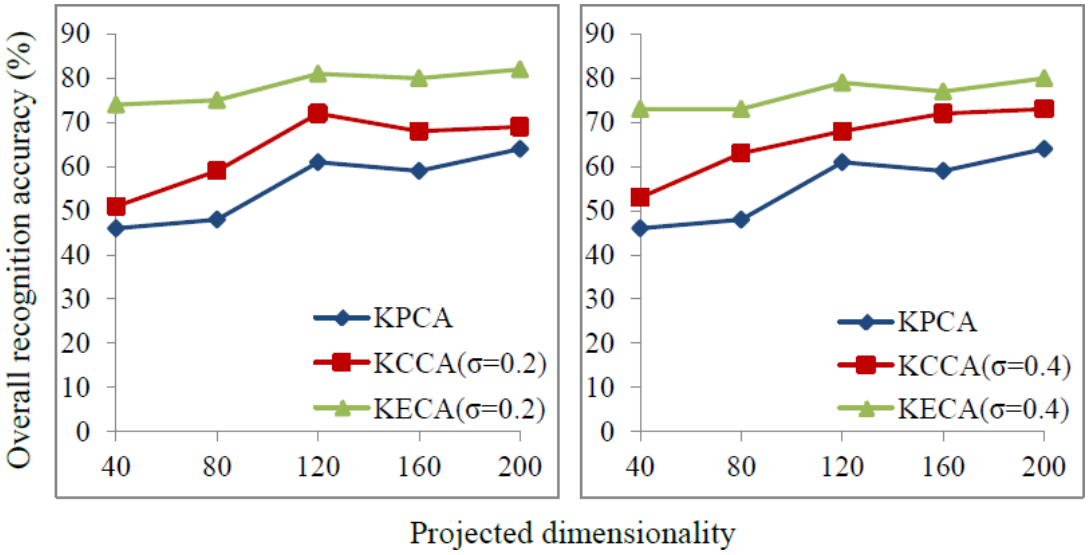

Figure 4.6: Experimental results of RML database. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.2; Right: σ=0.4

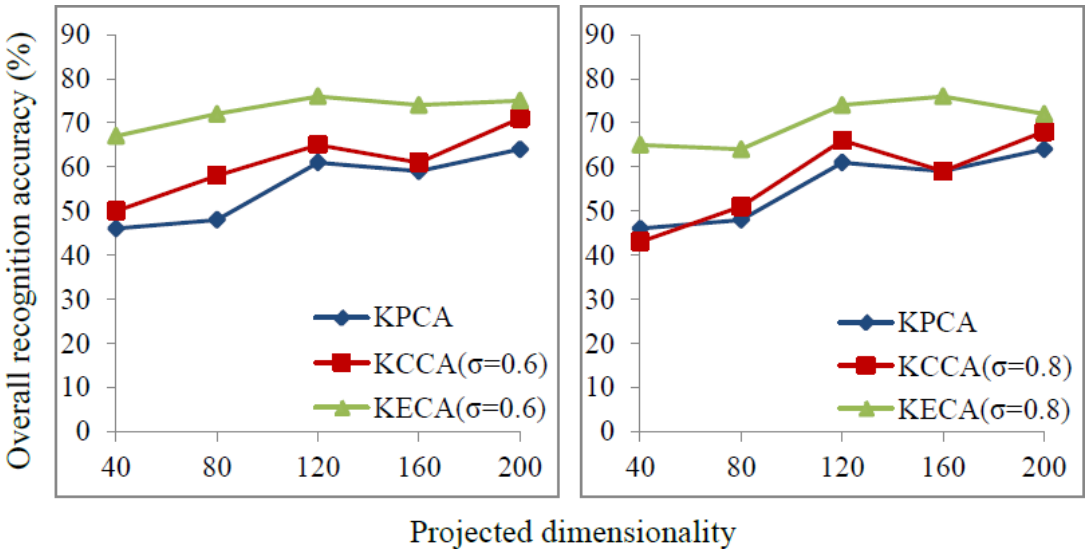

Figure 4.7: Experimental results of RML database. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.6; Right: σ=0.8

Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7 demonstrate the overall recognition accuracy of eNTERFACE database and RML database using KECA, KPCA and KCCA based feature level fusion at different dimensionality. The parameter σ stands for kernel size and it is set from 0.2 to 0.8 with a step size of 0.2 for comparison. The score level fusion strategy for all these experiments is based on MCC. The visual feature extraction method is based on EBS model. The recognition accuracy is calculated as the ratio of the number of correctly classified samples and the total number of samples in the data set. The improved performance of KECA is reasonably satisfactory and encouraging. The proposed method effectively captures the intrinsic relationship between two modalities and outperforms other methods in dimensionality reduction and accuracy performance. The overall recognition accuracy of the proposed solution levels off even if the projected dimensionality is low, which displays the dimensionality reduction property of KECA based fusion strategy. On the other hand, the accuracy of the traditional methods is largely decreased if the projected dimension is too small. The extracted features at the higher dimension might carry redundant or noisy data. Hence if the fusion model is not appropriately selected, the degraded performance may be generated. More stable and accurate results achieved by the KECA based fusion method result from the effective preservation of most useful information. The experiment results show that the proposed method outperforms the existing methods, like KPCA and KCCA. From these

Actual Emotion(%)	Detected Emotion(%)					
	Happiness	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Happiness	**84.13**	2.45	3.56	4.41	3.02	2.43
Disgust	4.16	**78.5**	4.44	3.12	4.09	5.69
Fear	2.51	3.29	**85.89**	3.35	2.67	2.29
Angry	2.31	2.3	3.03	**85.7**	2.77	3.89
Surprise	4.16	5.1	3.54	5.02	**77.87**	4.31
Sadness	1.7	4.31	6.61	2.88	5.21	**79.29**

Table 4.1: Confusion matrix of average performance on two databases. The feature level fusion is based on KECA. The score level fusion is based on MCC.

Actual Emotion(%)	Detected Emotion(%)					
	Happiness	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Happiness	**63.45**	10.33	9.34	5.98	9.11	1.79
Disgust	11.42	**66.22**	3.22	11.12	3.98	4.04
Fear	14.22	5.1	**59.98**	5.31	12.54	2.85
Angry	6.98	5.32	5.34	**64.41**	11.23	6.72
Surprise	10.01	9.34	1.23	2.1	**62.11**	15.21
Sadness	8.31	4.63	5.21	12.98	9.42	**59.45**

Table 4.2: Confusion matrix of average performance on two databases. The feature level fusion is based on KPCA. The score level fusion is based on MCC.

