Multimodal emotion recognition from low-level cues

Maja Pantic, George Caridakis, Elisabeth André, Jonghwa Kim, Kostas Karpouzis, Stefanos Kollias

Angaben zur Veröffentlichung / Publication details:

Pantic, Maja, George Caridakis, Elisabeth André, Jonghwa Kim, Kostas Karpouzis, and Stefanos Kollias. 2010. "Multimodal emotion recognition from low-level cues." In *Emotion-Oriented Systems*, edited by Roddy Cowie, Catherine Pelachaud, and Paolo Petta, 115–32. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-642-15184-2_8.

Multimodal Emotion Recognition from Low-Level Cues

Maja Pantic, George Caridakis, Elisabeth André, Jonghwa Kim, Kostas Karpouzis, and Stefanos Kollias

Abstract Emotional intelligence is an indispensable facet of human intelligence and one of the most important factors for a successful social life. Endowing machines with this kind of intelligence towards affective human–machine interaction, however, is not an easy task. It becomes more complex with the fact that human beings use several modalities jointly to interpret affective states, since emotion affects almost all modes – audio-visual (facial expression, voice, gesture, posture, etc.), physiological (respiration, skin temperature, etc.), and contextual (goal, preference, environment, social situation, etc.) states. Compared to common unimodal approaches, many specific problems arise from the case of multimodal emotion recognition, especially concerning fusion architecture of the multimodal information. In this chapter, we firstly give a short review for the problems and then present research results of various multimodal architectures based on combined analysis of facial expression, speech, and physiological signals. Lastly we introduce designing of an adaptive neural network classifier that is capable of deciding the necessity of adaptation process in respect of environmental changes.

1 Human Affect Sensing: The Problem Domain

The ability to detect and understand affective states and other social signals of someone with whom we are communicating is the core of social and emotional intelligence. This kind of intelligence is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life (Goleman, 1995). When it comes to computers, however, they are socially ignorant (Pelachaud et al., 2002). Current computing technology does not account for the fact that human–human communication is always socially situated and that

M. Pantic (✉)
Department of Computing, Imperial College, London, UK; Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands
e-mail: M.Pantic@imperial.ac.uk

discussions are not just facts but part of a larger social interplay. Not all computers will need social and emotional intelligence and none will need all of the related skills humans have. Yet, human–machine interactive systems capable of sensing stress, inattention, confusion, and heedfulness and capable of adapting and responding to these affective states of users are likely to be perceived as more natural, efficacious, and trustworthy (Picard, 1997; Picard, 2003; Pantic, 2005). For example, in education, pupils' affective signals inform the teacher of the need to adjust the instructional message. Successful human teachers acknowledge this and work with it; digital conversational embodied agents must begin to do the same by employing tools that can accurately sense and interpret affective signals and social context of the pupil, learn successful context-dependent social behaviour, and use a proper affective presentation language (e.g. Pelachaud et al., 2002) to drive the animation of the agent. Automatic recognition of human affective states is also important for video surveillance. Automatic assessment of boredom, inattention, and stress would be highly valuable in situations in which firm attention to a crucial but perhaps tedious task is essential (Pantic, 2005; Pantic et al., 2005). Examples include air traffic control, nuclear power plant surveillance, and operating a motor vehicle. An automated tool could provide prompts for better performance informed by assessment of the user's affective state. Other domain areas in which machine tools for analysis of human affective behaviour could expand and enhance scientific understanding and practical applications include specialized areas in professional and scientific sectors (Ekman et al., 1993). In the security sector, affective behavioural cues play a crucial role in establishing or detracting from credibility. In the medical sector, affective behavioural cues are a direct means to identify when specific mental processes are occurring. Machine analysis of human affective states could be of considerable value in these situations in which only informal, subjective interpretations are now used. It would also facilitate research in areas such as behavioural science (in studies on emotion and cognition), anthropology (in studies on cross-cultural perception and production of affective states), neurology (in studies on dependence between emotion dysfunction or impairment and brain lesions), and psychiatry (in studies on schizophrenia and mood disorders) in which reliability, sensitivity, and precision of measurement of affective behaviour are persisting problems.

While all agree that machine sensing and interpretation of human affective information would be widely beneficial, addressing these problems is not an easy task. The main problem areas can be defined as follows:

- *What is an affective state?* This question is related to psychological issues pertaining to the nature of affective states and the best way to represent them.
- *Which human communicative signals convey information about affective state?* This issue shapes the choice of different modalities to be integrated into an automatic analyzer of human affective states.
- *How are various kinds of evidence to be combined to optimize inferences about affective states?* This question is related to how best to integrate information across modalities for emotion recognition.

In this section, we briefly discuss each of these problem areas in the field. The rest of the chapter is dedicated to a specific domain within the second problem area – sensing and processing visual cues of human affective displays.

