J Multimodal User Interfaces
DOI 10.1007/s12193-009-0030-8

ORIGINAL PAPER

Multimodal user's affective state analysis in naturalistic interaction

George Caridakis · Kostas Karpouzis · Manolis Wallace · Loic Kessous · Noam Amir

Received: 30 March 2009 / Accepted: 9 November 2009

Abstract Affective and human-centered computing have attracted an abundance of attention during the past years, mainly due to the abundance of environments and applications able to exploit and adapt to multimodal input from the users. The combination of facial expressions with prosody information allows us to capture the users' emotional state in an unintrusive manner, relying on the best performing modality in cases where one modality suffers from noise or bad sensing conditions. In this paper, we describe a multi-cue, dynamic approach to detect emotion in naturalistic video sequences, where input is taken from nearly real world situations, contrary to controlled recording conditions of audiovisual material. Recognition is performed via a recurrent neural network, whose short term memory and approximation capabilities cater for modeling dynamic events in facial and prosodic expressivity. This approach also differs from existing work in that it models user expressivity using a dimensional representation, instead of detecting discrete 'universal emotions', which are scarce in everyday human-machine interaction. The algorithm is deployed on an audiovisual database which was recorded simulating human-human discourse and, therefore, contains less extreme expressivity and subtle variations of a number of emotion labels. Results show that in turns lasting more than a few frames, recognition rates rise to 98%.

Keywords Affective computing · Emotion recognition · Recurrent neural network · Emotion dynamics · Multimodal analysis

1 Introduction

The introduction of the term 'affective computing' by R. Picard [70] epitomizes the fact that computing is no longer considered a 'number crunching' discipline, but should be thought of as an interfacing means between humans and machines and sometimes even between humans alone. To achieve this, application design must take into account the ability of humans to provide multimodal input to computers, thus moving away from the monolithic window-mouse-pointer interface paradigm and utilizing more intuitive concepts, closer to human niches [43, 68]. A large part of this naturalistic interaction concept is expressivity [71], both in terms of interpreting the reaction of the user to a particular event or taking into account their emotional state and adapting presentation to it, since it alleviates the learning curve for conventional interfaces and makes less technology-savvy users feel more comfortable. For exhaustive surveys of the past work in the machine analysis of affective expressions, readers are referred to [20, 31, 62, 64, 67, 75, 81, 87], which were published between 1992 and 2007. Overviews of early work on facial expression analysis can be read at [31, 63, 75], of multimodal affect recognition methods at

G. Caridakis (✉) · K. Karpouzis
Image, Video and Multimedia Systems Laboratory, National Technical University of Athens, Athens, Greece
e-mail: gcari@image.ntua.gr

K. Karpouzis
e-mail: kkarpou@image.ntua.gr

M. Wallace
Department of Computer Science, University of Peloponnese, Tripolis, Greece
e-mail: wallace@uop.gr

L. Kessous
Institut des Systèmes Intelligents et de Robotique, Paris, France

N. Amir
Tel Aviv Academic College of Engineering, Tel Aviv, Israel

Published online: 15 December 2009

[20, 64, 67, 81, 101], while surveys of techniques for automatic facial muscle action recognition and facial expression analysis at [62, 87].

In this framework, both speech and facial expressions are of great importance, since they usually provide a comprehensible view of users' reactions; actually, Cohen commented on the emergence and significance of multimodality, albeit in a slightly different human-computer interaction (HCI) domain, in [12] and [13], while Oviatt [57] indicated that an interaction pattern constrained to mere 'speak-and-point' only makes up for a very small fraction of all spontaneous multimodal utterances in everyday HCI [58]. In the context of HCI, [42] defines a multimodal system as one that 'responds to inputs in more than one modality or communication channel' abundance, while Mehrabian [53] suggests that facial expressions and vocal intonations are the main means for someone to estimate a person's affective state [98], with the face being more accurately judged, or correlating better with judgments based on full audiovisual input than on voice input [42, 61]. This fact led to a number of approaches using video and audio to tackle emotion recognition in a multimodal manner [21, 32, 40, 44], while recently the visual modality has been extended to include facial, head or body gesturing ([35] and [46]). Also, it has been shown by several experimental studies that integrating the information from audio and video leads to an improved performance of affective behavior recognition. The improved reliability of audiovisual approaches in comparison to single-modal approaches can be explained as follows: Current techniques for the detection and tracking of facial expressions are sensitive to head pose, clutter, and variations in lighting conditions, while current techniques for speech processing are sensitive to auditory noise. Audiovisual fusion can make use of the complementary information from these two channels. In addition, many psychological studies have theoretically and empirically demonstrated the importance of the integration of information from multiple modalities (vocal and visual expression in this paper) to yield a coherent representation and inference of emotions [2, 74, 78]. As a result, an increased number of studies on audiovisual human affect recognition have emerged in recent years (e.g., [10, 100]).

Additional factors that contribute to the complexity of estimating expressivity in everyday HCI are the fusion of the information extracted from modalities [57], the interpretation of the data through time and the noise and uncertainty alleviation from the natural setting [66]. In the case of fusing multimodal information [95], systems can either integrate signals at the feature level [84] or, after coming up with a class decision at the feature level of each modality, by merging decisions at a semantic level (late identification, [84] and [72]), possibly taking into account any confidence measures provided by each modality or, generally, a mixture of experts mechanism [37]. Hence, while automatic detection of the six basic emotions in posed controlled audio or visual displays can be done with reasonably high accuracy, detecting these expressions or any expression of human affective behavior in less constrained settings is still a very challenging problem due to the fact that deliberate behavior differs in visual appearance, audio profile, and timing from spontaneously occurring behavior. Due to this criticism received from both cognitive and computer scientists, the focus of the research in the field started to shift to the automatic analysis of spontaneously displayed affective behavior. Several studies have recently emerged on the machine analysis of spontaneous facial expressions (e.g., [4, 5, 14, 90]) and vocal expressions (e.g. [47]). Most of the existing methods for audiovisual affect analysis are based on deliberately posed affect displays (e.g., [9, 34, 38, 82, 83, 91, 99, 102, 103] and [104]). Recently, a few exceptional studies have been reported toward audiovisual affect analysis in spontaneous affect displays (e.g., [10, 59, 69, 100]. Zeng et al. [100] used the data collected in psychological research interview (Adult Attachment Interview), Pal et al. [59] used recordings of infants, and Petridis and Pantic [69] used the recordings of people engaged in meetings AMI corpus. On the other hand, Fragopanagos and Taylor [32], Caridakis et al. [10], and Karpouzis et al. [46], used the data collected in Wizard of Oz scenarios. Since the available data were usually insufficient to build a robust machine learning system for the recognition of fine-grained affective states (e.g., basic emotions), the recognition of coarse affective states was attempted in most of the aforementioned studies. The studies of Zeng et al. focus on audiovisual recognition of positive and negative affect [100], while other studies report on the classification of audiovisual input data into the quadrants in the evaluation-activation space [10]. The studies reported in [10] and [46] applied the Feeltrace system that enables raters to continuously label changes in affective expressions. However, note that the study reported on a considerable labeling variation among four human raters due to the subjectivity of audiovisual affect judgment. More specifically, one of the raters mainly relied on audio information when making judgments, while another rater mainly relied on visual information. This experiment actually also reflects the asynchronization of audio and visual expression. In order to reduce this variation of human labels, the studies of Zeng et al. [100] made the assumption that facial expression and vocal expression have the same coarse emotional states (positive and negative) and then directly used FACS-based labels of facial expressions as audiovisual expression labels.

On the other hand, a growing body of research in cognitive sciences argues that the dynamics of human behavior are crucial for its interpretation (e.g., [26, 74, 76] and [78]). For example, it has been shown that temporal dynamics of facial behavior represent a critical factor for distinction between spontaneous and posed facial behavior (e.g., [14, 26, 90] and [89]) and for categorization of complex behaviors like pain, shame, and amusement (e.g., [4, 26, 94] and [51]). Based on these findings, we may expect that the temporal dynamics of each modality (facial and vocal) and the temporal correlations between the two modalities play an important role in the interpretation of human naturalistic audiovisual affective behavior. However, these are virtually unexplored areas of research. Regarding the dynamic nature of expressivity, Littlewort [51] states that while muscle-based techniques can describe the morphology of a facial expression, it is very difficult for them to illustrate in a measurable (and, therefore detectable) manner the dynamics, i.e. the temporal pattern of muscle activation and observable feature movement or deformation. She also makes a case of natural expressivity being inherently different in temporal terms than posed, presenting arguments from psychologists ([25] and [33]), proving the dissimilarity of posed and natural data, in addition to the need to tackle expressivity using mechanisms that capture dynamic attributes. As a general rule, the naturalistic data chosen as input in this work is closer to human reality since intercourse is not acted and expressivity is not guided by directives (e.g. Neutral expression to one of the six universal emotions and back to neutral). This amplifies the difficulty in discerning facial expressions and speech patterns. Nevertheless it provides the perfect test-bed for the combination of the conclusions drawn from each modality in one time unit and use as input in the following sequence of audio and visual events analyzed. Examples of affect-sensitive multimodal HCI systems include the following:

1. the system of Lisetti and Nasoz [49], which combines facial expression and physiological signals to recognize the user's emotions, like fear and anger, and then to adapt an animated interface agent to mirror the user's emotion,
2. the multimodal system of Duric et al. [23], which applies a model of embodied cognition that can be seen as a detailed mapping between the user's affective states and the types of interface adaptations,
3. the proactive HCI tool of Maat and Pantic [52], which is capable of learning and analyzing the user's context-dependent behavioral patterns from multisensory data and of adapting the interaction accordingly,
4. the automated Learning Companion of Kapoor et al. [45], which combines information from cameras, a sensing chair, and mouse, wireless skin sensor, and task state to detect frustration in order to predict when the user needs help, and
5. the multimodal computer-aided learning system in the Beckman Institute, University of Illinois, Urbana-Champaign (UIUC), where the computer avatar offers an appropriate tutoring strategy based on the information of the user's facial expression, keywords, eye movement, and task state.