Actual Emotion(%)	Detected Emotion(%)					
	Happiness	**Disgust**	**Fear**	**Angry**	**Surprise**	**Sadness**
Happiness	**74.45**	3.33	10.34	4.98	5.11	1.79
Disgust	9.42	**68.22**	3.22	10.12	3.98	5.04
Fear	4.22	6.1	**67.98**	8.31	8.54	4.85
Angry	1.98	2.32	5.34	**74.41**	7.23	8.72
Surprise	10.01	1.34	4.23	2.1	**73.11**	9.21
Sadness	3.31	5.63	5.21	12.98	3.42	**69.45**

Table 4.3: Confusion matrix of average performance on two databases. The feature level fusion is based on KCCA. The score level fusion is based on MCC.

experiments, the noticeable improvement of dimensionality reduction and the ability of preserving useful data of the proposed fusion strategy have been clearly demonstrated.

The comparison of different methods is also displayed as a two-dimensional confusion

matrix with a row and a column for each class. Table 4.1, Table 4.2 and Table 4.3 display three confusion matrices of the average performance on eNTERFACE database and RML database based on KECA, KPCA and KCCA at feature level and MCC at score level. The projected dimension is set to 120. It is observed that, in general, smaller σ values tend to lead to better performance. Hence the kernel size σ is set to 0.2. The rows of confusion matrix are associated with the actual classes of the samples and the columns of confusion matrix are associated with the detected classes of the samples. Each element of the matrix represents the percentage of correctly classified samples for which the actual class is the row and the detected class is the column. The diagonal elements are the percentage of correct classification, while the others are the percentage of incorrect classification. These tables confirm the result of proposed method is better than those obtained by the methods based on KPCA and KCCA. Moreover it is apparent that the proposed audiovisual fusion solution has better accuracy and stability. Overall, this confirms the effectiveness of the proposed fusion solution.

Comparison of score level fusion

We compare the recognition accuracy of six emotion states between audio modality only, visual modality only and audiovisual multimodal fusion in Figure 4.8 and Figure 4.9. The feature level fusion of these experiments is using the proposed method based

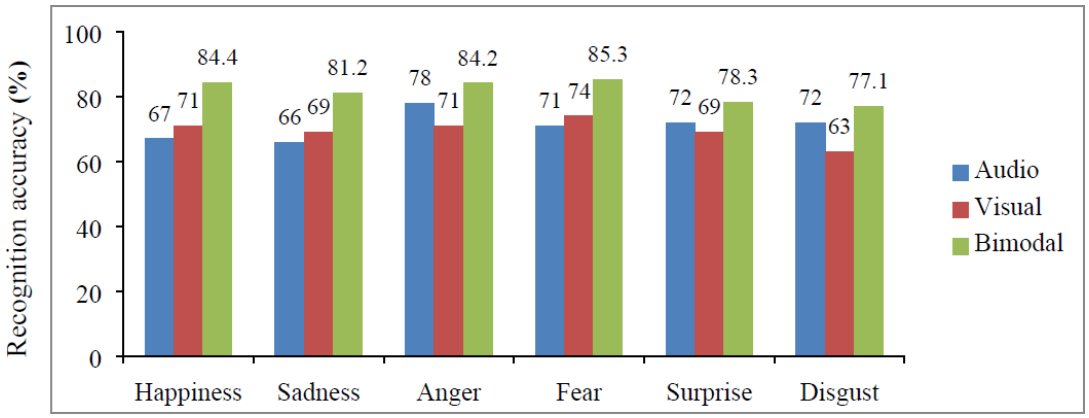

Figure 4.8: Comparison between audio modality only, visual modality only and audiovisual fusion of eNTERFACE database. The feature level fusion is based on KECA. The score level fusion is based on MCC.

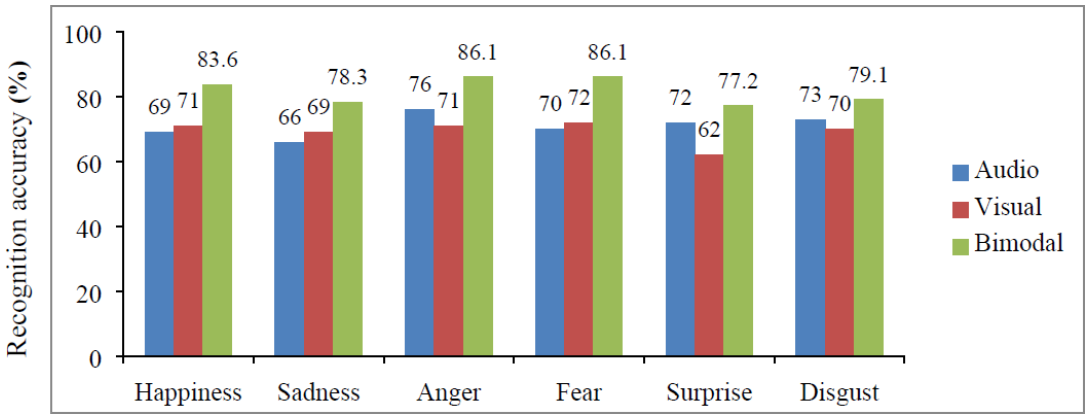

Figure 4.9: Comparison between audio modality only, visual modality only and audiovisual fusion of RML database. The feature level fusion is based on KECA. The score level fusion is based on MCC.

on kernel entropy component analysis. The score level fusion for audiovisual modalities is using the proposed solution based on maximum correntropy criterion. The continuous and spectral features are extracted for audio stream. The visual feature based on EBS model is processed for visual channel. The graphic comparison depicts that the fusion result outperforms those obtained by non-fusion methods using features from a single

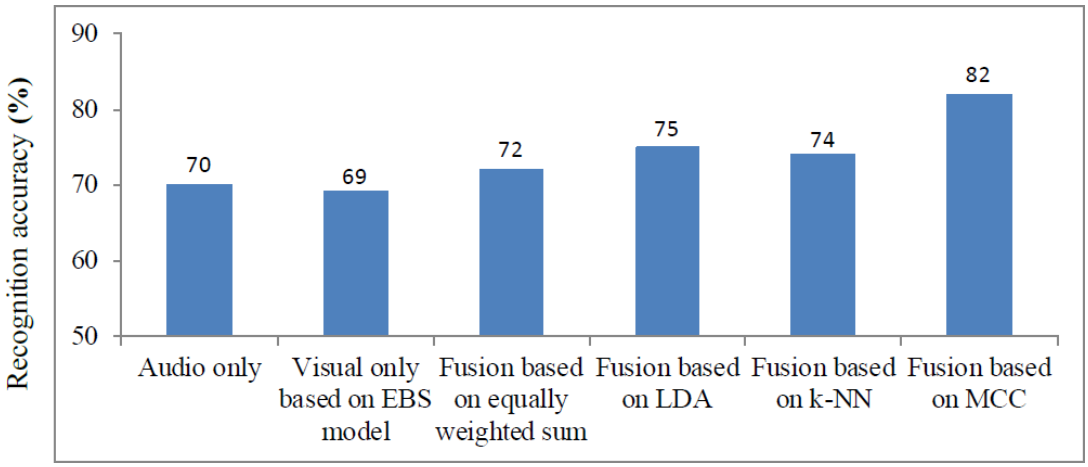

Figure 4.10: Average accuracy of two emotion databases using audio modality only, visual modality only and audiovisual fusion based on different score level fusion methods.