1.1 What Is an Affective State?

Traditionally, the terms "affect" and "emotion" have been used synonymously. Following Darwin, discrete emotion theorists propose the existence of six or more basic emotions that are universally displayed and recognized (Ekman and Friesen, 1969; Keltner and Ekman, 2000). These include happiness, anger, sadness, surprise, disgust, and fear. Data from both Western and traditional societies suggest that nonverbal communicative signals (especially facial and vocal expression) involved in these basic emotions are displayed and recognized cross-culturally. In opposition to this view, Russell (1994) among others argues that emotion is best characterized in terms of a small number of latent dimensions, rather than in terms of a small number of discrete emotion categories. Russell proposes bipolar dimensions of arousal and valence (pleasant versus unpleasant). Watson and Tellegen propose unipolar dimensions of positive and negative affect, while Watson and Clark proposed a hierarchical model that integrates discrete emotions and dimensional views (Larsen and Diener, 1992; Watson et al., 1995a, 1995b). Social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations. They argue further that emotion is culturally constructed and no universals exist. From their perspective, subjective experience and whether or not emotion is better conceptualized categorically or dimensionally is culture specific. Then there is lack of consensus on how affective displays should be labelled. For example, Fridlund argues that human facial expressions should not be labelled in terms of emotions but in terms of behavioural ecology interpretations, which explain the influence a certain expression has in a particular context (Fridlund, 1997). Thus, an "angry" face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. Yet, people still tend to use *anger* as the interpretation rather than *readiness-to-attack* interpretation. Another issue is that of culture dependency; the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent (Matsumoto, 1990; Watson et al., 1995a). In summary, previous research literature pertaining to the nature and suitable representation of affective states provides no firm conclusions that could be safely presumed and adopted in studies on machine analysis of human affective states and affective computing. Also, not only discrete emotional states like surprise or anger are of importance for the realization of proactive human–machine interactive systems, but also sensing and responding to behavioural cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signalling like empathy and antipathy are essential (Pantic et al., 2006). Hence, in contrast to traditional approach, we treat affective states as being correlated not only to discrete emotions but to other, aforementioned social signals as well. Furthermore, since it is not certain that each of us will express a particular affective state by modulating the same communicative

signals in the same way nor is it certain that a particular modulation of interactive cues will be interpreted always in the same way independently of the situation and the observer, we advocate that pragmatic choices (e.g. application- and user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback (Pantic and Rothkrantz, 2003; Pantic et al., 2005, 2006).

1.2 Which Human Behavioural Cues Convey Information About Affective State?

Affective arousal modulates all human communicative signals (Ekman and Friesen, 1969). However, the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioural cues (Ambady and Rosenthal, 1992). Human judges seem to be most accurate in their judgment when they are able to observe the face and the body. Ratings that were based on the face and the body were 35% more accurate than the ratings that were based on the face alone. Yet, ratings that were based on the face alone were 30% more accurate than ratings that were based on the body alone and 35% more accurate than ratings that were based on the tone of voice alone (Ambady and Rosenthal, 1992). These findings indicate that to interpret someone's behavioural cues, people rely on shown facial expressions and to a lesser degree on shown body gestures and vocal expressions. However, although basic researchers have been unable to identify a set of voice cues that reliably discriminate among emotions, listeners seem to be accurate in decoding emotions from voice cues (Juslin and Scherer, 2005). Thus, automated human affect analyzers should at least include facial expression modality and preferably they should also include (one or both) modalities for perceiving body gestures and tone of the voice. Finally, while too much information from different channels seem to be confusing to human judges, resulting in less accurate judgments of shown behaviour when three or more observation channels are available (e.g. face, body, and speech) (Ambady and Rosenthal, 1992), combining those multiple modalities (including speech and physiology) may prove appropriate for realization of automatic human affect analysis.

1.3 How Are Various Kinds of Evidence to Be Combined to Optimize Inferences About Affective States?

Humans simultaneously employ the tightly coupled modalities of sight, sound, and touch. As a result, analysis of the perceived information is highly robust and flexible. Thus, in order to accomplish a multimodal analysis of human behavioural signals acquired by multiple sensors, which resembles human processing of such information, input signals should not be considered mutually independent and should not be combined only at the end of the intended analysis as the majority of current studies do. The input data should be processed in a joint feature space and according to a

context-dependent model (Pantic and Rothkrantz, 2003). The latter refers to the fact that one must know the context in which the observed behavioural signals have been displayed (who the expresser is, what his or her current environment and task are, when and why did he or she display the observed behavioural signals) in order to interpret the perceived multi-sensory information correctly (Pantic et al., 2006).

2 Classification and Fusion Approaches

2.1 Short-Term, Low-Level Multimodal Fusion

The term multimodal has been used in many contexts and across several disciplines. In the context of emotion recognition, a multimodal system is simply one that responds to inputs in more than one modality or communication channel (e.g. face, gesture, and speech prosody in our case, writing, body posture, linguistic content, and others) (Kim and André, 2006; Pantic, 2005). Jaimes and Sebe use a human-centred approach in this definition; by modality we mean mode of communication according to human senses or type of computer input devices. In terms of human senses, the categories are sight, touch, hearing, smell, and taste. In terms of computer input devices, we have modalities that are equivalent to human senses: cameras (sight), haptic sensors (touch), microphones (hearing), olfactory (smell), and even taste (Taylor and Fragopanagos, 2005). In addition, however, there are input devices that do not map directly to human senses: keyboard, mouse, writing tablet, motion input (e.g. the device itself is moved for interaction), and many others.

Various multimodal fusion techniques are possible (Zeng et al., 2009). Feature-level fusion can be performed by merging extracted features from each modality into one cumulative structure and feeding them to a single classifier, generally based on multiple hidden Markov models or neural networks. In this framework, correlation between modalities can be taken into account during classifier learning. In general, feature fusion is more appropriate for closely coupled and synchronized modalities, such as speech and lip movements, but tends not to generalize very well if modalities differ substantially in the temporal characteristics of their features, as is the case between speech and facial expression or gesture inputs. Moreover, due to the high dimensionality of input features, large amounts of data must be collected and labelled for training purposes.

Taylor and Fragopanagos describe a neural network architecture in Taylor and Fragopanagos (2004, 2005) in which features, from various modalities, that correlate with the user's emotional state are fed to a hidden layer, representing the emotional content of the input message. The output is a label of this state. Attention acts as a feedback modulation onto the feature inputs, so as to amplify or inhibit the various feature inputs, as they are or are not useful for the emotional state detection. The basic architecture is thus based on a feedforward neural network, but with the addition of a feedback layer (IMC in Fig. 1 below), modulating the activity in the inputs to the hidden layer.

