Machine Audition:

Principles, Algorithms and Systems

Wenwu Wang
University of Surrey, UK

Information Science
REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Director of Book Publications:	Julia Mosemann
Acquisitions Editor:	Lindsay Johnston
Development Editor:	Joel Gamon
Publishing Assistants:	Casey Conapitski and Travis Gundrum
Typesetter:	Michael Brehm
Production Editor:	Jamie Snavely
Cover Design:	Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Copyright © 2011 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher. Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

Machine audition : principles, algorithms, and systems / Wenwu Wang, editor.
p. cm.
Includes bibliographical references and index.
Summary: "This book covers advances in algorithmic developments, theoretical frameworks, and experimental research findings to assist professionals who want an improved understanding about how to design algorithms for performing automatic analysis of audio signals, construct a computing system for understanding sound, and to learn how to build advanced human-computer interactive systems"--Provided by publisher. ISBN 978-1-61520-919-4 (hardcover) -- ISBN 978-1-61520-920-0 (ebook) 1. Computational auditory scene analysis. 2. Signal processing. 3. Auditory perception--Computer simulation. I. Wang, Wenwu, 1974- TK7881.4.M27 2010
006.4'5--dc22

2010010161

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 17
Multimodal Emotion Recognition

Sanaul Haq
University of Surrey, UK

Philip J.B. Jackson
University of Surrey, UK

ABSTRACT

Recent advances in human-computer interaction technology go beyond the successful transfer of data between human and machine by seeking to improve the naturalness and friendliness of user interactions. An important augmentation, and potential source of feedback, comes from recognizing the user's expressed emotion or affect. This chapter presents an overview of research efforts to classify emotion using different modalities: audio, visual and audio-visual combined. Theories of emotion provide a framework for defining emotional categories or classes. The first step, then, in the study of human affect recognition involves the construction of suitable databases. The authors describe fifteen audio, visual and audio-visual data sets, and the types of feature that researchers have used to represent the emotional content. They discuss data-driven methods of feature selection and reduction, which discard noise and irrelevant information to maximize the concentration of useful information. They focus on the popular types of classifier that are used to decide to which emotion class a given example belongs, and methods of fusing information from multiple modalities. Finally, the authors point to some interesting areas for future investigation in this field, and conclude.

INTRODUCTION

Speech is the primary means of communication between human beings in their day-to-day interaction with one another. Speech, if confined in meaning as the explicit verbal content of what is spoken, does not by itself carry all the information that is conveyed during a typical conversation, but is in fact nuanced and supplemented by additional modalities of information, in the form of vocalized emotion, facial expressions, hand gestures and body language. These supplementary sources of information play a vital role in conveying the emotional state of interacting human beings,

DOI: 10.4018/978-1-61520-919-4.ch017

Copyright © 2011, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

referred to as the "human affective state". The human affective state is an indispensable component of human-human communication. Some human actions are activated by emotional state, while in other cases it enriches human communication. Thus emotions play an important role by allowing people to express themselves beyond the verbal domain.

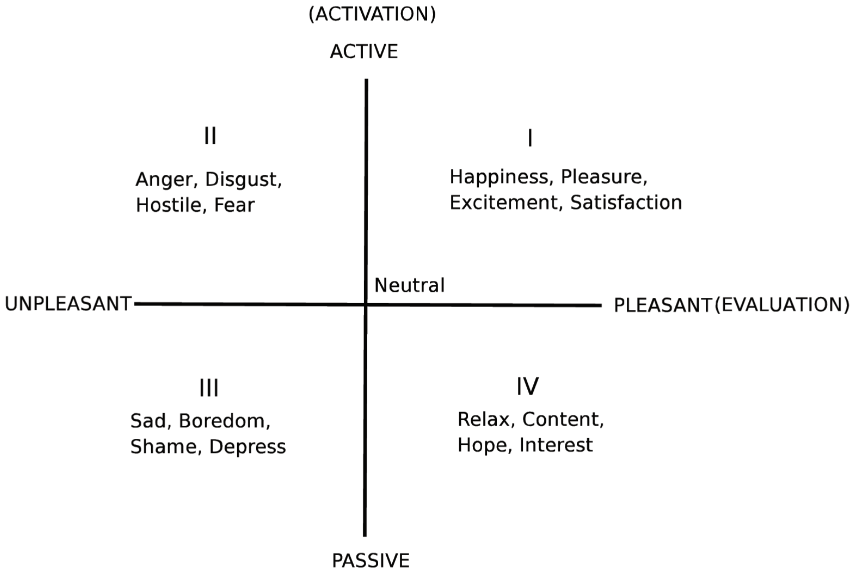
Most current state-of-the-art human-computer interaction systems are not designed to perceive the human affective state, and as such are only able to deliver or process explicit information (such as the verbal content of speech) and not the more subtle or latent channels of information indicative of human emotion; in effect, the information from the latter sources is lost. There are application domains within existing HCI technology where the ability of a computer to perceive and interpret human emotional state can be regarded as an extremely desirable feature. Consider, for example, that if intelligent automobile systems can sense the driver's emotional state and tune its behavior accordingly, it can react more intelligently in avoiding road accidents. Another example is that of an affect sensing system at a call center for emergency services which can perceive the urgency of the call based on the caller's perceived emotional state, allowing better response to the situation. We can also envision applications in the game and entertainment industries; in fact the ability of computers to interpret and possibly emulate emotion opens up potentially new territories in terms of applications that were previously out of bounds for computers. These considerations have activated investigation in the area of emotion recognition turning it into an independent and growing field of research within the pattern recognition and HCI communities.

There are two main theories that deal with the conceptualization of emotion in psychological research. The research into the structure and description of emotion is very important because it provides information about expressed emotion, and is helpful into affect recognition. Many psychologists have described emotions in terms of discrete theories (Ortony et al., 1990), which are based on the assumption that there exist some universal basic emotions, although their number and type varies from one theory to another. The most popular example of this description is the classification of basic emotions into anger, disgust, fear, happiness, sadness and surprise. This idea was mainly supported by cross-cultural studies conducted by Ekman (1971, 1994), which showed that emotion perception in different cultures is the same for some basic facial expressions. Most of the recent research in affect recognition, influenced by the discrete emotion theory, has focused on recognizing these basic emotions. The advantage of the discrete approach is that in daily life people normally describe observed emotions in terms of discrete categories, and the labeling scheme based on category is very clear. But the disadvantage is that it is unable to describe the range of emotions which occur in natural communication. There is another theory known as dimensional theory (Russell et al., 1981; Scherer, 2005), which describes emotions in terms of small sets of dimensions rather than discrete categories.

These dimensions include evaluation, activation, control, power, etc. Evaluation and activation are the two main dimensions to describe the main aspects of emotion. The evaluation dimension measures human feeling, from pleasant to unpleasant, while the activation dimension, from active to passive, measure how likely the human is going to take action under the emotional state. The emotion distribution in two dimensions is summarized in Figure 1, which is based on Russell et al. (1981) and Scherer (2005) research.

The first quadrant consists of happiness, pleasure, excitement and satisfaction, the second quadrant consists of anger, disgust, hostile, fear, the third quadrant consists of sad, boredom, shame, depress, and the fourth quadrant consist of relax, content, hope and interest. The point of intersection of the two dimensions represents neutral. The dimensional representation makes it possible for

Figure 1. Distribution of emotion in 2D space based on Russell et al., 1981 and Scherer, 2005 research. The evaluation dimension measures human feeling from pleasant to unpleasant, while the activation dimension measures how likely the human is going to take an action under the emotional state from active to passive. The intersection of the two dimensions provides the neutral state.


(ACTIVATION)
ACTIVE
II
Anger, Disgust, Hostile, Fear
I
Happiness, Pleasure, Excitement, Satisfaction
UNPLEASANT
Neutral
PLEASANT (EVALUATION)
III
Sad, Boredom, Shame, Depress
IV
Relax, Content, Hope, Interest
PASSIVE

the evaluators to label a range of emotions. In this method, since high dimensional emotional states are projected onto 2D space which result in some loss of information. It becomes difficult to differentiate between some emotions, e.g. anger and fear, while others lie outside 2D Space, e.g. surprise. The evaluators will need training to label the data because this representation is not very clear, e.g. Feeltrace system (Cowie et al., 2000). The results from different raters may be more inconsistent compared to the discrete approach.

The goal of this chapter is to provide a summary of the research work that has been done in the field of human affect recognition by using the audio, visual, and audio-visual information. We first discuss the different types of databases (audio, visual and audio-visual modalities) that have been recorded for the analysis of human affect behavior. The next section explores various kinds of audio and visual features which are investigated by researchers. The feature extraction is followed by feature selection and feature reduction techniques, which are used to reduce the dimensionality of data for computational efficiency and improved performance. In particular, we present two linear transformations, principal component analysis (PCA) and linear discriminant analysis (LDA). We then describe popular classification strategies, which is followed by fusion techniques, future research directions and the conclusion.

METHODOLOGY

Databases

In order to develop an automatic emotion recognizer, the first requirement is to have sufficient data that spans the variety and range of affective expressions. Spontaneous emotion data are difficult to collect because they are relatively rare, short lived and involve ethical issues. The other problem with these databases is that the data needs to be labeled, which can be expensive, time consuming and error-prone, making it really difficult to analyze the automatic spontaneous emotion recognition. Due to these problems, most of the research in this field is based on acted emotions. The acted databases are recorded by asking the actors or non-actors to express different emotions in front of a recording camera and/or microphone. The recording is performed in a controlled laboratory environment.