modality. It is noted that some emotion states, like anger, disgust and surprise are better distinguished in audio channel, while other emotion states, like fear, sadness and happiness are more discriminant in visual channel. By integrating audio and visual modalities, the recognition accuracies of most emotion states are improved. It is clearly observed that the complementary relationship of audio and visual information enhances the system performance, confirming the effectiveness of the proposed method.

Figure 4.10 shows the comparison of average recognition accuracy between different score level fusion methods on eNTERFACE database and RML database. The accuracy is measured by the percentage of correctly classified samples. In order to demonstrate the superior performance of the proposed method, we compare the emotion recognition results obtained from the unimodal and different fusion methods. We also make a

comparison among different late fusion schemes. We compare the performance between audio modality only and visual modality only. For the result of audio modality only, continuous and spectral features are utilized. For the result of visual modality only, EBS model based method is employed. We also compare the maximum correntropy criterion based solution with the existing methods, like pre-defined rule based algorithm, linear discriminant analysis (LDA) and k-nearest neighbors (k-NN). The pre-defined rule based algorithm is based on equally weighted sum rule. LDA is used to further discriminate the emotion features to their respective classes. The k-NN method is an instance-based learning classifier, which is widely used for classifying unknown instances based on some distance or similarity function. In the comparison of different score level methods, the visual feature extraction method is based on EBS model and the feature level fusion is based on KECA. Figure 4.10 demonstrates that the traditional methods could provide better performance than audio modality only and visual modality only, but the proposed fusion solution achieves the best overall recognition result. It is shown that our method successfully utilizes the characteristics of multimodal emotional data. It is more robust and stable for the practical environment. In summary, from the above experimental results, the effectiveness of the proposed fusion solution has been confirmed and we observe the superior performance of the proposed strategy when it is compared with the previous methods.

Comparison with Gabor filter based method

The experiments mentioned above are based on EBS model for visual extraction. In many existing applications, 2D Gabor filter is one of the most popular visual extraction methods. In this section, we present and compare the fusion schemes using 2D Gabor filter based method for video stream feature extraction. Figure 4.11, Figure 4.12, Figure 4.13 and Figure 4.14 demonstrate the overall recognition accuracy of eNTERFACE emotion database and RML emotion database using 2D Gabor filter based visual feature extraction method. KECA, KPCA and KCCA based feature level fusion strategies are used. The parameter σ stands for kernel size and it is set from 0.2 to 0.8 with a step size of 0.2 for comparison. The score level fusion is based on MCC. The visual feature extraction method is based on Gabor filter model. Comparison with Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7 shows that EBS model based visual extraction method provides better overall recognition accuracy and improves accuracy rate by about 4%. Hence EBS model outperforms Gabor filter in the visual extraction for emotion recognition.

Figure 4.15 compare the emotion recognition results obtained from unimodal methods and different score level fusion methods. For the audio modality, continuous and spectral features are utilized. For the video modality, Gabor filter based method is employed. Different score level fusion solutions are compared with the proposed MCC based method,

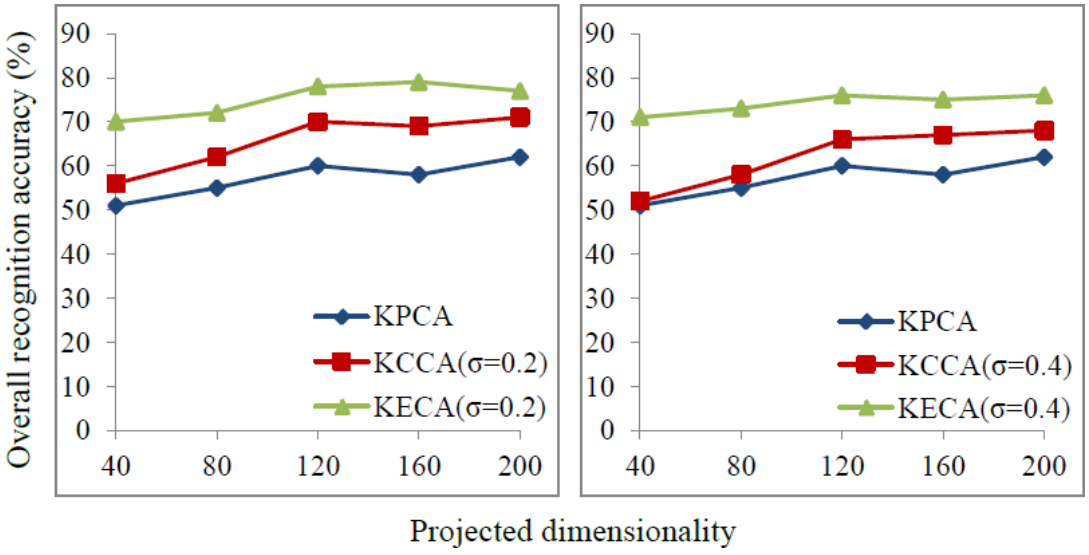

Figure 4.11: Experimental results of eNTERFACE database. The visual feature extraction is using Gabor filter based method. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.2; Right: σ=0.4