Current work aims to interpret sequences of events by modeling the user's behavior in a natural HCI setting through time. With the use of a recurrent neural network, the short term memory provided through its feedback connection, works as a memory buffer and the information remembered is taken under consideration in every next time cycle. Theory on this kind of network backs up the claim that it is suitable for learning to recognize and generate temporal patterns as well as spatial ones [28]. In addition to this, results show that this approach can capture the varying patterns of expressivity with a relatively low-scale network, which is not the case with other works operating on acted data.

The paper is structured as follows: in Sect. 2 we provide the fundamental notions upon which the remaining presentation is based. This includes the overall architecture of our approach as well as the running example which we will use throughout the paper in order to facilitate the presentation of our approach. In Sect. 3 we present our feature extraction methodologies, for both the visual and auditory modalities. In Sect. 4 we explain how the features extracted, although fundamentally different in nature, can be used to drive a recursive neural network, in order to acquire an estimation of the human's state. In Sect. 5 we present results from the application of our methodology on naturalistic data and in Section 6 we list our concluding remarks.

2 Fundamentals

2.1 Emotion representation

When it comes to recognizing emotions by computer, one of the key issues is the selection of appropriate ways to represent the user's emotional states. The most familiar and commonly used way of describing emotions is by using categorical labels, many of which are either drawn directly from everyday language, or adapted from it. This trend may be due to the great influence of the works of Ekman and Friesen who proposed that the archetypal emotions correspond to distinct facial expressions which are supposed to be universally recognizable across cultures [24, 27].

On the contrary psychological researchers have extensively investigated a broader variety of emotions. An extensive survey on emotion analysis can be found in [17]. The main problem with this approach is deciding which words qualify as genuinely emotional. There is, however, general agreement as to the large scale of the emotional lexicon, with most lists of descriptive terms numbering into the hundreds; the Semantic Atlas of Emotional Concepts lists 558 words with 'emotional connotations'. Of course, it is difficult to imagine an artificial systems being able to match the level of discrimination that is implied by the length of this list.

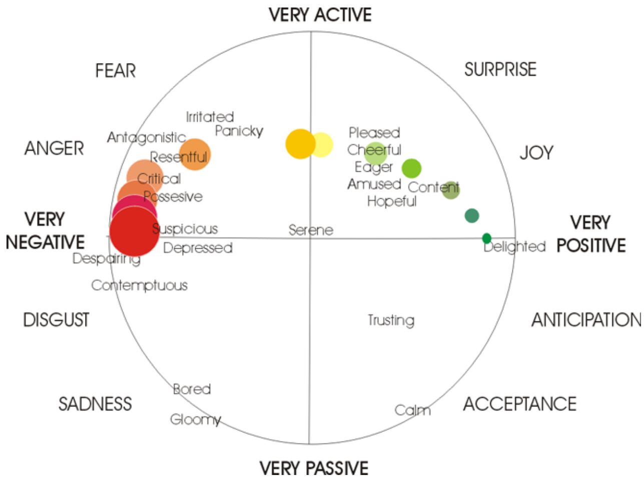


Fig. 1 The activation/valence dimensional representation [93]

Although the labeling approach to emotion representation fits perfectly in some contexts and has thus been studied and used extensively in the literature, there are other cases in which a continuous, rather than discrete, approach to emotion representation is more suitable. At the opposite extreme from the list of categories are dimensional descriptions, which identify emotional states by associating them with points in a multidimensional space. The approach has a long history, dating from Wundt's [17] original proposal to Schlossberg's reintroduction of the idea in the modern era [79]. For example, activation-emotion space as a representation has great appeal as it is both simple, while at the same time makes it possible to capture a wide range of significant issues in emotion [18]. The concept is based on a simplified treatment of two key themes:

1. Valence: The clearest common element of emotional states is that the person is materially influenced by feelings that are valenced, i.e., they are centrally concerned with positive or negative evaluations of people or things or events.
2. Activation level: Research from Darwin forward has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e., the strength of the person's disposition to take some action rather than none.

There is general agreement on these two main dimensions. Still, in addition to these two, there are a number of other possible dimensions, such as power-control, or approach-avoidance. Dimensional representations are attractive mainly because they provide a way of describing emotional states that is more tractable than using words. This is of particular importance when dealing with naturalistic data, where a wide range of emotional states occur. Similarly, they are much more able to deal with non discrete emotions and variations in emotional state over time.

Table 1 Emotional classes

Label	Location in FeelTrace [19] diagram
Q1	Positive activation, positive valence (+/+)
Q2	Positive activation, negative valence (+/−)
Q3	Negative activation, negative valence (−/−)
Q4	Negative activation, positive valence (−/+)
Neutral	Close to the center

In this work we have focused on the general area in which the human emotion lies, rather than on the specific point on the diagram presented in Fig. 1. One of the reasons that has lead us to this decision is that it is not reasonable to expect human annotators to be able to discriminate between an extra pixel to the left or to the right as being an indication of a shift in observed emotional state, and therefore it does not make sense to construct a system that attempts to do so either. Thus, as is also displayed in Fig. 1, we have segmented the emotion representation space in broader areas.

As we can see in the figure, labels are typically given for emotions falling in areas where at least one of the two axes has a value considerably different than zero. On the other hand, the beginning of the axes (the center of the diagram) is typically considered as the neutral emotion. For the same reasons as mentioned above, we find it is not meaningful to define the neutral state so strictly. Therefore, we have added to the more conventional areas corresponding to the four quadrants a fifth one, corresponding to the neutral area of the diagram, as is depicted in Table 1.

2.2 Methodology outline

As we have already explained, the overall approach is based on a multimodal processing of the input sequence. As we

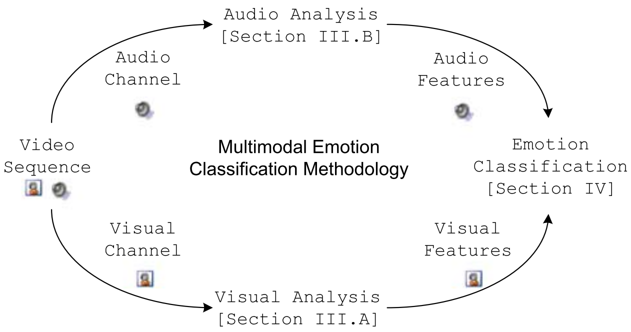


Fig. 2 Graphical outline of the proposed approach

have already explained, there are two different methodologies that fall under this general label; decision-level and feature-level. The decision-level approach has the benefit of being very easy to implement when the independent systems are already available in the literature. The feature-level approach has the drawback of being often difficult to implement, as different information queues are often different in nature and are thus difficult to incorporate in one uniform processing scheme, but, when successfully realized, produces systems that are able to achieve considerably better performances [8].

Our approach is of the latter type; the general architecture of our approach is depicted in Fig. 2.

The considered input sequence is split into the audio and visual sequences. The visual sequence is analyzed frame by frame using the methodology presented in Sect. 3 and the audio sequence is analyzed as outlined in Sect. 3.2 and further explained in [40]. Visual features of all corresponding frames are fed to a recurrent network as explained in Sect. 4, where the dynamics in the visual channel are picked up and utilized in classifying the sequence to one of the five considered emotional classes mentioned in Table 1. Due to the fact that the features extracted from the audio channel are fundamentally different in nature than those extracted from the visual channel, the recurrent network structure is altered accordingly in order to allow both inputs to be fed to the network at the same time, thus allowing for a truly multimodal classification scheme.

The evaluation of the performance of our methodology includes statistical analysis of application results, quantitative comparisons with other approaches focusing on naturalistic data and qualitative comparisons with other known approaches to emotion recognition, all listed in Sect. 5.

2.3 Running example

In developing a multimodal system one needs to integrate diverse components which are meant to deal with the different modalities. As a result, the overall architecture comprises a wealth of methodologies and technologies and can often be difficult to grasp in full detail. In order to facilitate the presentation of the multimodal approach proposed herein for the estimation of human emotional state we will use the concept of a running example.

Our example is a sample from the dataset on which we will apply our overall methodology in Sect. 5. In Fig. 3 we present some frames from the sequence of the running example.

3 Feature extraction

3.1 Visual modality

Automatic estimation of facial model parameters is a difficult problem and although plethora of research has been done on selection and tracking of features [88], relatively little work has been reported [86] on the necessary initialization step of tracking algorithms, which is required in the context of facial feature extraction and expression recognition. Most facial expression recognition systems use the Facial Action Coding System (FACS) model introduced by Ekman and Friesen [24] for describing facial expressions. FACS describes expressions using 44 Action Units (AU) which relate to the contractions of specific facial muscles.

Additionally to FACS, MPEG-4 metrics [85] are commonly used to model facial expressions and underlying emotions. They define an alternative way of modeling facial expressions and the underlying emotions, which is strongly influenced by neurophysiologic and psychological studies. MPEG-4, mainly focusing on facial expression synthesis and animation, defines the Facial Animation Parameters (FAPs) that are strongly related to the Action Units (AUs), the core of the FACS. A comparison and mapping between FAPs and AUs can be found in [30].