Results have been presented for the success levels of the trained neural system based on a multimodal database, including time series streams of text (from

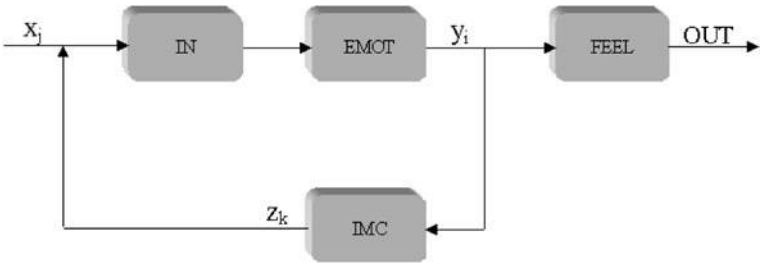

Fig. 1 Information flow in a multimodal emotion recognizer. IMC, inverse model controller; EMOT, hidden layer emotional state; FEEL, output state emotion classifier

an emotional dictionary), prosodic features (as determined by a prosodic speech feature extraction), and facial features (facial animation parameters). The obtained results are different for different viewers who helped to annotate the data sets. These results show high success levels on certain viewers while lower (but still good) levels on other ones. In particular, very high success was obtained using only prediction of activation values for one user who seemed to use mainly facial cues, whilst a similar, but slightly lower success level was obtained on an annotator who used predominantly prosodic cues.

Other two annotators appeared to use cues from all modalities, and for them, the success levels were still good but not so outstanding. This leads to the need for a further study to follow-up the spread of such cue extraction across the populace, since if this is an important component, then it would be important to know how broad is this spread, as well as to develop ways to handle such a spread (such as having a battery of networks, each trained on the appropriate subset of cues). It is, thus evident that adaptation to specific users and contexts is a crucial aspect in this type of fusion.

Decision-level fusion caters for integrating asynchronous but temporally correlated modalities. Here, each modality is first classified independently and the final classification is based on fusion of the outputs of the different modalities. Designing optimal strategies for decision-level fusion is still an open research issue. Various approaches have been proposed, e.g. sum rule, product rule, using weights, max/min/median rule, and majority vote. As a general rule, semantic fusion builds on individual recognizers, followed by an integration process; individual recognisers can be trained using unimodal data, which are easier to collect.

3 Cases of Multimodal Analysis

3.1 Recognition from Speech and Video Features

Visual sources can provide significant information about human communication activities. In particular, lip movement captured by stationary and steerable cameras can verify or detect that a particular person is speaking and help improve speech

recognition accuracy. The proposed approach is similar to human lip reading and consists of adding features like lip motion and other visual speech cues as additional inputs for recognition. This process is known as speech reading (Luettin et al., 1996; Potamianos et al., 2003), where most audio-visual speech recognition approaches consider the visual channel as a parallel and symmetric information source to the acoustic channel, resulting in the visual speech information being captured explicitly through the joint training of audio-visual phonetic models. As a result, in order to build a high-performance recognition system, large collections of audio-visual speech data are required. An alternative to the fusion approach is to use the visual and acoustic information in an asymmetric manner, where the tight coupling between auditory and visual speech in the signal domain is exploited and the visual cues used to help separate the speech of the target speaker from background speech and other acoustic events. Note that in this approach the visual channel is considered only up to the signal processing stage, and only the separated acoustic source is passed on to the statistical modelling level. In essence, the visual speech information here is used implicitly through the audio channel enhancement. This approach permits flexible and scalable deployment of audio-visual speech technology.

In the case of multimodal natural interaction (Caridakis et al., 2006), authors used earlier recordings during the FP5 IST Ermis project (FP5 IST ERMIS, 2007), where emotion induction was performed using the SAL approach. This material was labelled using FeelTrace (Cowie et al., 2000) by four labellers. The activation valence coordinates from the four labellers were initially clustered into quadrants and were then statistically processed so that a majority of decision could be obtained about the unique emotion describing the given moment. The corpus under investigation was segmented into 1,000 tunes of varying length. For every tune, the facial feature input vector consisted of the FAPs produced by the processing of the frames of the tune, while the acoustic input vector consisted of only one value per SBPF (segment-based prosodic feature) per tune. The fusion was performed on a frame basis, meaning that the values of the SBPFs were repeated for every frame of the tune. This approach was preferred because it preserved the maximum of the available information since SBPFs are meaningful only for a certain time period and cannot be calculated per frame.

In order to model the dynamic nature of facial expressivity, authors employed RNNs (recurrent neural networks – Fig. 2), where past inputs influence the processing of future inputs (Elman, 1990). RNNs possess the nice feature of modelling explicitly time and memory, catering for the fact that emotional states are not fluctuating strongly, given a short period of time. Additionally, they can model emotional transitions and not only static emotional representations, providing a solution for diverse feature variation and not merely for neutral to expressive and back to neutral, as would be the case for HMMs. The implementation of a RNN was based on an Elman network, with four output classes (three for the possible emotion quadrants, since the data for the positive/passive quadrant was negligible, and one for neutral affective state) resulting in a data set consisting of around 10,000 records. To cater for the fact that facial features are calculated per frame while speech prosody features are constant per tune, authors maintain the conventional input neurons met in

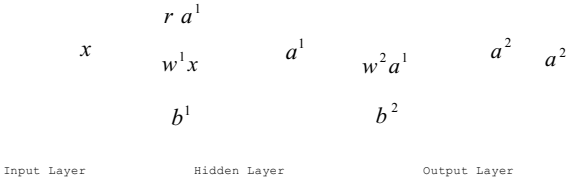

Fig. 2 Structure and functionality of a recursive neural network

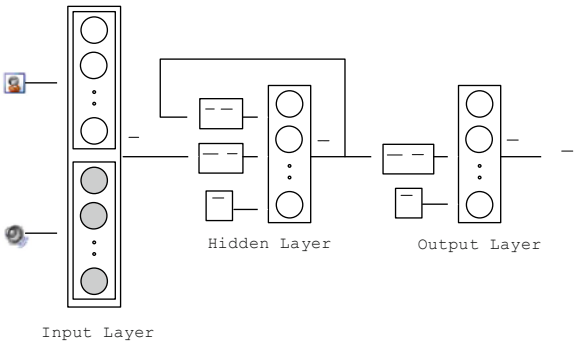

Fig. 3 Modified version of the Elman network

all neural networks, while for the auditory modality features, they use static value neurons (modified version shown in Fig. 3). The classification efficiency, for facial only and audio only, was measured at 67 and 73%, respectively, but combining the two modalities resulted in a recognition rate of 79%. This fact illustrates the ability of the proposed method to take advantage of multimodal information and the related analysis. After further processing by removing very short tunes (less than 10 frames or half a second), recognition rates in the naturalistic database rise to 98.55%.