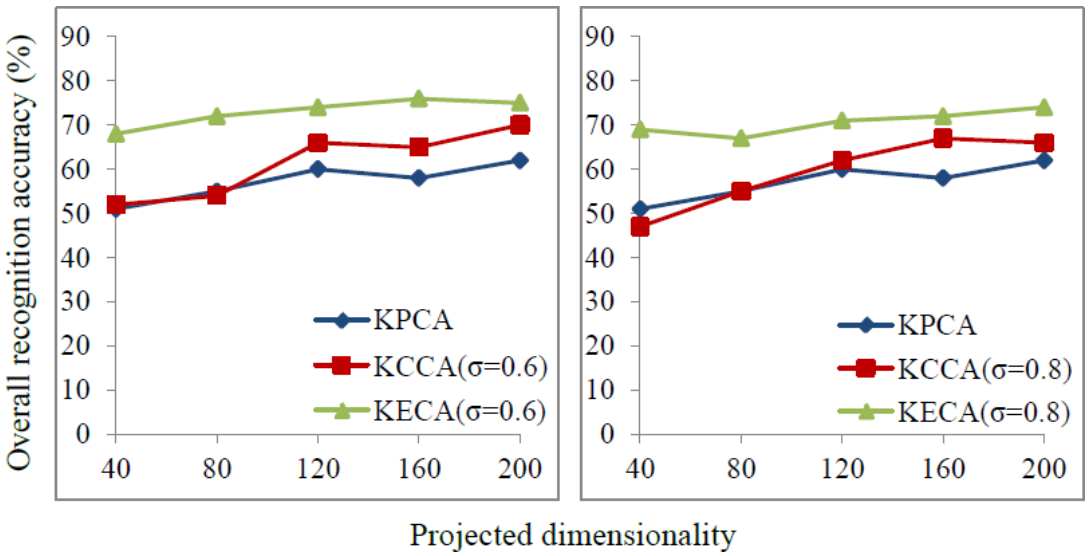

Figure 4.12: Experimental results of eNTERFACE database. The visual feature extraction is using Gabor filter based method. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.6; Right: σ=0.8

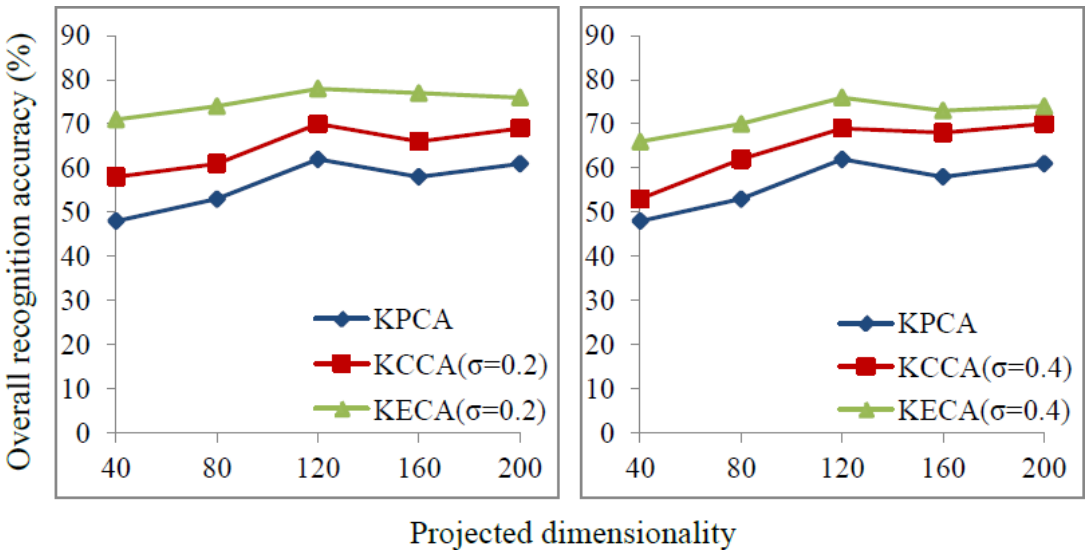

Figure 4.13: Experimental results of RML database. The visual feature extraction is using Gabor filter based method. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.2; Right: σ=0.4

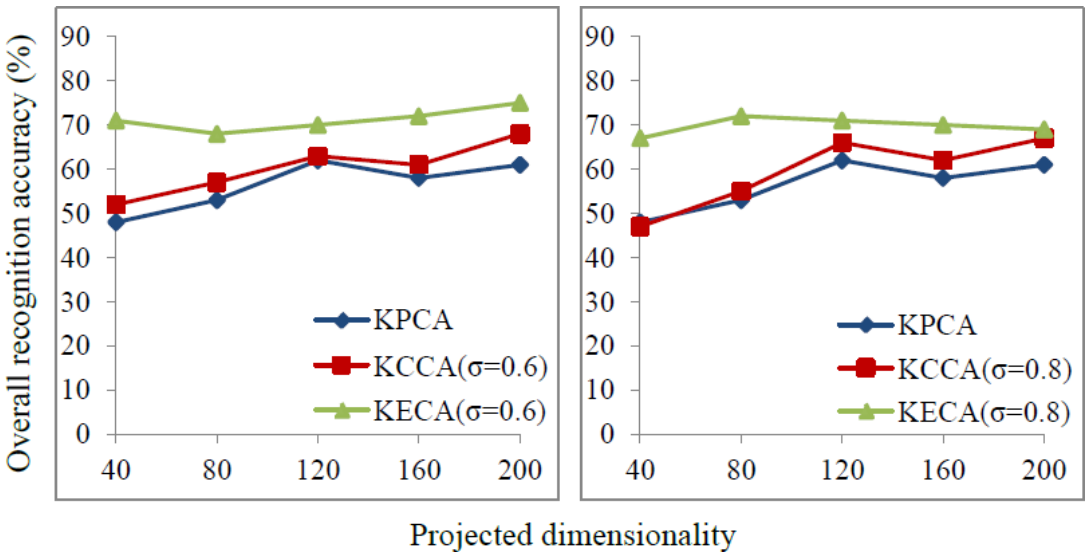

Figure 4.14: Experimental results of RML database. The visual feature extraction is using Gabor filter based method. The feature level fusion is based on KECA, KPCA and KCCA. The score level fusion is based on MCC. Left: σ=0.6; Right: σ=0.8

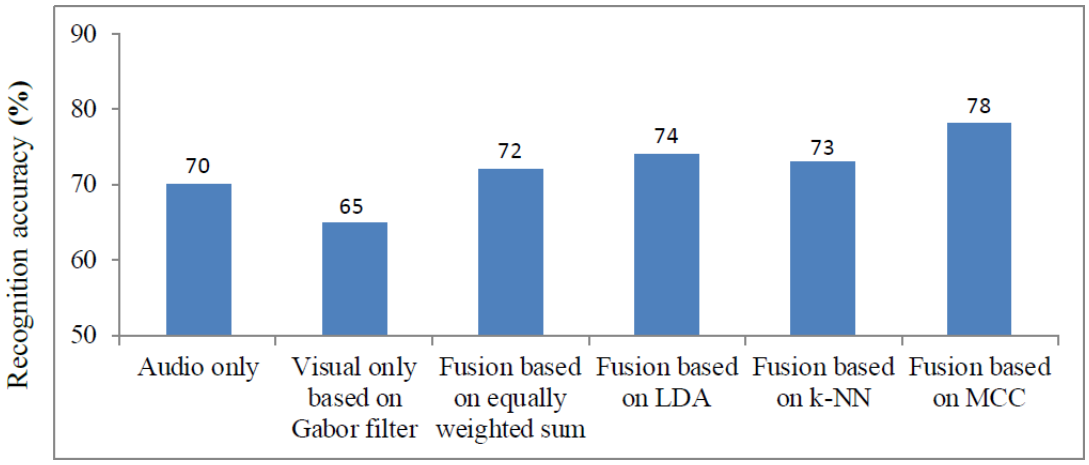

Figure 4.15: Average accuracy of two emotion databases using audio modality only, visual modality only and audiovisual fusion based on different score level fusion methods.

including pre-defined rule based algorithm, linear discriminant analysis (LDA) and k-nearest neighbors (k-NN). In the comparison of different score level methods, the feature level fusion is based on KECA and Gabor filter based method is used for visual extraction. As shown in Figure 4.15, combining the information of multiples modalities enhances the classification accuracy and the MCC based score level fusion achieves best performance.