Most existing approaches in facial feature extraction are either designed to cope with limited diversity of video characteristics or require manual initialization or intervention. Specifically [86] depends on optical flow, [48] depends on high resolution or noise-free input video, [80] depends on color information, [16] requires two head-mounted cameras and [65] requires manual selection of feature points on the first frame. Additionally very few approaches can perform in near-real time. In this work we combine a variety of feature detection methodologies in order to produce a robust FAP estimator, as outlined in the corresponding section.

Our approach, as described in detail in [41], initiates with the face detection and alignment (Fig. 4) procedure to cope with the head's possible roll rotation. In the following using anthropometric criteria regions of interest are determined (Fig. 5) in order to minimize the search area for candidate feature points. Finally masks are constructed using different techniques for different feature points which are then

Fig. 3 Frames from the running example

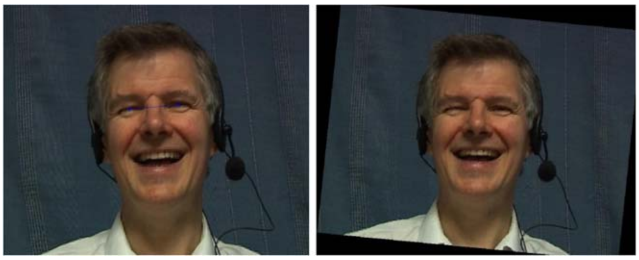


Fig. 4 Frame rotation based on eye locations

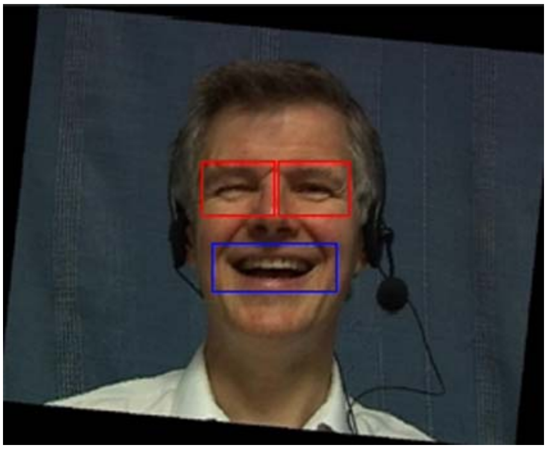


Fig. 5 Regions of interest for facial feature extraction

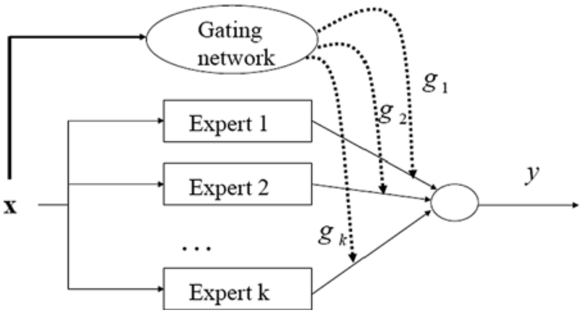


Fig. 6 Mixture of experts architecture

fused (Fig. 6) and provide a robust extraction of facial feature points for nose, eyebrows, eyes and mouth. This overall process detects 21 feature points (Figs. 7 and 8) which in turn are used to calculate FAPs.

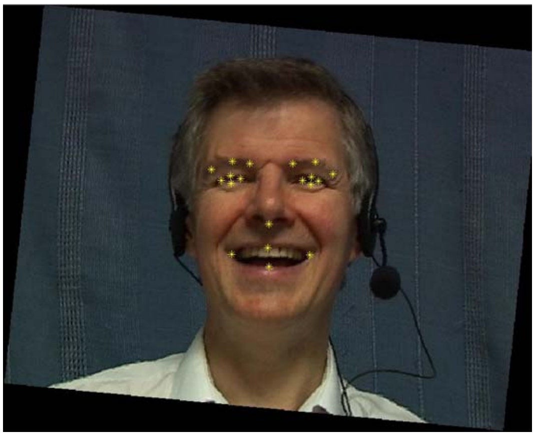


Fig. 7 Feature points detected on the input frame

3.2 Auditory modality

Starting as early as 2000, speech is considered as a modality that can be used as input in order to recognize human emotion [7, 56]. In these early works speech was used to make a two-way distinction between negative (encompassing user states such as anger, annoyance, or frustration) vs. the complement, i.e. neutral, neutral, annoyed, frustrated, tired, amused, and other. The main reason for this mapping onto negative valence vs. neutral/positive valence was that in the intended application, what was desired was to detect 'trouble in communication' [6]. More recent studies have managed to extend the classification categories to three [1] or even eight [15], thus indicating the speech is a modality that can truly provide important information regarding the emotional state.

The set of features used to quantify the prosodic variations in speech, serving as the basis for classification, is also continuously evolving. While earlier studies were biased heavily towards regarding F0 as the main bearer of emotional content, recent studies use a much more varied feature set, based on pitch, intensity, duration, spectrum, stability measures and lexical properties. Nevertheless, there is no standardization on this topic, with different researchers experimenting with quite different features. Most studies include basic statistics of the F0 and intensity contour such as min, max, range, standard deviation [3, 6, 22], though even here details of the normalization schemes may differ. Studies often diverge on the more evolved features taken from these contours, employing various higher order moments,

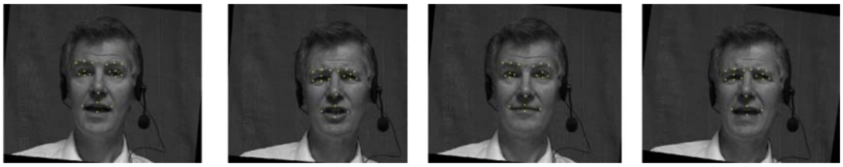


Fig. 8 Feature points detected from frames belonging to different sequences

Table 2 Audio features selected by discriminant function analysis

ptsegno	pfl2	ifqrange	ifmins
pfmean	pfmaxs	ifstart	ittmax
pfstd	pfmins	ifend	vfract
pfmax	pfmicro1	ifl1	vshimapq3
pfrange	pfmicro2	ifl2	vnhr
pfqrange	ifmean	ifpar3	vhnr
pfstart	ifstd	ifdct2	ltasslp
pfend	ifmax	ifmaxs	ltasfmax

curve fitting schemes, etc. As yet very few studies have employed features taken from perceptual or production models.

Overall, no large scale comparison among different feature groups has been published, evaluating their relative importance, though some recent works have begun to doubt whether F0 based features are actually more important than others.

An important difference between the visual and audio modalities is related to the duration of the sequence that we need to observe in order to be able to gain an understanding of the sequence's content. In case of video, a single frame is often enough in order for us to understand what the video displays and always enough for us to be able to process it and extract information. On the other hand, an audio signal needs to have a minimum duration for any kind of processing to be able to be made.

Therefore, instead of processing different moments of the audio signal, as we did with the visual modality, we need to process sound recordings in groups. Obviously, the meaningful definition of these groups will have a major role in the overall performance of the resulting system. In this work we consider sound samples grouped as tunes, i.e. as sequences demarcated by pauses. The basis behind this is that although expressions may change within a single tune, the underlying human emotion does not change dramatically enough to shift from one quadrant to another. For this reason, the tune is not only the audio unit upon which we apply our audio feature detection techniques but also the unit considered during the operation of the overall emotion classification system.

Initially we extract an extensive set of 377 audio features. This comprises features based on intensity, pitch, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment characteristics and pause length. We analyzed each tune with a method employing prosodic representation based on perception called Prosogram [54]. Prosogram is based on a stylization of the fundamental frequency data (contour) for vocalic (or syllabic) nuclei. It gives globally for each voiced nucleus a pitch and a length. According to a 'glissando threshold' in some cases we don't get a fixed pitch but one or more lines to define the evolution of pitch for this nucleus. This representation is in a way similar to the 'piano roll' representation used in music sequencers. This method, based on the Praat environment, offers the possibility of automatic segmentation based both on voiced part and energy maxima. From this model—representation stylization we extracted several types of features: pitch interval based features, nucleus length features and distances between nuclei.

Given that the classification model used in this work, as we will see in Sect. 4, is based on a neural network, using such a wide range of features as input to the classifies means that the size of the annotated data set as well as the time required for training will be huge. In order to overcome this we need to statistically process the acoustic feature, to discriminate the more prominent ones, thus performing feature reduction. In our work we achieve this by combining two well known techniques: analysis of variance (ANOVA) and Pearson product-moment correlation coefficient (PMCC). ANOVA is used first to test the discriminative ability of each feature. This resulting in a reduced feature set, containing about half of the features tested. To further reduce the feature space we continued by calculating the PMCC for all of the remaining feature pairs; PMCC is a measure of the tendency of two variables measured on the same object to increase or decrease together. Groups of highly correlated (>90%) features were formed, and a single feature from each group was selected. The overall process results in reducing the number of audio features considered during classification from 377 to only 32. All selected features are numerical and continuous.

4 Multimodal expression classification

4.1 The Elman net

In order to consider the dynamics of displayed expressions we need to utilize a classification model that is able to model and learn dynamics, such as a Hidden Markov Model or a recursive neural network. In this work we are using a recursive

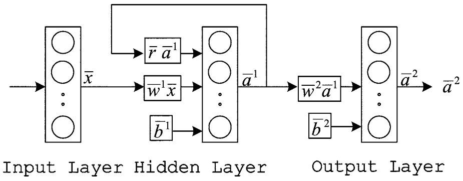


Fig. 9 The recursive neural network

neural network; see Fig. 9. This type of network differs from conventional feed-forward networks in that the first layer has a recurrent connection. The delay in this connection stores values from the previous time step which can be used in the current time step, thus providing the element of memory.

Although we are following an approach that only comprises a single layer of recurrent connections, in reality the network has the ability to learn patterns of a greater length as well, as current values are affected by all previous values and not only by the last one.

