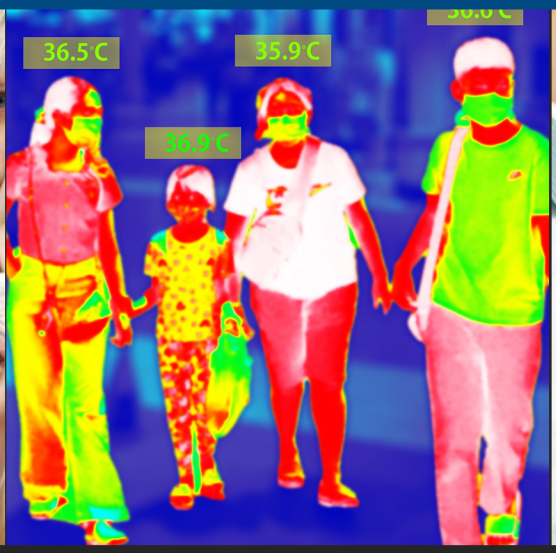
Multimodal Affective Computing

Affective Information Representation, Modelling, and Analysis

Gyanendra K. Verma



Bentham Books

Multimodal Affective Computing: Affective Information Representation, Modelling, and Analysis

Authored By

Gyanendra K. Verma
Department of Information Technology, National Institute of Technology Raipur,
Chhattisgarh, India

Multimodal Affective Computing:
Affective Information Representation, Modelling, and Analysis

Author: Gyanendra K. Verma

ISBN (Online): 978-981-5124-45-3

ISBN (Print): 978-981-5124-46-0

ISBN (Paperback): 978-981-5124-47-7

First published in 2023.

need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.

3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Pte. Ltd.
80 Robinson Road #02-00
Singapore 068898
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

FOREWORD

Affective Computing is a new area aiming to create intelligent computers that recognize, understand, and process human emotions. Affective Computing is an interdisciplinary area that encompasses a variety of disciplines, such as computer science, psychology, and cognitive science, among others. Emotion may be communicated in various ways, including gestures, postures, and facial expressions, as well as physiological signs, including brain activity, heart rate, muscle activity, blood pressure, and skin temperature. Humans can perceive emotion through facial expressions in general. However, not all emotions, particularly complex ones such as pride, love, mellowness, and sorrow, can be identified only through facial expressions. Physiological signals can therefore be utilized to represent complex emotions effectively.

This book aims to provide the audience with a basic understanding of Affective Computing and its application in many research fields. This state-of-the-art review of existing emotion theory and modeling approaches will help the readers explore various aspects of Affective Computing. By the end of the book, I hope that the readers will be able to understand emotion recognition methods based on audio, video, and physiological signals. Moreover, they will learn the fusion framework and familiarity to implement for emotion recognition.

Shitala Prasad
Institute for Infocomm Research
A*Star
Singapore

PREFACE

Affective Computing is an emerging field with the prime focus on developing intelligent systems that can perceive, interpret, process human emotions and act accordingly. Affective Computing incorporates interdisciplinary research areas like Computer Science, Psychology, Cognitive Science, Machine Learning, *etc*. Machines must perceive and interpret emotions in real-time and act accordingly for intelligent communication with human beings. Emotion plays a significant role in communication and can be expressed through many ways, like facial or auditory expression, gesture or sign language, *etc*. Brain activity, heart rate, muscular activity, blood pressure, and skin temperature are a few examples of physiological signals. It plays a crucial role in affect recognition compared to other emotion modalities. Humans perceive emotion primarily through facial expressions; yet, complex emotions such as pride, love, mellowness, and sorrow cannot be identified just by facial expressions. Physiological signals can thus be employed to recognize complex emotions.

The objective of this book is mainly three-fold: (1) Provide in-depth knowledge about affective Computing, affect information representation, models, and theories of emotions. (2) Emotion recognition from different affective modalities, such as audio, facial expression, and physiological signals, and (2) Multimodal fusion framework for emotion recognition in three-dimensional Valence, Arousal, and Dominance space.

Human emotions can be captured from various modalities, such as speech, facial expressions, physiological signals, *etc*. These modalities provide critical information that may be utilised to infer a user's emotional state. The primary emotions can be captured easily by facial and vocal expressions. However, facial expressions or audio information cannot detect complex emotions. Therefore, an efficient emotion model is required to predict complex emotions. The dimensional model of emotion can effectively model and recognize complex emotions.

Most emotion recognition work is based on facial and vocal expressions. However, the existing literature completely lacks emotion modeling in a continuous space. This book contributes in this direction by proposing an emotion model to predict a large number (more than fifteen) of complex emotions in a three-dimensional continuous space. We have implemented various systems to recognize emotion from speech, facial expression, physiological signals, and multimodal fusion of the above modalities. Our emphasis is on emotion modeling in a continuous space. Emotion prediction from physiological signals as complex emotions is better captured by physiological signals rather than facial or vocal expressions. The main contributions of this book can be summarized as follows:

1. This book presents a state-of-the-art review of Affective Computing and its application in various areas like gaming, medicine, virtual reality, *etc*.
2. A detailed review of multimodal fusion techniques is presented to assimilate multiple modalities to accomplish multimodal fusion tasks. The fusion methods are provided from the perspective of the requirement of multimodal fusions, the level of information fusion, and their applications in various domains, as reported in the literature. Moreover, significant challenges in multimodal fusions are also highlighted. Further, we present the evaluation measures for evaluating multimodal fusion techniques.
3. The significant contribution of this book is the three-dimensional emotion model based on valence, arousal, and dominance. The emotion prediction in three-dimensional space based on valence, arousal, and dominance is also presented.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

Gyanendra K. Verma
Department of Information Technology
National Institute of Technology Raipur
Chhattisgarh
India

ACKNOWLEDGEMENTS

I acknowledge the guidance of Prof. U. S. Tiwary, IIIT Allahabad India, who motivated me to choose Affective Computing as a research topic many years ago. Some of his insights are still present in this book. I want to thank Dr. Shitala Prasad, Scientist, Institute for Infocomm Research, A*Star, Singapore, for his valuable suggestions. And last but not least, Ms. Humaira Hashmi, Editorial Manager Publications, Bentham Books, for extending her kind cooperation to complete this book project.

CHAPTER 1

Affective Computing

Abstract: With the invention of high-power computing systems, machines are expected to show intelligence at par with human beings. A machine must be able to analyze and interpret emotions to demonstrate intelligent behavior. Affective computing not only helps computers to improve performance intelligently but also helps in decision-making. This chapter introduces affective computing and related issues that influence emotions. This study also provides an overview of human-computer interaction (HCI) and the possible use of different modalities for HCI. Further, challenges in affective computing are also discussed, along with the application of affective computing in various areas.

Keywords: Arousal, DEAP database, Dominance, EEG, Multiresolution analysis, Support vector machine, Valence.

1.1. INTRODUCTION

The cognitive, affective, and emotional information is crucial in HCI to improve user-computer connection [1]. It significantly enhances the learning environment. Emotion recognition is crucial since it has several applications in HCI and Human-Robot Interaction (HRI) [2] and many other new fields. Affective computing is a hot topic in the field of human-computer interaction. “Affective Computing is the research and development of systems and technologies that can identify, understand, process, and imitate human emotions,” according to the definition.

Affective computing is an interdisciplinary area that encompasses a variety of disciplines, such as computer science, psychology, and cognitive science, among others. Emotions can be exhibited in various ways, such as gestures, postures, facial expressions, and physiological signs, including brain activity, heart rate, muscular activity, blood pressure, and skin temperature [1].

People generally perceive emotion through facial expressions; nevertheless, complex emotions such as pride, gorgeousness, mellowness, and sadness cannot be identified through facial expressions [3]. Physiological signals can therefore be utilized to represent complicated effects.

Gyanendra K. Verma

1.2. WHAT IS EMOTION?

"Everyone knows what an emotion is until asked to give a definition." [4].