3.2 Recognition from Physiological and Speech Features

Kim and Andre in Kim et al. (2005) and Kim and André (2006) studied various methods for fusing physiological and voice data at the feature level and the decision level, as well as a hybrid integration scheme. The results of the integrated recognition approach were then compared with the individual recognition results from each modality using the multimodal corpus of speech and physiological data we recorded within the FP6 NoE Humaine for three subjects. After synchronized segmentation of bimodal signals, we obtained a total of 138 features, 77 features from the five-channel biosignals (EMG, BVP, SC, RSP, Temp), and 61 features from the speech segments.

Feature-level fusion is performed by merging the calculated features from each modality into one cumulative structure, selecting the relevant features, and feeding

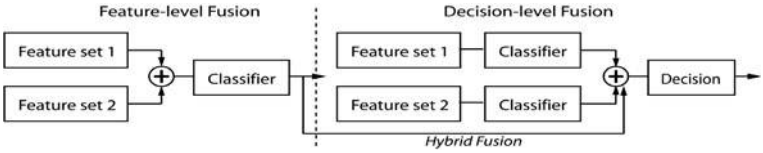

Fig. 4 Feature, decision and hybrid fusion of affective cues

them to a classifier. Decision-level fusion caters for integrating asynchronous but temporally correlated modalities (Fig. 4). Each modality is first classified independently by the classifier, and the final decision is obtained by fusing the output from the modality-specific classification processes. Three criteria, maximum, average, and product, were applied to evaluate the posterior probabilities of the unimodal classifiers at the decision stage. As a further variation of decision-level fusion, we employed a new hybrid scheme of the two fusion methods in which the output of feature-level fusion is also fed as an auxiliary input to the decision-level fusion stage. In Table 1 the best results are summarized that we achieved by the classification schemes described above. We classified the bimodal data subject

Table 1 Recognition results in rates (1.0 = 100% accuracy) achieved by using SBS, LDA, and leave-one-out cross-validation

System	High/pos	High/neg	Low/neg	Low/pos	*Average*
Subject A					
Biosignal	0.95	0.92	0.86	0.85	0.90
Speech signal	0.64	0.75	0.67	0.78	0.71
Feature fusion	0.91	0.92	1.00	0.85	**0.92**
Decision fusion	0.64	0.54	0.76	0.67	0.65
Hybrid fusion	0.86	0.54	0.57	0.59	0.64
Subject B					
Biosignal	0.50	0.79	0.71	0.45	0.61
Speech signal	0.76	0.56	0.74	0.72	0.70
Feature fusion	0.71	0.56	0.94	0.79	**0.75**
Decision fusion	0.59	0.68	0.82	0.69	0.70
Hybrid fusion	0.65	0.64	0.82	0.83	0.73
Subject C					
Biosignal	0.52	0.79	0.70	0.52	0.63
Speech signal	0.55	0.77	0.66	0.71	0.67
Feature fusion	0.50	0.67	0.84	0.74	**0.69**
Decision fusion	0.32	0.77	0.74	0.64	0.62
Hybrid fusion	0.40	0.73	0.86	0.71	0.68
All: subject independent					
Biosignal	0.43	0.53	0.54	0.52	0.51
Speech signal	0.40	0.53	0.70	0.53	0.54
Feature fusion	0.46	0.57	0.63	0.56	**0.55**
Decision fusion	0.34	0.50	0.70	0.54	0.52
Hybrid fusion	0.41	0.51	0.70	0.55	0.54

dependently (subjects A, B, and C) and subject independently (All) since this gave us a deeper insight on what terms the multimodal systems could improve the results of unimodal emotion recognition. We performed both feature-level fusion and decision-level fusion using LDA (linear discriminant analysis) in combination with SBS (sequential backward searching).

The results show that the performance of the unimodal systems varies not only from subject to subject but also for the single modalities. During our experiment, we could observe individual differences in the physiological and vocal expressions of the three test subjects. As shown in Table 1, the emotions of user A were more accurately recognized by using biosignals (90%) than by his voice (71%), whereas it was inverse for users B and C (70 and 67% for voice and 61 and 63% for biosignals). In particular for subject A, the difference between the accuracies of the two modalities is sizable. However, no suggestively dominant modality could be observed in the results of subject-dependent classification in general, which may be used as a decision criterion in the decision-level fusion process to improve the recognition accuracy. Overall, we obtained the best results for feature-level fusion. For instance, we got an acceptable recognition accuracy of 92% for subject A when using feature-level fusion which considerably went down, however, when using decision-level or hybrid fusion. Generally, feature-level fusion is more appropriate for combining modalities with analogous characteristics. As the data for subject A show, a high accuracy obtained for one modality may be declined by a relatively low accuracy from another modality when fusing data at the decision level. This observation indicates the limitations of the decision-level fusion scheme we used, which is based on a pure arithmetic evaluation of the posterior probabilities at the decision stage rather than a parametric assessment process (Fig. 5).