It has been found that the acted emotions are different in audio profile, visual appearance and timing from spontaneous emotions. Whissell (1989) concluded that acted emotions in spoken language may differ in timing and choice of words from spontaneous emotions. In the case of facial expressions, differences exist between acted and spontaneous expressions in terms of dynamics and muscle movement (Ekman et al., 2005). Many types of spontaneous smiles, e.g. polite smile, are smaller in amplitude, longer in total duration and slower in onset and offset times than the acted smile (Cohn et al., 2004; Ekman et al., 2005; Valstar et al., 2007). It has been found that spontaneous brow actions are different in morphology and temporal structure from acted brow actions (Valstar et al., 2006). In general, acted emotion expressions are more exaggerated than natural ones, and due to these reasons, a system trained on acted emotion expressions may fail to generalize properly to spontaneously occurring emotions. The other issue is that current emotion recognition systems are evaluated on clear noise free data which has high quality audio and frontal face visual data. However, in natural environment the data may be noisy and the face may not be ideally posed with respect to the camera. There is also a problem of emotion categories; in actual human-computer interaction scenarios the emotions are normally non-basic (Cowie et al., 2005), but still most of the existing emotion recognition systems classify expressions from basic emotion categories.

Despite the existence of differences between acted expressions and natural expressions, databases of acted emotions are still useful and have been recorded for the analysis of emotions. The main advantage to this method is that it allows more control over the design of database. A phonetically balanced set of sentences can be recorded in different emotions, which is difficult to achieve in real environment. Since the acted database is normally recorded in a controlled lab environment, this results in high quality noise-free data.

Emotional behavior databases (audio, visual and audio-visual) have been recorded for investigation of emotion, some natural, while others acted or elicited, as shown in Table 1. Many audio emotional databases have been recorded for the analysis of vocal expressions of emotions. The AIBO database (Batliner et al., 2004) is a natural database which consists of recording from children while interacting with robot. The data consist of 110 dialogues and 29200 words. The emotion categories include anger, bored, emphatic, helpless, ironic, joyful, motherese, reprimanding, rest, surprise and touchy. The data labeling is based on listeners' judgment. The Berlin Database of Emotional Speech (Burkhardt et al., 2005) is a German acted database, which consists of recordings from 10 actors (5 male, 5 female). The data consist of 10 German sentences recorded in anger, boredom, disgust, fear, happiness, sadness and neutral. The final database consists of 493 utterances after listeners' judgment. The Danish Emotional Speech Database (Engberg, 1996) is another audio database recorded from 2 actors and 2 actresses. The recorded data consist of 2 words, 9 sentences and 2 passages, resulting in 10 minutes of audio data. The recorded emotions are anger, happiness, sadness, surprise and neutral. The ISL meeting corpus (Burger et al., 2002) is a natural audio database which consists of recordings from 18 meetings with 5 persons, on average, per meeting. There are three emotion categories: negative, positive and neutral. The data are labeled based on listeners' judgment.

Some facial expressions databases have been recorded for the analysis of facial emotional behavior. The BU-3DFE (Yin et al., 2006) is another

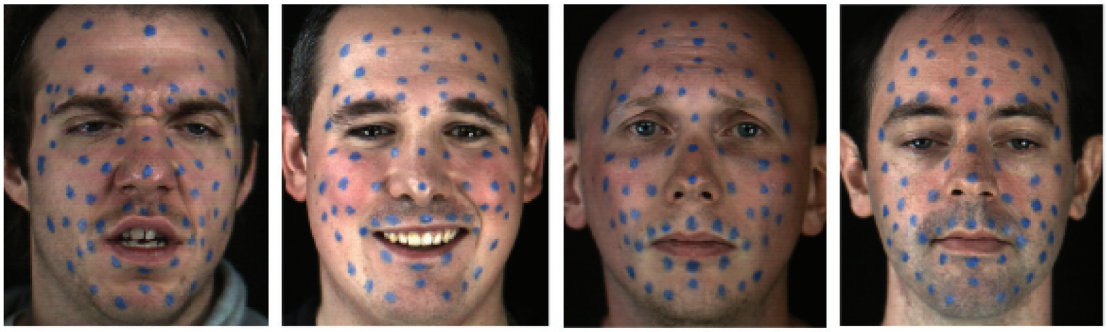
Table 1. Audio and/or Visual Emotional Databases: where A: Audio, V: Visual, AV: Audio-Visual, 6 basic emotions: anger, disgust, fear, happiness, sad, surprise

Database	A/V	Elicitation method	Size	Emotion categories
AIBO database (Batliner et al., 2004)	A	Natural: children interaction with robot	110 dialogues, 29200 words	anger, bored, emphatic, helpless, ironic, joyful, motherese, reprimanding, rest, surprise, touchy
Berlin Database (Burkhardt et al., 2005)	A	Acted	493 sentences; 5 actors & 5 actresses	anger, boredom, disgust, fear, happiness, sadness, neutral
Danish Emotional Speech Database (Engberg, 1996)	A	Acted	10 minutes ; 2 actors & 2 actresses; 2 words, 9 sentences, 2 passages	anger, happiness, sadness, surprise, neutral
ISL meeting corpus (Burger et al., 2002)	A	Natural: meeting corpus	18 meetings; average 5 persons per meeting	negative, positive, neutral
BU-3DFE database (Yin et al., 2006)	V	Acted	100 adults	6 basic emotions with four levels of intensity
Cohn-Kanade database (Kanade et al., 2000)	V	Acted	210 adults; 480 videos	6 basic emotions, Action Units (AUs)
FABO face and body gesture database (Gunes et al., 2006)	V	Acted	23 adults; 210 videos	6 basic emotions, anxiety, boredom, neutral, uncertainty
MMI database (Pantic et al., 2007; Pantic et al., 2005)	V	Acted: static images, and videos in frontal and profile view Natural: Children interacted with a comedian, adults watched emotion inducting videos	Acted: 61 adults Natural: 11 children and 18 adults Total: 1250 videos, 600 static images	6 basic emotions, single Action Unit and multiple Action Units activation
UT Dallas database (O'Toole et al., 2005)	V	Natural: subjects watched emotion inducing videos	229 adults	6 basic emotions, boredom, disbelief, laughter, puzzle
Adult Attachment Interview database (Roisman, 2004)	AV	Natural: subjects were interviewed to describe the childhood experience	60 adults: each interview was 30-60 minutes long	6 basic emotions, contempt, embarrassment, shame, general positive and negative emotions
Belfast database (Douglas-Cowie et al., 2003)	AV	Natural: clips taken from television and realistic interviews with research team	125 subjects; 209 clips from TV and 30 from interviews	Dimensional labeling/categorical labeling
Busso-Narayanan database (Busso et al., 2007)	AV	Acted	612 sentences; an actress	anger, happiness, sadness, neutral
Chen-Huang database (Chen, 2000)	AV	Acted	100 adults; 9900 visual and audio-visual expressions	6 basic emotions, boredom, frustration, interest, puzzle
Haq-Jackson database (Haq & Jackson, 2009)	AV	Acted: emotion stimuli were shown on screen	480 sentences; 4 male subjects	6 basic emotions, neutral
RU-FACS database (Bartlett et al., 2005)	AV	Natural: subjects tried to convince the interviewers about their truth	100 adults	33 Action Units

acted database which consists of 3D range data of 6 basic emotions expressed in four different intensity levels. The data consist of recordings from 100 adults. The Cohn-Kanade facial expression database (Kanade et al., 2000) is a popular acted database of facial expressions, with recordings from 210 adults, in 6 basic emotions and Action Units (AUs). The data is labeled using Facial Action Coding System (FACS). The FABO acted database (Gunes et al., 2006) consists of videos of facial expressions and body gestures from 23 adults in 6 basic emotions along with some non-basic emotions (uncertainty, anxiety, boredom and neutral). The MMI database is a very comprehensive data set of facial behavior (Pantic et al., 2007; Pantic et al., 2005). It consists of facial data for both the acted expressions and spontaneous expressions. The recorded data comprise of both static images and videos, where large parts of the data are recorded in both the frontal and the profile views of the face. For the natural data, children interacted with a comedian, while adults watched emotion-inducing videos. The database consists of 1250 videos and 600 static images in 6 basic emotions, single AU and multiple AUs. The data labeling is done by FACS and observers' judgment. The UT Dallas database (O'Toole et al., 2005) is a natural visual database which is recorded by asking subjects to watch emotion-inducing videos. The database consists of data from 229 adults in 6 basic emotions, along with puzzle, laughter, boredom and disbelief. The data labeling is based on observers' judgment.