Although emotion is prevalent in human communication, the term has no universally agreed meaning. Kleinginna and Kleinginna [5], on the other hand, gave the following definition of emotion:

"Emotion is a complex set of interactions between subjective and objective factors mediated by neural/hormonal systems that can:

1. Generate compelling experiences such as feelings of arousal, pleasure/ displeasure;

2. Generate cognitive processes such as emotionally relevant perceptual effects, appraisals, and labeling processes;

3. Activate widespread physiological adjustments to arousing conditions; and

4. Lead to behavior that is often, but not always, expressive."

1.2.1. Affective Human-Computer Interaction

The researchers described two ways to analyze emotion. The first method divides emotions into joy, fun, love, surprise, grief, *etc*. Another option is to display emotion on a multidimensional or continuous scale. Valence, arousal, and dominance are the three most prevalent aspects. How does a valence scale determine how happy or sad a person is? The arousal scale assesses how relaxed, bored, aroused, or thrilled [6]. The dominance scale depicts submissive (in control) or dominant (empowered) behavior. Emotion identification from facial expressions and voice signals is part of affective HCI. As a result, we will concentrate on the first two modalities, particularly concerning emotion perception. One of the essential needs of MMHCI is that multisensory data be processed individually before being merged.

A multi-modal system may be used in case of insufficient or noisy data. The system may use complementary information from other modalities if one modality's information is absent. If one modality fails to make a decision, the other must do so. Multi-modal HCI (MMHCI) incorporates several domains, such as Artificial Intelligence, Computer Vision, Psychology and others, according to Jaimes A. *et al*. [7]. People communicate frequently using facial expressions, bodily movement, sign language, and other non-verbal communication techniques [8].

Audio and video modalities are commonly employed in man-machine interaction; hence they are vital for HCI. At the feature or choice level, MMHCI focuses on merging several modalities of emotion. Probabilistic graphical models such as the Hidden Markov Model (HMM) and Bayesian Networks are beneficial, according to the study [9]. As a result of its ability to deal with missing values *via* probabilistic inference, Bayesian networks are widely used for data fusion. Vision methods are another option that may be employed for MMHCI [9]. The vision techniques categorize using a human-centered approach and decide how people may engage with the system.

1.3. BACKGROUND

Most emotion recognition research focuses on facial expression and voice emotion [10, 11, 12, 13]. Our book contributed to this approach by presenting an emotion model to predict many complicated emotions in a three-dimensional continuous space, lacking in the previous literature [14]. Even though we have created systems that identify emotion from speech, facial expression, physiological data, and multi-modal fusion of the modalities mentioned above, we focus on emotion modeling in a continuous space and emotion prediction using multi-modal cues.

People usually gather information from various sensory modalities, such as vision (sight), audition (hearing), tactile stimulation (touch), olfaction (smell), and gustation (taste). Then, this information is processed by integrating it into a single cohesive stream of information to communicate with others. In order to integrate numerous complementary and supplemental information, the human brain receives information from multiple communication modalities (such as reading text).

Multi-modal information fusion can be employed in effective systems to integrate related information from different modalities/cues to improve performance [15] and decrease ambiguity in decision-making by reducing data categorization uncertainty. Multi-modal information fusion is necessary for many applications where information from a single modality is inadequate and may contain noise or be insufficient to make conclusions. Consider a visual surveillance system where an object is monitored using visual information. If the object gets occluded, the surveillance system will have no way of tracking it.

Consider a surveillance system that takes information from two modalities (audio and visual information). The object can be tracked even if one of the modalities is unavailable; the system can process the information obtained from other modalities.

CHAPTER 2

Affective Information Representation

Abstract: This chapter presents a brief overview of Affective computing and a formal definition of emotion given by various researchers. Human-computer interaction aims to enhance communication between man and machine so that machines can acquire, analyze, interpret and act on par with human beings. At the same time, Affective human-computer interaction focuses on enhancing communication between man and machines using affective information. Moreover, this chapter deals with Human emotional expression and perception through various modalities such as speech, facial expressions, physiological signals, *etc*. It also detailed the overview of Action Units and Techniques for classifying facial expressions as reported in the literature.

Keywords: Action units, Affective computing, Affective HCI, Emotion expression, Facial expression, HCI.

2.1. INTRODUCTION

A review of multimodal emotional information extraction and processing is presented in this chapter. Affective computing may be thought of as an issue of automatic emotion perception for improved human-machine connection. It entails the detection and interpretation of human emotion as well as the prediction of the user's mental state. It also entails the examination of a person's emotional data in order to determine his or her mental state. Facial emotion identification is explored in the chapter, along with the principles of facial expression and emotion modeling. With the system's limitations, facial expression representation is also available.

2.2. AFFECTIVE COMPUTING AND EMOTION

Affective computing is an interdisciplinary area that encompasses a variety of disciplines, such as computer science, psychology, and cognitive science, among others. Emotion may be exhibited in various ways, including gestures, postures, facial expressions, and physiological signs, heart rate, muscular activity, blood pressure , and skin temperature [1].

Gyanendra K. Verma

According to the definition, "Affective Computing is the research and development of systems and technologies that can identify, understand, process, and imitate human emotions".

People generally perceive emotion through facial expressions; nevertheless, complex emotions such as pride, gorgeousness, mellowness, and sadness cannot be identified through facial expressions [2]. Physiological signals can therefore be utilized to represent complicated effects.

2.2.1. Affective Human-Computer Interaction

The researchers described two ways to analyze emotion. The first method divides emotions into categories, such as joy, fun, love, surprise, grief, *etc*. Another option is to display emotion on a multidimensional or continuous scale. Valence, arousal and dominance are the three most prevalent aspects. How does a valence scale determine how happy or sad a person is? The arousal scale assesses how relaxed, bored, aroused, or thrilled you are [3]. The dominance scale depicts submissive (in control) or dominant (empowered) behavior. Emotion identification from facial expressions and voice signals is part of affective HCI. As a result, we will concentrate on the first two modalities, particularly concerning emotion perception. One of the essential needs of MMHCI is that multisensory data be processed individually before being merged.

Multimodal HCI (MMHCI) incorporates several domains, such as Artificial Intelligence, Computer Vision, Psychology, and others, according to Jaimes A. *et al*. [4]. When people connect, they frequently employ both verbal and nonverbal communication. Non-verbal communication includes facial expressions, bodily movement, sign language, and other non-verbal techniques [5].

The most commonly used modalities in HCI are audio and video; hence, they are vital for HCI. Multimodal HCI focuses on merging several modalities of emotion at the feature or decision level. Probabilistic graphical models such as the Hidden Markov Model (HMM) and Bayesian Networks are beneficial, according to [6]. As a result of its ability to deal with missing values, Bayesian networks are commonly utilized for data fusion through probabilistic inference. Vision methods are another option that may be employed for MMHCI [6]. The vision techniques categorize using a human-centered approach and decide how people may engage with the system.

2.2.2. Human Emotion Expression and Perception

"Everyone knows what an emotion is, until asked to give a definition." [7].

Although emotion is prevalent in human communication, the term has no universally agreed meaning. (Kleinginna and Kleinginna) [8].

On the other hand, it gave the following definition of emotion:

1. "Emotion is a complex set of interactions between subjective and objective factors mediated by neural/hormonal systems that can:

2. Generate compelling experiences, such as feelings of arousal, pleasure/displeasure; b) Generate cognitive processes, such as emotionally relevant perceptual effects, appraisals, and labeling processes;

3. Activate widespread physiological adjustments to arousing conditions; and

4. Lead to behavior that is often, but not always, expressive."

Automatic Human Emotion Recognition captures and extracts information from numerous emotional modalities. We have a wide range of sensors and devices to gather voice data, visual signals, linguistic contents, and physiological signals, among other things, in order to capture emotional information. Although emotional information may be expressed in various ways, Fig. (**2.1**) depicts the most commonly utilized signals reported in the literature.

2.2.2.1. Facial Expressions

The majority of the study has focused on detecting emotion through facial expressions. One of the most dependable and natural ways to communicate emotion is through facial expression. We may quickly detect another person's enjoyment, grief, disagreement, and intentions during the conversation by facial expression. Another advantage of facial expression is that anybody may show it, regardless of age or gender. As a result, affective computing's primary source/channel is facial expression.