Figure 4.16 compares the average emotion recognition results obtained from unimodal methods and different score level fusion methods based on Gabor filter and EBS model. The score level fusion solutions include equally weighted sum rule, linear discriminant analysis (LDA), k-nearest neighbors (k-NN), and the proposed MCC based method. As shown in Figure 4.16, the MCC based score level fusion outperforms the traditional methods. It is worthwhile comparing the left columns based on EBD model with the

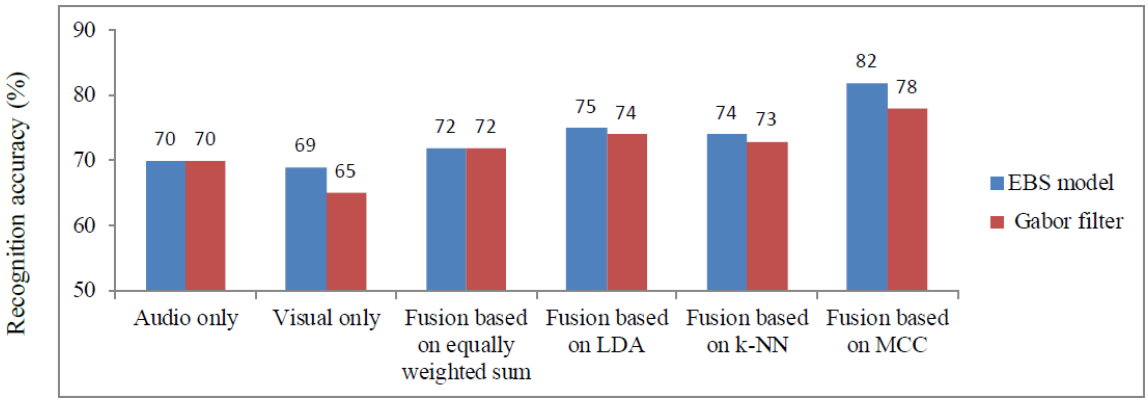

Figure 4.16: Average performance comparison of two emotion databases using audio modality only, visual modality only and audiovisual fusion based on different score level fusion methods. Left columns are using EBS model for visual feature extraction. Right columns are using Garbor filter for visual feature extraction.

right columns based on Gabor filter. The comparison reveals that using EBS features leads to 4 percent improvement over using Gabor features in visual only recognition. The elaborated EBS features extracted from 3D model apparently provide more accurate information in representing the visual characteristics of human facial expressions than the features extracted using 2D Gabor filter. It is quite certain that this is also the reason for better performance in bimodal emotion recognition using EBS visual features.

4.7 Summary

This chapter presents the advantage and limitations of information fusion at score level. It also discusses three categories of previous methods including rule based fusion, classifier based fusion, and density based fusion. Chapter 4 describes the issue of data structure

analyzed by similarity metric with information theoretic learning principle. The information theoretic concepts, such as correntropy and maximum correntropy criterion (MCC), are presented and discussed in details. To overcome the existing drawbacks of score level fusion, this chapter introduces a novel approach for score level fusion based on MCC. The proposed methods of feature level fusion and score level fusion are implemented in the emotion recognition application based on integration of audio and visual information. Visual feature extraction is based on two methods including Gabor filter and EBS model. At score level, the fusion solution based on maximum correntropy criterion is employed to combine the corresponding matching scores and obtain the detected emotion state. This application has been evaluated through extensive experiments conducted on eNTERFACE and RML emotion databases. Experimental results demonstrate the effectiveness of the proposed frameworks in terms of both accuracy and reliability.

Chapter 5

Conclusions and Future Work

Within the last couple of years, multimodal recognition of human emotional states has gained a considerable interest from the research community. Technologies for human emotion perception including speech, facial expression and language have expanded the interaction modalities between humans and computers. With the growing usage of human computer interaction, emotion recognition technology provides an opportunity to promote effective communication. Since understanding emotional expression can refer to verbal and non-verbal channels, constructing an emotion recognition system from multimodal information is desirable. In this thesis, the integration of audio and visual information with the application to emotion recognition has been studied. The proposed methods of feature level fusion and score level fusion are implemented in audiovisual

emotion recognition. We explore possibilities for enhancing the generality and robustness of emotion recognition system by using the techniques of multimodal information fusion based on entropy estimation. Extensive experiments demonstrate the feasibility of the proposed multimodal emotion recognition frameworks based on integrated analysis of speech and facial expression.

This thesis presents an overview of emotion recognition applications and information fusion strategies. It also covers several critical issues of emotion recognition and information fusion including the challenges, the techniques of the existing methods and so on. The strategies of information fusion have been employed to accomplish various emotion analysis tasks. Considering the drawbacks of the previous methods, this thesis introduces novel strategies for multimodal information fusion which implement the techniques of information theoretic learning into the area of information fusion. A novel information theoretic tool, kernel entropy component analysis, has been applied to feature level fusion for emotion recognition, which aims at building up a close connection between multimodal information fusion and information theoretical tools. The thesis also proposes a new method of score level fusion based on maximum correntropy criterion estimation. Unlike the existing rule-based score level fusion algorithms, the proposed method depends on entropy estimation which reveals the nature of information. These methods have been applied to audio and visual modalities at both feature level and score level in

the application of emotion recognition. The experimental results demonstrate the feasibility of the proposed solutions, and shows that the proposed solutions outperform the existing methods. Therefore, this thesis offers us novel audiovisual emotion recognition frameworks using multimodal information fusion based on entropy estimation.

