

Multimodal Approaches for Emotion Recognition: A Survey

Nicu Sebe^a, Ira Cohen^b, Theo Gevers^a, and Thomas S. Huang^c

^a Faculty of Science, University of Amsterdam, The Netherlands;

^bHP Labs, USA;

^cBeckman Institute, University of Illinois at Urbana-Champaign, USA

ABSTRACT

Recent technological advances have enabled human users to interact with computers in ways previously unimaginable. Beyond the confines of the keyboard and mouse, new modalities for human-computer interaction such as voice, gesture, and force-feedback are emerging. Despite important advances, one necessary ingredient for natural interaction is still missing—emotions. Emotions play an important role in human-to-human communication and interaction, allowing people to express themselves beyond the verbal domain. The ability to understand human emotions is desirable for the computer in several applications. This paper explores new ways of human-computer interaction that enable the computer to be more aware of the user’s emotional and attentional expressions. We present the basic research in the field and the recent advances into the emotion recognition from facial, voice, and physiological signals, where the different modalities are treated independently. We then describe the challenging problem of multimodal emotion recognition and we advocate the use of probabilistic graphical models when fusing the different modalities. We also discuss the difficult issues of obtaining reliable affective data, obtaining ground truth for emotion recognition, and the use of unlabeled data.

Keywords: emotion recognition, multimodal approach, human-computer interaction

1. INTRODUCTION

Maybe no movie of modern time has explored the definition of what it means to be human better than *Blade Runner*. The Tyrell Corporation’s motto, “More human than human”, serves as the basis for exploring the human experience through true humans and created humans, or Replicants. Replicants are androids that were built to look like humans and to work or fight their wars. In time, they began to acquire emotions (so much like humans) and it became difficult to tell them apart. With emotions, they began to feel oppressed and many of them became dangerous and committed acts of extreme violence to be free. Fortunately, Dr. Eldon Tyrell, the creator of the Replicants, installed a built-in safety feature in these models: a four-year life span.

It is evident from the above story that it is not sufficient for a machine (computer) to look like a human (e.g., have skin, face and facial features, limbs, etc). Something else is also essential: the ability to acquire and show the emotions. Moreover, the machine should learn to recognize faces and to understand the emotions to be able to have a human-like interaction with its human counterpart. Machines may never need all of the emotional skills that people need but they will inevitably require some of these skills to appear intelligent when interacting with people. It is argued that to truly achieve effective human-computer intelligent interaction (HCII), there is a need for the computer to be able to interact naturally with the user, similar to the way human-human interaction takes place. For example, if a machine talks to you but never listens to you, then it is likely to be annoying, analogous to the situation where a human talks to you but never listens. Reeves and Nass¹ have conducted several experiments of classical human-human interaction, taking out one of the humans and putting in a computer. Their conclusion is that for an intelligent interaction, the basic human-human issues should hold.

Humans interact with each other mainly through speech, but also through body gestures, to emphasize a certain part of the speech and display of emotions. As a consequence, the new interface technologies are steadily driving toward accommodating information exchanges via the natural sensory modes of sight, sound, and touch. In face-to-face exchange, humans employ these communication paths simultaneously and in combination, using one to complement and enhance another. The exchanged information is largely encapsulated in this natural, multimodal format. Typically, conversational interaction bears a central burden in human communication, with vision, gaze, expression, and manual gesture often contributing critically, as well as frequently embellishing

attributes such as emotion, mood, attitude, and attentiveness. But the roles of multiple modalities and their interplay remain to be quantified and scientifically understood. What is needed is a science of human-computer communication that establishes a framework for multimodal “language” and “dialog”, much like the framework we have evolved for spoken exchange.

In some applications, it may not be necessary for computers to recognize emotions. For example, the computer inside an automatic teller machine or an airplane probably does not need to recognize emotions. However, in applications where computers take on a social role such as an “instructor,” “helper,” or even “companion,” it may enhance their functionality to be able to recognize users’ emotions. In her recent book, Picard² suggested several applications where it is beneficial for computers to recognize human emotions. For example, knowing the user’s emotions, the computer can become a more effective tutor. Synthetic speech with emotions in the voice would sound more pleasing than a monotonous voice. Computer “agents” could learn the user’s preferences through the users’ emotions. Another application is to help the human users monitor their stress level. In clinical settings, recognizing a person’s inability to expression certain facial expressions may help diagnose early psychological disorders.

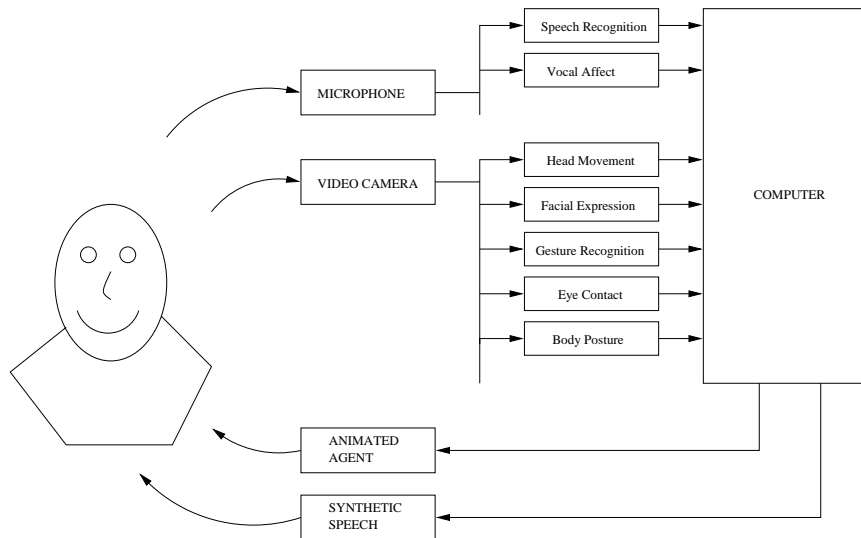


Figure 1. Multimodal human-computer interaction.