2.2.2.2. Audio and Speech

For spoken communication, audio is the most common channel. A speech reflects the style of communication by conveying emotion through linguistic and paralinguistic signals. In any event, a person's affective state may be deduced directly from speech. Emotion theory shows that a physiological response accompanies the emotional state. When we are walking through a jungle, we are

CHAPTER 3

Models and Theory of Emotion

Abstract: This chapter presents a state-of-the-art review of existing emotion theory, modeling approaches, and affective information extraction and processing methods. The basic theory of emotions deals with Darwin's evolutionary theory, Schechter's theory of emotion, and James–Lange's theory. These theories are fundamental building blocks of Affective Computing research. Emotion modeling approaches can be categorized into categorical, appraisal, and dimensional models. Noticeable contributions to Affect recognition systems in terms of modality, database, and dimensionality are also discussed in this chapter.

Keywords: Appraisal model, Categorical model, Dimensional model, Emotion modeling, Emotion theory.

3.1. INTRODUCTION

Modeling emotion is essential for a better understanding of emotions. Efficient modeling of emotions is still challenging due to the involvement of various emotional modalities. As each modality has a different pattern, emotion modeling depends upon the type of input signals. Emotion theories and models are the basis for gaining in-depth knowledge about the induction of emotions. Researchers have proposed many emotion theories; among them, the James-Lange theory, Canon-Bard theory, and Schachter-Singer theory of emotion contributed significantly.

The emotional state of a human being at a particular moment is the combination of the user's physiological, psychological, and subjective experience [1]. The appraisal experience depends on various parameters, such as growing environment, background, and culture. Thus, people feel different experiences of similar phenomena or events.

3.2. EMOTION THEORY

The three primary emotion models are the category, appraisal-based, and dimensional models [2].

The category model is concerned with universally recognized basic emotions. The number of these basic emotions is modest, yet they are all linked to our brains [3]. The appraisal-based approach deals with modelling emotions as a physiological response to stimuli or events that leads to the emergence of emotion and associated action. On a continuous or discrete scale, dimensional methods describe emotion through some independent dimensions. Fig. (**3.1**) depicts various emotion theories.

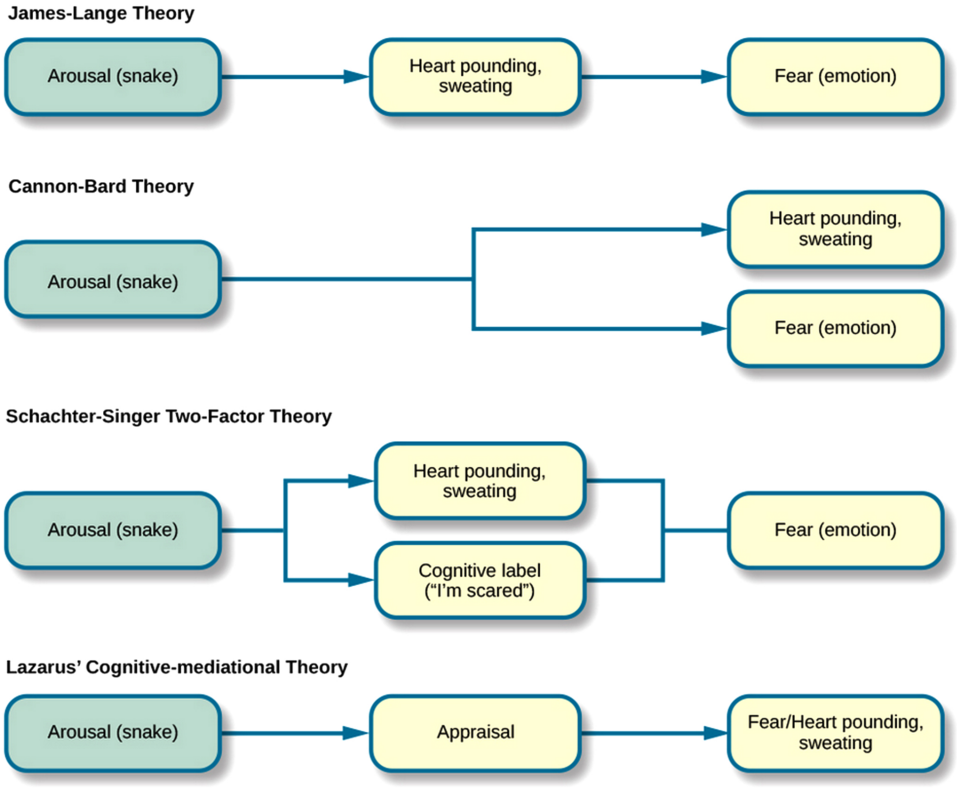


Fig.(3.1). Major emotion theories.

3.2.1. Categorical Approach

Prof. P. Ekman made significant contributions in the area of Emotion Recognition. He was the first to introduce six basic emotions visible in facial expressions [4]. Fig. (**3.2**) illustrates the six primary emotions proposed by Ekman. In 1980, Robert Plutchik introduced the "Wheel of emotion," a new notion of emotion [5]. Fig. (**3.2**) classified a few feelings as core emotions: joy, sorrow, anger, fear, trust, contempt, surprise, and anticipation. In 2001, Parrot offered a classification of emotion, which Robert Plutchik followed.

Fig. (3.2). Ekman's six universal emotions.

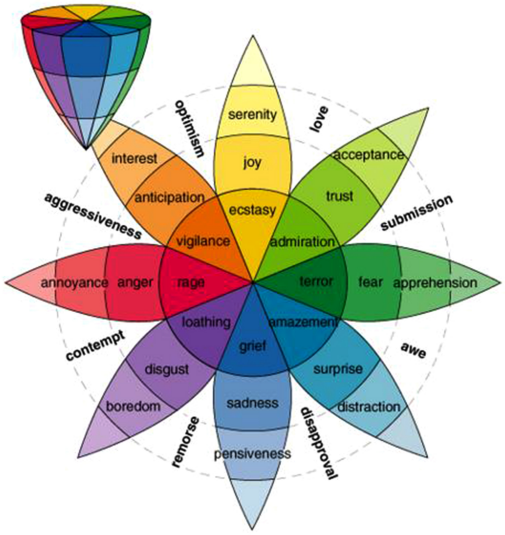


Fig. (3.3). Robert Plutchik's Wheel of emotion [8].

The idea of Parrot was to divide emotions into three categories: primary, secondary, and tertiary [6]. There are six significant emotions, twenty-five secondary emotions, and more than tertiary ones in Parrots' hypothesis. Among the different theories of emotions published in the literature, Plutchik and Conte's [7] study is the most noteworthy. The "Psycho-evolutionary Theory of Emotions" was created by them. Robert Plutchik's wheel of emotion [8] is shown in Figure (**3.3**)

3.2.2. Evolutionary Theory of Emotion by Darwin

Ekman [9] defined the six primary emotions as "happy," "sad," "angry," "fear," "surprise," and "disgust." These six fundamental emotions are based on Darwin's

CHAPTER 4

Affective Information Extraction, Processing and Evaluation

Abstract: This chapter presents a state-of-the-art review of existing affective information extraction and processing approaches. Various evaluation criteria, such as Evaluation matrices like ROC, F1 measure, Mean Square Error, Mean Average Error, Threshold criteria, and Performance criteria are also reported in this chapter.

Keywords: Evaluation measures, F1 measures, Information extraction, ROC curve.

4.1. INTRODUCTION

Information extraction from multimodal cues is usually used as complementary. When getting the information from one cue is terminated, the system may shift to other traits and continue its functioning. There are three primary sources of affective information: speech, images or videos, and text. The first two cues, *i.e.*, speech and images, are widely used for affective information processing. However, text-based affective information analysis targets specific applications and comes under Natural Language Processing (NLP) domain [1]. Where information processing is concerned, the time-series data is easy to process and extract. Usually, the audio and image information is synchronized in a time-dependent manner.