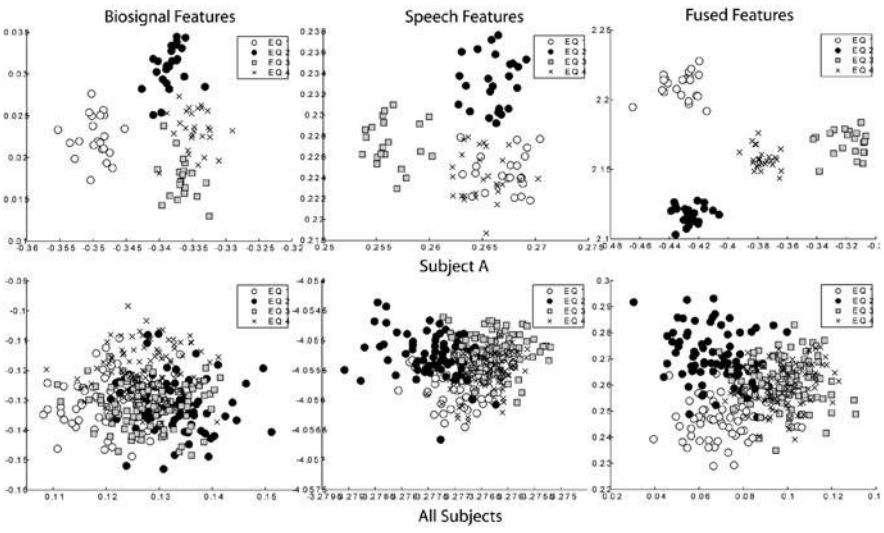

Fig. 5 Fisher projection examples of the bimodal features

3.3 Recognition from Facial Expressions and Hand Gesturing

In Karpouzis et al. (2004) and Balomenos et al. (2006), gestures are utilized to support the outcome of the facial expression analysis subsystem, since in most cases they are too ambiguous to indicate a particular emotion. However, in a given context of interaction, some gestures are obviously associated with a particular expression – e.g. hand clapping of high frequency expresses joy, satisfaction – while others can provide indications for the kind of the emotion expressed by the user. In particular, quantitative features derived from hand tracking, like speed and amplitude of motion, fortify the position of an observed emotion; for example, satisfaction turns to joy or even to exhilaration, as the speed and amplitude of clapping increases.

Given a particular context of interaction, gesture classes corresponding to the same emotional state are combined in a "logical OR" form. Table 2 shows that a particular gesture may correspond to more than one gesture classes carrying different affective meaning. For example, if the examined gesture is clapping, detection of high frequency indicates joy, but a clapping of low frequency may express irony and can reinforce a possible detection of the facial expression disgust.

Although face is the main 'demonstrator' of user's emotion (Ekman and Friesen, 1975), the recognition of the accompanying gesture increases the confidence of the result of facial expression subsystem. Further research is necessary to be carried out in order to define how powerful the influence of a gesture in the recognition of an emotion actually is. It would also be helpful to define which, face or gesture, is more useful for a specific application and change the impact of each subsystem on the final result.

In the current implementation the two subsystems are combined as a weighted sum. Let b_k be the degree of belief that the observed sequence presents the kth emotional state, obtained from the facial expression analysis subsystem, and EI_k be the corresponding emotional state indicator, obtained from the affective gesture analysis subsystem, then the overall degree of belief d_k is given by

$$d_k = w_1 \cdot b_k + w_2 \cdot EI_k$$

Table 2 Correlation between gestures and emotional states

Emotion	Gesture class
Joy	Hand clapping – high frequency
Sadness	Hands over the head – posture
Anger	Lift of the hand – high speed Italianate gestures
Fear	Hands over the head – gesture Italianate gestures
Disgust	Lift of the hand – low speed Hand clapping – low frequency
Surprise	Hands over the head – gesture

where the weights w_1 and w_2 are used to account for the reliability of the two subsystems as far as the emotional state estimation is concerned. In their implementation, the authors used w_1=0.75 and w_2=0.25. These values enables the affective gesture analysis subsystem to be important in cases where the facial expression analysis subsystem produces ambiguous results while at the same time leaves the latter subsystem to be the main contributing part in the overall decision system.

4 The Need for Adaptivity

In Caridakis et al. (2008), Caridakis builds on Balomenos et al. (2006) and Karpouzis et al. (2004), to provide adaptivity characteristics for decision- and feature-level fusion. In this case, a fuzzy logic-based system was derived, based on the formulation shown in Fig. 6.

While the multimodal system outperforms both unimodal (face and gesture) ones, it is clear that the ability of the system to adapt to the specific characteristics and user/situation contexts of the interaction is crucial. This approach still fails to model the interplay between the different modalities, a fact which one can exploit to fortify the results obtained from an individual modality (e.g. correlation between visemes and phonemes) or resolve uncertainty in cases where one or more modalities are not dependable (e.g. speech analysis in the presence of noise can be assisted by visually extracting visemes and mapping them to possible phonemes). The latter approach is termed dominant modality recoding model. Nevertheless, identification

Fig. 6 Overall architecture of the multimodal emotion recognition process in Caridakis et al. (2008)

of dominant modalities is another open issue, which could be resolved if (performance) confidence levels could be estimated in each unimodal case and used thereafter.

4.1 Detecting the Need for Adaptation

The purpose of this mechanism is to detect when the output of the neural network classifier is not appropriate and consequently to activate the adaptation algorithm at those time instances when a change of the environment occurs.