Psychologists and engineers alike have tried to analyze facial expressions, vocal emotions, gestures, and physiological signals in an attempt to understand and categorize emotions. This knowledge can be used to teach computers to recognize human emotions from video images acquired from built-in cameras, and from speech waveforms gathered from on-board microphones. A natural two-way interaction between the human and the computer through multiple modalities is depicted in Figure 1. In this diagram, one of the inputs to the computer is vision (video), from which gaze, posture, gestures, and facial and lip movements can be extracted. Computers may learn to recognize gestures, postures, facial expressions, eye contact, etc. Likewise, speech and voice (audio) through the microphone may convey linguistic as well as paralinguistic information. On the output side, the computer may appear in the form of an “agent”—a computer-animated face or a personified animated character. This “agent” can speak to the human through a synthesized speech and display corresponding facial and mouth movements on the screen. Even if they are not explicitly presented in the figure, some other modalities such as tactile or physiological signals can also be used in conjunction with the video and audio signals.

The goal of this paper is to explore new ways of human-computer interaction by enabling the computer to be more aware of the human user’s emotional and attentional expressions. In particular, we concentrate on the problem of integrating audiovisual inputs for the detection the users’ facial and vocal emotional expressions and attentive states. By “emotional expression” we mean any outward expression that arises as a response to some stimulus event. These may include typical expressions such as a “smile” to show that one is happy, or to show one likes what one sees.

Throughout the paper we explore and try to provide answers to the following questions:

- What clues are there on the face and in the voice that reveal a person’s emotions, preferences, and attentional states?
- How well can we use these clues to train the computer to recognize human emotion from audio and from video?
- Does the use of joint audiovisual input allow for more accurate or efficient emotion recognition than using a single modality?
- In realistic scenarios, can the two modalities be treated separately?
- How to collect multimodal data with emotional expressions and how should it be labeled?
- Can we get away with labeling only small amounts of data and use unlabeled data to help in training the models that recognize emotional expressions?
- What data should be collected? Spontaneous or posed data?

2. EMOTIONAL EXPRESSION RECOGNITION FOR HUMAN-COMPUTER INTERACTION

The mounting evidence of the importance of emotions in human-human interaction provided the basis for researchers in the engineering and computer science communities to develop automatic ways for computers to recognize emotional expression, as a goal towards achieving human-computer intelligent interaction. The labeling of emotions into different states led most research to use pattern recognition approaches for recognizing emotions, using different modalities as inputs to the emotion recognition models. Next we review some of these works.

2.1. Facial Expression Recognition Studies

Since the early 1970s, Paul Ekman and his colleagues have performed extensive studies of human facial expressions.³ They found evidence to support universality in facial expressions. These “universal facial expressions” are those representing happiness, sadness, anger, fear, surprise, and disgust. They studied facial expressions in different cultures, including preliterate cultures, and found much commonality in the expression and recognition of emotions on the face. However, they observed differences in expressions as well, and proposed that facial expressions are governed by “display rules” in different social contexts. For example, Japanese subjects and American subjects showed similar facial expressions while viewing the same stimulus film. However, in the presence of authorities, the Japanese viewers were more reluctant to show their real expressions.

Ekman and Friesen⁴ developed the Facial Action Coding System (FACS) to code facial expressions where movements on the face are described by a set of action units (AUs). Each AU has some related muscular basis. Each facial expression may be described by a combination of AUs. This system of coding facial expressions is done manually by following a set prescribed rules. The inputs are still images of facial expressions, often at the peak of the expression. This process is very time-consuming.

Ekman’s work inspired many researchers to analyze facial expressions by means of image and video processing. By tracking facial features and measuring the amount of facial movement, they attempt to categorize different facial expressions. Recent work on facial expression analysis and recognition^{5–14} has used these “basic expressions” or a subset of them. The recent surveys in the area^{15–17} provide an in-depth review of many of the research done in automatic facial expression recognition in recent years.

Recent work in computer-assisted quantification of facial expressions did not start until the 1990s. Mase⁵ used optical flow (OF) to recognize facial expressions. He was one of the first to use image processing techniques to recognize facial expressions. Lanitis et al.⁷ used a flexible shape and appearance model for image coding, person identification, pose recovery, gender recognition and facial expression recognition. Black and Yacoob⁸ used local parameterized models of image motion to recover non-rigid motion. Once recovered, these parameters are fed