4.2. AFFECTIVE INFORMATION EXTRACTION AND PROCESSING

4.2.1. Information Extraction from Audio

Diverse speech aspects, such as mood, speaker audio, and both, are used to represent different speech information. As a result, experts are interested in learning more about the speech characteristics of various emotions. Vocal tract, prosodic, and excitation source characteristics are the three types of speech features. Short segments of voice signals are used to extract Vocal Tract characteristics. The energy distribution for a variety of speech frequencies is

Gyanendra K. Verma

represented by these properties. Various vocal tract elements and their combinations are employed by various studies for emotion identification. T. Long [2] employed a mixture of Perceptual Linear Prediction (PLP), MFCC, and LPCC to distinguish emotions, such as angry, happy, sad, bored, and neutral, using the log frequency power coefficient (LFPC) vocal tract feature. As pitch has a higher discriminating power than other prosodic variables, it is the most extensively employed prosodic feature for emotion identification. Aside from pitch, log energy is a popular metric for analysing speaking styles and emotions.

4.2.2. Information Extraction from Video

The appearance of changes owing to lighting and position fluctuations complicates geometric feature-based face analysis. Hence, spatio-temporal characteristics can be utilised to identify minor changes in a face. The interest spots in picture sequences are detected using Dollar's approach [3]. This approach was created to identify minor changes in the spatial and temporal domains in which the human action recognition community is most interested. Multiresolution techniques may also be used to extract visual information from photographs.

4.2.3. Information Extraction from Physiological Signals

Physiological signals are vital for emotion, according to several emotion theories. Ekman established that a certain emotion may be linked to a specific physiological pattern. The frequency and amplitude of an EEG signal can be used to characterize it.

To split the physiological signal into distinct frequency bands, a bandpass filter can be utilised. Discrete Wavelet Transform is another approach for decomposing EEG signals into distinct frequency bands (DWT). Galvanic Skin Response (GSR), respiration amplitude, electrocardiogram (ECG), electromyograms (EMG), electrooculogram (EOG), Electro dermal Activity (EDA), Galvanic Skin Response (GSR), Skin Conductance Response (SCR), and skin temperature are examples of peripheral biosignals. EDA and GSR are skin conductance measurements that are extensively utilised for automated emotion identification. GSR is a fairly reliable physiological marker of human arousal, according to J. Kim [4].

4.3. STUDIES ON AFFECT INFORMATION PROCESSING

People use various modes of communication in day-to-day interaction with others. We can broadly categorize these communication modes into verbal and non-verbal communication. Verbal communication involves speech and audio; however, non-verbal communication uses facial expressions, body gestures, and sign language. Both communication modes are vital and play a significant role in communication among human beings. Unfortunately, we lack a better human-computer interface to use these vital communication channels. Facial expressions are essential to recognize sentiments in communications. Koelstra *et al*. [5] discovered substantial changes in N400 ERP responses when placing relevant and irrelevant tags on short videos. They started by extracting features from audio and video channels. Then, the correlation between audio and visual properties was investigated using the method described above. After that, a hidden Markov model is utilized to characterize statistical dependency across time segments and discover the features in the altered domain's fundamental temporal structure. Extensive system testing is used to assess the resilience of our suggested solution.

Emotion identification can be unimodal or multimodal. The unimodal-based method uses a single modality to recognize the emotion. In contrast, the multimodel technique collects emotional information from several inputs, such as audio, video, image, physiological signals, *etc*. We have dealt with multimodel emotion recognition in this study. Several researchers also used this multimodel strategy. S. Koelstra *et al*. [5] suggested a facial expression and EEG signal fusion system for emotion identification using multimodal fusion.

Mamalis A. Nikolaou [6] demonstrated a multimodal emotion identification system that combines facial expressions, shoulder gestures, and aural signals. They used two-dimensional spaces of valence and arousal to map multimodal emotions on a continuous scale. Further, an associative fusion framework was presented using Support Vector Regression (SVR) and Long Short Term Memory neural networks (BLSTM-NNs).

Y. Wang *et al*. [7] used the kernel approach to examine multimodal information extraction and analysis. For modeling the nonlinear interaction between two multidimensional variables, they used Kernel cross-modal factor analysis. They have also developed a method for determining the best transformations to describe patterns.

M. Paleri *et al*. [8] published a paper on feature selection for automated audio-visual person-independent emotion identification. They employed a neural network to compare the performance of different characteristics in an emotion identification system.

CHAPTER 5

Multimodal Affective Information Fusion

Abstract: The multimodal information can be assimilated at three levels 1) early fusion, 2) intermediate fusion, and 3) late fusion. Early fusion can be performed at the sensor or signal level. Intermediate fusion can be at the feature level, and late fusion may be done at the decision level. Apart from that, some more fusion techniques are rank-based, adaptive, *etc*. This chapter provides an extensive review of studies based on fusion and reported noticeable work herewith. Eventually, we discussed the challenges associated with multimodal fusion.

Keywords: Decision fusion, Feature fusion, Multi-modal fusion, Sensor fusion.

5.1. INTRODUCTION

Affective information plays a significant role in emotion recognition. The information acquired from multiple sources or modalities, such as audio, video, and text, is known as Multimodal information. Multimodal information fusion is described as merging information from many sources/modalities to produce superior performance than a single source/modality [1]. There have been several fusion categories mentioned in the literature. The fusion might happen at the signal, feature, or decision level. Early fusion occurs when the fusion is accomplished at the signal or feature level. Late fusion refers to fusion that occurs after a choice has been made. It is not required for distinct modalities to give complementary information in the fusion process, according to P. K. Atrey *et al*. [2], hence it is essential to know which modalities contribute the most. The appropriate number of modalities in the fusion process is a critical component.

5.2. MULTIMODAL INFORMATION FUSION

The early fusion integrates the information acquired from different modalities before applying learning models. Late fusion (also known as decision-level fusion) integrates the information acquired from the output of different algorithms or models. According to L. Hoste [3], Late fusion is based on the semantic information fusion obtained from different modalities. Each modality should be processed first, then integrated at the end to handle multimodal data. Several st-

Gyanendra K. Verma

udies [4 - 6] discussed fusion architectures and multimodal data processing. The data in multimodal processing is not necessarily mutually independent and, therefore, cannot be merged in a context-free fashion. However, the information must be processed in a combined space using a context-dependent model. The critical challenges in multimodal fusion are different feature formats. The dimensionality of joint feature space with temporal synchronization is another issue [4].

The two most significant aspects of multimodal information fusion are i) level of fusion and ii) the type of fusion. The fusion process must be synchronized. Time complexity and cost-effectiveness are the other performance factors. The fusion of two or more modalities must be done methodically. The primary difficulties in multimodal fusion are the number of modalities and information synchronization. The primary difficulties are a correlation between information collected from multiple modalities and the level at which the information is fused [7, 8]. During fusion, many modalities may not necessarily give supplementary information. As a result, it is critical to comprehend the contributions of each modality. The prime role of Multimodal information fusion is to combine data from different modalities/cues to eliminate ambiguity and uncertainty.

Information can be derived from various sources/modalities, such as text, images, and speech. The information can have low and high-level characteristics depending on the input source. The features are fused at either a low (feature fusion) or a high (decision level) level. The fusion techniques and associated works documented in the literature are also presented in this chapter.

5.2.1. Early Fusion

Information can be fused early, *i.e*., at the sensor or signal level. A three-dimensional image, for example, can be created by fusing two or more two-dimensional images [9]. An example of early fusion is audio-visual information fusion, which integrates the audio and video information into a single feature vector. Dimensional reduction methods like principal component analysis or linear discriminate analysis can be used for dimensionality reduction. The two modalities are combined even before classification using Adaptive fusion. In Adaptive fusion, the weights are assigned to each modality, which is one of the essential advantages of early fusion. When we wish to give a modality greater weight, adaptive fusion comes in handy.

Numerous approaches have been reported in the literature to execute fusion, including Bayesian inference, Dempster–Shafer fusion, Maximum Entropy model, and Nave Bay's algorithms. Pitsikalis *et al*. [10] suggested a Bayesian

inference-based technique integrating audio-visual information. They calculated the joint probability of integrated MFCC and texture analysis characteristics.