Recent work in the field of emotion recognition involves combining the audio and visual modalities to improve the performance of emotion recognition systems. This has resulted in the recording of audio-visual databases, where the facial expressions of the emoting performers are captured simultaneously with speech. The Adult Attachment Interview (AAI) database (Roisman, 2004) is a natural audio-visual database where the subjects are interviewed to describe their childhood experiences. The data consist of recordings from 60 adults and each interview lasts for 30-60 minutes. The database consists of the 6 basic emotions along with embarrassment, contempt, shame, in addition to general kinds of positive and negative emotion. The data labeling is performed by using FACS. The Belfast database (Douglas-Cowie et al., 2003) is another natural audio-visual database which consists of clips taken from television and realistic interviews conducted by a research team. The database consists of data from 125 subjects, which consists of 209 sequences from TV and 30 from interviews. The data are labeled with both dimensional and categorical approaches using Feeltrace system. The Busso-Narayanan acted database (Busso et al., 2007) consists of recordings from an actress, who is asked to read a phoneme-balanced corpus four times, expressing anger, happiness, sadness and neutral state. A detailed description of the actress' facial expression and rigid head motion are acquired by attaching 102 markers to her face. A VICON motion capture system with three cameras is used to capture the 3D position of each marker. The markers' motion and aligned audio is captured simultaneously in a quiet room. The total data consist of 612 sentences. Chen-Huang audio-visual database (Chen, 2000) is one of the largest acted databases, which consists of acted audio and visual expressions in the 6 basic emotions and 4 cognitive states: boredom, interest, frustration and puzzlement. The database consists of recordings from 100 adults with 9900 visual and audio-visual expressions. Haq & Jackson (2009) recorded an audio-visual database from four English male actors in seven emotions in a controlled environment (see Figure 2). The data are recorded in six basic emotions: anger, disgust, fear, happiness, sadness, surprise and in neutral mode. The database consists of 120 utterances per actor, which resulted in 480 sentences in total. To track the visual features, the actors' face is painted with 60 markers. Recordings consist of 15 phonetically-balanced TIMIT database sentences per emotion: 3 common sentences, 2 emotion specific sentences, and 10 generic sentences that are different for each emotion. The Emotion to be simulated and text prompts are displayed on a monitor in front of the actor during the recordings. The 3dMD dynamic face capture system provided 2D frontal color video and Beyer dynamics microphone signals. The data evaluation is performed by 10 subjects, of which 5 are native English speakers and the remaining subjects lived in UK for more than a year. It has been found in some studies that female experience emotion more intensively than male (Swerts et al., 2008), to avoid

Figure 2. Haq & Jackson (2009) audio-visual emotional database (from left): Displeased (anger, disgust), Excited (happy, surprise), Gloomy (sad, fear), and Neutral (neutral), reproduced from Haq & Jackson (2009).



gender biasing half of the evaluators are female. Three types of human evaluation experiments are designed: audio, visual, and audio-visual. Slides are used to show audio, visual and audio-visual clips of each utterance. The data are randomized to remove systematic bias from the responses of human evaluators. For each of the evaluators, a different data set is created by using the Balanced Latin Square method (Edwards, 1962). The portrayed emotion is easier to correctly identify via the visual data alone, compared to the audio data alone, and the overall performance improves by combining the two modalities. The RU-FACS is a natural database (Bartlett et al., 2005) where subjects are tried to convince the interviewers that they are telling the truth. The database consists of data from 100 adults in 33 AUs, and data are labeled by using FACS.

Feature Extraction

It has been found that audio signals follow certain patterns for different kind of emotions. The relationship between audio and emotion is summarized by Cowie et al. (2001). For example anger is characterized by faster speech rate, higher energy and pitch values compared to sadness. The important audio features for emotion recognition are pitch, intensity, duration, spectral energy distribution, formants, Mel Frequency Cepstral Coefficients (MFCCs), jitter and shimmer. These features are identified as important both at utterance level (Luengo et al., 2005; Ververidis et al., 2005; Vidrascu et al., 2005; Borchert et al., 2005; Haq et al., 2008; Haq & Jackson, 2009) and at frame level (Nogueiras et al., 2001; Lin et al., 2005; Kao et al., 2006; Neilberg et al., 2006).

New research on spontaneous emotion analysis suggests the use of only paralinguistic audio features may not be enough for emotion recognition. It is indicated by Batliner et al. (2003) that the reliability of prosody features for affect recognition degraded in real scenarios. In the initial experiments of Devillers et al. (2006) aimed at recognizing anger, fear, relief and sadness in medical call conversations between humans, it was found that lexical cues performed better than paralinguistic cues. Other studies have been performed to investigate using a combination of acoustic features and linguistic features to improve the performance of audio emotion recognition systems. Litman et al. (2004) and Schuller et al. (2005) used spoken words and acoustic features to recognize emotions. Lee et al. (2005) performed emotion recognition by using prosodic features along with spoken words and information of repetition. Graciarena et al. (2006) combined prosodic, lexical and cepstral features to achieve higher performance. Batliner et al. (2003) used prosodic features, part of speech, dialogue act, repetitions, corrections and syntactic-prosodic boundary to detect the emotions. The role of context information (e.g. subject, gender and turn-level features representing local and global aspects of the dialogue) has also been investigated by Litman et al. (2004) and Forbes-Riley et al. (2004). The above studies showed improvement in performance by using information related to language, discourse and context, but the automatic extraction of these features is a difficult task. First, the automatic speech recognition systems are unable to reliably recognize the verbal content of emotional speech (Athanaselis et al., 2005), and Second, the extraction of semantic discourse information is even more difficult. These features are normally extracted manually or directly from transcripts.

Since facial expressions plays an important role to convey and perceive emotions, most of the vision-based emotion recognition methods focus on the analysis of facial expressions. The machine analysis of facial expression can be divided into two main groups: the recognition of emotions and the recognition of facial muscle actions (facial AUs) (Cohn, 2006; Pantic et al., 2007). The facial AUs are descriptions of facial signals which can be mapped to emotion categories by using high level mapping, like EMFACS and FACSAID (Hager, 2003). The current facial expression based emotion recognition systems use different pattern recognition methods and are based on various 2D spatiotemporal facial features. There are mainly two types of facial features which are used for affect recognition: geometric and appearance features. The examples of geometric features are shape of facial components (eyes, mouth, etc.) and the location of facial salient points (corners of eyes, mouth, etc.). The appearance features represent facial texture which includes wrinkles, bulges, and furrows. The examples of methods based on geometric features are those of Chang et al. (2006), who used shape model defined by 58 facial points, of Pantic et al. (2007, 2006, 2004) and Valstar et al. (2007, 2006), who used a set of facial points around the mouth, eyes, eyebrows, nose and chin, and of Kotsia et al. (2007), who used the Candide grid. Other examples are the systems developed by Busso et al. (2004), who used 102 facial markers, and by Haq & Jackson (2009), who used 60 frontal face markers. The examples of appearance-feature-based methods are those of Bartlett et al. (2003, 2005, 2006), Littlewort et al. (2007) and Guo et al. (2005), who used Gabor wavelet, Whitehill et al. (2006), who used Haar features, Anderson et al. (2006), who used a holistic spatial ratio face template, and Valstar et al. (2004), who used temporal templates.

It has been suggested in some studies (Pantic et al., 2006), that using both geometric and appearance features may be the best choice for designing an automatic affect recognizer. The examples of hybrid geometric and appearance based features are those of Tian et al. (2005), who used facial component shapes and the transient components (like crow's feet wrinkles and nasal-labial furrows) and that of Zeng et al. (2005), who used 26 facial points around the eyes, eyebrows, and mouth, and the transient features proposed by Tian et al.(2005). A similar method was proposed by Lucey et al. (2007), who used the Active Appearance Model (AAM) to capture the characteristics of facial appearance and shape of facial expressions. Most of the existing 2D feature based methods are suitable for the analysis of facial expressions under small head motions.

There are few studies of automatic facial affect recognition which are based on 3D face models. Cohn et al. (2007) worked on analysis of brow AUs and head movement based on a cylindrical head model (Xiao et al., 2003). Huang and colleagues (Cohen et al., 2003; Sebe et al., 2004; Wen et al., 2003; Zeng et al., 2007) used feature extracted by a 3D face tracker called the Piecewise Bezier Volume Deformation Tracker (Tao et al., 1999). Chang et al. (2005) and Wang et al. (2006) used 3D expression data for facial expression recognition. The progress of the methodology based on 3D face models may be helpful for view-independent facial expression recognition, which is really important in natural settings due to the unconstrained environment.

Feature Selection

Appropriate feature selection is essential for achieving good performance with both global utterance-level and instantaneous frame-level features. This process helps to remove uninformative, redundant or noisy information. In audio-based emotion recognition, Lin and Wei (2005) reported higher recognition rate for 2 prosodic and 3 voice quality instantaneous level features selected by the Sequential Forward Selection (SFS) method from fundamental frequency (f0), energy, formants, MFCCs and Mel sub-band energies features. Kao & Lee (2006) investigated multilevel features for emotion recognition, and found that frame-level features are better than syllable and word-level features. The best performance is achieved with an ensemble of three feature levels. In phoneme based emotion recognition, it is found that some phonemes, particularly semivowels and vowels, are more important than others (Sethu et al., 2008). Schuller et al. (2003) halved the error rate with 20 global pitch and energy features compared to that of 6 instantaneous pitch and energy features.

For Vision-based emotion recognition, Ashraf et al. (2007) used an AAM to decouple shape and appearance parameters from the digitized facial images, to distinguish between pain and no-pain expressions. Bartlett et al. (2005) used Gabor wavelets features for classification of facial expressions and facial Action Units. The feature selection was performed by PCA and AdaBoost before classification. The performance of both LDA and linear kernel SVM classifier was lower without feature selection. The feature selection by PCA improved the performance of LDA classifier but degraded that of SVM classifier. The use of AdaBoost technique for feature selection improved the performance of both classifiers compared to that of PCA. The AdaBoost feature selection along with SVM classification gave the best results. Gunes et al. (2005) performed visual emotion recognition from face and body. They fused facial expression and body gestures first at feature-level by combining the features from both modalities, and later at decision-level by integrating the outputs of individual systems with suitable criteria. In the feature level fusion, they applied feature selection on combined data with Best-first search method using Weka tool (Witten et al., 2000). The Best-first method can start from an empty set of features and search forward, or start with the full feature set and search backward, or start at any point and search in both directions. The feature-level fusion performed better than decision-level fusion, and the best performance was achieved with 45 features selected out of a pool of 206 features. Valstar et al. (2007) performed experiments to distinguish between posed and spontaneous smiles by fusing head, face and shoulder modalities. They performed fusion at three levels: early, mid-level and late. They used the GentleSVM-Sigmoid classifier for classification, which perform feature selection using GentleBoost and classification using SVM. Whitehill et al. (2006) used Haar features with an AdaBoost classifier to recognize FACS AUs. They compared both the recognition accuracy and processing time of the system with that of Gabor features with SVM classifier. The recognition accuracy of the two systems was comparable, but the AdaBoost classification system was at least 2 orders of magnitude faster than SVM system. They used AdaBoost to select the top 500 Haar features for each AU before classification.