to a rule-based classifier to recognize the six basic facial expressions. Yacoob and Davis¹⁸ computed optical flow and used similar rules to classify the six facial expressions. Rosenblum et al.⁹ also computed optical flow of regions on the face, then applied a radial basis function network to classify expressions. Essa and Pentland¹⁰ also used an optical flow region-based method to recognize expressions. Otsuka and Ohya¹¹ first computed optical flow, then computed their 2D Fourier transform coefficients, which were then used as feature vectors for a hidden Markov model (HMM) to classify expressions. The trained system was able to recognize one of the six expressions near realtime (about 10 Hz). Chen¹³ used a suite of static classifiers to recognize facial expressions, reporting on both person-dependent and person-independent results. Cohen et al.¹² describe classification schemes for facial expression recognition in two types of settings: dynamic and static classification. In the static setting, the authors learn the structure of Bayesian networks classifiers using as input 12 motion units given by a face tracking system for each frame in a video. For the dynamic setting, they used a multi-level HMM classifier that combines the temporal information and allows not only to perform the classification of a video segment to the corresponding facial expression, as in the previous works on HMM based classifiers, but also to automatically segment an arbitrary long sequence to the different expression segments without resorting to heuristic methods of segmentation.

These methods are similar in the general sense that they first extract some features from the images, then these features are fed into a classification system, and the outcome is one of the preselected emotion categories. They differ mainly in the features extracted from the video images or the processing of video images to classify emotions. The video processing falls into two broad categories. The first is “feature-based,” where one tries to detect and track specific features such as the corners of the mouth, eyebrows, etc.; the other approach is “region-based” in which facial motions are measured in certain regions on the face such as the eye/eyebrow and mouth regions. People have used different classification algorithms to categorize these emotions. In Table 1, we compare several facial expression recognition algorithms. In general, these algorithms perform well compared to trained human recognition of about 87% as reported by Bassili.¹⁹

Table 1. Comparisons of facial expression recognition algorithms.

Author	Processing	Classification	Number of Categories	Number of Subjects	Performance
Mase	optical flow	kNN	4	1	86%
Black & Yacoob	parametric model	rule-based	6	40	92%
Yacoob & Davis	optical flow	rule-based	6	32	95%
Rosenblum et al.	optical flow	neural networks	2	32	88%
Essa & Pentland	optical flow	distance-based	5	8	98%
Otsuka & Ohya	2D FT of optical flow	HMM	6	4	93%
Lanitis et al.	appearance model	distance-based	7	-	74%
Chen	appearance model	Winnow	6	5	86%
Cohen et al.	appearance model	Bayesian networks	7	5+53	83%

Another interesting thing to point out is the problem of the commonly confused categories in the six basic expressions. As reported by Ekman, *anger* and *disgust* are commonly confused in judgment studies. Also, *fear* and *surprise* are commonly confused. The reason why these confusions occur is because they share many similar facial actions.⁴ *Surprise* is sometimes mistaken for *interest*, but not the other way around. In the computer recognition studies, some of these confusions are observed.^{8, 12, 18}

2.2. Vocal Emotion Recognition Studies

The vocal aspect of a communicative message carries various kinds of information. If we disregard the manner in which the message was spoken and consider the verbal part (e.g., words) only, we might miss the important aspects of the pertinent utterance and we might even completely misunderstand what was the meaning of the message. Nevertheless, in contrast to spoken language processing, which has recently witnessed significant advances, the processing of emotional speech has not been widely explored.

Starting in the 1930s, quantitative studies of vocal emotions have had a longer history than quantitative studies of facial expressions. Traditional as well as most recent studies in emotional contents in speech^{20–23} have used “prosodic” information which includes the pitch, duration, and intensity of the utterance.²⁴ Williams and Stevens²⁵ studied the spectrograms of real emotional speech and compared them with acted speech. They found similarities which suggest the use of acted data. Murray and Arnott²⁰ reviewed findings on human vocal emotions. They also constructed a synthesis-by-rule system to incorporate emotions in synthetic speech.²⁶ To date, most works have concentrated on the analysis of human vocal emotions. Some studied human abilities to recognize vocal emotions. These findings are very useful for the present work.

There has been less work on recognizing human vocal emotions by computers than there has been on recognizing facial expressions by machine. Chiu et al.²¹ extracted five features from speech and used a multilayered neural network for the classification. For 20 test sentences, they were able to correctly label all three categories. Dellaert et al.²² used 17 features and compared different classification algorithms and feature selection methods. They achieved 79.5% accuracy with 4 categories and 5 speakers speaking 50 short sentences per category. Petrushin²⁷ compared human and machine recognition of emotions in speech and achieved similar rates for both (around 65%). In that work, 30 subjects spoke 4 sentences, with each sentence repeated 5 times, once for each emotion category. Scherer²³ performed a large-scale study using 14 professional actors. In this study, he extracted as many as 29 features from the speech. According to Scherer, human ability to recognize emotions from purely vocal stimuli is about 60%. He pointed out that “sadness and anger are best recognized, followed by fear and joy. Disgust is the worst.”

Chen¹³ proposed a rule-based method for classification of input audio data into one of the following emotions categories: happiness, sadness, fear, anger, surprise, and dislike. The input data contained 2 speakers, one speaking Spanish and the other one Sinhala. The choice of these languages was such that the subjective judgments were not influenced by the linguistic content as the observers did not comprehend either language. Each speaker was asked to speak 6 different sentences for each emotion and the contents of the sentences were related in most of the cases to one category and some of them could be applied to two different categories. From the audio signals pitch, intensity, and pitch contours were estimated as acoustic features which were then classified using some predefined rules.