Mena and Malpica employed a Dempster–Shafer fusion technique for color picture segmentation [11]. For semantic multimedia indexing, Magalhaes and Ruger [12] employed the maximum entropy approach. For image retrieval, they integrated text and picture characteristics.

5.2.2. Intermediate Fusion

The drawback of the early fusion technique is its inability to deal with erroneous data that avoids explicit modeling of the various modalities. Early fusion methods suffer from relative dependability, fluctuations, and asynchrony. Such issue may be resolved by comparing the time instance feature to the time scale dimension of the relevant stream. As a result, by comparing previously seen data instances with current data transmitted *via* a working observation channel, one may create a statistical forecast with a derivable probability value for erroneous instances owing to sensor failures, *etc*.

Probabilistic graphical models are best suited for fusing numerous sources of data in this context [13]. Probabilistic inference also handles noisy features and missing feature values. A hierarchical HMM to identify facial expressions was presented by Cohen *et al*. [14]. The capability to fuse multiple sources of information of dynamic Bayesian networks and HMM variations was demonstrated by Minsky [15]. Carpenter R [16]. proposed a fusion approach to detect office activity and events in video utilizing audio and video signals.

5.2.3. Late Fusion

A multimodal system combines many modalities in order to conclude. It demands a standard framework for representing shared meaning across all modalities and a well-defined method for information assimilation [4]. Late fusion models typically employ distinct classifiers for each stream that are trained separately. The outcome is derived by fusing the outcome of individual classifiers. Only at the integration stage is the correspondence between the channels recognized. Late fusion has several clear advantages.

The inputs may be recognized independently, they do not have to coincide. The late fusion system employs classifiers that can be trained on a single data set but are scalable in terms of the number of modes and vocabulary. As with audio-visual recognition, we need to find a decent heuristic for extracting features from

CHAPTER 6

Multimodal Fusion Framework and Multiresolution Analysis

Abstract: This chapter presents a multi-modal fusion framework for emotion recognition using multiresolution analysis. The proposed framework consists of three significant steps: (1) feature extraction and selection, (2) feature level fusion, and (3) mapping of emotions in three-dimensional VAD space. The proposed framework considers subject-independent features that can incorporate many more emotions. It is possible to handle many channel features, especially synchronous EEG channels and feature-level fusion works. This framework of representing emotions in 3D space can be extended for mapping emotion in three-dimensional spaces with three specific coordinates for a particular emotion. In addition to the fusion framework, we have explained multiresolution approaches, such as wavelet and curvelet transform, to classify and predict emotions.

Keywords: Curvelet transform, Emotion recognition, Wavelet transform.

6.1. INTRODUCTION

Multimodal fusion combines multiple cues that may act as complementary information to improve the system's performance. Several fusion approaches are reported in the literature; nonetheless, early, middle and late fusion are three primary categories. Before feeding into the learning phase, the features gathered from diverse modalities must be integrated into a single representation in an early fusion. Intermediate fusion can deal with insufficient data and asynchrony across distinct modalities. Decision-level fusion deals with semantic information as the decision is made by considering the outcome of different modalities after feature extraction. One of the essential criteria for multimodal data processing is that the data be processed individually before joining.

6.2. THE BENEFITS OF MULTIMODAL FEATURES

The challenging issues in multimodal fusion are 1) data types of modalities, 2) the synchronization of different types of modalities , and 3) the level of the fusion [1, 2]. The choice of fusion level may be easy for a similar type of data. For example,

if the two modalities are temporal, the fusion may be easy; however, if one is temporal and the other is special, fusion becomes challenging. The contribution of each modality is unique, and each fusion process does not need to enhance the system's performance.

Multiple modalities, such as text, picture, audio, and physiological inputs, can all have features extracted individually. The multimodal element not only adds to the information available but also has the potential to improve the system's performance. The following are some of the reasons for not relying on a single method of features:

6.2.1. Noise In Sensed Data

Sensor, channel, and modality-specific noise are the three forms of noise in sensed data. Sensor noise is the noise produced by the sensor. Each pixel of a camera sensor, for example, is made up of one or more light-sensitive photodiodes that convert incoming light into an electrical signal. The color value of the final image pixel represents the signal. Even if the same pixel were exposed to the same quantity of light numerous times, the resultant color value would not be equal. However, it would have a slight variance known as “noise.”

Channel noise, on the other hand, is the result of the data transmission or medium deteriorating. Under slightly varied lighting conditions, for example, the same HCI modality may change. Person identification is perhaps the most well-known example. Under varied lighting circumstances, the same face looks different from two separate faces recorded under the same illumination conditions. Finally, modality-specific noise is noise induced by a disagreement between the acquired data and the standard interpretation of the modality.

6.2.2. Non-Universality

A system may not be able to get valuable data from only one modality. For example, complex emotions, such as pride, joy, excitement, sorrow, *etc*., cannot be identified by facial expressions only. Thus we must rely on other methods, such as physiological signals, to recognize complex emotions. Similarly, iris recognition biometrics may fail due to different eye conditions like long eyelashes, sloping eyelids, or certain eye pathologies issues. On the other hand, a face recognition system may nevertheless be a valuable biometric modality. While no one modality is ideal, combining them should provide more excellent user coverage, enhancing accessibility, particularly for the impaired.

6.2.3. Complementary Information

The information gained from the other modality can be utilized as a supplement. A single modality-based algorithm may fail if the input signal is lost or corrupted. A unimodal system may stop functioning in case of inputs from one modality interrupt. However, a multimodal system can continuously perform by taking inputs from other modalities. An object-tracking method based on visual modality works perfectly in the usual scenario. However, if the object occluded, it will stop tracking; in this instance, the voice modality may provide complementary information.

6.3. FEATURE LEVEL FUSION

Multimodal information fusion is the job of combining corresponding data from different modalities/cues in order to eliminate ambiguity and uncertainty. Depending on the applications, information can be acquired from various sources/modalities, such as text, pictures, and speech. The low and high-level features extracted from various modalities can be fused at either the feature level or a decision level.

Sensor level fusion combines raw data or data generated from several sources from sensory (raw) data. The finest example of sensor-level fusion is creating a 3D picture from two 2D images. The type of sensors and information sources are two essential concerns in sensor-level fusion. Other important characteristics for sensor level fusion are as follows [3]:

- Sensors' computational capability
- Topology, communication structure, computing resources, and
- System goals and optimization
- Improved detection, tracking, and identification are some of the benefits of sensor-level data fusion.
- Improved situational awareness and assessment
- Increased sturdiness
- Coverage that is both spatially and temporally extensive
- Reduced communication and computing costs, as well as a faster reaction time.

Feature level fusion is achieved by extracting features from several modalities/sources individually and then combining them after normalization. The benefit of feature-level fusion is that it can find the correlation between feature vectors, which can help the system perform better. The following parameters affect feature-level fusion.

CHAPTER 7

Emotion Recognition From Facial Expression In A Noisy Environment

Abstract: This study presents emotion recognition from facial expressions in a noisy environment. The challenges addressed in this study are noise in the images and illumination changes. Wavelets have been extensively used for noise reduction; therefore, we have applied wavelet and curvelet analysis from noisy images. The experiments are performed with different values of Gaussian noise (mean: 0.01, 0.03) and (variance: 0.01, 0.03). Similarly, for experimentation with illumination changes, we have considered different dynamic ranges (0.1, 0.9). Three benchmark databases, Cohn-Kanade, JAFFE, and In-house, are used for all experimentation. The five best machine learning algorithms are used for classification purposes. Experimental results show that SVM and MLP classifiers with wavelet and curvelet-based coefficients yield better results for emotion recognition. We can conclude that Wavelet coefficients-based features perform well, especially in the presence of Gaussian noise and illumination changes for facial expression recognition.

Keywords: Cohn-Kanade, Curvelet transform, Facial expression recognition, JAFFE, MLP, SVM, Wavelet transform.