Multi-modal emotion recognition is proposed by Chen et al. (2005). The facial features consisted of 27 features related to eyes, eyebrows, furrows and lips, and the acoustic features consisted of 8 features related to pitch, intensity and spectral energy. The performance of the visual system was better than the audio system, and the overall performance improved for the bimodal system.

Busso et al. (2004) performed emotion recognition using an audio, visual and bimodal system. The audio system used 11 prosodic features selected by the Sequential Backward Selection (SBS) technique, and the visual features were obtained from 102 markers on the face by applying PCA to each of the five parts of face: forehead, eyebrow, low eye, right cheek and left cheek. The visual system performed better than audio system and overall performance improved for the bimodal system. Schuller et al. (2008) reported that emotion recognition in noisy conditions improves with noise and speaker adaptation, and further improvement is achieved with feature selection. The experiments on audio-visual data showed that the performance for both audio and visual features improved with feature selection, and combining the two modalities before feature selection further improved the performance. Haq & Jackson (2009), and Haq et al. (2008) performed feature selection (audio, visual) by Plus *l*-Take Away *r* algorithm (Chen, 1978) based on the Bhattacharyya distance criterion (Campbell, 1997). The algorithm is a combination of SFS and SBS algorithms. The SFS algorithm is a bottom up search method that starts out with an empty feature set, and at each step adds one new feature chosen from a set of candidate features, which performs best in combination with the already chosen features. The problem with the SFS algorithm is that once a feature is added, it cannot be removed. The SBS on the other hand is a top down process. It starts from complete feature set and at each step the worst feature is discarded such that the reduced set gives maximum value of the criterion function. The SBS gives better results but is computationally more complex. Sequential Forward Backward Selection (SFBS) offers benefits of both SFS and SBS, via Plus *l*-Take Away *r* algorithm. At each step, *l* features are added to the current feature set and *r* features are removed. The process continues until the required feature set size is achieved.

Feature Reduction

One of the problems faced by pattern recognition is the dimensionality of data. It is difficult to deal with high dimensional data because it is computationally more expensive. To overcome this problem various techniques have been developed to reduce the dimensionality of data such that most of the useful information is retained. The dimensionality of a feature set can be reduced by using statistical methods to maximize the relevant information preserved. This can be done by applying a linear transformation, Y=WX, where Y is a feature vector in the reduced feature space, X is the original feature vector, and W is the transformation matrix. Principal Component Analysis (PCA) (Shlens, 2005) and Linear Discriminant Analysis (LDA) (Duda et al., 2001) are the examples of such techniques.

Principal Component Analysis

PCA is a simple and non-parametric method to extract useful information from noisy data, and is widely used in statistical analysis of data. PCA is capable of reducing the dimensionality of data to extract the hidden, simple structure of the complex data and remove noise.

The PCA method is described below in detail. Let X be an $m \times n$ matrix, where m is the number of features and n is the number of samples. First, the mean value of each feature is subtracted and each feature is divided by its standard deviation so that each feature variation is contained in the same range, since different types of features have different range of variation. Second, let us define a new mwhere each column of Y has zero mean. It can be shown that

$$Y^TY=C_X \qquad (2)$$

i.e. Y^TY is equal to covariance of X. The principal components of X are the eigenvectors of C_X. After calculating the *SVD* of Y, the columns of matrix V (eigenvector matrix) contain the eigenvectors of $Y^TY=C_X$. Thus the columns of V are the principal components of X. Matrix V rotates the row space of matrix Y, therefore it must rotate matrix X.

The *SVD* decomposition of a matrix M is given by equation,

$$M=U\Sigma V^T \quad (3)$$

Here U and V are orthogonal matrices, where elements of V are the eigenvectors, and U is a set of vectors defined by $\hat{\mathbf{u}}_i \equiv \frac{1}{\sigma_i}\mathbf{X}\hat{\mathbf{v}}_i$. The Σ is a diagonal matrix with rank-ordered set of singular values, $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$. Singular values are positive real and are obtained by taking the square root of eigenvalues of a matrix. Equation (3) states that any arbitrary matrix M can be decomposed into an orthogonal matrix, a diagonal matrix and another orthogonal matrix (or rotation, stretch and another rotation).

The steps for performing PCA can be summarized as follows.

1. Organize the data set as an $m \times n$ matrix, where m is the number of features and n is the number of trials.
2. Subtract off the mean of each feature, or row of matrix X.
3. Calculate the *SVD*.

Linear Discriminant Analysis

LDA is another feature reduction technique, which maximizes the ratio of between-class variance to within-class variance to optimize the separability between classes. The criterion function for the LDA is given by

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|} \quad (4)$$

where

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \quad (5)$$

$$\mathbf{S}_W = \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \quad (6)$$

$$S_T = S_W + S_B \quad (7)$$

$$\mathbf{S}_T = \sum_{x}(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T \quad (8)$$

where S_B is between-class scatter, S_W is within-class scatter, and S_T is total scatter matrix. The m is the total mean vector, m_i is the mean vector for class *i*, and *c* is the total number of classes. The transformation matrix W in equation (4) maximizes the ratio of between-class variance to the within-class variance.

PCA is non-parametric and the answer is unique and independent of any hypothesis about data probability distribution. These two properties are the weakness as well as strength of PCA. Since it is non-parametric, no prior knowledge can be incorporated and also there is loss of information due to PCA compression. The applicability of PCA is limited by the assumptions made in its derivation, which are linearity, statistical importance of mean and covariance, and that larger variances have important information. To resolve the linearity problem of PCA other non-linear methods, e.g. kernel PCA, have been developed. PCA uses a simple criterion for selection of bases, i.e. it chooses bases that maximize the variance of the observed data points, and consider the new dimensions one at a time. An Independent Component Analysis (ICA) is another technique which uses a finer criterion that looks at the relationship between the projections of data into the new dimensions, and optimizes some criterion based on two or more dimensions at once.

LDA is closely related to PCA in that both are linear feature reduction techniques. The difference is that PCA does not take into account any difference in classes, while the LDA explicitly attempts to model the difference between the classes of data. Some other generalizations of LDA for multiple classes have also been defined to address the problem of heteroscedastic distributions, one such method is Heteroscedastic LDA. The other subspace methods include Factor Analysis (FA), Curvilinear Component Analysis (CCA), Principal Manifold, MLP based method, etc.

To classify among facial expressions, Bartlett et al. (2005) used PCA for feature selection which substantially improved the performance of LDA classifier. Petridis & Pantic (2008) used PCA to reconstruct the positions of 20 facial points for the audio-visual based discrimination between laughter and speech. Busso et al. (2004) used audio and visual information for emotion recognition, and they divided the face into five parts: forehead, eyebrow, low eye, right cheek and left cheek. PCA was applied to each part of the face for dimensionality reduction of facial features (3D markers' coordinate). Haq & Jackson (2009) and Haq et al. (2008) used PCA and LDA to reduce the dimensionality of selected audio and visual features for audio-visual emotion recognition.

Classification

The choice of classifier can also significantly affect the recognition accuracy. In the field of emotion recognition various classifiers have been used, among commonly used approaches are Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Neural Network (NN), Support Vector Machine (SVM) and AdaBoost.

Gaussian Mixture Model (Bishop, 1995) models the probability density function of observed variables using a multivariate Gaussian mixture density. Given a series of inputs, it refines the weights of each distribution through expectation-maximization algorithms. The Hidden Markov Model (Young & Woodland, 2009) is a finite set of states, each of which is associated with a probability distribution which is generally multidimensional. The transitions among the states are governed by a set of probabilities known as transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. AdaBoost (Adaptive Boosting) (Freund & Schapire, 1999) is a machine learning algorithm, which is used for pattern recognition and feature selection. AdaBoost is adaptive in the sense that subsequent classifiers built by assigning more weights to those samples which are misclassified by the previous classifiers. AdaBoost calls a weak classifier repeatedly in a series of rounds, where weak classifier is the base learning algorithm that can predict better than a chance. For each call a distribution of weights is updated that indicates the importance of examples in the data set for the classification. On each round, the weights of each incorrectly classified example are increased, so that the new classifier is built with more focus on wrongly classified examples.

A Neural Network (Bishop, 1995) consists of units known as neurons, arranged in layers, which convert an input vector into some output. Neural Network consists of three layers: input, hidden and output. Each unit takes an input, applies a function to it and then passes the output on to the next layer. Generally the networks are feed-forward, where a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and these weightings are tuned in the training phase to adapt NN to a specific problem. A single sweep forward through the network results in the assignment of a value to each output node, and data is assigned to that class' node which has the highest value. Support Vector Machine (Burges, 1998) performs classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories. Consider that the input data are two sets of vectors in an *n*-dimensional space, an SVM will construct a separating hyperplane in that space such that it maximizes the margin between the two data sets. To calculate the margin, two hyperplanes are constructed, one on each side of the separating hyperplane, which are pushed up against the two data sets. A good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes. When the data points are separated by a nonlinear region, it is difficult to separate them by simply constructing an *N*-dimensional hyperplane. SVM handles this by using a kernel function to map the data onto a high dimensional space where it becomes possible for a hyperplane to do the separation. The different kernel functions of SVM are Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid.