5.1 Future Work

As an extension of this thesis, we propose the following possible directions for future research. In this thesis, a connection between information theory and information fusion has been built up, but the concepts of information theory covered in the thesis are only entropy and correntropy. There are more sophisticated tools in information theory such as joint entropy, mutual entropy, mutual information, etc., which can help explore more reasonable and efficient relationship between different modalities in information fusion. The application of information theory enables us to consider the problem of machine learning 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.

Despite the fact that a great number of multimodal information analysis tasks have been successfully performed, there are some issues which have not yet been explored

sufficiently. In fact there is little in-depth research on the characteristics of emotion databases. How may the changing context influence the fusion and classification process? Is there any model or feature most suitable to simulate the varying nature of emotion databases? How to use the tools of information theory instead of statistics to describe the model of emotional data? The above questions require greater attention from researchers in emotion recognition and information fusion.

Most of the existing research works in the field of affective computing focus on recognition of basic emotions mainly using acted databases. The speech and facial expressions are usually captured in predefined controlled lab. The training and evaluation of emotion recognition typically take place on databases obtained from the laboratory environment. Although some advances have been achieved, automatic emotion recognition occurring in natural environment is not fully explored. The nature of emotions is subtle, mixed and complex. And it is difficult to map the human emotional states in realistic settings into a single label. These facts lead to a challenging problem; How to accurately and realistically describe the intensity of emotions. Further work is necessary for spontaneous emotion estimation in practical situations.

Moreover, the current research work is restricted to dedicated databases which have limited samples. However, the increasingly popularity of networking websites has led to the explosive growth of multimedia data. High volumes of multimedia, such as audio,

video and images are being generated daily. Hence we should consider the information fusion of multimedia data as a problem of big data. The challenge is that the big data of audio, video and images requires more sophisticated algorithms for content analysis than previous databases with limited data. There is a tremendous amount of research works on efficient and effective techniques for large-scale multimedia information processing and storage based on the techniques of big data. With the rapid development of cloud computing technologies, it is possible for us to achieve complicated algorithms on massive multimedia data, like information fusion algorithms on large-scale audiovisual data, since the technologies of cloud computing are utilized to meet the requirements on infrastructure for big data, for example elasticity, cost efficiency, and smooth upgrading or downgrading [135]. Therefore, we should pay attention on how to realize information fusion for massive multimedia data from widely distributed data sources.

Cloud computing techniques realize distributed computing by utilizing multiple computers to execute computing simultaneously on the service side. Moreover, in order to process the increasing quantity of multimedia data, numerous cloud computing frameworks and data storage techniques have been developed [136]. The key core techniques for large-scale data computing techniques are Hadoop and MapReduce. Hadoop can conduct the parallel computing of big data through a computing cluster formed by low-priced hardware. MapReduce is a parallel computing algorithm and provides integrated

computing resources in distributed computing to reduce the computing time. Other techniques like NoSQL are currently taking the role of traditional database techniques to enhance the processing efficiency and flexibility for multimedia data, especially for audio and video signals. Popular NoSQL databases are MongoDB, Cassandra, Redis and CouchDB [137]. The impressive development of cloud computing brings us more potentials and possibilities in processing large-scale multimedia data. We are now able to analyze large-scale dataset about users' emotional states by utilizing higher-speed multimedia data storage and processing frameworks for the first time. Millions of emotion data points present a unique opportunity to improve the ability of machine learning and emotion recognition is emerging as the next great source for us to learn about the users. Hence there are opportunities to further extend and improve our information fusion algorithms on emotion recognition. We believe that this research direction is worth special attention.

Bibliography

[1] A. Houjeij, L. Hamieh, and N. Mehdi, "A novel approach for emotion classification based on fusion of text and speech," in *Proceedings of the 19th International Conference on Telecommunications (ICT)*, 2012, pp. 1–6.

[2] C. Maaoui, F. Abdat, and A. Pruski, "Physio visual data fusion for emotion recognition," *IRBM*, vol. 35, no. 3, pp. 109–118, 2014.

[3] T. Joshi, S. Dey, and D. Samanta, "Multimodal biometrics: state of the art in fusion techniques," *International Journal of Biometrics*, vol. 1, no. 4, pp. 393–417, 2009.

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

[5] 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.

[6] L. Guan, Y. Wang, and Y. Tie, "Toward natural and efficient human computer interaction," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 1560–1561.

[7] D. Sauter, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2408–2412, 2010.

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

[9] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[10] Z. Zeng, M. Pantic, G. Roisman, and T. 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.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[12] H. X. Hua, Y. Jian, and G. Jue, "Application of speech emotion recognition in intelligent household robot," in *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence (CSCI)*, vol. 1, 2010, pp. 537–541.

[13] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[14] M. Mansoorizadeh and M. C. Nasrollah, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications*, vol. 49, no. 2, pp. 277–297, 2010.

[15] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, pp. 1–11, 2013.

[16] A. Milton and S. Tamil Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," *Computer Speech and Language*, vol. 28, no. 3, pp. 727–742, 2014.

[17] E. Vayrynen, T. Juhani, and S. Tapio, "Classification of emotion in spoken Finnish using vowel-length segments: increasing reliability with a fusion technique," *Speech Communication*, vol. 53, no. 3, pp. 269–282, 2011.

[18] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.

[19] C. S. Ooi, "A new approach of audio emotion recognition," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[20] M. Bejani, G. Davood, and M. C. Nasrollah, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Computing and Applications*, vol. 24, no. 2, pp. 399–412, 2014.

[21] C. Xu, T. Cao, Z. Feng, and C. Dong, "Multi-modal fusion emotion recognition based on HMM and ANN," *Contemporary Research on E-business Technology and Strategy*, pp. 541–550, 2012.

[22] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDE)*, 2006, p. 8.

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

[24] P. Over, G. Awad, and J. Fiscus, "TRECVID 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TREC Video Retrieval Evaluation Online (TRECVID)*, 2011, pp. 1–56.

[25] B. B. Enrique, B. Samy, B. Frederic, H. Miroslav, and K. Josef, "The BANCA database and evaluation protocol," in *Proceedings of Audio and Video-based Biometric Person Authentication*, 2003, pp. 625–638.

[26] M. Kieron, K. Josef, S. Mohammad, M. Sebastien, and M. Christine, "Face verification competition on the XM2VTS database," in *Proceedings of Audio and Video-based Biometric Person Authentication*, 2003, pp. 964–974.

[27] G. S. Sonia, B. Charles, C. Gerard, D. Bernadette, and L. J. Jean, "BIOMET: A multimodal person authentication database including face, voice, fingerprint,

hand and signature modalities," in *Proceedings of Audio and Video-based Biometric Person Authentication*, 2003, pp. 845–853.

[28] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT landmarked face database," *Pattern Recognition Association of South Africa*, vol. 1, p. 4, 2010.