Recent studies seem to use the “Ekman six” basic emotions, although others in the past have used many more categories. The reasons for using these basic six categories are often not justified. It is not clear whether there exist “universal” emotional characteristics in the voice for these six categories. Table 2 shows a summary of human vocal affects as reported by Murray and Arnott.²⁰ This table describes mostly *qualitative* characteristics associated with these emotions. These are listed in relation to the neutral voice.

Table 2. Summary of human vocal affects described relative to neutral speech.

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	slightly faster	faster or slower	slightly slower	much faster	very much slower
Pitch Average	very much higher	much higher	slightly lower	very much higher	very much lower
Pitch Range	much wider	much wider	slightly narrower	much wider	slightly wider
Intensity	higher	higher	lower	normal	lower
Voice Quality	breathy	blaring	resonant	irregular	grumbled

2.3. Emotion Recognition from Physiological Signals

Emotion consists of more than outward physical expression; it also consists of internal feelings and thoughts, as well as other internal processes of which the person having the emotion may not be aware. Still, these physiological processes can be naturally recognized by people. A stranger shaking your hand can feel its clamminess (related to skin conductivity); a friend leaning next to you may sense your heart pounding, etc.

Physiological pattern recognition of emotion has important applications in medicine, entertainment, and human-computer interaction.²⁸ Physiological pattern recognition can potentially help in assessing and quantifying stress, anger, and other emotions that influence health. Affective states of depression, anxiety, and chronic anger have been shown to impede the work of the immune system, making people more vulnerable to infections, and slowing healing from surgery or disease. Changes in physiological signals can also be examined for signs of stress arising while users interact with the technology, helping detect where the product causes unnecessary irritation or frustration. This information may help developers to redesign and improve their technology.

One of the big questions in emotion theory is whether distinct physiological patterns accompany each emotion.²⁹ The physiological muscle movements comprising what looks to an outsider to be a facial expression may not always correspond to a real underlying emotional state. This relation between the bodily feelings and externally observable expression is still an open research area, with a history of controversy. Historically, James was the major proponent of emotion as an experience of bodily changes, such as the perspiring hands or a pounding heart.³⁰ This view was challenged by Cannon³¹ and by Schachter³² who argued that the experience of physiological changes was not sufficient to discriminate emotions. According to Schachter,³² physiological responses such as sweaty hands and a rapid heart beat inform our brain that we are aroused and then the brain must analyze the situation we are in before it can label the state with an emotion such as fear or love.

Since these classic works, there has been a debate about whether or not emotions are accompanied by specific physiological changes other than simply arousal level. Winton et al.³³ provided some of the first findings showing significant differences in autonomic nervous system signals according to a small number of emotional categories or dimensions, but there was no exploration of automated classification. Fridlund and Izard³⁴ appear to have been the first to apply pattern recognition (linear discriminants) to classification of emotion from physiological features, attaining rates of 38-51 percent accuracy (via cross-validation) on subject-dependent classification of four different facial expressions (happy, sad, anger, fear) given four facial electromyogram signals. Picard et al.²⁸ classified physiological patterns for a set of eight emotions (including neutral) by applying pattern recognition techniques and by focusing on felt emotions of a single subject over sessions spanning many weeks.

3. MULTIMODAL APPROACH TO EMOTION RECOGNITION

The studies in facial expression recognition and vocal affect recognition have been done largely independent of each other. The aforementioned works in facial expression recognition used still photographs or video sequences where the subject exhibits only facial expression without speaking any words. Similarly, the works on vocal emotion detection used only the audio information. There are situations where people would speak and exhibit facial expressions at the same time. For example, "he said hello with a smile." Pure facial expression recognizers may fail because the mouth movements may not fit the description of a pure "smile." For computers to be able to recognize emotional expression in practical scenarios, these cases must be handled.

3.1. Related Research

Combining audio and visual cues has been studied in recent years for speech recognition.³⁵ It has been shown that in situations when background noise makes the speech waveforms very noisy, cues from the lip movements improve speech recognition accuracy a great deal. In speech recognition, the lip movements and speech sounds are tightly coupled. For emotional expression recognition, the coupling is not so tight. Very little has been done to utilize both modalities for recognizing emotions.

Pelachaud et al.³⁶ constructed a system that generated animated facial expressions for synthetic speech. Again, this work only emphasized the synthetic aspect and not the recognition of emotions.

De Silva and Ng³⁷ proposed a rule-based method for singular classification of audiovisual input data into one of the six emotion categories: happiness, sadness, fear, anger, surprise, and dislike. Each of their subjects

was asked to portray 12 emotion outbursts per category by displaying the related prototypical facial expression while speaking a single English word of his choice. The audio and visual material has been processed separately. They used optical flow for detecting the displacement and velocity of some key facial features (e.g., corners of the mouth, inner corners of the eye brows). From the audio signal, pitch and pitch contours were estimated by using the method proposed by Medan et al.³⁸ A nearest neighbor method has been used to classify the extracted facial features and an HMM has been used to classify the estimated acoustic features into one of the emotion categories. Per subject, the results of the classification were plotted in two graphs and based upon these graphs the rules for emotion classification of the audiovisual input material were defined.