Let us index images or video frames in time, denoting by $\boldsymbol{x}(k, N)$, the feature vector of the kth image or image frame, following the image at which the adaptation of the Nth network occurred. Index k is therefore reset each time adaptation takes place, with $\boldsymbol{x}(0, N)$ corresponding to the feature vector of the image, where the nth adaptation of the network was accomplished. Adaptation of the network classifier is accomplished at time instances where its performance deteriorates, i.e. the current network output deviates from the desired one. Let us recall that vector $\boldsymbol{c}$ expresses the difference between the desired and the actual network outputs based on weights $\boldsymbol{w}_b$ and is applied to the current data set $\boldsymbol{S_c}$. As a result, if the norm of vector $\boldsymbol{c}$ increases, network performance deviates from the desired one and adaptation should be applied. On the contrary, if vector $\boldsymbol{c}$ takes small values, then no adaptation is required. In the following we denote this vector as $\boldsymbol{c}(k, N)$ depending upon feature vector $\boldsymbol{x}(k, N)$.

Let us assume that the Nth adaptation phase of the network classifier has been completed. If the classifier is then applied to all instances $\boldsymbol{x}(0, N)$, including the ones used for adaptation, it is expected to provide classification results of good quality. The difference between the output of the adapted network and that produced by the initially trained classifier at feature vector $\boldsymbol{x}(0, N)$ constitutes an estimate of the level of improvement that can be achieved by the adaptation procedure. Let us denote this difference by $e(0, N)$ and let $e(k, N)$ denote the difference between the corresponding classification outputs when the two networks are applied to the feature set of the kth image or image frame (or speech segment) following the Nth network adaptation phase. It is anticipated that the level of improvement expressed by $e(k, N)$ will be close to that of $e(0, N)$ as long as the classification results are good. This will occur when input images are similar to the ones used during the adaptation phase. An error $e(k, N)$, which is quite different from $e(0, N)$, is generally due to a change of the environment. Thus, the quantity $\alpha(k, N) = |e(k, N) - e(0, N)|$ can be used for detecting the change of the environment or equivalently the time instances where adaptation should occur. Thus, no adaptation is needed if $\alpha(k, N) > T$, where T is a threshold which expresses the max tolerance, beyond which adaptation is required for improving the network performance. In case of adaptation, index k is reset to zero, while index N is incremented by one.

Such an approach detects with high accuracy the adaptation time instances in cases of both abrupt and gradual changes of the operational environment since the comparison is performed between the current error difference $e(k, N)$ and the one

obtained right after adaptation, i.e. $e(0, N)$. In an abrupt operational change, error $e(k, N)$ will not be close to $e(0, N)$; consequently, $\alpha(k, N)$ exceeds threshold T and adaptation is activated. In the case of a gradual change, error $e(k, N)$ will gradually deviate from $e(0, N)$ so that the quantity $\alpha(k, N)$ gradually increases and adaptation is activated at the frame where $\alpha(k, N) > T$.

Network adaptation can be instantaneously executed each time the system is put in operation by the user. Thus, the quantity $\alpha(0, 0)$ initially exceeds threshold T and adaptation is forced to take place.

4.2 The Adaptive Neural Network Architecture

Let us assume that we seek to classify, to one of, say, p available emotion classes ω, each input vector $\boldsymbol{x}_i$ containing the features extracted from the input signal. A neural network produces a p-dimensional output vector $\boldsymbol{y}(\boldsymbol{x}_i)$

$$\underline{y}(\underline{x}_i) = \left[p^i_{\omega_1}\ p^i_{\omega_2} \cdots p^i_{\omega_p} \right]^{\mathrm{T}} \tag{1}$$

where $p^i_{\omega_j}$ denotes the probability that the ith input belongs to the jth class.

Let us first consider that the neural network has been initially trained to perform the classification task using a specific training set, say,

$$S_b = \left\{ \left(\underline{x}'_1,\ \underline{d}'_1 \right), \ldots, \left(\underline{x}'_{m_b},\ \underline{d}'_{m_b} \right) \right\}$$

where vectors $\underline{x}'_i$ and $\underline{d}'_i$ with $i = 1, 2, \cdots, m_b$ denote the ith input training vector and the corresponding desired output vector consisting of p elements, respectively.

Then, let $\boldsymbol{y}(\boldsymbol{x}_i)$ denote the network output when applied to a new set of inputs, and let us consider the ith input outside the training set, possibly corresponding to a new user or to a change of the environmental conditions. Based on the above described discussion, slightly different network weights should probably be estimated in such cases, through a network adaptation procedure.

Let $\boldsymbol{w}_b$ include all weights of the network before adaptation, and $\boldsymbol{w}_\alpha$ the new weight vector which is obtained after adaptation is performed. To perform the adaptation, a training set $\boldsymbol{S}_c$ has to be extracted from the current operational situation composed of (one or more), say, $\boldsymbol{m}_c$ inputs:

$$S_c = \left\{ \left(\underline{x}_1,\ \underline{d}_1 \right), \ldots, \left(\underline{x}_{m_c},\ \underline{d}_{m_c} \right) \right\}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{d}_i$ with $i = 1, 2, \ldots, \boldsymbol{m}_c$ similarly correspond to the ith input and desired output data used for adaptation, respectively. The adaptation algorithm that is activated, whenever such a need is detected, computes the new network weights $\boldsymbol{w}_\alpha$, minimizing the following error criteria with respect to weights:

$$E_a = E_{c,a} + \eta E_{f,a}, E_{c,a} = \frac{1}{2}\sum_{i=1}^{m_c} \left\| \underline{z}_a(\underline{x}_i) - \underline{d}_i \right\|_2, E_{f,a} = \frac{1}{2}\sum_{i=1}^{m_b} \left\| \underline{z}_a(\underline{x'}_i) - \underline{d'}_i \right\|_2 \quad (2)$$

where $E_{c,\alpha}$ is the error performed over training set $\boldsymbol{S}_c$ ("current" knowledge), $E_{f,\alpha}$ is the corresponding error over training set $\boldsymbol{S}_b$ ("former" knowledge); $\boldsymbol{z}_\alpha(\boldsymbol{x}_i)$ and $\boldsymbol{z}_\alpha(\boldsymbol{x'}_i)$ are the outputs of the adapted network, corresponding to input vectors $\boldsymbol{x}_i$ and $\boldsymbol{x'}_i$, respectively, of the network consisting of weights $\boldsymbol{w}_\alpha$. Similarly $\boldsymbol{z}_b(\boldsymbol{x}_i)$ would represent the output of the network consisting of weights $\boldsymbol{w}_b$ when accepting vector $\boldsymbol{x}_i$ at its input; when adapting the network for the first time $\boldsymbol{z}_b(\boldsymbol{x}_i)$ is identical to $\boldsymbol{y}(\boldsymbol{x}_i)$. Parameter η is a weighting factor accounting for the significance of the current training set compared to the former one and $||.||_2$ denotes the L_2 norm.