7.1. INTRODUCTION

Emotions are essential for machines to make an intelligent decision. We are witnessing exponential growth in computing power, however, lacking robust algorithms that can enhance the intelligence of the machines. Emotions play a significant role in an intelligent behavior by machines at par with human beings [1, 2, 3, 4]. Emotions can be exhibited through various modes, such as facial expression, auditory expression, physiological expression, gesture, body language, sign language, *etc*. Facial expression is the most widely used modality among the above modalities due to quick presentation and recognition. Moreover, facial expressions are more accessible to acquire, process, and analyze than other modalities.

Multiresolution Analysis (MRA) proved useful in various applications, including medical imaging, satellite imaging, biometrics, *etc*.Wavelet and Curvelet transforms are two classical algorithms used for MRA. In [5, 6], wavelet

Gyanendra K. Verma

transform-based multiresolution analysis was performed. However, the application of MRA in a noisy environment is relatively new for emotion recognition.

Some of the research [7] works are based on applying multiple modalities rather than a single modality to analyze emotions. Mansoorizadeh *et al.* [8] proposed a multimodal fusion framework for human emotion recognition. Some of the work is based on curvelet analysis [9, 10]. Lee and Shih [11] presented contourlet analysis with regularized discriminate analysis for facial emotion recognition. Shan *et al.* [12] presented a facial expression recognition system using local binary patterns. M. Yeasin *et al.* [13] proposed an approach for the measurement of levels of interest from video for facial expression recognition.

Generally, raw images are degraded due to various phenomena such as varying lighting conditions, environmental effects, high/ low brightness, contrast, *etc*. Thus, facial expression recognition became more challenging in a noisy environment. This chapter deals with the recognition of facial expressions in a noisy environment. We have experimented with three benchmark databases to prove the proposed algorithm's usefulness and robustness. Various types of noise have been added with different mean and variance values. Then, the multiresolution approaches were applied to extract the most prominent features of noisy facial expression images. Getting the desired accuracy of the system in a noisy environment is still a challenge [14]. This chapter proposes multiresolution approaches based on wavelet and curvelet analysis to improve emotion recognition performance.

7.2. THE CHALLENGES IN FACIAL EMOTION RECOGNITION

The major challenge in facial expression recognition is noise and illumination change. The performance of most of the algorithms degraded due to the presence of these two parameters: 1) the presence of noise in facial expression and 2) varying illumination. We have recreated a noisy database by adding Gaussian white noise with different mean and variance values. At the same time, we have also modified the dynamic range of the test images by 0.1 to 0.9. The sample database images having different noise and illumination changes are shown in Fig. (**7.1**).

The images may be degraded during acquisition due to environmental conditions such as lighting, illumination, handshake, *etc*. The noise may also be added due to the sensor error during raw data acquisition. Many systems cannot preprocess or remove such added noise in the images. Therefore, a robust system is required to handle and analyze such noisy data efficiently. In an image, the edge information

is degraded due to noise. It may happen even due to contrast reversals in some cases.

We have prepared the database for the experiments by adding Gaussian white noise in the images at different degrees. The values of mean and variance were kept in the range of 0.01- 0.03. For the first iteration of the experiment, the mean and variance values are set as 0.1 and 0.1. For the second test, the values are 0.03 and 0.03. Fig. (**7.2**) illustrates some sample images having additive noise. The research revealed that emotional signals could withstand noise-induced distortion. As all multiresolution approaches are inherently scale-invariant, we did not test for scale variation.

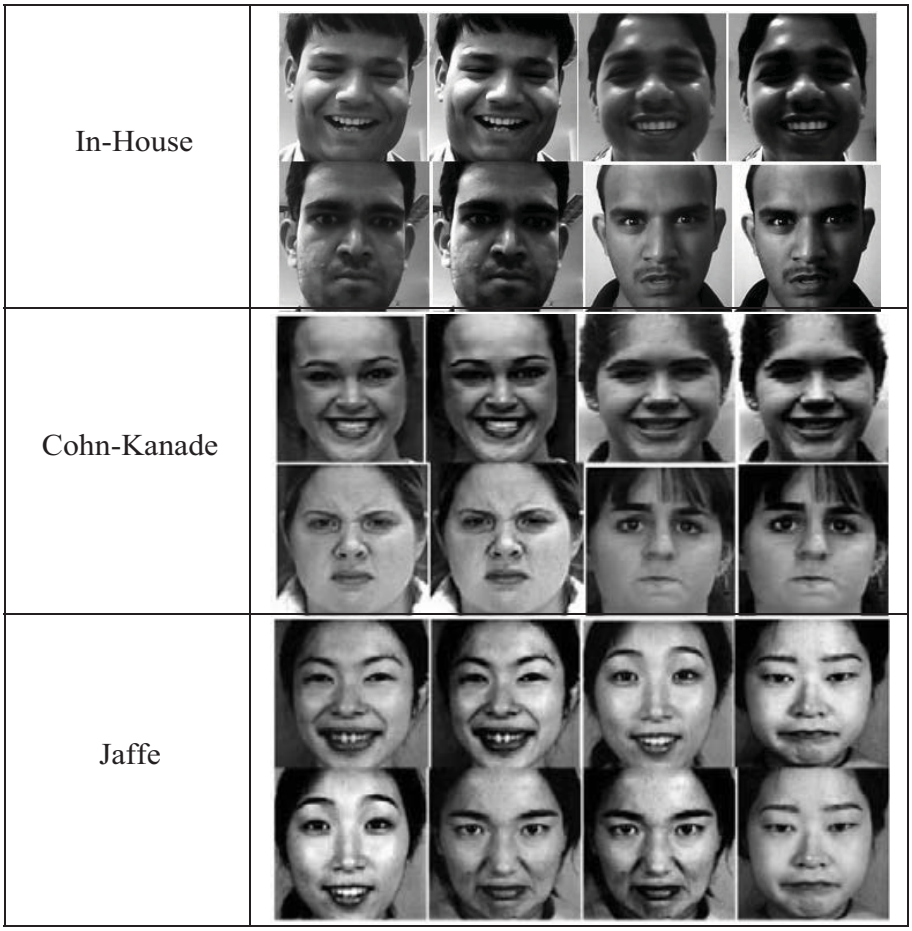


Fig. (7.1). Sample images with varying illumination conditions.

CHAPTER 8

Spontaneous Emotion Recognition From Audio-Visual Signals

Abstract: This chapter introduces an emotion recognition system based on audio and video cues. For audio-based emotion recognition, we have explored various aspects of feature extraction and classification strategy and found that wavelet analysis is sound. We have shown comparative results for discriminating capabilities of various combinations of features using the Fisher Discriminant Analysis (FDA). Finally, we have combined the audio and video features using a feature-level fusion approach. All the experiments are performed with eNTERFACE and RML databases. Though we have applied multiple classifiers, SVM shows significantly improved performance with a single modality and fusion. The results obtained using fusion outperformed in contrast results based on a single modality of audio or video. We can conclude that fusion approaches are best as it is using complementary information from multiple modalities.

Keywords: Emotion recognition system, eNTERFACE, Feature fusion, Machine learning, RML database, Support vector machine.

8.1. INTRODUCTION

Most emotion recognition work is based on facial expression and vocal emotion [1, 2, 3, 4]. However, emotion modeling in a continuous space is completely lacking in the existing literature [5]; our book contributed in this direction by proposing an emotion model to predict a large number (more than fifteen) of complex emotions in a three-dimensional continuous space. Although we have implemented various systems to recognize emotion from speech, facial expression, physiological signals, and multimodal fusion of the above modalities, our emphasis is on emotion modeling in a continuous space.

This chapter investigates the use of audio-visual signals for spontaneous affect recognition. The proposed approach is based on Multiresolution analysis (MRA), which can analyze a signal at many resolutions. An MRA is the design process for the most practically important discrete wavelet transforms (DWT) and the reasoning for the Fast Wavelet Transform algorithm (FWT). MRA uses DWT to

Gyanendra K. Verma

extract the features. Different classifiers are used to classify the data, including the Support Vector Machine (SVM), Multilayer Perceptron (MLP), and K mean classifier. The MRA-based method in this book was created by building on prior work in emotion recognition using MRA [6]. Emotion recognition using audio and visual modalities is covered in this chapter.