Chen and Huang³⁹ proposed a set of methods for singular classification of input audiovisual data into one of the basic emotion categories: happiness, sadness, disgust, fear, anger, and surprise. They collected data from five subjects which displayed 6 basic emotions 6 times by producing the appropriate facial expression right before or after speaking a sentence with the appropriate vocal emotion. Each of these single-emotion sequences started and ended with a neutral expression. Considering the fact that in the recorded data a pure facial expression occurred right before or after the sentence spoken with the appropriate vocal emotion, the authors applied a single-modal classification method in a sequential manner.

To conclude, the most surprising issue regarding the multimodal emotion recognition problem, is that although the recent advances in video and audio processing could make the multimodal analysis of human affective state tractable, there were only a few research efforts which tried to implement a multimodal emotion analyzer. Further, there is no record of a research effort that aims at integrating all nonverbal modalities into a single system for affect-sensitive analysis of human behavior.

3.2. Fusing Multimodal Information Using Probabilistic Graphical Models

A typical issue of multimodal data processing so far is that the multisensory data are typically processed separately and only combined at the end. Yet this is almost certainly incorrect; people display audio and visual communicative signals in a complementary and redundant manner. Chen et al.⁴⁰ have shown this experimentally. In order to accomplish a human-like multimodal analysis of multiple input signals acquired by different sensors, the signals cannot be considered mutually independent and cannot be combined in a context-free manner at the end of the intended analysis but, on the contrary, the input data should be processed in a joint feature space and according to a context-dependent model. In practice, however, besides the problems of context sensing and developing context-dependent models for combining multisensory information, one should cope with the size of the required joint feature space, which can suffer from large dimensionality, different feature formats, and timing. A potential way to achieve the target tightly coupled multisensory data fusion is to develop context-dependent versions of a suitable method such as the Bayesian inference method proposed by Pan et al.⁴¹

If we consider the state of the art in audio and visual signal processing, noisy and partial input data should also be expected. A multimodal system should be able to deal with these imperfect data and generate its conclusion so that the certainty associated with it varies in accordance to the input data. A way of achieving this is to consider the time-instance versus time-scale dimension of human nonverbal communicative signals as suggested by Pantic and Rothkrantz.¹⁷ By considering previously observed data (time scale) with respect to the current data carried by functioning observation channels (time instance), a statistical prediction and its probability might be derived about both the information that have been lost due to malfunctioning/inaccuracy of a particular sensor and the currently displayed action/reaction. Probabilistic graphical models, such as hidden Markov Models (including their hierarchical variants), Bayesian networks, and Dynamic Bayesian networks are very well suited for fusing such different sources of information. These models can handle noisy features, temporal information, and missing values of features all by probabilistic inference. Hierarchical HMM-based systems¹² have been shown to work well for facial expression recognition. Dynamic Bayesian Networks and HMM variants⁴² have been shown to fuse various sources of information in recognizing user intent, office activity recognition, and event detection in video using both audio and visual information.

The success of these research efforts has shown that fusing audio and video for detection of discrete events using probabilistic graphical models is possible. Therefore, we propose the Bayesian network topology for recognizing emotions from audio and facial expressions presented in Figure 2. While the network shown is static, it can be extended to be a dynamic Bayesian network in a straightforward manner. The network topology combines the

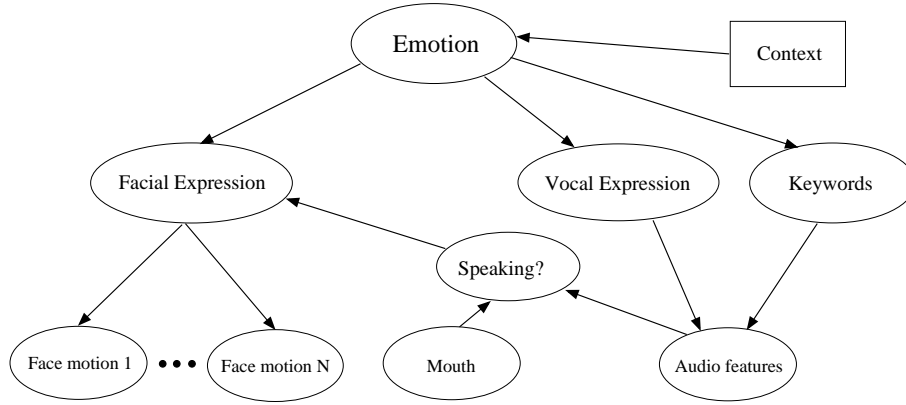


Figure 2. Bayesian network topology for bimodal emotion expression recognition.

two modalities in a probabilistic manner. The top node is the class variable (recognized emotional expression). It is affected by recognized facial expressions, recognized vocal expressions, recognized keywords that have an affective meaning, and by the context in which the system operates (if that is available). Vocal emotions are recognized from audio features extracted from the person’s audio track. Facial expressions are recognized by facial features tracked using video, but the recognition is also affected by a variable that indicates whether the person is speaking or not. Recognizing whether a person is speaking uses both visual cues (mouth motion) and audio features (using similar techniques as Garg et al.⁴²). The parameters of the proposed network can be learned from data, or manually set for some variables. Inferring the human emotional expression can be performed even when some pieces of information are missing, e.g., when audio is too noisy, or the face tracking loses the face.