Out of all possible recurrent implementations we have chosen the Elman net for our work [28, 29]. This is a two-layer network with feedback from the first layer output to the first layer input. This recurrent connection allows the Elman network to both detect and generate time-varying patterns.

The transfer functions of the neurons used in the Elman net are tan-sigmoid for the hidden (recurrent) layer and purely linear for the output layer. More formally

$$a_i^1 = \tan sig\left(k_i^1\right) = \frac{2}{1+\mathrm{e}^{-2k_i^1}} - 1$$

$$a_j^2 = k_j^2$$

where a_i^1 is the activation of the i-th neuron in the first (hidden) layer, k_i^1 is the induced local field or activation potential of the i-th neuron in the first layer, a_j^2 is the activation of the j-th neuron in the second (output) layer and k_j^2 is the induced local field or activation potential of the j-th neuron in the second layer.

The induced local field in the first layer is computed as:

$$k_i^1 = \bar{w}_i^1 \cdot \bar{x} + \bar{r}_i \cdot \bar{a}^1 + b_i^1$$

where $\bar{x}$ is the input vector, $\bar{w}_i^1$ is the input weight vector for the i-th neuron, $\bar{a}^1$ is the first layer's output vector for the previous time step, $\bar{r}_i$ is the recurrent weight vector and b_i^1 is the bias. The local field in the second layer is computed in the conventional way as:

$$k_j^2 = \bar{w}_j^2 \cdot \bar{a}^1 + b_j^2$$

where $\bar{w}_i^2$ is the input weight and b_j^2 is the bias.

This combination of activation functions is special in that two-layer networks with these transfer functions can approximate any function (with a finite number of discontinuities) with arbitrary accuracy. The only requirement is that the hidden layer must have enough neurons ([77] and [36]).

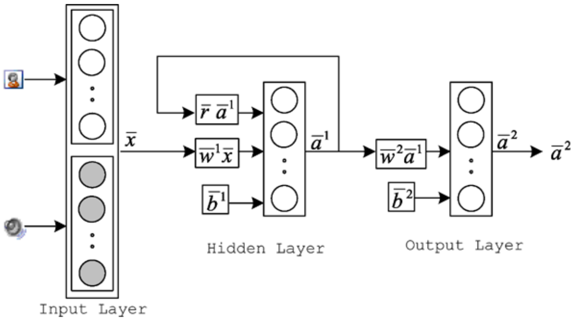


Fig. 10 The modified Elman net

As far as training is concerned, the truncated back-propagation through time (truncated BPTT) algorithm is used [37].

The input layer of the utilized network has 57 neurons (25 for the FAPs and 32 for the audio features). The hidden layer has 20 neurons and the output layer has 5 neurons, one for each one of five possible classes: Neutral, Q1 (first quadrant of the Feeltrace [19] plane), Q2, Q3 and Q4. The network is trained to produce a level of 1 at the output that corresponds to the quadrant of the examined tune and levels of 0 at the other outputs.

4.2 Dynamic and non dynamic inputs

In order for the network to operate we need to provide as inputs the values of the considered features for each frame. As the network moves from one frame to the next it picks up the dynamics described by the way these features are changed and thus manages to provide a correct classification in its output.

One issue that we need to consider, though, is that not all of the considered inputs are dynamic. Specifically, as we have already seen in Sect. 3.2, as far as the auditory modality is concerned the tune is processed as a single unit. Thus, the acquired feature values are referring to the whole tune and cannot be allocated to specific frames. As a result, a recurrent neural network cannot be used directly and unchanged in order to process our data.

In order to overcome this, we modify the simple network structure of Fig. 9 as shown in Fig. 10. In this modified version input nodes of two different types are utilized. For the visual modality features we maintain the conventional type of input neurons that are encountered in recurrent neural networks. For the auditory modality features we use static value neurons. These maintain the same value throughout the operation of the neural network.

The auditory feature values that have been computed for a tune are fed to the network as the values that correspond to the first frame. In the next time steps, while visual features

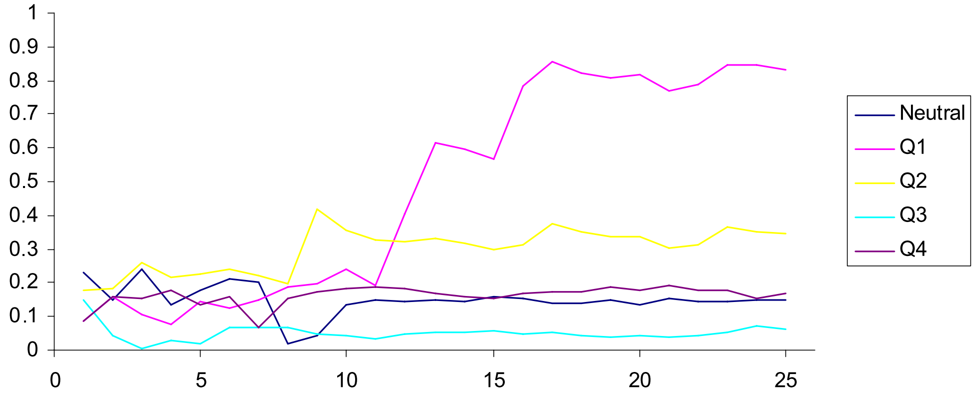


Fig. 11 Individual network outputs after each frame

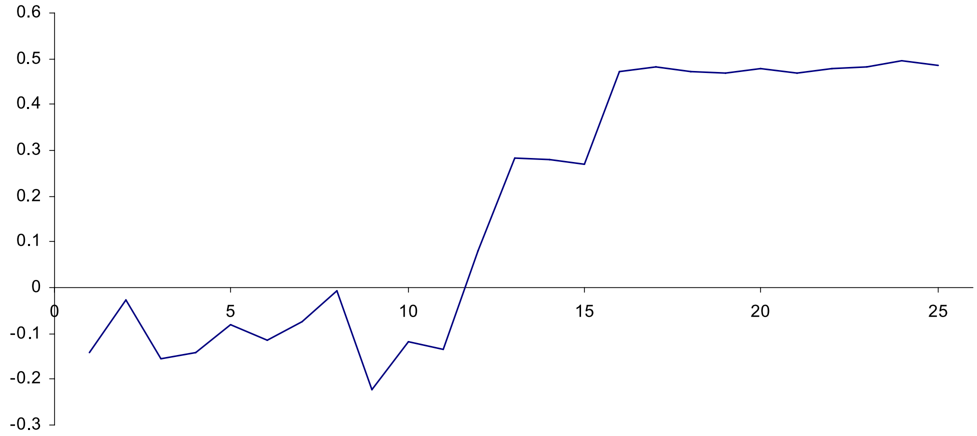


Fig. 12 Margin between correct and next best output

corresponding to the next frames are fed to the first input neurons of the network, the static input neurons maintain the original values for the auditory modality features, thus allowing the network to operate normally.

One can easily notice that although the network has the ability to pick up the dynamics that exist in its input, it cannot learn how to detect the dynamics in the auditory modality since it is only fed with static values. Still, we should comment that the dynamics of this modality are not ignored. Quite the contrary, the static feature values computed for this modality, as has been explained in Sect. 3.2, are all based on the dynamics of the audio channel of the recording.

4.3 Classification

The most common applications of recurrent neural networks include complex tasks such as modeling, approximating, generating and predicting dynamic sequences of known or unknown statistical characteristics. In contrast to simpler neural network structures, using them for the seemingly easier task of input classification is not equally simple or straight forward.

The reason is that where simple neural networks provide one response in the form of a value or vector of values at their output after considering a given input, recurrent neural networks provide such inputs after each different time step. So, one question to answer is at which time step the network's output should be read for the best classification decision to be reached.

As a general rule of thumb, the very first outputs of a recurrent neural network are not very reliable. The reason is that a recurrent neural network is typically trained to pick up the dynamics that exist in sequential data and therefore needs to see an adequate length of the data in order to be able to detect and classify these dynamics. On the other hand, it is not always safe to utilize the output of the very last time step as the classification result of the network because:

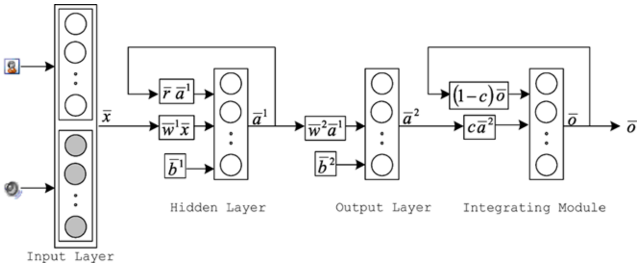


Fig. 13 The Elman net with the output integrator

- the duration of the input data may be a few time steps longer than the duration of the dominating dynamic behavior and thus the operation of the network during the last time steps may be random
- a temporary error may occur at any time step of the operation of the network

For example, in Fig. 11 we present the output levels of the network after each frame when processing the tune of the running example. We can see that during the first frames the output of the network is quite random and changes swiftly. When enough length of the sequence has been seen by the network so that the dynamics can be picked up, the outputs start to converge to their final values. But even then small changes to the output levels can be observed between consecutive frames.

Although these are not enough to change the classification decision (see Fig. 12) for this example where the classification to Q1 is clear, there are cases in which the classification margin is smaller and these changes also lead to temporary classification decision changes.