Various results have been reported in emotion recognition literature that uses audio, visual and audio-visual information with these different kinds of classifiers, as shown in Table 2. Borchert et al. (2005) reported an accuracy of 76% with SVM classifier, and 75% with AdaBoost classifier for speaker-dependent case, and 70% with each of the two classifiers for speaker-independent case. The classification was performed for 7 emotions using 63 prosody and quality features. Lin and Wei (2005) achieved recognition rates of 100% with 5-state HMM, 89% with SVM, and 85% with KNN (K=21) for the speaker dependent case using 5 best audio features. There were 5 emotion categories and the extracted audio features were prosody, MFCC, Mel frequency sub-band energies. Luengo et al. (2005) reported 92% recognition rate for SVM classifier compared to 87% for GMM classifier with best six pitch and energy related features. A recognition rate of 98% was achieved for the GMM classifier (512 mixtures) with MFCC features for seven emotions. Schuller et al. (2003) achieved 87% accuracy with 4 components GMM for 7 emotions, compared to 78% with 64-state continuous HMM using pitch and energy related features.

With regard to visual classification, Ashraf et al. (2007) used SVM classifier with several representations from AAM. They were able to achieve an equal error rate of 19% using canonical appearance and shape features to classify between pain and no-pain. Bartlett et al. (2005) used SVM, AdaBoost and LDA with Gabor wavelet features to classify between 7 facial expressions. They were able to achieve 90% accuracy with AdaBoost, 88% with SVM (linear kernel), 89% with SVM (RBF kernel) without feature selection. The performance improved by using AdaBoost and PCA as feature selection techniques. The best performance was achieved with SVM classifier, and using AdaBoost for feature selection. The recognition rate increased to 93 for SVM (linear and RBF kernel) and for LDA to 88% with AdaBoost feature selection. Gunes et al. (2005) performed affect recognition from face and body by combining the two types of features at feature-level and at decision-level. They used C4.5 decision tree and BayesNet classifiers for classification and Best-first search method for feature selection using Weka tool (Witten et al., 2000). The feature-level fusion performed better than decision-level fusion, and best performance was achieved with BayesNet classifier using 45 features selected out of total 206 combined features. For eight emotion categories, C4.5 decision tree classifier achieved a best performance of 94% with 206 features, and BayesNet classifier achieved a best performance of 96% with 45 selected features. Valstar et al. (2007) fused head, face and shoulder modalities to distinguish between posed and spontaneous smiles. They used GentleSVM-Sigmoid classifier for classification, which perform feature selection using GentleBoost and classification using SVM. Since the output of SVM is not a good measure for the posterior probability of its prediction, they pass the output of SVM to a sigmoid function that is a reasonable measure for the posterior probability. The features were fused at three levels: early, mid-level and late. In late fusion the head, face and shoulder

Table 2. Emotion classification using audio, visual and audio-visual data: where A: Audio, V: Visual, AV: Audio-Visual, MFCC: Mel Frequency Cepstral coefficient, AAM: Active Appearance Model, SVM: Support Vector Machine, GMM: Gaussian Mixture Model, AdaBoost: Adaptive Boosting, HMM: Hidden Markov Model, KNN: K Nearest Neighbor, LDA: Linear Discriminant Analysis, SD: Speaker-Dependent, SI: Speaker-Independent, GI: Gender-Independent.

Reference	Data	Features	Classifier	Test paradigm	Classes	Accuracy
Borchert et al., 2005	A; Berlin Database (Burkhardt et al., 2005); 493 sentences; 5 male, 5 female	Prosody, quality	SVM	Training: 7 speakers data, testing: 3 speakers data	7	70% (SI)
			AdaBoost			70% (SI)
Lin & Wei, 2005	A; DES Database (Engberg, 1996); 10 min; 2 actors & 2 actresses; 2 words, 9 sentences, 2 passages	Prosody, MFCC, Mel freq. sub-band energies	HMM	4-fold leave-one-out cross-validation	5	100% (GI)
		Mel energy spectrum dynamics coefficients	SVM			89% (GI)
			KNN			85% (GI)
Luengo et al., 2005	A; 97 samples per emotion; 21 number, 21 words, 55 sentences; single actress	Prosody	SVM	5-fold leave-one-out cross-validation	7	92% (SD)
			GMM			87% (SD)
		MFCC	GM			98% (SD)
Schuller et al., 2003	A; 5250 phrases in German and English; 5 speakers	Prosody	GMM	Training: 100 utterances per emotion and speaker, testing: 50 utterances per emotion and speaker	7	87% (SD)
			HMM			78% (SD)
Ashraf et al., 2007	V (face); shoulder pain expressions data from 21 subjects	AAM	SVM	Leave-one-subject-out cross-validation	2	Equal Error Rate: 19% (SI)
Bartlett et al., 2005	V (face); Cohn-Kanade database (Kanade et al., 2000); 210 adults; 480 videos	Gabor wavelets	SVM	Leave-one-subject-out cross-validation	7	93% (SI)
			LDA			88% (SI)
Gunes et al., 2005	V (face and body); 206 instances; 3 subjects	Shape features, optical flow	BayesNet	Training: 156 instances, testing: 50 instances	8	96% (SD) (feature-level fusion)
			C4.5 decision tree			94% (SD) (feature-level fusion)
Valstar et al., 2007	V (face, head and shoulder); MMI database (Pantic et al., 2007); 100 videos of posed smile and 102 videos of spontaneous smile	12 facial points, 5 shoulder points, and 6 degrees of freedom of head motion	Gentle SVM-Sigmoid	10-fold cross-validation	2	94% (decision-level fusion)
						89% (feature-level fusion)
						88% (mid-level fusion)
Whitehill et al., 2006	V (face); Cohn-Kanade database (Kanade et al., 2000); 210 adults; 480 videos	Haar features	AdaBoost	Training: 580 images, testing: on all AUs for which at least 40 training images were present; 10-fold cross-validation	11 AUs	92% (SI)
		Gabor features	SVM			91% (SI) (AdaBoost system was at least 2 times faster than SVM system)

continued on following page

Table 2. continued

Reference	Data	Features	Classifier	Test paradigm	Classes	Accuracy
Busso et al., 2004	AV; 612 phonetically balanced sentences; an actress	Prosody, 102 marker points	SVM	Leave-one-out cross-validation	4	71% (A) (SD) 85% (V) (SD) 89% (AV fused at feature-level and at decision-level) (SD)
Haq & Jackson, 2009	AV; 480 sentences; four male subjects	Prosody, MFCC, 60 facial marker	Gaussian	4-fold leave-one-out cross-validation	7	56% (A) (SD) 95% (V) (SD) 98% (AV fused at decision level) (SD)
					4	69% (A) (SD) 98% (V) (SD) 98% (AV fused at decision level) (SD)
Haq, Jackson & Edge, 2008	AV; 120 sentences; a male subject	Prosody, MFCC, 60 facial marker	Gaussian	6-fold leave-one-out cross-validation	7	53% (A) (SD) 98% (V) (SD) 98% (AV fused at decision level) (SD)
Pal et al., 2006	AV; Infant's cry face and sound data	Fundamental frequency, first two formants, vertical grey level	Rules, k-means	Not available	5	64% (A) 74% (V) 75% (AV fused at decision level)
Schuller et al., 2007	AV; 10.5 hours of spontaneous human-to-human conversation; 11 male and 10 female	Prosody, articulatory, voice quality and linguistic information, AAM, movement activity	SVM	Trainig: 14 subjects, testing: 7 subjects; 3-fold subject independent SCV	3	overall recall: 64% (SI) (audio + activity), (feature-level fusion) 59% (SI) (audio + AAM) 42% (SI) (AAM + activity)
Song et al., 2004	AV; 1384 samples	Prosody, 54 facial animation parameters	Tripled HMM	Training; 700 samples, testing: 684 samples	7	85.0% (model-based fusion)
Wang & Guan, 2005	AV; 500 videos; 8 subjects, 6 different languages	Prosody, MFCC, formants, Gabor wavelets	Fisher's LDA	Training: 360 samples, testing: 140 samples	6	82% (SI) (decision-level fusion)
Zeng et al., 2005a	AV; 660 video sequences; 10 male and 10 female subjects	Prosody, motion units	Multi-stream Fused HMM	Leave-one-subject-out cross-validation	11	81% (SI) (model based fusion)

modalities were combined using three criteria: sum, product and weight. The best recognition accuracy of 94% was achieved with late fusion (product). The recognition rates for the early and the mid-level fusions were 89% and 88%. Whitehill et al. (2006) recognized FACS Action Units by using two systems: first, AdaBoost classifier with Haar features, and second, SVM classifier with Gabor features. The AdaBoost system used AdaBoost to select top 500 Haar features for each

AU before classification. For 11 AUs, an average recognition accuracy of 91% was achieved with SVM classifier using Gabor features, and 92% with AdaBoost classifier using Haar features. The recognition accuracy was comparable for both systems, but AdaBoost classifier was at least 2 times faster than SVM classifier.