Another issue which makes the problem of emotional expression recognition even more difficult to solve in a general case is the dependency of a person’s behavior on his/her personality, cultural, and social vicinity, current mood, and the context in which the observed behavioral cues were encountered. One source of help for these problems is machine learning: rather than using a priori rules to interpret human behavior, we can potentially learn application-, user-, and context-dependent rules by watching the user’s behavior in the sensed context.⁴³ This leads to another advantage of probabilistic graphical models: well known algorithms exist to adapt the models, and it is possible to use prior knowledge when learning new models. For example, a prior model of emotional expression recognition trained based on a certain user can be used as a starting point for learning a model for another user, or for the same user in a different context. Though context sensing and the time needed to learn appropriate rules are significant problems in their own right, many benefits could come from such an adaptive affect-sensitive HCI tool.

While fusing multimodal information for emotion recognition is an important issue, there are some other aspects that are of equal importance. Difficult problems are those of obtaining the ground truth of the data and getting data that genuinely corresponds to a particular emotional state. Even though there are cases when the data can be easily labeled (e.g., a singular strong emotion is captured, such as an episode of rage), in most of the cases the ground truth — which emotion was present? — is difficult to establish. We discuss these issues in detail in the following sections.

3.3. Collecting Multimodal Data for Emotion Recognition

In general, the goal of emotional expression is to detect the emotional state of the person in a natural situation. However, as any photographer can attest, getting a real smile can be challenging. Asking someone to smile often does not create the same picture as an authentic smile. The fundamental reason of course is that the subject often does not feel happy so his smile is artificial and in many subtle ways quite different than a genuine smile.

Picard et al.²⁸ outlined five factors that influence the affective data collection:

- *Spontaneous* versus *posed*: Is the emotion elicited by a situation or stimulus that is outside the subject’s control or the subject is asked to elicit the emotion?

- *Lab setting* versus *real-world*: Is the data recording taking place in a lab or the emotion is recorded in the usual environment of the subject?
- *Expression* versus *feeling*: Is the emphasis on external expression or on internal feeling?
- *Open recording* versus *hidden recording*: Is the subject aware that he is being recorded?
- *Emotion-purpose* versus *other-purpose*: Does the subject know that he is a part of an experiment and the experiment is about emotion?

Note that these factors are not necessarily independent. The most natural setup would imply that the subject feels the emotion internally (*feeling*), the emotion occurs *spontaneous*, while the subject is in his usual environment (*real-world*). Also, the subject should not know that he is being recorded (*hidden recording*) and that he is a part of an experiment (*other-purpose*). Such data are usual impossible to obtain because of privacy and ethics concerns. As a consequence, several researchers^{28,44} were trying to use a setup that resembled as much as possible the natural setup. Picard et al.²⁸ collected data using a *posed*, closed to *real-world* (subject's comfortable usual workplace), *feeling*, *open-recording*, and *emotion-purpose* methodology. The key factor that made their data unique is that the subject tried to elicit an internal *feeling* of each emotion. Sebe et al.⁴⁴ were more interested in collecting *spontaneous* emotion data. They created a video kiosk (*lab setting*) with a hidden camera (*hidden-recording*) which displayed segments from recent movie trailers. This setup had the main advantage that it naturally attracted people to watch and could potentially elicit emotions through different genres of video footage — i.e. horror films for shock, comedy for joy, etc.

The issue of whether to use *posed* or *spontaneous* expressions in selecting facial stimuli, has been hotly debated.⁴⁵ Experimentalists and most emotion theorists argue that spontaneous expressions are the only “true” expressions of facial emotion and therefore such stimuli are the only ones of merit.

When recording authentic (*spontaneous*) emotions several aspects should be considered.⁴⁴ Not all people express emotion equally well; many individuals have idiosyncratic methods of expressing emotion as a result of personal, familial, or culturally learned display rules. Situations in which authentic emotions are usually recorded (e.g., *lab setting*) are often unusual and artificial. If the subject is aware of being photographed or filmed (*open-recording*), his emotional response may not be spontaneous anymore. Even if the subject is unaware of being filmed (*hidden-recording*), the laboratory situation may not encourage natural or usual emotion response. In interacting with scientists or other authorities (*emotion-purpose*), subjects will attempt to act in appropriate ways so that emotion expression may be masked or controlled. Additionally, there are only a few universal emotions and only some of these can be ethically stimulated in the laboratory.

On the other hand, posed expressions may be regarded as an alternative, provided that certain safeguards are followed. Increased knowledge about the face, based in large part on observation of spontaneous, naturally occurring facial expressions, has made possible a number of methods of measuring the face. The same situation stands for voice analysis. These measurement techniques can be used to ascertain whether or not emotional behavior has occurred and what emotion is shown in a given instance. Such facial scoring provides a kind of stimulus criterion validity that is important in this area. Additionally, posers can be instructed, not to act or pose a specific emotion, but rather to move certain muscles so as to effect the desired emotional expression. In this way, experimental control may be exerted on the stimuli and the relationship between the elements of the expression and the responses of observers may be analyzed and used as a guide in item selection.

It should be noted that the distinction between posed and spontaneous behavior is not directly parallel to the distinction between artificial and natural occurrences. Though posing is by definition artificial, spontaneous behavior may or may not be natural.⁴⁵ Spontaneous behavior is natural when some part of life itself leads to the behavior studied. Spontaneous behavior elicited in the laboratory may be representative of some naturally occurring spontaneous behavior, or conceivably it could be artificial if the eliciting circumstance is unique and not relevant to any known real life event.

From the above discussion, it is clear that the authentic emotion analysis should be performed whenever is possible. Posed expression may be used as an alternative only in restricted cases and they can be mostly used for benchmarking the authentic expressions.