The following is a breakdown of the chapter's structure. We review the many techniques for emotion recognition that have been reported in the literature. Then, using a development set of eNTERFACE [7] and RML [8] multimodal emotion databases, we present an emotion detection system based on auditory and visual cues and some of its main components. The use of multimodal fusion for emotion recognition is then investigated. Finally, experimental findings for audio and visual cues, as well as their combination utilizing feature level fusion, are shown on the eNTERFACE and RML databases.

8.2. RECOGNITION OF SPONTANEOUS EFFECTS

Affect recognition has been an emerging study subject within Human-Computer Interaction (HCI). Several affective states, including thinking, shame, and despair, are considered complex affective states and communicated *via* hundreds of distinct facial expressions. Researchers employed multimodal cues to identify emotion as a single cue may not correctly describe complex emotional states [9]. Several methods have been reported in the literature to extract characteristics from audio and visual cues.

8.3. THE DATABASE

Many academicians have been inspired to construct an emotion database due to recent breakthroughs in emotion identification.

- MIT [10],
- MMI [11],
- HUMAINE [12],
- VAM [13],
- SEMAINE [14],
- MAHNOB-HCI [15], and
- DEAP [16].

We used the eNTERFACE and RML audio-visual databases in this study. Fig. (**8.1**) depicts the facial expressions of various emotions for the eNTERFACE and RML databases. At the same time, Table **8.1** summarizes the database content for both databases.

8.3.1. eNTERFACE Database

The eNTERFACE05 is an audio and video emotion database that serves as a benchmark. Happy, Angry, Disgust, Sadness, Surprise, and Fear are the six primary emotions in the database. Forty-two people were involved in the collection of speech samples.

Fig. (8.1). Facial expressions for different emotions first row - eNTERFACE database, second row- RML database.

Table 8.1. Database content summary of eNTERFACE and RML database.

eNTERFACE	
Database type	Audiovisual
No. of subjects	44
Language	English
# Emotion	6 Universal emotions
#Video	1320
Image frame size	720*576
Frame rate	25
Audio sampling rate	48000
RML	
Database type	Audiovisual
No. of subjects	8
Language	6 (English, Mandarin, Urdu, Punjabi, Persian, Italian)

CHAPTER 9

Multimodal Fusion Framework: Emotion Recognition From Physiological Signals

Abstract: This study presents a multimodal fusion framework for emotion recognition from physiological signals. In contrast to emotion recognition through facial expression, a large number of emotions can be recognized accurately through physiological signals. The DEAP database, a benchmark multimodal database with many collections of EEG and peripheral signals, is employed for experimentation. The proposed method takes into account those features that are subject-independent and can incorporate many more emotions. As it is possible to handle many channel features, especially synchronous EEG channels, feature-level fusion is applied in this study. The features extracted from EEG and peripheral signals include relative, logarithmic, and absolute power energy of Alpha, Beta, Gamma, Delta, and Theta. Experimental results demonstrate that physiological signals' Theta and Beta bands are the most significant contributor to the performance. On the other hand, SVM performs outstandingly.

Keywords: 3D emotion model, EEG, Facial expression, Fusion framework, Physiological signal.

9.1. INTRODUCTION

Physiological cues are essential in determining whether or not a person's behavior or emotional state has changed. Physiological signals are biosignals such as Galvanic Skin Response (GSR), Respiration amplitude, Electrocardiogram (ECG), Electro-mayograms (EMG), Electrooculogram (EOG), Electro Dermal Activity (EDA), Galvanic Skin Response (GSR), Skin Conductance Response (SCR), and skin temperature [1]. EDA and GSR are skin conductance measurements extensively utilized for automated emotion identification. According to Kim J., GSR is a relatively reliable physiological marker of human arousal [2]. EEG, EMG, Skin conductance, BVP, and other physiological signals can be employed for emotion identification.

Electroencephalogram (EEG) is used in this study. EEG measures voltage changes in the brain's neurons induced by ionic current flows. For a brief length of time, 20–40 minutes, several electrodes put on the scalp capture the brain's spont-

Gyanendra K. Verma

aneous electrical activity. A brief introduction of various bio signals is given. A typical architecture of emotion recognition from physiological signals is illustrated in Fig. (**9.1**).

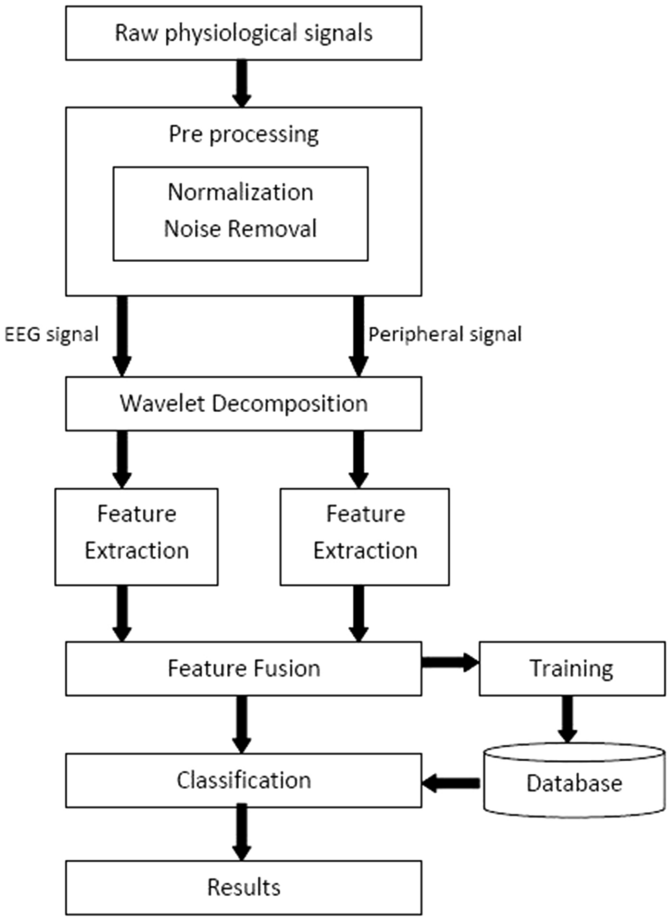


Fig. (9.1). Architecture of emotion recognition from physiological signal.

9.1.1. Electrical Brain Activity

EEG is a technique for detecting continuous electrical brain activity with amplitudes ranging from 1 to 200 microvolts. The 10-20 electrode placement approach is the most prevalent. Hundreds of electrodes (roughly) are implanted in the skull depending on its geometric dimensions.

9.1.2. Muscle Activity

Electromyography measures the electrical potentials that muscle fibers generate from muscular contraction (EMG). Ekman *et al*. [3] stated that facial EMG could give a sensitive and objective metric for emotion recognition as some facial muscles cannot be engaged voluntarily.

9.1.3. Skin Conductivity

The skin's electrical resistance is measured by skin conductance and regulated by the activity of the preparatory glands. The greater the skin conductance, the more active the preparation glands are.

9.1.4. Skin Temperature

According to McFarland [4], stimulating negative emotions causes skin temperature to drop, but calm, pleasant emotions cause skin temperature to rise. Emotion identification may also be made using heart rate and respiratory rate. Different physiological signals may be integrated to increase the accuracy of emotion identification.

Many Physiological signals are vital for emotion, according to several emotion theories. Ekman [5] proved that a particular physiological pattern might be linked to a specific emotion. For feature extraction, two types of data are employed in this study: EEG and peripheral. Frequency and amplitude are two parameters that may be used to define EEG. The EEG signal may be categorized into different bands based on frequency: Alpha, Beta, Gamma, Delta, and Theta. The frequency range of the above bands are 1 - 4 Hz., 4 - 8 Hz., 8 -12.5 Hz., 12.5 - 28 Hz. and 30-40 Hz, respectively.

9.2. MULTIMODAL EMOTION DATABASE

The availability of multimodal datasets is an essential factor in evaluating the performance of a pattern recognition system. The following are the multimodal databases that are currently available:

- HUMAINE database: This is an audiovisual database with gestures.
- Audiovisual database of Belfast naturalists.
- Smartkom combines audiovisual and gestural input.
- Salas audiovisual database.