In order to arm our classification model with robustness we have added a weighting integrating module to the output of the neural network which increases its stability (see Fig. 13). Specifically, the final outputs of the model are computed as:

$$o_j(t) = c \cdot a_j^2 + (1 - c) \cdot o_j(t - 1)$$

where $o_j(t)$ is the value computed for the j-th output after time step t, $o_j(t-1)$ is the output value computed at the previous time step and c is a parameter taken from the $(0, 1]$ range that controls the sensitivity/stability of the classification model. When c is closer to zero the model becomes very stable and a large sequence of changed values of k_j^2 is required to affect the classification results while as c approaches one the model becomes more sensitive to changes in the output of the network. When $c = 1$ the integrating module is disabled and the network output is acquired as overall classification result. In our work, after observing the models performance for different values of c, we have chosen $c = 0.5$.

In Fig. 14 we can see the decision margin when using the weighting integration module at the output of the network. When comparing to Fig. 12 we can clearly see that the progress of the margin is more smooth, which indicates that we have indeed succeeded in making the classification performance of the network more stable and less dependent on frame that is chosen as the end of a tune.

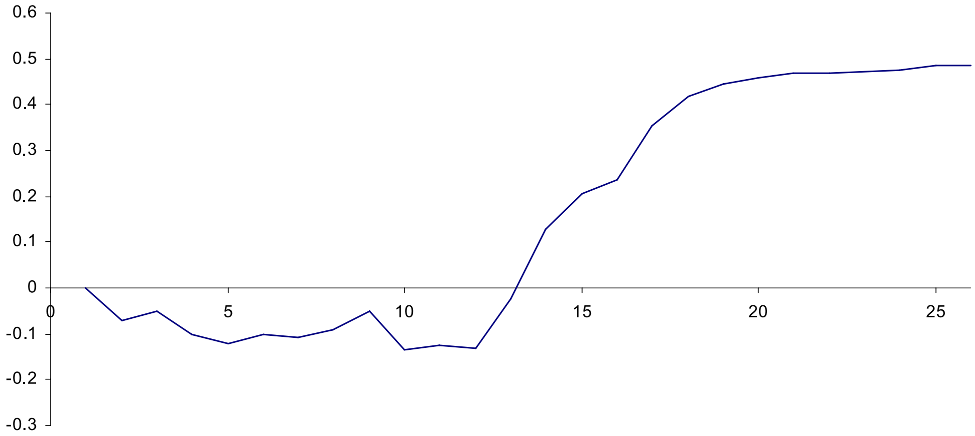
Of course, in order for this weighted integrator to operate, we need to define output values for the network for time step 0, i.e. before the first frame. It is easy to see that due to the way that the effect of previous outputs wares off as time steps elapse due to c, this initialization is practically indifferent for tunes of adequate length. On the other hand, this value may have an important affect on tunes that are very short. In this work, we have chosen to initialize all initial outputs at

$$\bar{o}(0) = 0$$

Another meaningful alternative would be to initialize $\bar{o}(0)$ based on the percentages of the different output classes in the ground truth data used to train the classifier. We have avoided doing this in order not to add a bias towards any of the outputs, as we wanted to be sure that the performance acquired during testing is due solely to the dynamic and multimodal approach proposed in this work.

It is worth noting that from a modeling point of view it was feasible to include this integrator in the structure of the network rather than having it as an external module, simply by adding a recurrent loop at the output layer as well. We have decided to avoid doing so, in order not to also affect the training behavior of the network, as an additional recurrent loop would greatly augment the training time and size and average length of training data required.

Fig. 14 Decision margin when using the integrator



5 Experimental results

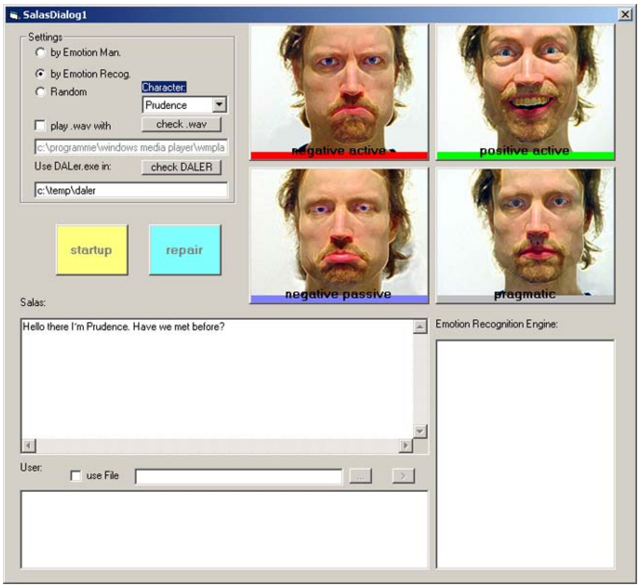
5.1 Ground truth

Since the aim of this work is to emphasize on the ability to classify sequences with naturalistic expressions, we have chosen to utilize the SAL database for training and testing purposes [39]. Recordings were based on the notion of the "Sensitive Artificial Listener", where the SAL simulates what some interviewers and good listeners do, i.e. engages a willing person in emotionally colored interaction on the basis of stock lines keyed in a broad way to the speaker's emotions. Although the final goal is to let the SAL automatically assess the content of the interaction and select the line with which to respond, this had not yet been fully implemented at the time of the creation of the SAL database and thus a "Wizard of Oz" approach was used for the selection of the SAL's answers.

Our test data have been produced using the SAL testbed application developed within the ERMIS and HUMAINE projects, which is an extension of one of the highlights of AI research in the 1960s, Weizenbaum's ELIZA [92]. The ELIZA framework simulates a Rogerian therapy, during which clients talk about their problems to a listener that provides responses that induces further interaction without passing any comment or judgment.

Recording is an integral part of this challenge. With the requirement both audio and visual inputs, the need to compromise between demands of psychology and signal processing is imminent. If one is too cautious about the recording quality, subjects may feel restrained and are unlikely to show the everyday, relaxed emotionality that would cover most of the emotion representation space. On the other hand, visual and audio analysis algorithms cannot be expected to cope with totally unconstrained head and hand movement, subdued lighting, and mood music. Major issues may also arise from the different requirements of the individual modalities: while head mounted microphones might suit analysis of speech, they can have devastating consequences for visual analysis. Eventually arrangements were developed to ensure that on the visual side, the face was usually almost frontal and well and evenly lit to the human eye; that it was always easy for a human listener to make out what was being said; and that the setting allowed most human participants to relax and express emotion within a reasonable time

The implementation of SALAS is mainly a software application designed to let a user work through various emotional states. It contains four 'personalities' shown in shown in Fig. 15 that listen to the user and respond to what he/she says, based on the different emotional characteristics that each of the 'personalities' possesses. The user controls the emotional tone of the interaction by choosing which 'personality' they will interact with, while still being able to change the tone at any time by choosing a different personality to talk to.


SalasDialog1

Settings
○ by Emotion Man.
● by Emotion Recog.
○ Random
Character: Prudence
□ play .wav with
c:\programme\windows media player\wmpla
Use DALer.exe in:
c:\temp\daler

Salas:
Hello there I'm Prudence. Have we met before?

User:
□ use File

Emotion Recognition Engine:

Fig. 15 SALAS interaction interface

The initial recording took place with 20 subjects generating approximately 200 minutes of data. The second set of recordings comprised 4 subjects recording two sessions each, generating 160 minutes of data, providing a total of 360 minutes of data from English speakers; both sets are balanced for gender, 50/50 male/female. These sets provided the input to facial feature extraction and expression recognition system of this paper.

A point to consider in natural human interaction is that each individual's character has an important role on the human's emotional state; different individuals may have different emotional responses to similar stimuli. Therefore, the annotation of the recordings should not be based on the intended induced emotion but on the actual result of the interaction with the SAL. Towards this end, FeelTrace was used for the annotation of recordings in SAL [19]. This is a descriptive tool that has been developed at Queen's University Belfast using dimensional representations, which provides time-sensitive dimensional representations. It lets observers track the emotional content of a time-varying stimulus as they perceive it. Figure 15, illustrates the kind of display that FeelTrace users see.

The space is represented by a circle on a computer screen, split into four quadrants by the two main axes. The vertical axis represents activation, running from very active to very passive and the horizontal axis represents evaluation, running from very positive to very negative. It reflects the popular view that emotional space is roughly circular. The centre of the circle marks a sort of neutral default state, and putting the cursor in this area indicates that there is no real emotion being expressed.

A user uses the mouse to move the cursor through the emotional space, so that its position signals the levels of activation and evaluation perceived by her/him, and the system automatically records the co-ordinates of the cursor at any time.

For reasons outlined in Sect. 2.1 the x–y coordinates of the mouse movements on the two-dimensional user interface are mapped to the five emotional categories presented in Table 1. Applying a standard pause detection algorithm on the audio channel of the recordings in examination, the database has been split into 477 tunes, with lengths ranging from 1 frame up to 174 frames (approximately 7 seconds). A bias towards Q1 exists in the database, as 42.98% of the tunes are classified to Q1, as shown in Table 3.

5.2 Statistical results

From the application of the proposed methodology on the data set annotated as ground truth we acquire a measurement of 81.55% for the system's accuracy. Specifically, 389 tunes were classified correctly, while 88 were misclassified. Clearly, this kind of information, although indicative, is not sufficient to fully comprehend and assess the performance of our methodology.

Table 3 Class distribution in the SAL dataset

	Neutral	Q1	Q2	Q3	Q4
Tunes	47	205	90	63	72
Percentages	9.85%	42.98%	18.87%	13.21%	15.09%

Table 4 Overall confusion matrix

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	34	1	5	3	0	43
Q1	1	189	9	12	6	217
Q2	4	3	65	2	1	75
Q3	4	6	7	39	3	59
Q4	4	6	4	7	62	83
Totals	47	205	90	63	72	477

Towards this end, we provide in Table 4 the confusion matrix for the experiment. In the table rows correspond to the ground truth and columns to the system's response. Thus, for example, there were 5 tunes that were labeled as neutral in the ground truth but were misclassified as belonging to Q2 by our system. Additionally, unimodal recognition rates for each of the two modalities were 72.19% and 74.39% for facial and acoustic features respectively.