Busso et al. (2004) performed emotion classification using both audio and visual features. The audio based system used 11 features selected by SBS. The visual based system used facial marker related features by applying PCA to each of the five parts of face: forehead, eyebrow, low eye, right cheek and left cheek. For the bimodal system, the audio and visual information were fused at two different levels: feature-level and decision-level. The SVM classifier was used for classification of 4 emotion categories. The overall recognition rate of audio system was 71%, and of visual system was 85%. The overall performance for the bimodal system improved to 89% for both of the fusion at feature-level and at decision-level. Haq & Jackson (2009) performed audio-visual emotion recognition using an English database from four male speakers. The audio and visual features were fused at decision level for the audio-visual experiments. They performed speaker dependent experiments using a single mixture Gaussian classifier. For seven emotion classes, average recognition rates of 56%, 95% and 98% were achieved for the audio, visual and audio-visual features compared to 67%, 88% and 92% recognition rates of human. For four emotion categories, average recognition rates of 69%, 98% and 98% were achieved for the audio, visual and audio-visual features compared to 76%, 91% and 95% of humans. Haq, Jackson & Edge (2008) performed audio-visual emotion recognition using single subject audio-visual data. Their recognition system was consisted of four stages: feature extraction, feature selection, feature reduction and classification. A single mixture Gaussian classifier was used for classification. In experiments, audio and visual features were combined at four different stages: feature level, after feature selection, after feature reduction and at decision level. The fusion at decision level and after feature reduction performed better than the fusion at feature level and after feature selection. A maximum recognition rate of 53% was achieved with audio features alone, 98% with visual features alone, and 98% with audio-visual feature fused at decision level. Emotion recognition from infant facial expressions and cries were investigated by Pal et al. (2006). The facial features were related to eyebrow, mouth and eyes positions. The audio features consisted of fundamental frequency and first two formants. For five classes, the overall accuracy of audio, visual and audio-visual systems were 64%, 74% and 75% respectively. The audio-visual experiments were performed with decision level fusion. Schuller et al. (2007) worked on recognition of three levels of interest in a spontaneous conversation by using the audio-visual information. The audio features consisted of prosody, articulatory, voice quality and linguistic information, and visual features consisted of AAM and movement activity detection which was derived from eye positions. The feature selection was performed for each of the audio and visual features before feature-level fusion and SVM was used for classification. The overall recall for combining the audio and activity features was 64%, for the audio and AAM was 59%, and for the AAM and activity features was 42%. Song et al. (2004) reported 85% accuracy for 7 emotions with tripled HMM classifier using both audio and visual features. The facial feature points were tracked with an AAM based instance which were segmented into two groups: expression and visual speech. For a video frame sequence, express vector stream and visual speech vector stream were generated. The audio feature vector stream was extracted based on low level acoustic features. The three streams were feed to HMM system and higher performance was achieved compared to single modality. Wang and Guan (2005) performed classification experiments using an audio-visual database, which consisted of data from 8 speakers in 6 different languages. The visual features consisted of Gabor wavelets, and audio features were prosody, MFCC and formants. A step wise method based on Mahalanobis distance was used for feature selection. The proposed classification scheme was based on analysis of each individual class and combinations of different classes. An overall accuracy of 82% was achieved over a language and race independent data. Zeng et al. (2005a) used Multi-stream Fused HMM (MFHMM) to detect 11 emotions using both audio and visual information. They used composite facial features, speech energy and pitch as three tightly coupled streams. The MFHMM allows building of an optimal connection among multiple streams based on maximum entropy principle and maximum mutual information criterion. An overall accuracy of 81% was achieved with MFHMM which outperformed face-only HMM, pitch-only HMM, energy-only HMM and independent HMM fusion which assume independence among audio and visual streams.

FUSION TECHNIQUES

Audio-visual emotion recognition is based on three types of fusion techniques: feature-level, decision-level and model-level. Feature-level fusion is performed by combining the features of audio and visual modalities into a single feature vector. Examples of methods based on feature-level fusion are those of Zeng et al. (2005b), Busso et al. (2004), Schuller et al. (2007) and Haq et al. (2008). Feature-level fusion may involve feature selection of individual modalities either before or after combining them. Feature-level fusion has the disadvantage of combining the two different kinds of modalities, which have different time scales and metric levels. The other problem with feature-level fusion is high dimensionality of resulting feature vector, which can degrade the performance of emotion recognition system.

In decision-level fusion, the data from audio and visual modalities are treated independently and the single-modal recognition results are combined at decision level. The results from different modalities are combined by using some criterion (e.g. sum, product, and weighted sum or product). Many researchers have combined audio and visual modalities at decision-level (Busso et al., 2004; Wang et al., 2005; Zeng et al., 2007a; Zeng et al., 2007b; Pal et al., 2006; Petridis et al., 2008; Haq et al., 2008; Haq & Jackson, 2009). Decision-level fusion overcomes the problem of different time scales and metric levels of audio and visual data, plus high dimensionality of the concatenated vector resulted in case of feature-level fusion. Decision-level fusion is based on the assumption that audio and visual data are independent, but in reality humans produce audio and visual expressions in a complementary and redundant manner. The assumption of independence results in loss of mutual correlation information between audio and visual modality.

A model-level fusion technique is proposed by some researchers (Fragopanagos et al., 2005; Zeng et al., 2005a; Sebe et al., 2006; Caridakis et al., 2006; Song et al., 2004) to make use of the correlation between audio and visual information with a relaxed synchronization of the two modalities. Song et al. (2004) used a tripled HMM to model the correlation properties of three component HMMs based on one audio and two visual streams. Zeng et al. (2005a) proposed MFHMM for audio-visual affect recognition. The MFHMM builds an optimal connection between different streams based on maximum entropy and maximum mutual information criterion. Caridakis et al. (2006) and Petridis et al. (2008) proposed neural networks to combine the audio and visual modalities for audio-visual emotion recognition. Sebe et al. (2006) proposed Bayesian network topology to recognize emotions from audio and visual modalities. The Bayesian network topology combines the two modalities in a probabilistic manner.

FUTURE RESEARCH DIRECTIONS

The number of efforts that have been put into improve the automatic emotion recognition have resulted some promising achievements in terms of realistic emotional databases recording, audio and visual modalities analysis, feature extraction, feature selection and fusion of two modalities to improve the classification performance. But there are some potential areas that need to be explored for improvement in automatic emotion recognition systems.

Many audio, visual and audio-visual emotional databases have been recorded for the analysis of emotions, but there is no emotional database which can be used as a benchmark. The emotion research community needs to do collective efforts towards recording a larger emotional database that can be used as a benchmark. Most of the recent methods are developed based on high quality lab recorded data, but for realistic natural environment, methods need to be developed which are robust to arbitrary human movement, occlusion, and noisy conditions. The temporal correlation between audio and visual modalities needs to be explored and techniques need improvement to incorporate temporal behavior of each modality, their correlation and contextual information. The development of various audio-visual fusion techniques to improve the performance of affect recognizers is one of important research areas.

CONCLUSION

The field of emotion recognition has come a long way since its modest beginnings. Significant strides have been made in several areas: acquisition of emotion data for research and experimentation, extraction and selection of feature sets, and techniques of classification. The initial studies on emotion recognition were mostly based on small data sets of acted audio or visual expressions, with the classification categories generally restricted to the six basic emotions. Data were not shared among researchers. Studies on multimodal emotion recognition were rare; most of the studies were based on either audio or visual modality, but not both. Recent studies have progressed to recording large emotional databases (audio, visual and audio-visual) of different kinds (acted, natural), and with a greater number and of emotion categories. Moreover, several audio, visual and audio-visual databases are publicly available for the research. Despite the progress that has been made, there are still some issues related to the emotional data acquisition that need to be addressed. The compilation of naturally-occurring databases is quite a difficult task, it is hard to acquire data in the natural environment, and the databases so obtained are normally unbalanced and the quality of the data is not as good. Although it is comparatively easy to record data in a controlled lab environment, the resulting loss of "naturalness" in the data can have some disadvantages. In addition, some emotions, such as happiness, are relatively easy to induce in a laboratory environment, by showing the subjects some example clips chosen to induce the desired emotion, but other emotions, such as fear, sadness and disgust, pose greater difficulty. Another significant problem with natural databases is that of labeling, which becomes quite difficult for the data that lies outside the range of the six basic emotions. While facial expressions can be labeled using FACS Action Units, which are objective descriptors (that can be used for high-level decision-making processes including emotion recognition), there is no similar coding system available to label emotional audio data. The data recorded is influenced by culture and context (stimuli, data recording environment, and the presence of people), and information about these aspects should be recorded as well. It is proposed that the labeling process can be made more reliable if the data is labeled by many subjects, and the subjects are trained before data labeling. A system designed with this kind of data is expected to be more reliable. Another problem, besides the issue of subjectiveness of human-labeled data mentioned, is that it is very time consuming and expensive to manually label the training data. A possible solution to this problem is to use a semi-supervised method, which involves automatic labeling followed by human labeling. The systems developed by Pantic et al. (2007) and Tian et al. (2005) can recognize the AUs in frontal face images, which can be used for automatic data labeling. Although many efforts have been made in compiling emotional databases, there is a need for more collective effort to develop large and comprehensive emotional databases that can be used as a benchmark for the evaluation of emotion recognition techniques. An example of similar kind of database is that of MMI facial expression database (Pantic et al., 2005; Pantic et al., 2007), which provide easy access and search to the facial images.

Various audio and visual features have been identified as being important for emotion recognition. Some of the important audio features for emotion recognition are pitch, intensity, duration, spectral energy distribution, formants, MFCCs, jitter and shimmer. These features are identified as being significant both at utterance level and at frame level. Some studies showed the improvement in performance by using information related to language, discourse and context. However it is difficult to extract these features automatically: it is difficult to recognize the verbal content of emotional speech, and even harder than that is the problem of extracting semantic discourse information. Vision-based emotion recognition is based primarily on facial expressions, as obviously face plays the most important role in conveying emotions. There are two types of facial features - geometric features and appearance features - which are used for affect recognition. Examples of geometric features are shapes of facial components (eyes, mouth, etc.) and the location of salient facial points (corners of eyes, mouth, etc.). The appearance features represent facial texture which includes wrinkles, bulges, and furrows. Some studies suggested that using both geometric and appearance features may be the best choice for designing an automatic affect recognizer. There are other studies which are based on 3D face models, and are capable to incorporate head movement in the direction of camera, which is not possible with 2D techniques. Audio and visual features need to be explored that are robust to noise, occlusion and arbitrary human movement. For view-independent facial expression recognition, which is important in natural environments, developments in 3D face modeling techniques may be helpful.