The goal of the training procedure is to minimize the above equation and estimate the new network weights $\boldsymbol{w}_\alpha$, i.e. ${\boldsymbol{w}^0}_\alpha$ and $\boldsymbol{w}_\alpha$, respectively. Let us first assume that a small perturbation of the network weights (before adaptation) $\boldsymbol{w}_b$ is enough to achieve good classification performance. Then

$$\underline{w}_a = \underline{w}_b + \Delta\underline{w}$$

where $\Delta\boldsymbol{w}$ are small increments. This assumption leads to an analytical and tractable solution for estimating $\boldsymbol{w}_\alpha$, since it permits linearization of the non-linear activation function of the neuron, using a first-order Taylor series expansion.

Equation (2) indicates that the new network weights are estimated taking into account both the current and the previous network knowledge. To stress, however, the importance of current training data, one can replace the first term by the constraint that the actual network outputs are equal to the desired ones, that is

$$z_a(\underline{x}_i) = d_i \quad i = 1, ..., m_c, \quad \text{for all data in } S_c \quad (3)$$

This equation indicates that the first term of (2), corresponding to error $E_{c,\alpha}$, takes values close to 0, after estimating the new network weights. Through linearization, solution of (3) with respect to the weight increments is equivalent to a set of linear equations

$$\underline{c} = \mathbf{A} \cdot \Delta\underline{w}$$

where vector $\boldsymbol{c}$ and matrix $\mathbf{A}$ are appropriately expressed in terms of the previous network weights. In particular

$$\underline{c} = \left[z_a(\underline{x}_1) \cdots z_a(\underline{x}_{m_c})\right]^{\mathrm{T}} - \left[z_b(\underline{x}_1) \cdots z_b(\underline{x}_{m_c})\right]^{\mathrm{T}}$$

expressing the difference between network outputs after and before adapting to all input vectors in $\boldsymbol{S}_c$. $\boldsymbol{c}$ can be written as

$$\underline{c} = \left[d_1 \cdots d_{m_c}\right]^{\mathrm{T}} - \left[z_b(\underline{x}_1) \cdots z_b(\underline{x}_{m_c})\right]^{\mathrm{T}} \quad (4)$$

Equation (4) is valid only when weight increments $\Delta \boldsymbol{w}$ are small quantities. It can be shown (Doulamis et al., 2000) that given a tolerated error value, proper bounds θ and φ can be computed for the weight increments and input vector $\boldsymbol{x}_i$ in $\boldsymbol{S}_c$.

Let us assume that the network weights before adaptation, i.e. $\boldsymbol{w}_b$, have been estimated as an optimal solution over data of set $\boldsymbol{S}_b$. Furthermore, the weights after adaptation are considered to provide a minimal error over all data of the current set $\boldsymbol{S}_c$. Thus, minimization of the second term of (2), which expresses the effect of the new network weights over data set $\boldsymbol{S}_b$, can be considered as minimization of the absolute difference of the error over data in $\boldsymbol{S}_b$ with respect to the previous and the current network weights. This means that the weight increments are minimally modified, resulting in the following error criterion:

$$E_S = \left\| E_{f,a} - E_{f,b} \right\|_2$$

with $E_{f,b}$ defined similarly to $E_{f,a}$, with z_α replaced by z_b in (2) (Park et al., 1991) shows that the above equation takes the form of

$$E_S = \frac{1}{2} (\Delta \underline{w})^{\mathrm{T}} \cdot \mathbf{K}^{\mathrm{T}} \cdot \mathbf{K} \cdot \Delta \underline{w} \tag{5}$$

where the elements of matrix $\mathbf{K}$ are expressed in terms of the previous network weights $\boldsymbol{w}_b$ and the training data in $\boldsymbol{S}_b$. The error function defined by (5) is convex since it is of squared form. The constraints include linear equalities and inequalities. Thus, the solution should satisfy the constraints and minimize the error function in (5). The gradient projection method is adopted to estimate the weight increments.

Each time the decision mechanism ascertains that adaptation is required, a new training set $\boldsymbol{S}_c$ is created, which represents the current condition. Then, new network weights are estimated, taking into account both the current information (data in $\boldsymbol{S}_c$) and the former knowledge (data in $\boldsymbol{S}_b$). Since the set $\boldsymbol{S}_c$ has been optimized over the current condition, it cannot be considered suitable for following or future states of the environment. This is due to the fact that data obtained from future states of the environment may be in conflict with data obtained from the current one. On the contrary, it is assumed that the training set $\boldsymbol{S}_b$, which is in general based on extensive experimentation, is able to roughly approximate the desired network performance at any state of the environment. Consequently, in every network adaptation phase, a new training set $\boldsymbol{S}_c$ is created and the previous one is discarded, while new weights are estimated based on the current set $\boldsymbol{S}_c$ and the old one $\boldsymbol{S}_b$, which remains constant throughout network operation.