Given the fact that our ground truth is biased towards Q1, we also provide in Table 5 the confusion matrix in the form of percentages so that the bias is removed from the numbers. There we can see that the proposed methodology performs reasonably well for most cases, with the exception of Q3, for which the classification rate is very low. What is more alarming is that more than 10% of the tunes of Q3 have been classified as belonging to the exactly opposite quadrant, which is certainly a major mistake.

Still, in our analysis of the experimental results so far we have not taken into consideration a very important factor: that of the length of the tunes. As we have explained in Sect. 5, in order for the Elman net to pick up the expression dynamics of the tune an adequate number of frames needs to be available as input. Still, there is a number of tunes in the ground truth that are too short for the network to reach a point where its output can be read with high confidence.

In order to see how this may have influence our results we present in the following separate confusion matrices for short (see Tables 8 and 9) and normal length tunes (see Tables 6 and 7). In this context we consider as normal tunes that comprise at least 10 frames (approximately 0.5 seconds) and as short tunes with length from 1 up to 9 frames.

First of all, we can see right away that the performance of the system, as was expected is quite different in these two cases. Specifically, there are 83 errors in just 131 short tunes while there are only 5 errors in 346 normal tunes. Moreover, there are no severe errors in the case of long tunes, i.e. there are no cases in which a tune is classified in the exact opposite quadrant than in the ground truth.

Overall, the operation of our system in normal operating conditions (as such we consider the case in which tunes have a length of at least 10 frames) is accompanied by a classification rate of 98.55%, which is certainly very high, even for controlled data, let alone for naturalistic data.

Table 5 Overall confusion matrix expressed in percentages

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	79.07%	2.33%	11.63%	6.98%	0.00%	100.00%
Q1	0.46%	87.10%	4.15%	5.53%	2.76%	100.00%
Q2	5.33%	4.00%	86.67%	2.67%	1.33%	100.00%
Q3	6.78%	10.17%	11.86%	66.10%	5.08%	100.00%
Q4	4.82%	7.23%	4.82%	8.43%	74.70%	100.00%
Totals	9.85%	42.98%	18.87%	13.21%	15.09%	100.00%

Of course, the question still remains of whether it is meaningful for the system to fail so much in the case of the short tunes, or if the information contained in them is sufficient for a considerably better performance and the system needs to be improved. In order to answer this question we will use Williams' index [7].

This index was originally designed to measure the joint agreement of several raters with another rater. Specifically, the index aims to answer the question: "Given a set of raters and one other rater, does the isolated rater agree with the set of raters as often as a member of that set agrees with another member in that set?" which makes it ideal for our application.

In our context, the examined rater is the proposed system. In order to have an adequate number of raters for the application of Williams' methodology we have asked three additional humans to manually classify the 131 short tunes into one of the five considered emotional categories. In our case, Williams' index for the system with respect to the four human annotators is reduced to the following:

The joint agreement between the reference annotators is defined as:

$$P_g = \frac{2}{4(4-1)} \sum_{a=1}^{3} \sum_{b=a+1}^{4} P(a,b)$$

where $P(a,b)$ is the proportion of observed agreements between annotators a and b

$$P(a,b) = \frac{|\{s \in S : R_a(s) = R_b(s)\}|}{131}$$

In the above S is the set of the 131 annotated tunes and $R_a(s)$ is the classification that annotator a gave for tune s. The observed overall group agreement of the system with the reference set of annotators is measured by

$$P_0 = \frac{\sum_{a=1}^{4} P(0,a)}{4}$$

where we use 0 to denote the examined annotator, i.e. our system. Williams' index for the system is the ratio:

$$I_0 = \frac{P_0}{P_g}$$

The value of I_0 can be interpreted as follows: Let a tune be selected at random and rated by a randomly selected reference annotator. This rating would agree with the system's rating at a rate $\frac{I_0}{100}$ percent of the rate that would be obtained by a second randomly selected reference annotator. Applying this methodology for the 131 short tunes in the ground truth data set with reference to the one original and 3 additional human annotators we have calculated $I_0 = 1, 12$. A rate of I_0 that is larger than one, as we have computed in our example, indicates that the system agrees with the human annotators more often than they agree with each other. As our system does not disagree with the human annotators more than they disagree with each other, we can conclude that the system performs at least as well as humans do in the difficult and uncertain task of classifying so short tunes. Consequently, the poor classification performance of the system in the case of short tunes is fully understandable and should not be taken as an indication of a systematic failure or weakness.

Table 6 Confusion matrix for normal tunes

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	29	0	0	0	0	29
Q1	0	172	3	0	0	175
Q2	1	1	54	0	0	56
Q3	0	0	0	30	0	30
Q4	0	0	0	0	56	56
Totals	30	173	57	30	56	346

Table 8 Confusion matrix for short tunes

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	5	1	5	3	0	14
Q1	1	17	6	12	6	42
Q2	3	2	11	2	1	19
Q3	4	6	7	9	3	29
Q4	4	6	4	7	6	27
Totals	17	32	33	33	16	131

Table 7 Confusion matrix for normal tunes expressed in percentages

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	100.00%	0.00%	0.00%	0.00%	0.00%	100.00%
Q1	0.00%	98.29%	1.71%	0.00%	0.00%	100.00%
Q2	1.79%	1.79%	96.43%	0.00%	0.00%	100.00%
Q3	0.00%	0.00%	0.00%	100.00%	0.00%	100.00%
Q4	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%
Totals	8.67%	50.00%	16.47%	8.67%	16.18%	100.00%

Table 9 Confusion matrix for short tunes expressed in percentages

	Neutral	Q1	Q2	Q3	Q4	Totals
Neutral	35.71%	7.14%	35.71%	21.43%	0.00%	100.00%
Q1	2.38%	40.48%	14.29%	28.57%	14.29%	100.00%
Q2	15.79%	10.53%	57.89%	10.53%	5.26%	100.00%
Q3	13.79%	20.69%	24.14%	31.03%	10.34%	100.00%
Q4	14.81%	22.22%	14.81%	25.93%	22.22%	100.00%
Totals	12.98%	24.43%	25.19%	25.19%	12.21%	100.00%

5.3 Quantitative comparative study

In a previous work we have proposed a different methodology to process naturalistic data with the goal of estimating the human's emotional state [40]. In that work a very similar approach is followed in the analysis of the visual component of the video with the aim of locating facial features. FAP values are then fed into a rule based system which provides a response concerning the human's emotional state.

In a later version of this work, we evaluate the likelihood of the detected regions being indeed the desired facial features with the help of anthropometric statistics acquired from [96] and produce degrees of confidence which are associated with the FAPs; the rule evaluation model is also altered and equipped with the ability to consider confidence degrees associated with each FAP in order to minimize the propagation of feature extraction errors in the overall result. When compared to our current work, these systems have the extra advantages of considering expert knowledge in the form of rules in the classification process being able to cope with feature detection deficiencies. On the other hand, they are lacking in the sense that they do not consider the dynamics of the displayed expression and they do not consider other modalities besides the visual one.

Thus, they make excellent candidates to compare our current work against in order to evaluate the practical gain from the proposed dynamic and multimodal approach. In Table 10 we present the results from the two former and the current approach. Since dynamics are not considered, each frame is treated independently in the preexisting systems. Therefore, statistics are calculated by estimating the number of correctly classified frames; each frame is considered to belong to the same quadrant as the whole tune.

It is worth mentioning that the results are from the parts of the data set that were selected as expressive for each methodology. But, whilst for the current work this refers to 72.54% of the data set and the selection criterion is the length of the tune, in the previous works only about 20% of the frames was selected with a criterion of the clarity with which the expression is observed, since frames close to the beginning or the end of the tune are often too close to neutral to provide meaningful visual input to a system. The performance of all approaches on the complete data set is presented in Table 11, where it is obvious that the dynamic and multimodal approach is by far superior.

Table 10 Classification rates on parts of the naturalistic data set

Methodology	Rule based	Possibilistic rule based	Dynamic and multimodal
Classification rate	78.4%	65.1%	98.55%

Table 11 Classification rates on the naturalistic data set

Methodology	Rule based	Possibilistic rule based	Dynamic and multimodal
Classification rate	27.8%	38.5%	81.55%

5.4 Qualitative comparative study

As we have already mentioned, during the recent years we have seen a very large number of publications in the field of the estimation of human expression and/or emotion. Although the vast majority of these works is focused on the six universal expressions and use sequences where extreme expressions are posed by actors, it would be an omission if not even a qualitative comparison was made to the broader state of the art.

In Table 12 we present the classification rates reported in some of the most promising and well known works in the current state of the art. Certainly, it is not possible or fair to compare numbers directly, since they come from the application on different data sets. Still, it is possible to make qualitative comparisons base on the following information:

The Cohen2003 is a database collected of subjects that were instructed to display facial expressions corresponding to the six types of emotions. In the Cohn–Kanade database subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units and combinations of action units.

In the MMI database subjects were asked to display 79 series of expressions that included either a single AU or a combination of a minimal number of AUs or a prototypic combination of AUs (such as in expressions of emotion). They were instructed by an expert (a FACS coder) on how to display the required facial expressions, and they were asked to include a short neutral state at the beginning and at the end of each expression. The subjects were asked to display the required expressions while minimizing out-of-plane head motions.

In the Chen2000 subjects were asked to display 11 different affect expressions. Every emotion expression sequence lasts from 2 s to 6 s with the average length of expression sequences being 4 s; even the shortest sequences in this dataset are by far longer than the short tunes in the SAL database.