As is true for most classification problems, the performance of emotion recognition system depends on three factors: feature selection, dimensionality reduction and choice of classifier. Feature selection is used to discard uninformative, redundant or noisy information. The process of feature selection improves both classification performance and computational efficiency. Different methods have been used for feature selection, which include SFS, SBS, SFBS, AdaBoost, GentleBoost, PCA, and Best-first search method. In general it is difficult to deal with high-dimensional data and is computationally expensive. To overcome this problem various techniques have been developed to reduce the dimensionality of data, while at the same time retaining the most useful information. The dimensionality of the feature set is reduced by using statistical methods that minimize redundancy and noise while still retaining relevant information. PCA, Kernel PCA, ICA, LDA and Heteroscedastic LDA are the examples of such techniques. In addition to feature selection and feature reduction, the choice of classifier plays an important role in the performance of affect recognizer. In the field of emotion recognition different kind of classifiers have been used among which GMM, HMM, NN, SVM, and AdaBoost being the most common.

The human emotion recognition is a complex problem, and so far many individual efforts have been made to resolve this issue. This is a multidiscipline's problem and in order to truly understand the human affect behavior, researchers from different disciplines, e.g. psychology, linguistic, engineering, computer science and related fields, need to develop a wider network for collective efforts.

ACKNOWLEDGMENT

We are thankful to Kevin Lithgow, James Edge, Joe Kilner, Darren Cosker, Nataliya Nadtoka, Samia Smail, Idayat Salako, Affan Shaukat and Aftab Khan for help with the data capture, evaluation and as subjects, to James Edge for his marker tracker, to Adrian Hilton for use of his equipment, to Akiel Khan and Tim Sheerman-Chase for corrections, and to University of Peshawar, Pakistan for funding.

REFERENCES

Anderson, K., & McOwan, P. W. (2006). A Real-Time Automated System for Recognition of Human Facial Expressions. *IEEE Trans. Systems, Man, and Cybernetics Part B*, *36*(1), 96–105. doi:10.1109/TSMCB.2005.854502

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K., et al. (2007). The Painful Face: Pain Expression Recognition Using Active Appearance Models. *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces* (pp. 9-14).

Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance. *Neural Networks*, *18*, 437–444. doi:10.1016/j.neunet.2005.03.008

Bartlett, M. S., Littlewort, G., Braathen, P., Sejnowski, T. J., & Movellan, J. R. (2003). A Prototype for Automatic Recognition of Spontaneous Facial Actions. *Advances in Neural Information Processing Systems*, *15*, 1271–1278.

Bartlett, M. S., Littlewort, G., Frank, M., et al. (2005). Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition* (pp. 568-573).

Bartlett, M. S., Littlewort, G., Frank, M. G., et al. (2006). Fully Automatic Facial Action Recognition in Spontaneous Behavior. *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition* (pp. 223-230).

Batliner, A., et al. (2004). You Stupid Tin Box—Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech. *Proc. Fourth Int'l Conf. Language Resources and Evaluation.*

Batliner, A., Fischer, K., Hubera, R., Spilkera, J., & Noth, E. (2003). How to Find Trouble in Communication. *Speech Communication*, *40*, 117–143. doi:10.1016/S0167-6393(02)00079-1

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Borchert, M., & Düsterhöft, A. (2005). Emotions in Speech – Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments. *Proc. IEEE Int'l Conf. on Natural Language Processing and Knowledge Engineering* (pp. 147–151).

Burger, S., MacLaren, V., & Yu, H. (2002). The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. *Proc. Eighth Int'l Conf. Spoken Language Processing.*

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167. doi:10.1023/A:1009715923555

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). *A Database of German Emotional Speech* (pp. 1517–1520). Proc. Interspeech.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004). Analysis of Emotion Recognition Using Facial Expressions. Speech and Multimodal Information. *Proc. ACM Int'l Conf. Multimodal Interfaces (*pp. 205-211).

Busso, C., & Narayanan, S. S. (2007). Interrelation between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study. *IEEE Trans. on Audio, Speech, and Language Processing*, *15*(8), 2331–2347. doi:10.1109/TASL.2007.905145

Campbell, J. (1997). Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, *85*(9), 1437–1462. doi:10.1109/5.628714

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaiou, A., & Karpouzis, K. (2006). Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition. *Proc. ACM Int'l Conf. Multimodal Interfaces,* (pp. 146-154).

Chang, Y., Hu, C., Feris, R., & Turk, M. (2006). Manifold Based Analysis of Facial Expression. *J. Image and Vision Computing*, *24*(6), 605–614. doi:10.1016/j.imavis.2005.08.006

Chang, Y., Vieira, M., Turk, M., & Velho, L. (2005). Automatic 3D Facial Expression Analysis in Videos. *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures* (*V*ol. 3723, pp. 293-307).

Chen, C. (1978). *Pattern Recognition and Signal Processing*. The Netherlands: Sijthoff & Noordoff.

Chen, C., Huang, Y., & Cook, P. (2005). Visual/Acoustic emotion recognition. *Proc. Int'l Conf. on Multimedia & Exp*o (pp. 1468-1471).

Chen, L. S. (2000). *Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction*. (PhD dissertation, Univ. of Illinois, Urbana-Champaign).

Cohen, L., Sebe, N., Garg, A., Chen, L., & Huang, T. (2003). Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. *Computer Vision and Image Understanding*, *91*(1-2), 160–187. doi:10.1016/S1077-3142(03)00081-X

Cohn, J. F. (2006). Foundations of Human Computing: Facial Expression and Emotion. *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces, (*pp. 233-238).

Cohn, J. F., Reed, L. I., Ambadar, Z., Xiao, J., & Moriyama, T. (2004). Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior. *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics* (vol. 1, pp. 610-616).

Cohn, J. F., & Schmidt, K. L. (2004). The Timing of Facial Motion in Posed and Spontaneous Smiles. *International Journal of Wavelets, Multresolution, and Information Processing*, *2*, 1–12. doi:10.1142/S021969130400041X

Cowie, R., Douglas-Cowie, E., et al. (2000). Feeltrace: An Instrument for Recording Perceived Emotion in Real Time. *Proc. ISCA Workshop Speech and Emotion* (pp. 19-24).

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond Emotion Archetypes: Databases for Emotion Modeling Using Neural Networks. *Neural Networks*, *18*, 371–388. doi:10.1016/j.neunet.2005.03.002

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 32–80. doi:10.1109/79.911197

Devillers, L. & Vidrascu, L. (2006). Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs. *Proc. Interspeech,* 801-804.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional Speech: Towards a New Generation of Database. *Speech Communication*, *40*(1/2), 33–60. doi:10.1016/S0167-6393(02)00070-5

Duda, R., Hart, P., & Stork, D. (2001). *Pattern Classification*. New York: John Wiley & Sons, Inc.

Edwards, A. L. (1962). *Experimental Design in Psychological Research*. New York: Holt, Rinehart and Winston.

Ekman, P. (1971). Universal and Cultural Differences in Facial Expressions of Emotion. *Proc. Nebraska Symp. Motivation*, 207-283.

Ekman, P. (1994). Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique. *Psychological Bulletin*, *115*(2), 268–287. doi:10.1037/0033-2909.115.2.268

Ekman, P., & Rosenberg, E. L. (2005). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*,(Eds.). Oxford Univ. Press.

Engberg, I. S., & Hansen, A. V. (1996). *Documentation of the Danish Emotional Speech Database (DES)*. Denmark: Aalborg University.

Forbes-Riley, K., & Litman, D. (2004). Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *Proc. Human Language Technology Conf. North Am. Chapter of the Assoc. Computational Linguistics (HLT/NAACL).*

Fragopanagos, F., & Taylor, J. G. (2005). Emotion Recognition in Human-Computer Interaction. *Neural Networks*, *18*, 389–405. doi:10.1016/j.neunet.2005.03.006

Freund, Y., & Schapire, R. E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*(5), 771–780.

Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., & Kajarekar, S. (2006). Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection. *Proc. Int'l Conf. Acoustics, Speech and Signal Processing, 1*, 1033-1036.

Gunes, H., & Piccardi, M. (2005). Affect Recognition from Face and Body: Early Fusion versus Late Fusion. *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics* (pp. 3437-3443).

Gunes, H., & Piccardi, M. (2006). A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior. *Proc. 18th Int'l Conf. Pattern Recognition*, *1*, 1148–1153.

Guo, G., & Dyer, C. R. (2005). Learning from Examples in the Small Sample Case: Face Expression Recognition. *IEEE Trans. Systems, Man, and Cybernetics Part B*, *35*(3), 477–488. doi:10.1109/TSMCB.2005.846658

Hager, J. C. (2003). *Date Face.* Retrieved from http://face-and-emotion.com/dataface/general/homepage.jsp

Haq, S., & Jackson, P. J. B. (2009). *Speaker-Dependent Audio-Visual Emotion Recognition*. Proc. Auditory-Visual Speech Processing.

Haq, S., Jackson, P. J. B., & Edge, J. (2008). *Audio-visual feature selection and reduction for emotion Classification* (pp. 185–190). Proc. Auditory-Visual Speech Processing.

Kanade, T., Cohn, J., & Tian, Y. (2000). Comprehensive Database for Facial Expression Analysis, *Proc. IEEE Int'l Conf. Face and Gesture Recognition* (pp. 46-53).

Kao, Y., & Lee, L. (2006). *Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language* (pp. 1814–1817). Proc. Interspeech.

Kotsia, I., & Pitas, I. (2007). Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Transactions on Image Processing*, *16*(1), 172–187. doi:10.1109/TIP.2006.884954

Lee, C. M., & Narayanan, S. S. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293–303. doi:10.1109/TSA.2004.838534

Lin, Y., & Wei, G. (2005). Speech Emotion Recognition Based on HMM and SVM. *Proc. 4th Int'l Conf. on Mach. Learn. and Cybernetics* (pp.4898-4901).