[29] S. Koelstra, "Deap: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

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

[31] B. Khaleghia, A. Khamisa, F. Karraya, and S. Razavib, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[32] A. O. Faisal, M. Bennamoun, and A. Mian, "Spatially optimized data-level fusion of texture and shape for face recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 859–872, 2012.

[33] Z. Gilula and R. McCulloch, "Multi-level categorical data fusion using partially fused data," *Quantitative Marketing and Economics*, vol. 11, no. 3, pp. 353–377, 2013.

[34] A. Noore, R. Singh, and M. Vatsa, "Robust memory-efficient data level information fusion of multi-modal biometric images," *Information Fusion*, vol. 8, no. 4, pp. 337–346, 2007.

[35] A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Proceedings of SPIE Coference on Biometric Technology for Human Identification*, vol. 5779, 2005, pp. 196–204.

[36] 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.

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

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

[39] A. Jaina, K. Nandakumara, and A. Rossb, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[40] N. Karthik and Y. Chen, "Quality-based score level fusion in multibiometric systems," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 473–476.

[41] M. Hanmandlua, J. Grovera, and A. Gurejab, "Score level fusion of multimodal biometrics using triangular norms," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1843–1850, 2011.

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

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

[44] J. Zhou, T. Xu, and J. Gan, "Facial expression recognition based on local directional pattern using SVM decision-level fusion," in *Proceedings of the tenth International Conference on Computability and Complexity in Analysis*, 2013, pp. 8–10.

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

[46] S. Kumar and J. Ghosh, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Analysis and Applications*, vol. 5, no. 2, pp. 210–220, 2002.

[47] R. Singh, M. Vatsa, and A. Noore, "Hierarchical fusion of multi-spectral face images for improved recognition performance," *Information Fusion*, vol. 9, no. 2, pp. 200–210, 2008.

[48] M. S. Hussain, R. Calvo, and P. A. Pour, "Hybrid fusion approach for detecting affects from multichannel physiology," *Affective Computing and Intelligent Interaction*, pp. 568–577, 2011.

[49] R. Snelick, M. Indovina, J. Yen, and A. Mink, "Multimodal biometrics: issues in design and testing," in *Proceedings of the 5th international conference on Multimodal Interfaces (ICMI)*, 2003, pp. 68–72.

[50] M. T. Yang, S. C. Wang, and Y. Y. Lin, "A multimodal fusion system for people detection and tracking," *International Journal of Imaging Systems and Technology*,

vol. 15, no. 2, pp. 131–142, 2005.

[51] B. Moslem, M. Khalil, M. Diab, and C. Marque, "Classification of multichannel uterine EMG signals by using a weighted majority voting decision fusion rule," in *Proceedings of the 16th IEEE Mediterranean Electrotechnical Conference*, 2012, pp. 331–334.

[52] T. W. Chua, K. Leman, and N. T. Pham, "Human action recognition via sum-rule fusion of fuzzy K-Nearest Neighbor classifiers," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2011, pp. 484–489.

[53] N. Pfleger, "Context based multimodal fusion," in *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)*, 2004, pp. 265–272.

[54] A. M. Aziz, "A new multiple decisions fusion rule for targets detection in multiple sensors distributed detection systems with data fusion," *Information Fusion*, vol. 18, pp. 175–186, 2014.

[55] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 3858–3866, 2007.

[56] A. Stephane, G. Quenot, and J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," *Advances in Information Retrieval*, pp. 494–504, 2007.

[57] W. H. Adams, G. Iyengar, C. Y. Lin, M. R. Naphade, and C. Neti, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Advances in Signal Processing*, vol. 2, pp. 170–185, 2003.

[58] R. C. Luo, C. C. Yih, and K. L. Su, "Multisensor fusion and integration: approaches, applications, and future research directions," *Sensors Journal*, vol. 2, no. 2, pp. 107–119, 2002.

[59] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.

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

[61] A. Pradeep, M. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems," *Multimedia Systems*, vol. 12,

no. 3, pp. 239–253, 2006.

[62] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor fusion using Dempster-Shafer theory [for context-aware HCI]," in *Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (I2MTC)*, vol. 1, 2002, pp. 7–12.

[63] Q. Chen and U. Aickelin, "Anomaly detection using the Dempster-Shafer method," in *Proceedings of International Conference on Data Mining (ICDM)*, 2006, pp. 232–240.

[64] M. Guironnet, D. Pellerin, and M. Rombaut, "Video classification based on low-level feature fusion model," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2005, pp. 1–4.

[65] R. Singh, M. Vatsa, and A. Noore, "DS theory based fingerprint classifier fusion with update rule to minimize training time," *IEICE Electronics Express*, vol. 3, no. 20, pp. 429–435, 2006.

[66] X. Li, A. Dick, C. Shen, and Z. Zhang, "Visual tracking with spatio-temporal Dempster-Shafer information fusion," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3028–3040, 2013.

[67] T. Choudhury, J. M. Rehg, V. Pavlovic, and A. Pentland, "Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR)*, vol. 3, 2002, pp. 789–794.

[68] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 11, pp. 1274–1288, 2002.

[69] Y. Zhang and Q. Ji, "Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 2, pp. 467–472, 2006.

[70] B. Dumas, B. Signer, and D. Lalanne, "Fusion in multimodal interactive systems: an HMM-based algorithm for user-induced adaptation," in *Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 2012, pp. 15–24.

[71] A. Nefian, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 11, no. 1900, pp. 1274–1288, 2002.

[72] J. Pinquier, S. Karaman, L. Letoupin, and P. Guyot, "Strategies for multiple feature fusion with Hierarchical HMM: application to activity recognition from wearable audiovisual sensors," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 3192–3195.

[73] H. ElAskary, A. Agarwal, T. ElGhazawi, M. Kafatos, and J. LeMoigne, "Enhancing dust storm detection using PCA based data fusion," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, 2005, pp. 1424–1427.

[74] G. Marcialis and F. R. Luca, "Fusion of LDA and PCA for face verification," *Biometric Authentication*, pp. 30–37, 2002.

[75] Y. Shin and C. Park, "Analysis of correlation based dimension reduction methods," *International Journal of Applied Mathematics and Computer Science*, vol. 21, no. 3, pp. 549–558, 2011.

[76] Y. Wang, L. Guan, and A. Venetsanopoulos, "Audiovisual emotion recognition via cross-modal association in kernel space," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[77] B. Schölkopf, A. Smola, and K. R. Müller, "Kernel principal component analysis," *Artificial Neural Networks*, pp. 583–588, 1997.

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

[79] Y. Wang, L. Guan, and A. 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.