The original instruction given to the actors has been taken as the actual displayed expression in all above mentioned databases, which means that there is an underlying assumption is that there is no difference between natural and acted expression.

As we can see, what is common among the datasets most commonly used in the literature for the evaluation of facial expression and/or emotion recognition is that expressions are solicited and acted. As a result, they are generally displayed clearly and to their extremes. In the case of natural human interaction, on the other hand, expressions are typically more subtle and often different expressions are mixed. Also, the element of speech adds an important degree of deformation to facial features which is not associated with the displayed expression and can be misleading for an automated expression analysis system.

Consequently, we can argue that the fact that the performance of the proposed methodology when applied to a naturalistic dataset is comparable to the performance of other works in the state of the art when applied to acted sequences is an indication of its success. Additionally, we can observe that when extremely short tunes are removed from the data set the classification performance of the proposed approach exceeds 98%, which, in current standards, is very high for an emotion recognition system.

The Multistream Hidden Markov Model approach is probably the one that is most directly comparable to the work presented herein. This is the alternative dynamic multimodal approach, where HMMs are used instead of RNNs for the modeling of the dynamics of expressions. Although a different data set has been utilized for the experimental evaluation of the MHMMs approach, the large margin between the performance rates of the two approaches indicates that the utilization of RNNs for dynamic multimodal classification of human emotion is a promising direction.

Table 12 Classification rates reported in the broader state of the art [11, 60, 97]

Methodology	Classification rate	Data set
TAN	83.31%	Cohen2003
Multi-level HMM	82.46%	Cohen2003
TAN	73.22%	Cohn–Kanade
PanticPatras2006	86.6%	MMI
Multistream HMM	72.42%	Chen2000
Proposed methodology	81.55%	SAL Database
Proposed methodology	98.55%	Partial SAL

6 Conclusions

In this work we have focused on the problem of human emotion recognition in the case of naturalistic, rather than acted and extreme, expressions. The main elements of our approach are that (i) we use multiple algorithms for the extraction of the “difficult” facial features in order to make the overall approach more robust to image processing errors, (ii) we focus on the dynamics of facial expressions rather than on the exact facial deformations they are associated with, thus being able to handle sequences in which the interaction is natural or naturalistic rather than posed or extreme and (iii) we follow a multimodal approach where audio and visual modalities are combined, thus enhancing both performance and stability of the system.

From a more technical point of view, our contributions include: (i) A modified input layer that allows the Elman net to process both dynamic and static inputs at the same time. This is used to fuse the fundamentally different visual and audio inputs in order to provide for a truly multimodal classification scheme. (ii) A modified output scheme that allows the Elman that integrates previous values, with value significance decreasing exponentially through time. This allows the network to display augmented stability. (iii) A modified mixture of experts module that, additionally to characteristics drawn from the experts’ input, can also draw information from the experts’ output in order to drive the output mediation step. This is used in order to incorporate the results from the statistical anthropometric evaluation of the acquired masks in the operation of the output combiner module.

Practical application of our methodology in a ground truth data set of naturalistic sequences has given a performance of 98.55% for tunes that are long enough for dynamics to be able to be picked up in both the visual and the audio channel.

For our future work, we intend to further extend our work in multimodal naturalistic expression recognition by considering more modalities such as posture and gestures and by incorporating uncertainty measuring and handling modules in order to maximize the system's performance and stability in difficult and uncontrolled environments.

References

1. Ai H, Litman D, Forbes-Riley K, Rotaru M, Tetreault J, Purandare A (2006) Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proceedings of interspeech ICSLP, Pittsburgh, PA
2. Ambady A, Rosenthal R (1992) Thin slices of expressive B predictors of interpersonal consequences: a meta-analysis. Psychol Bull 111(2):256–274
3. Ang J, Dhilon R, Krupski A, Shriberg E, Stolcke A (2002) Prosody based automatic detection of annoyance and frustration in human computer dialog. In: Proc of ICSLP, pp 2037–2040
4. Ashraf AB, Lucey S, Cohn JF, Chen T, Ambadar Z, Prkachin K, Solomon P, Theobald BJ (2007) The painful face: pain expression recognition using active appearance models. In: Proc ninth ACM int'l conf multimodal interfaces (ICMI'07), pp 9–14
5. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: Proc IEEE int'l conf computer vision and pattern recognition (CVPR'05), pp 568–573
6. Batliner A, Fischer K, Huber R, Spilker J, Noeth E (2003) How to find trouble in communication. Speech Commun, 40:117–143
7. Batliner A, Huber R, Niemann H, Noeth E, Spilker J, Fischer K (2000) The recognition of emotion. In: Wahlster W: Verbmobil: foundations of speech-to-speech translations. Springer, New York, pp 122–130
8. Bertolami R, Bunke H Early feature stream integration versus decision level combination in a multiple classifier system for text line recognition. In: 18th international conference on pattern recognition (ICPR'06)
9. Busso C et al. (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proc sixth ACM int'l conf multimodal interfaces (ICMI'04), pp 205–211
10. Caridakis G, Malatesta L, Kessous L, Amir N, Paouzaiou A, Karpouzis K (2006) Modeling naturalistic affective states via facial and vocal expression recognition. In: Proc eighth ACM int'l conf multimodal interfaces (ICMI'06), pp 146–154
11. Cohen I, Sebe N, Garg A, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: temporal and static modeling. Comput Vis Image Underst 91:160–187
12. Cohen PR (2001) Multimodal interaction: a new focal area for AI. In: IJCAI, pp 1467–1473
13. Cohen PR, Johnston M, McGee D, Oviatt S, Clow J, Smith I (1998) The efficiency of multimodal interaction: A case study. In: Proceedings of international conference on spoken language processing, ICSLP'98, Australia
14. Cohn JF, Schmidt KL (2004) The timing of facial motion in posed and spontaneous smiles. Int J Wavelets Multiresolut Inf Process 2:1–12
15. Cohn JF (2006) Foundations of human computing: facial expression and emotion. In: Proc eighth ACM int'l conf multimodal interfaces (ICMI'06), pp 233–238
16. Cootes T, Edwards G, Taylor C (2001) Active appearance models. IEEE PAMI 23(6):681–685
17. Cowie R, Cornelius R (2003) Describing the emotional states that are expressed in speech. Speech Commun 40:5–32
18. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. IEEE Signal Process Mag
19. Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schroder M (2000) 'FeelTrace': an instrument for recording perceived emotion in real time. In: Proceedings of ISCA workshop on speech and emotion, pp 19–24
20. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. In: IEEE Signal Process Mag, 32–80
21. De Silva LC, Ng PC (2000) Bimodal emotion recognition. In: Proc face and gesture recognition conf, pp 332–335
22. Devillers L, Vidrascu L (2007) Real-life emotion recognition human-human call center data with acoustic and lexical cues. In: Müller C, Schötz S (eds) Speaker characterization. Springer, Berlin (to appear)
23. Duric Z, Gray WD, Heishman R, Li F, Rosenfeld A, Schoelles MJ, Schunn C, Wechsler H (2002) Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. In: Proc IEEE, vol 90(7), pp 1272–1289
24. Ekman P, Friesen WV (1978) The facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, San Francisco
25. Ekman P, Friesen WV (1982) Felt, false, and miserable smiles. J Nonverbal Behav, 6:238–252
26. Ekman P, Rosenberg EL (2005) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system, 2nd edn. Oxford University Press, London
27. Ekman P (1993) Facial expression and emotion. Am Psychol 48:384–392
28. Elman JL (1990) Finding structure in time. Cogn Sci 14:179–211
29. Elman JL (1991) Distributed representations, simple recurrent networks, and grammatical structure. Mach Learn 7:195–224
30. Essa IA, Pentland AP (1997) Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 19(7):757–763
31. Fasel B, Luttin J (2003) Automatic facial expression analysis: survey. Pattern Recogn 36(1):259–275
32. Fragopanagos N, Taylor JG (2005) Emotion recognition in human computer interaction. Neural Netw 18:389–405
33. Frank MG, Ekman P (1993) Not all smiles are created equal: differences between enjoyment and other smiles. Humor: Int J Res Humor 6:9–26
34. Go HJ, Kwak KC, Lee DJ, Chun MG (2003) Emotion recognition from facial image and speech signal. In: Proc int'l conf soc of instrument and control engineers, pp 2890–2895
35. Gunes H, Piccardi M (2005) Fusing face and body gesture for machine recognition of emotions. In: 2005 IEEE international workshop on robots and human interactive communication, pp 306–311
36. Hammer A, Tino P (2003) Recurrent neural networks with small weights implement definite memory machines. Neural Comput 15(8):1897–1929
37. Haykin S (1999) Neural networks: a comprehensive foundation. Prentice Hall, New York
38. Hoch S, Althoff F, McGlaun G, Rigoll G (2005) Bimodal fusion of emotional data in an automotive environment. In: Proc 30th int'l conf acoustics, speech, and signal processing (ICASSP '05), vol II, pp 1085–1088
39. http://emotion-research.net/toolbox/toolboxdatabase Humaine
40. Ioannou S, Raouzaiou A, Tzouvaras V, Mailis T, Karpouzis K, Kollias S (2005) Emotion recognition through facial expression analysis based on a neurofuzzy network. Neural Netw 18(4):423–435. Special issue on emotion: understanding & recognition
41. Ioannou S, Caridakis G, Karpouzis K, Kollias S (2007) Robust feature detection for facial expression recognition. EURASIP J Image Video Process 2007(2)