Litman, D. J., & Forbes-Riley, K. (2004). Predicting Student Emotions in Computer-Human Tutoring Dialogues. *Proc. 42nd Ann. Meeting of the Assoc. Computational Linguistics.*

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2007). Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain, *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces,* pp. 15-21.

Lucey, S., Ashraf, A. B., & Cohn, J. F. (2007). Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. In Delac, K., & Grgic, M. (Eds.), *Face Recognition* (pp. 275–286). New York: I-Tech Education and Publishing.

Luengo, I., & Navas, E. (2005). *Automatic Emotion Recognition using Prosodic Parameters* (pp. 493–496). Proc. Interspeech.

Neilberg, D., & Elenius, K. (2006). *Emotion Recognition in Spontaneous Speech Using GMMs* (pp. 809–812). Proc. Interspeech.

Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). *Speech Emotion Recognition Using Hidden Markov Models* (pp. 2679–2682). Proc. Eurospeech.

O'Toole, A. J. (2005). A Video Database of Moving Faces and People. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 812–816. doi:10.1109/TPAMI.2005.90

Ortony, A., & Turner, T. J. (1990). What's Basic About Basic Emotions? *Psychological Review*, *97*(3), 315–331. doi:10.1037/0033-295X.97.3.315

Pal, P., Iyer, A. N., & Yantorno, R. E. (2006). Emotion Detection from Infant Facial Expressions and Cries, *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing, 2*, 721-724.

Pantic, M., & Bartlett, M. S. (2007). Machine Analysis of Facial Expressions. In Delac, K., & Grgic, M. (Eds.), *Face Recognition* (pp. 377–416). New York: I-Tech Education and Publishing.

Pantic, M., & Patras, I. (2006). Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form Face Profile Image Sequences. *IEEE Trans. Systems, Man, and Cybernetics Part B*, *36*(2), 433–449. doi:10.1109/TSMCB.2005.859075

Pantic, M., & Rothkrantz, L. J. M. (2004). Facial Action Recognition for Facial Expression Analysis from Static Face Images. *IEEE Trans. Systems, Man, and Cybernetics Part B*, *34*(3), 1449–1461. doi:10.1109/TSMCB.2004.825931

Pantic, M., Valstar, M. F., Rademaker, R., & Maat, L. (2005). Web-Based Database for Facial Expression Analysis. *Proc. 13th ACM Int'l Conf. Multimedia,* (pp. 317-321).

Petridis, S., & Pantic, M. (2008). Audiovisual Discrimination between Laughter and Speech. *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (pp. 5117-5120).

Roisman, G. I., Tsai, J. L., & Chiang, K. S. (2004). The Emotional Integration of Childhood Experience: Physiological, Facial Expressive, and Self-Reported Emotional Response during the Adult Attachment Interview. *Developmental Psychology*, *40*(5), 776–789. doi:10.1037/0012-1649.40.5.776

Russell, J., Ward, L., & Pratt, G. (1981). Affective Quality Attributed to Environments: A Factor Analytic Study. *Environment and Behavior*, *13*(3), 259–288. doi:10.1177/0013916581133001

Scherer, K. (2005). What are emotions? and how can they be measured? *Social Sciences Information. Information Sur les Sciences Sociales*, *44*(4), 695–729. doi:10.1177/0539018405058216

Schuller, B., Muller, R., Hornler, B., Hothker, A., Konosu, H., & Rigoll, G. (2007). Audiovisual Recognition of Spontaneous Interest within Conversations *Proc. ACM Int'l Conf. Multimodal Interfaces (* pp. 30-37).

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov Model-Based Speech Emotion Recognition. *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, 2*, 1-4.

Schuller, B., Villar, R. J., Rigoll, G., & Lang, M. (2005). Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition. *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* (pp. 325-328).

Schuller, B., & Wimmer, M. (2008). *Detection of security related affect and behavior in passenger Transport* (pp. 265–268). Proc. Interspeech.

Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2006). Emotion Recognition Based on Joint Visual and Audio Cues. *Proc. Int'l Conf. Pattern Recognition* (pp. 1136-1139).

Sebe, N., Lew, M. S., Cohen, I., Sun, Y., Gevers, T., & Huang, T. S. (2004). Authentic Facial Expression Analysis. *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition.*

Sethu, V., Ambikairajah, E., & Epps, J. (2008). *Phonetic and speaker variations in automatic emotion Classification* (pp. 617–620). Proc. Interspeech.

Shlens, J. (2005). *A Tutorial on Principal Component Analysis*. Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla.

Song, M., Bu, J., Chen, C., & Li, N. (2004). Audio-Visual-Based Emotion Recognition: A New Approach. *Proc. Int'l Conf. Computer Vision and Pattern Recognition,* (pp. 1020-1025).

Swerts, M., & Krahmer, E. (2008). *Gender-related differences in the production and perception of Emotion* (pp. 334–337). Proc. Interspeech.

Tao, H., & Huang, T. S. (1999). Explanation-Based Facial Motion Tracking Using a Piecewise Bezier Volume Deformation Mode. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 1*, 611-617.

Tian, Y. L., Kanade, T., & Cohn, J. F. (2005). Facial Expression Analysis. S.Z. Li and A.K. Jain (Eds.). *Handbook of Face Recognition* (pp. 247-276). New York: Springer.

Valstar, M., Pantic, M., Ambadar, Z., & Cohn, J. F. (2006). Spontaneous versus Posed Facial Behavior: Automatic Analysis of Brow Actions. *Proc. Int'l Conf. Multimodal Interfaces* (pp.162-170).

Valstar, M., Pantic, M., & Patras, I. (2004). Motion History for Facial Action Detection from Face Video. *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics, 1*, 635-640.

Valstar, M. F., Gunes, H., & Pantic, M. (2007). How to Distinguish Posed from Spontaneous Smiles Using Geometric Features. *Proc. ACM Int'l Conf. Multimodal Interfaces* (pp. 38-45).

Ververidis, D., & Kotropoulos, C. (2005). Emotional speech classification using Gaussian mixture Models. *Proc. ISCAS* (pp. 2871-2874).

Vidrascu, L., & Devillers, L. (2005). *Detection of real-life emotions in call centers* (pp. 1841–1844). Proc. Interspeech.

Wang, J., Yin, L., Wei, X., & Sun, Y. (2006). 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2*, 1399-1406.

Wang, Y., & Guan, L. (2005). Recognizing Human Emotion from Audiovisual Information. *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing* (pp. 1125-1128).

Wen, Z., & Huang, T. S. (2003). Capturing Subtle Facial Motions in 3D Face Tracking. *Proc. Ninth IEEE Int'l Conf. Computer Vision* (pp. 1343-1350).

Whissell, C. M. (1989). The Dictionary of Affect in Language, Emotion: Theory, Research and Experience. In Plutchik, R., & Kellerman, H. (Eds.), *The Measurement of Emotions* (*Vol. 4*, pp. 113–13). Academic Press.

Whitehill, J., & Omlin, C. W. (2006). Haar Features for FACS AU Recognition, *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition* (pp. 217-222).

Witten, I. H., & Frank, E. (2000). *Data Mining: Practical machine learning tools with java implementations*. San Francisco: Morgan Kaufmann.

Xiao, J., Moriyama, T., Kanade, T., & Cohn, J. F. (2003). Robust Full-Motion Recovery of Head by Dynamic Templates and Re-Registration Techniques. *International Journal of Imaging Systems and Technology*, *13*(1), 85–94. doi:10.1002/ima.10048

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3D Facial Expression Database for Facial Behavior Research. *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*(pp.211-216).

Young, S., & Woodland, P. (2009). *Hidden Markov Model Toolkit*. Cambridge University Engineering Department (CUED), UK. Online: http://htk.eng.cam.ac.uk/.

Zeng, Z., Fu, Y., Roisman, G. I., Wen, Z., Hu, Y., & Huang, T. S. (2006). Spontaneous Emotional Facial Expression Detection. *J. Multimedia*, *1*(5), 1–8.

Zeng, Z., Hu, Y., Roisman, G. I., Wen, Z., Fu, Y., & Huang, T. S. (2007a). Audio-Visual Spontaneous Emotion Recognition. In Huang, T. S., Nijholt, A., Pantic, M., & Pentland, A. (Eds.), *Artificial Intelligence for Human Computing* (pp. 72–90). New York: Springer. doi:10.1007/978-3-540-72348-6_4

Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth, D., & Levinson, S. (2007b). Audio-Visual Affect Recognition. *IEEE Transactions on Multimedia*, *9*(2), 424–428. doi:10.1109/TMM.2006.886310

Zeng, Z., Tu, J., Pianfetti, P., Liu, M., Zhang, T., Zhang, Z., et al. (2005a). Audio-Visual Affect Recognition through Multi-Stream Fused HMM for HCI, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition* (pp. 967-972).

Zeng, Z., Zhang, Z., Pianfetti, B., Tu, J., & Huang, T. S. (2005b). Audio-Visual Affect Recognition in Activation-Evaluation Space. *Proc. 13th ACM Int'l Conf. Multimedia* (pp. 828-831).

Zhang, Y., & Ji, Q. (2005). Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 699–714. doi:10.1109/TPAMI.2005.93

ADDITIONAL READING

Since this chapter provides a limited understanding of the human affect recognition, the following list of additional references has been included. Readers are encouraged to refer to these sources which are really important for their personal understanding of the human affect recognition.

Fasel, B., & Luttin, J. (2003). Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, *36*(1), 259–275. doi:10.1016/S0031-3203(02)00052-3

Murrey, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097–1108. doi:10.1121/1.405558

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *Int'l J. Human-Computer Studies*, *59*, 157–183. doi:10.1016/S1071-5819(02)00141-6

Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390. doi:10.1109/JPROC.2003.817122

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58. doi:10.1109/TPAMI.2008.52