CHAPTER 10

Emotions Modelling in 3D Space

Abstract: In this study, we have discussed emotion representation in two and three-dimensional space. The three-dimensional space is based on the three emotion primitives, *i.e*., valence, arousal, and dominance. The multimodal cues used in this study are EEG, Physiological signals, and video (under limitations). Due to the limited emotional content in videos from the DEAP database, we have considered only three classes of emotions, *i.e*., happy, sad, and terrible. The wavelet transforms, a classical transform, were employed for multi-resolution analysis of signals to extract features. We have evaluated the proposed emotion model with standard multimodal datasets, DEAP. The experimental results show that SVM and MLP can predict emotions in single and multimodal cues.

Keywords: Arousal, DEAP database, Dominance, EEG, Multiresolution analysis, Support vector machine, Valence.

10.1. INTRODUCTION

Emotion modelling is an effective mechanism to implement emotions in various domains like Artificial Intelligence, Robotics, Psychology and Human-Computer Interaction. Modelling emotions enables machines to exhibit human being like behavior. With the advancement of technology, affect recognition is shifting from distinct emotions toward continuous two or three-dimensional space. Now, researchers focus on the investigation of the dimensional model of emotion. The Dimensional model deals with complex emotions that can be represented by two or three-dimension space. Generally, a 2D model has two dimensions: valence and arousal (in positive and negative axes). A 2D model is typically represented by four quadrants. Some scholars have used various terminologies for distinct emotion dimensions.

Whissell [1] developed a two-dimensional emotion model based on appraisal and activation. We looked at some studies on emotion representation in two-dimensional spaces. We reviewed the two-dimensional emotion model in this chapter. The limitations of the two-dimensional model and the needs of the three-dimensional emotion model are discussed. We will examine how emotions are represented in three-dimensional VAD spaces. After that, a 3D emotion graph is

Gyanendra K. Verma

presented by expressing many emotions in three-dimensional space. In separate portions of this chapter, emotion predictions from multimodal signals are also defined. Finally, we conclude this chapter by presenting the essential findings and contributions.

To represent emotions in three-dimensional space, we used the DEAP database [2], an extensive database encompassing more than twenty-five emotions [3, 4, 5] have all effectively employed the DEAP database in emotion recognition experiments.

The significant contributions of this work for predicting multimodal emotion in VAD space are as follows:

1. It presents the approach for affect prediction regarding valence, arousal, and dominance based on facial expression, EEG, and peripheral cues.
2. It also predicts the correlation between emotion dimensions and demonstrates significant performance improvement.
3. It compares state-of-the-art machine learning techniques *i.e*., Support Vector Machine and Multilayer Perceptron for continuous affect prediction.

10.2. AFFECT REPRESENTATION IN 2D SPACE

Whissell [1] proposed a two-dimensional emotion model based on the appraisal and activation of two values. They depict the location of emotive words in a (-3; +3) Evaluation-Activation (EA) space. Each of the six Ekman's fundamental emotions is represented as a 2D EA space. Each emotion has specific affective weights to be determined by the emotion analyzer.

Fig. (**10.1**) illustrates the 2D valence-arousal space as proposed by Whissell. It is fascinating that the emotion of 'joy' exists in isolation, even though it should be associated with cheerfulness and pleasure. Furthermore, the emotions' surprise' and 'joy' belong to the same quadrant 2. Furthermore, intermediary states between 'joy' and the rest of the emotions are rarely detected by the Whissell space. As a result, we may conclude that the existing theory cannot be validated using this two- dimensional model. Now the open question is whether two-dimension space suffices to represent all emotions or if there is a need for more dimensions to represent emotions efficiently.

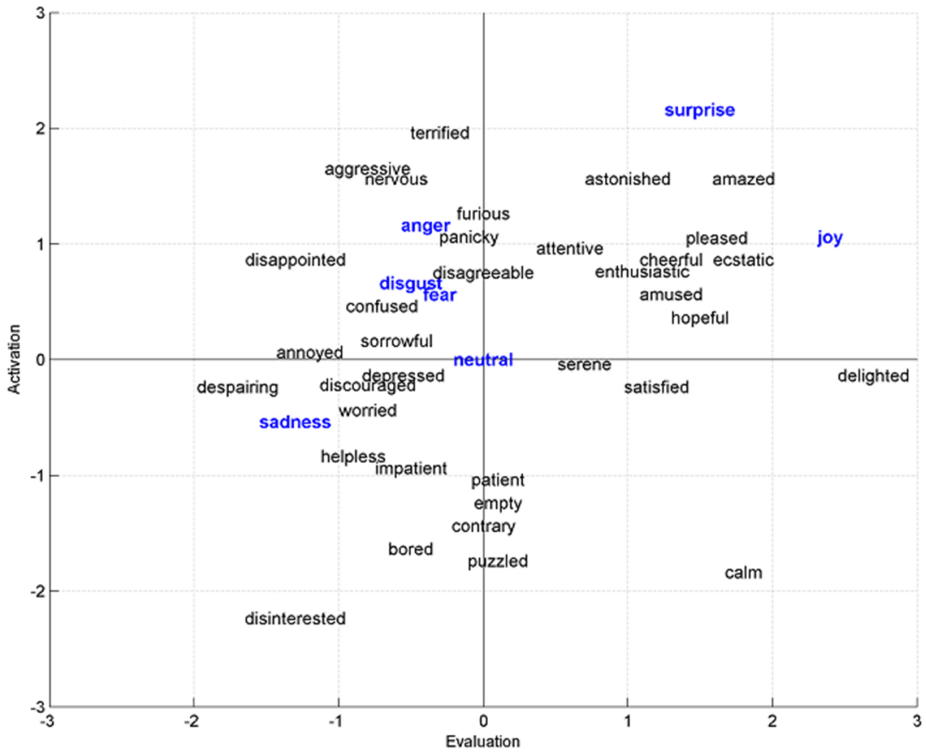


Fig (10.1). EV space as proposed by Whissell.

In an article [6], an anonymous researcher described a mood and emotion monitoring experiment. The resultant map of emotions shows that the two-dimensional mood measure does an excellent job of capturing emotional states. The positive and powerful feelings, such as pride, excitement, and joy, ended up in the space's top right quadrant. Irritation, anguish, and worry were in the third quadrant, which was the reverse of the first. The weak emotions are around the center of the arousal scale. However, stronger emotions are in the higher and lower regions of the plane, depending on valence.

As can be seen from Fig. (**10.2**), 'contentment,' 'affection,' and 'sadness,' are in the same valence axis in the VA space, which is not feasible according to existing theory. However, because the feeling 'affection' lies in the cheerful group, it must be close to the word 'joy.' This model also incorporates restricted emotions.

In valence-arousal space, this model is unable to reflect emotions appropriately. So, once again, the issue arises: do we need more dimensions to express emotions correctly?

SUBJECT INDEX

Gyanendra K. Verma

D

E

F

G

H

I

K

L

M

N

T

V

W

Gyanendra K. Verma

Gyanendra K. Verma is currently an Assistant Professor (senior grade) in the Department of Information Technology, National Institute of Technology Raipur, India. He completed his B. Tech. in 2006 from Harcourt Butler Technical University (formerly HBTI) Kanpur, India and M. Tech. & Ph.D. from the Indian Institute of Information Technology (IIITA) Allahabad, India, in 2009 and 2016, respectively. He has teaching and research experience of more than 10 years in the area of Computer Science and Information Technology, with a special interest in Image Processing, Speech and Language Processing and Human-Computer Interaction. He has collaborated actively with researchers in several other disciplines of computer science, particularly on Medical Imaging problems at the SILP laboratory, IIIT Allahabad. He has served on roughly fifty conference and workshop program committees and served as the Organizing Chair for MIND 2019. His research on the application of Wavelet Transform in Medical Imaging and Computer Vision problems have been cited extensively. He is a senior member of Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM); reputed International Technical Societies.