42. Jaimes A, Sebe N (2005) Multimodal human computer interaction: a survey. In: IEEE international workshop on human computer interaction, ICCV 2005, Beijing, China
43. Jaimes A (2006) Human-centered multimedia: culture, deployment, and access. IEEE Multimedia Mag 13(1)
44. Kapoor A, Picard RW, Ivanov Y (2004) Probabilistic combination of multiple modalities to detect interest. In: Proc of IEEE ICPR
45. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Human-Comput Stud 65(8):724–736
46. Karpouzis K, Caridakis G, Kessous L, Amir N, Raouzaiou A, Malatesta L, Kollias S (2007) Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In: Huang T, Nijholt A, Pantic M, Pentland A (eds) Lecture notes in artificial intelligence, vol 4451. Springer, Berlin. pp 91–112. Special Volume on AI for Human Computing
47. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303
48. Leung SH, Wang SL, Lau WH (2004) Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. IEEE Trans Image Process 13(1)
49. Lisetti CL, Nasoz F (2002) MAUI: a multimodal affective user interface. In: Proc 10th ACM int'l conf multimedia (Multimedia '02), pp 161–170
50. Littlewort G, Bartlett MS, Fasel I, Susskind J, Movellan J (2006) Dynamics of facial expression extracted automatically from video. Image Vis Comput 24:615–625
51. Littlewort GC, Bartlett MS, Lee K (2007) Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In: Proc ninth ACM int'l conf multimodal interfaces (ICMI'07), pp 15–21
52. Maat L, Pantic M (2006) Gaze-X: adaptive affective multimodal interface for single-user office scenarios. In: Proc eighth ACM int'l conf multimodal interfaces (ICMI'06), pp 171–178
53. Mehrabian A (1968) Communication without words. Psychol. Today 2(4):53–56
54. Mertens P (2004) The prosogram: semi-automatic transcription of prosody based on a tonal perception model. In: Bel B, Marlien I (eds) Proc of speech Prosody, Japan
55. Neiberg D, Elenius K, Karlsson I, Laskowski K (2006) Emotion recognition in spontaneous speech. In: Proceedings of fonetik 2006, pp 101–104
56. Oudeyer PY (2003) The production and recognition of emotions in speech: features and algorithms. Int J Human-Comput Interact 59(1–2):157–183
57. Oviatt S (1999) Ten myths of multimodal interaction. Commun ACM 42(11):74–81
58. Oviatt S, DeAngeli A, Kuhn K (1997) Integration and synchronization of input modes during multimodal human-computer interaction. In: Proceedings of conference on human factors in computing systems CHI'97. ACM, New York, pp 415–422
59. Pal P, Iyer AN, Yantorno RE (2006) Emotion detection from infant facial expressions and cries. In: Proc IEEE int'l conf acoustics, speech and signal processing (ICASSP'06), vol 2, pp 721–724
60. Pantic M, Patras I (2006) Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans Syst Man Cybern, Part B 36(2):433–449
61. Pantic M, Rothkrantz LJM (2003) Towards an affect-sensitive multimodal human-computer interaction. Proc IEEE 91(9):1370–1390
62. Pantic M, Bartlett MS (2007) Machine analysis of facial expressions. In: Delac K, Grgic M (eds) Face recognition, I-Tech Education and Publishing, pp 377–416
63. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
64. Pantic M, Rothkrantz LJM (2003) Toward an affect-sensitive multimodal human-computer interaction. Proc IEEE 91(9):1370–1390
65. Pantic M, Rothkrantz LJM (2000) Expert system for automatic analysis of facial expressions. Image Vis Comput 18:881–905
66. Pantic M (2005) Face for interface. In: Pagani M (ed) The encyclopedia of multimedia technology and networking. Idea Group Reference, Hershey, vol 1, pp 308–314
67. Pantic M, Sebe N, Cohn JF, Huang T (2005) Affective multimodal human-computer interaction. In: Proc 13th ACM int'l conf multimedia (Multimedia '05), pp 669–676
68. Pentland A (2005) Socially aware computation and communication. Computer 38(3):33–40
69. Petridis S, Pantic M (2008) Audiovisual discrimination between laughter and speech. In: IEEE int'l conf acoustics, speech, and signal processing (ICASSP), pp 5117–5120
70. Picard RW (1997) Affective computing. MIT Press, Cambridge
71. Picard RW (2000) Towards computers that recognize and respond to user emotion. IBM Syst J 39(3–4):705–719
72. Rogozan A (1999) Discriminative learning of visual data for audiovisual speech recognition. Int J Artif Intell Tools 8:43–52
73. Russell JA, Mehrabian A (1977) Evidence for a three-factor theory of emotions. J Res Pers 11:273–294
74. Russell JA, Bachorowski J, Fernandez-Dols J (2003) Facial and vocal expressions of emotion. Ann Rev Psychol 54:329–349
75. Samal A, Iyengar PA (1992) Automatic recognition and analysis of human faces and facial expressions: a survey. Pattern Recogn 25(1):65–77
76. Sander D, Grandjean D, Scherer KR (2005) A system approach to appraisal mechanisms in emotion. Neural Netw 18:317–352
77. Schaefer M, Zimmermann HG (2006) Recurrent neural networks are universal approximators, ICANN 2006, pp 632–640
78. Scherer KR (1999) Appraisal theory. In: Dalgleish T, Power MJ (eds) Handbook of cognition and emotion, pp 637–663. Wiley, New York
79. Schlosberg H (1954) A scale for judgment of facial expressions. J Exp Psychol 29:497–510
80. Sebe N, Lew MS, Cohen I, Sun Y, Gevers T, Huang TS (2004) Authentic facial expression analysis. In: International conference on automatic face and gesture recognition (FG'04), Seoul, Korea, May 2004, pp 517–522
81. Sebe N, Cohen I, Huang TS (2005) Multimodal emotion recognition. Handbook of pattern recognition and computer vision. World Scientific, Singapore
82. Sebe N, Cohen I, Gevers T, Huang TS (2006) Emotion recognition based on joint visual and audio cues. In: Proc 18th int'l conf pattern recognition (ICPR'06), pp 1136–1139
83. Song M, Bu J, Chen C, Li N (2004) Audio-visual-based emotion recognition: a new approach. In: Proc int'l conf computer vision and pattern recognition (CVPR'04), pp 1020–1025
84. Teissier P, Robert-Ribes J, Schwartz JL (1999) Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Trans Speech Audio Process 7:629–642
85. Tekalp M, Ostermann J (2000) Face and 2-D mesh animation in MPEG-4. Signal Process Image Commun 15:387–421
86. Tian YL, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans PAMI 23(2)
87. Tian YL, Kanade T, Cohn JF (2005) Facial expression analysis. In: Li SZ, Jain AK (eds) Handbook of face recognition, pp 247–276. Springer, Berlin
88. Tomasi C, Kanade T (1991) Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991

89. Valstar MF, Gunes H, Pantic M (2007) How to distinguish posed from spontaneous smiles using geometric features. In: Proc ninth ACM int'l conf multimodal interfaces (ICMI'07), pp 38–45
90. Valstar M, Pantic M, Ambadar Z, Cohn JF (2006) Spontaneous versus posed facial behavior: automatic analysis of Brow actions. In: Proc eight int'l conf multimodal interfaces (ICMI'06), pp 162–170
91. Wang Y, Guan L (2005) Recognizing human emotion from audiovisual information. In: Proc int'l conf acoustics, speech, and signal processing (ICASSP '05), pp 1125–1128
92. Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–35
93. Whissel CM (1989) The dictionary of affect in language. In: Plutchnik R, Kellerman H (eds) Emotion: theory, research and experience: the measurement of emotions. Academic Press, New York, vol 4, pp 113–131
94. Williams A (2002) Facial expression of pain: an evolutionary account. Behav Brain Sci 25(4):439–488
95. Wu L, Oviatt SL, Cohen PR (1999) Multimodal integration—a statistical view. IEEE Trans Multimedia 1(4)
96. Young JW (1993) Head and face anthropometry of adult U.S. civilians. FAA Civil Aeromedical Institute, 1963–1993 (final report 1993)
97. Zeng Z, Tu J, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S (2007) Audio-visual affect recognition. IEEE Trans Multimedia 9(2):424–428
98. Zeng Z, Tu J, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S (2007) Audio-visual affect recognition. IEEE Trans Multimedia 9(2)
99. Zeng Z, Hu Y, Liu M, Fu Y, Huang TS (2006) Training combination strategy of multi-stream fused hidden Markov model for audio-visual affect recognition. In: Proc 14th ACM int'l conf multimedia (Multimedia'06), pp 65–68
100. Zeng Z, Hu Y, Roisman GI, Wen Z, Fu Y, Huang TS (2007) Audio-visual spontaneous emotion recognition. In: Huang TS, Nijholt A, Pantic M, Pentland A (eds) Artificial intelligence for human computing, pp 72–90. Springer, Berlin
101. Zeng Z, Pantic M, Roisman GI, Huang TS (2007) A survey of affect recognition methods: audio, visual, and spontaneous expressions. In: Proc ninth ACM int'l conf multimodal interfaces (ICMI'07), pp 126–133
102. Zeng Z, Tu J, Liu M, Zhang T, Rizzolo N, Zhang Z, Huang TS, Roth D, Levinson S (2004) Bimodal HCI-related emotion recognition. In: Proc sixth ACM int'l conf multimodal interfaces (ICMI'04), pp 137–143
103. Zeng Z, Tu J, Pianfetti P, Liu M, Zhang T, Zhang Z, Huang TS, Levinson S (2005) Audio-visual affect recognition through multi-stream fused HMM for HCI. In: Proc IEEE int'l conf computer vision and pattern recognition (CVPR'05), pp 967–972
104. Zeng Z, Tu J, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S (2007) Audio-visual affect recognition. IEEE Trans Multimedia 9(2):424–428