3.4. Leveraging Unlabeled Data for Emotion Recognition

As pointed out in the previous section, collecting emotional expression data is a difficult task. Labeling those data adds an additional challenge, as it is time-consuming, error prone, and expensive. In addition, an emotion expression recognition system that is deployed in a realistic setting would easily obtain an abundance of emotion expressions, but would not be able to obtain manual labeling of that data — if a computer constantly asks a user for his/her emotion, we can be quite sure that eventually the response would be that of anger or annoyance. Therefore, it would be very beneficial to construct methods that utilize both scarcely available labeled data and abundance of unlabeled data — where the labels are the emotional state (or expression) of a user.

Again, probabilistic graphical models are ideal candidates for such data, as efficient and convergent algorithms exist for handling missing data in general and unlabeled data in particular. Cohen et al.¹⁴ have shown that unlabeled data can be used for recognizing facial expressions using Bayesian networks with a combination of labeled and unlabeled data. However, they have shown that care must be taken when attempting such schemes. While in the purely supervised case (with only labeled data), adding more labeled data always improves the performance of the classifier, adding more unlabeled data can be detrimental to performance. As shown by Cohen et al.¹⁴ such detrimental effects occur when the assumed classifier’s model does not match the distribution generating data. They propose an algorithm for stochastically searching the space of Bayesian networks, converging on a classifier which does utilize positively unlabeled data.

To conclude, further research is required to achieve maximum utilization of unlabeled data for the problem of emotion recognition, but it is clear that such methods would provide great benefit.

4. DISCUSSION AND CONCLUSION

As remarked by Goleman⁴⁶ emotional skills are an essential part of what is called “intelligence”. This is based on recent scientific findings about the role of emotional abilities in human intelligence and on the way human-machine interaction imitates human-human interaction. As a consequence, emotions, largely overlooked in early efforts to develop machine intelligence, are increasingly regarded as an area of important research.

Emotion modulates almost all modes of human communication — facial expression, gestures, posture, tone of voice, choosing of words, respiration, skin temperature and clamminess, etc. Emotions can significantly change the message: sometimes it is not what was said that is the most important, but how it was said. Faces tend to be the most visible form of emotion communication, but they are also most easily controlled in response to different social situations when compared to the voice and other ways of expression. As noted by Picard² affect recognition is most likely to be accurate when it combines multiple modalities, information about the user’s context, situation, goal, and preferences. A combination of low-level features, high-level reasoning, and natural language processing is likely to provide the best emotion inference. Considering all these aspects, Reeves and Nass¹ and Pentland⁴³ believe that multimodal context-sensitive human-computer interaction is likely to become the single most widespread research topic of the artificial intelligence research community. Advances in this area could change not only how professionals practice computing, but also how mass consumers interact with the technology.

As we discussed in this paper and pointed out by Pantic and Rothkrantz,¹⁷ although there were significant advances in the fields of video and audio processing, pattern recognition, computer vision, and affective computing, the realization of a robust, multimodal, adaptive, context-sensitive analyzer of human nonverbal affective state is far from being a reality. Currently, the researchers have to cope with the lack of a better understanding of individual- and context-dependent human behavior and with a better integration of multiple sensors and pertinent modalities according to the model of human sensory system. Besides these problems there are other social and ethical issues that should be considered. The context-sensitive multimodal system that is supposed to interact with the human should not invade the user’s privacy. Computer technology and especially affect-sensitive monitoring tools might be regarded as “big brother” tools. As remarked by Schneiderman,⁴⁷ a large proportion of the population would be terrified by the vision of the universal use of computers in the coming era of ubiquitous computing. Another important aspect is related to teaching the HCI systems our interaction patterns and related behavior and our social and cultural profile. It is obvious that is inefficient and annoying for the human user to train separately all the HCI systems that will be all around us in the future. One way

of dealing with this problem is the incorporation of unlabeled data as we pointed out in this paper. Moreover, the system itself should be able to monitor human nonverbal behavior and to adapt to the current user, to his context, and to the current scenario and environment.

By taking all of these aspects into account, we hope to be able to develop into the near future multimodal context-sensitive systems that are smart, perceptually aware, recognize the context in which they act, can adapt to their users, and can understand how they feel, and respond appropriately. In some sense, these systems will be the friendly variants of the Replicants from the Blade Runner.