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

[81] S. Watanabe, *Pattern Recognition: Human and Mechanical*. John Wiley and Sons, Inc., 1985.

[82] C. E. Shannon, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

[83] J. Wua, J. Suna, L. Lianga, 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.

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

[85] P. Li and C. H. Zhang, "A new algorithm for compressed counting with applications in Shannon entropy estimation in dynamic data," *Journal of Machine Learning Research-Proceedings Track*, vol. 19, pp. 477–496, 2011.

[86] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

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

[88] L. M. Martyushev and V. D. Seleznev, "Maximum entropy production principle in physics, chemistry and biology," *Physics Reports*, vol. 436, no. 1, pp. 1–45, 2006.

[89] A. Renyi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

[90] E. Lenzia, R. Mendesb, and L. Silva, "Statistical mechanics based on Renyi entropy," *Physica A: Statistical Mechanics and its Applications*, vol. 280, no. 3, pp. 337–345, 2000.

[91] R. Jenssen, "Information theoretic learning and kernel methods," *Information Theory and Statistical Learning*, pp. 209–230, 2009.

[92] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[93] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 551–556.

[94] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[95] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[96] R. Jenssen, T. Eltoft, M. Girolami, and D. Erdogmus, "Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm," *Advances in Neural Information Processing Systems*, vol. 19, pp. 633–640, 2007.

[97] R. Jenssen, "Kernel entropy component analysis: New theory and semi-supervised learning," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2011, pp. 1–6.

[98] L. Gomez-Chova, R. Jenssen, and G. Camps-Vails, "Kernel entropy component analysis in remote sensing data clustering," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 3728–3731.

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

[100] Z. Xie and L. Guan, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, 2012, pp. 1–8.

[101] R. Cowie, E. Douglas-Cowie, and N. Tsapatsoulis, "Emotion recognition in human computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80,

2011.

[102] Y. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 8, 2005, pp. 4898–4901.

[103] T. S. Nwe, L. Foo, and D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[104] M. Suzuki, S. Nakagawa, and K. Kita, "Emotion recognition method based on normalization of prosodic features," in *Proceedings of the IEEE Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–5.

[105] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.

[106] K. Krishna and S. Krishna, "Emotion recognition in speech using MFCC and wavelet features," in *Proceedings of the IEEE 3rd International Advance Computing Conference (IACC)*, 2013, pp. 842–847.

[107] N. J. Nalini, S. Palanivel, and M. Balasubramanian, "Speech emotion recognition using residual phase and MFCC features," *International Journal of Engineering and Technology*, vol. 5, no. 6, pp. 4515–4527, 2013.

[108] 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. 1, pp. 25–42, 2013.

[109] R. Aggarwal and J. Singh, "Noise reduction of speech signal using wavelet transform with modified universal threshold," *International Journal of Computer Applications*, vol. 20, no. 5, pp. 14–19, 2011.

[110] A. Ross, A. K. Jain, and N. Karthik, "Score level fusion," *Handbook of Multibiometrics*, pp. 91–142, 2006.

[111] S. Chitroub, "Classifier combination and score level fusion: concepts and practical aspects," *International Journal of Image and Data Fusion*, vol. 1, no. 2, pp. 113–135, 2010.

[112] J. Fierrez-Aguilar, "A comparative evaluation of fusion strategies for multimodal biometric verification," in *Proceedings of Audio and Video-based Biometric Person Authentication*, vol. 2688, 2003, pp. 830–837.

[113] K. A. Toh, J. Kim, and S. Lee, "Biometric scores fusion based on total error rate minimization," *Pattern Recognition*, vol. 41, no. 3, pp. 1066–1082, 2008.

[114] F. Wang and J. Han, "Multimodal biometric authentication based on score level fusion using support vector machine," *Opto-Electronics Review*, vol. 17, no. 1, pp. 59–64, 2009.

[115] M. He, S. J. Horng, P. Fan, and R. S. Run, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010.

[116] L. Nanni, A. Lumini, and S. Brahnam, "Likelihood ratio based features for a trained biometric score fusion," *Expert Systems with Applications*, vol. 38, no. 1, pp. 58–63, 2011.

[117] K. Nandakumar, Y. Chen, S. Dass, and A. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 342–347, 2008.

[118] Y. Makihara, D. Muramatsu, Y. Yagi, and M. Hossain, "Score-level fusion based on the direct estimation of the Bayes error gradient distribution," in *Proceedings of the International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–8.

[119] M. He, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010.

[120] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: A localized similarity measure," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2006, pp. 4919–4924.

[121] A. Singh and J. C. Principe, "Information theoretic learning with adaptive kernels," *Signal Processing*, vol. 91, no. 2, pp. 203–213, 2011.

[122] A. Gunduz and J. C. Principe, "Correntropy as a novel measure for nonlinearity tests," *Signal Processing*, vol. 89, no. 1, pp. 14–23, 2009.

[123] A. Singh and J. C. Principe, "Using correntropy as a cost function in linear adaptive filters," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2009, pp. 2950–2955.

[124] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: properties and applications in non-Gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[125] W. Zhou and S. Kamata, "Linear discriminant analysis with maximum correntropy criterion," *Computer Vision ACCV*, pp. 500–511, 2013.

[126] R. He, W. S. Zheng, and B. G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2011.

[127] S. Shivappa, M. Trivedi, and B. 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.

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

[129] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognition*, vol. 47, no. 3, pp. 1282–1293, 2014.

[130] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2474–2477.

[131] J. Yu and B. Bir, "Evolutionary feature synthesis for facial expression recognition," *Pattern Recognition Letters*, vol. 27, no. 11, pp. 1289–1298, 2006.

[132] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," *International Journal of Multimedia Data Engineering and Management*, vol. 4, no. 4, pp. 1–14, 2013.

[133] S. Bashyal and G. K. Venayagamoorthy, "Recognition of facial expressions using Gabor wavelets and learning vector quantization," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 7, pp. 1056–1064, 2008.

[134] Y. Tie and L. Guan, "A deformable 3D facial expression model for dynamic human emotional state recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 1, pp. 142–157, 2013.

[135] R. Catherine and E. Bijolin, "A survey on recent trends in cloud computing and its application for multimedia," *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, vol. 2, no. 1, pp. 304–309, 2013.

[136] W. K. Lai, "Towards a framework for large-scale multimedia data storage and processing on Hadoop platform," *The Journal of Supercomputing*, vol. 68, no. 1,

pp. 488–507, 2014.

[137] D. Hammes, H. Medero, and H. Mitchell, "Comparison of NoSQL and SQL databases in the cloud," *SAIS Proceedings*, pp. 12–21, 2014.