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,
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
Related Content

The Role of Textual Graph Patterns in Discovering Event Causality
Bryan Rink, Cosmin Adrian Bejan and Sanda Harabagiu (2012). *Applied Natural Language Processing: Identification, Investigation and Resolution* (pp. 334-350).
[www.igi-global.com/chapter/role-textual-graph-patterns-discovering/61057?camid=4v1a](www.igi-global.com/chapter/role-textual-graph-patterns-discovering/61057?camid=4v1a)

Corpus Linguistics: An Exploration of the Possibility of Improving ELS Learning and Teaching in the Zimbabwean High School
[www.igi-global.com/chapter/corpus-linguistics/108767?camid=4v1a](www.igi-global.com/chapter/corpus-linguistics/108767?camid=4v1a)

Attitudes toward Computer Synthesized Speech
[www.igi-global.com/chapter/attitudes-toward-computer-synthesized-speech/40867?camid=4v1a](www.igi-global.com/chapter/attitudes-toward-computer-synthesized-speech/40867?camid=4v1a)

Lip Contour Extraction from Video Sequences under Natural Lighting Conditions
[www.igi-global.com/chapter/lip-contour-extraction-video-sequences/31068?camid=4v1a](www.igi-global.com/chapter/lip-contour-extraction-video-sequences/31068?camid=4v1a)