REFERENCES

1. B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places*, Cambridge Univ. Press, 1996.
2. R. W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 1997.
3. P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique," *Psychological Bulletin* **115**(2), pp. 268–287, 1994.
4. P. Ekman and W. Friesen, *Facial Action Coding System: Investigator's Guide*, Consulting Psychologists Press, 1978.
5. K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.* **E74**(10), pp. 3474–3483, 1991.
6. N. Ueki, S. Morishima, H. Yamada, and H. Harashima, "Expression analysis/synthesis system based on emotion space constructed by multilayered neural network," *Systems and Computers in Japan* **25**(13), pp. 95–103, 1994.
7. A. Lanitis, C. Taylor, and T. Cootes, "A unified approach to coding and interpreting face images," in *Proc. International Conf. on Computer Vision*, pp. 368–373, 1995.
8. M. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Proc. International Conf. on Computer Vision*, pp. 374–381, 1995.
9. M. Rosenblum, Y. Yacoob, and L. Davis, "Human expression recognition from motion using a radial basis function network architecture," *IEEE Trans. on Neural Network* **7**(5), pp. 1121–1138, 1996.
10. I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **19**(7), pp. 757–763, 1997.
11. T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences," in *Proc. International Conf. on Image Processing*, pp. 546–549, 1997.
12. I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Computer Vision and Image Understanding* **91**(1-2), pp. 160–187, 2003.
13. L. Chen, *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.
14. I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang, "Semi-supervised learning of classifiers: Theory, algorithms, and applications to human-computer interaction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, to appear, 2004.
15. B. Fasel and J. Luetttin, "Automatic facial expression analysis: A survey," *Pattern Recognition* **36**, pp. 259–275, 2003.
16. M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22**(12), pp. 1424–1445, 2000.
17. M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE* **91**(9), pp. 1370–1390, 2003.
18. Y. Yacoob and L. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **18**(6), pp. 636–642, 1996.
19. J. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality and Social Psychology* **37**(11), pp. 2049–2058, 1979.
20. I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion," *Journal of the Acoustic Society of America* **93**(2), pp. 1097–1108, 1993.

21. C. Chiu, Y. Chang, and Y. Lai, "The analysis and recognition of human vocal emotions," in *Proc. International Computer Symposium*, pp. 83–88, 1994.
22. F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. International Conf. on Spoken Language Processing*, pp. 1970–1973, 1996.
23. K. Scherer, "Adding the affective dimension: A new look in speech analysis and synthesis," in *Proc. International Conf. on Spoken Language Processing*, pp. 1808–1811, 1996.
24. Y. Sagisaka, N. Campbell, and N. Higuchi, eds., *Computing Prosody*, Springer-Verlag, New York, NY, 1997.
25. C. Williams and K. Stevens, "Emotions and speech: Some acoustical correlates," *Journal of the Acoustic Society of America* **52**(4), pp. 1238–1250, 1972.
26. I. Murray and J. Arnott, "Synthesizing emotions in speech: Is it time to get excited?," in *Proc. International Conf. on Spoken Language Processing*, pp. 1816–1819, 1996.
27. V. Petrushin, "How well can people and computers recognize emotions in speech?," in *Proc. AAAI Fall Symposium*, pp. 141–145, 1998.
28. R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **23**(10), pp. 1175–1191, 2001.
29. J. Cacioppo and L. Tassinary, "Inferring psychological significance from physiological signals," *American Psychologist* **45**, pp. 16–28, 1990.
30. W. James, *The Principles of Psychology*, Henry Holt, New York, NY, 1890.
31. W. Cannon, "The James-Lange theory of emotion: A critical examination and an alternative theory," *American Journal of Psychology* **39**, pp. 106–124, 1927.
32. S. Schachter, "The interaction of cognitive and physiological determinants of emotional state," in *Advances in Experimental Psychology*, L. Berkowitz, ed., **1**, pp. 49–80, 1964.
33. W. Winton, L. Putman, and R. Krauss, "Facial and autonomic manifestations of the dimensional structure of the emotion," *Journal of Experimental Social Psychology* **20**, pp. 195–216, 1984.
34. A. Fridlund and C. Izard, "Electromyographic studies of facial expressions of emotions and patterns of emotion," in *Social Psychophysiology: A Sourcebook*, J. Cacioppo and R. Petty, eds., pp. 243–286, 1983.
35. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE* **91**(9), pp. 1306–1326, 2003.
36. C. Pelachaud, N. Badler, and M. Steedman, "Generating facial expression for speech," *Cognitive Science* **20**, pp. 1–46, 1996.
37. L. De Silva and P. Ng, "Bimodal emotion recognition," in *Proc. Automatic Face and Gesture Recognition*, pp. 332–335, 2000.
38. Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. on Signal Processing* **39**, pp. 40–48, 1991.
39. L. Chen and T. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proc. International Conference on Multimedia and Expo (ICME)*, pp. 423–426, 2000.
40. L. Chen, H. Tao, T. Huang, T. Miyasato, and R. Nakatsu, "Emotion recognition from audiovisual information," in *Proc. IEEE Workshop on Multimedia Signal Processing*, pp. 83–88, 1998.
41. H. Pan, Z. Liang, T. Anastasio, and T. Huang, "Exploiting the dependencies in information fusion," in *Proc. Conf. on Computer Vision and Pattern Recognition*, **2**, pp. 407–412, 1999.
42. A. Garg, V. Pavlovic, and J. Rehg, "Boosted learning in dynamic Bayesian networks for multimodal speaker detection," *Proceedings of the IEEE* **91**(9), pp. 1355–1369, 2003.
43. A. Pentland, "Looking at people," *Communications of the ACM* **43**(3), pp. 35–44, 2000.
44. N. Sebe, M. Lew, I. Cohen, Y. Sun, T. Gevers, and T. Huang, "Authentic facial expression analysis," in *Automatic Face and Gesture Recognition*, pp. 517–522, 2004.
45. P. Ekman, ed., *Emotion in the Human Face*, Cambridge University Press, New York, NY, 2nd ed., 1982.
46. D. Goleman, *Emotional Intelligence*, Bantam Books, 1995.
47. B. Schneiderman, "Human values and the future of technology: A declaration of responsibility," in *Sparks of Innovation in Human-computer Interaction*, B. Schneiderman, ed., 1993.