References

Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychol Bull 111(2):256–274

Balomenos T, Raouzaiou A, Ioannou S, Drosopoulos A, Karpouzis K, Kollias S (2006) Emotion analysis in man–machine interaction systems. In: Bengio S, Bourlard H (eds) Machine learning

for multimodal interaction. Lecture notes in computer science, vol 3361. Springer, Berlin, pp 318–328
Caridakis G, Karpouzis K, Kollias S (2008) User and context adaptive neural networks for emotion recognition. Neurocomputing, Elsevier, 71(13–15):2553–2562
Caridakis G, Malatesta L, Kessous L, Amir N, Raouzaiou A, Karpouzis K (2006) Modeling naturalistic affective states via facial and vocal expressions recognition. In: International conference on multimodal interfaces (ICMI'06), Banff, AB, 2–4 Nov 2006
Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schröder M (2000) FEELTRACE: an instrument for recording perceived emotion in real time. In: ISCA workshop on speech and emotion, Northern Ireland, pp 19–24
Doulamis N, Doulamis A, Kollias S (2000) On-line retrainable neural networks: improving performance of neural networks in image analysis problems. IEEE Trans Neural Netw 11(1):1–20
Ekman P, Friesen WF (1969) The repertoire of nonverbal behavioral categories – origins, usage, and coding. Semiotica 1:49–98
Ekman P, Friesen W (1975) Unmasking the face. Prentice-Hall, Englewood Cliffs, NJ<loc>
Ekman P, Huang TS, Sejnowski TJ, Hager JC (eds) (1993) NSF understanding the face. A Human Face eStore, Salt Lake City (see Library)
Elman JL (1990) Finding structure in time. Cogn Sci 14:179–211
FP5 IST ERMIS (2007) http://www.image.ntua.gr/ermis. Accessed 30 Oct 2007
Fridlund AJ (1997) The new ethology of human facial expression. In: Russell JA, Fernandez-Dols JM (eds) The psychology of facial expression. Cambridge University Press, Cambridge, MA, pp 103–129
Goleman D (1995) Emotional intelligence. Bantam Books, New York, NY
Juslin PN, Scherer KR (2005) Vocal expression of affect. In: Harrigan J, Rosenthal R, Scherer K (eds) The new handbook of methods in nonverbal behavior research. Oxford University Press, Oxford
Karpouzis K, Raouzaiou A, Drosopoulos A, Ioannou S, Balomenos T, Tsapatsoulis N, Kollias S (2004) Facial expression and gesture analysis for emotionally-rich man–machine interaction. In: Sarris N, Strintzis M (eds) 3D modeling and animation: synthesis and analysis techniques. Idea Group, Hershey, PA, pp 175–200
Keltner D, Ekman P (2000) Facial expression of emotion. In: Lewis M, Haviland-Jones JM (eds) Handbook of emotions. Guilford Press, New York, NY, pp 236–249
Kim J, André E (2006) Emotion recognition using physiological and speech signal in short-term observation. In: Perception and interactive technologies, LNAI 4201. Springer, Berlin, Heidelberg, pp 53–64
Kim J, André E, Rehm M, Vogt T, Wagner J (2005) Integrating information from speech and physiological signals to achieve emotional sensitivity. In: Proceedings of the 9th European conference on speech communication and technology, Lisbon, Portugal
Larsen RJ, Diener E (1992) Promises and problems with the circumplex model of emotion. In: Clark MS, (ed) Review of personality and social psychology, vol 13. Sage, Newbury Park, CA, pp 25–59
Luettin J, Thacker N, Beet S (1996) Active shape models for visual speech feature extraction. In: Storck DG, Hennecke ME (eds) Speechreading by humans and machines. Springer, Berlin, pp 383–390
Matsumoto D (1990) Cultural similarities and differences in display rules. Motiv Emot 14:195–214
Pantic M (2005) Affective computing. In: Pagani M (ed) Encyclopedia of multimedia technology and networking, vol 1. Idea Group Reference, Hershy, PA, pp 8–14
Pantic M, Pentland A, Nijholt A, Huang TS (2006) Human computing and machine understanding of human behaviour: a survey. In: Proceedings of the ACM international conference on multimodal interfaces, Banff, Alberta, Canada, pp 239–248
Pantic M, Rothkrantz LJM (2003) Toward an affect-sensitive multimodal human–computer interaction. Proc IEEE 91(9):1370–1390

Pantic M, Sebe N, Cohn JF, Huang TS (2005) Affective multimodal human–computer interaction. In: Proceedings of the 13th annual ACM international conference on Multimedia, pp 669–676
Park D, EL-Sharkawi MA, Marks RJ II (1991) An adaptively trained neural network. IEEE Trans Neural Netw 2:334–345
Pelachaud C, Carofiglio V, De Carolis B, de Rosis F, Poggi I (2002) Embodied contextual agent in information delivering application. In: Proceedings of the international conference on autonomous agents and multi-agent systems. Bologna, Italy
Picard RW (1997) Affective computing. The MIT Press, Cambridge, MA
Picard RW (2003) Affective computing: challenges. Int J Human–Comput Stud 59(1–2):55–64
Potamianos G, Neti C, Gravier G, Garg A (2003 Sept) Automatic recognition of audio-visual speech: recent progress and challenges. Proc IEEE 91(9):1306–1326
Russell JA (1994) Is there universal recognition of emotion from facial expression? Psychol Bull 115(1):102–141
Taylor J, Fragopanagos N (2004) Modelling human attention and emotions. Proc 2004 IEEE Int Joint Conf Neural Netw 1:501–506.
Taylor J, Fragopanagos N (2005) The interaction of attention and emotion. Neural Netw 18(4): 353–369
Watson D, Weber K, Assenheimer JS, Clark LA, Strauss ME, McCormick RA (1995a) Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. J Abnorm Psychol 104:3–14
Watson D, Clark LA, Weber K, Smith-Assenheimer J, Strauss ME, McCormick RA (1995b) Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. J Abnorm Psychol 104:15–25
Zeng Z, Pantic M, Roisman G, Huang T (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1)