Neurocomputing 71 (2008) 2553–2562

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

User and context adaptive neural networks for emotion recognition

George Caridakis *, Kostas Karpouzis, Stefanos Kollias

School of Electrical and Computer Engineering, National Technical University of Athens, Greece

ARTICLE INFO

Available online 9 May 2008

Keywords:
Neural networks
Emotion recognition
User and context adaptation

ABSTRACT

Recognition of emotional states of users in human–computer interaction (HCI) has been shown to be highly dependent on individual human characteristics and way of behavior. Multimodality is a key issue in achieving more accurate results; however, fusing different modalities is a difficult issue in emotion analysis. Emotion recognition systems are generally either rule-based or extensively trained through emotionally colored HCI data sets. In either case, such systems need to take into account, i.e., adapt their knowledge to, the specific user or context of interaction. Neural networks fit well with the adaptation requirement, by collecting and analyzing data from specific environments. An effective approach is presented in this paper, which uses neural network architectures to both detect the need for adaptation of their knowledge, and adapt it through an efficient adaptation procedure. An experimental study with emotion datasets generate in the framework of the EC IST Humaine Network of Excellence.

1. Introduction

1.1. Motivation and objectives

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 [12]. When it comes to computers, however, they are socially ignorant [29]. Current computing technology does not account for the fact that human–human communication is always socially situated and that 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 (see [24,30,31]). 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 behavior, and use a proper affective presentation language (e.g. [28]) 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 [24,25]. 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 behavior could expand and enhance scientific understanding and practical applications include specialized areas in professional and scientific sectors [7]. In the security sector, affective behavioral cues play a crucial role in establishing or detracting from credibility. In the medical sector, affective behavioral 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 behavioral 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 behavior are persisting problems.

While all agree that machine sensing and interpretation of human affective information would be widely beneficial,

* Corresponding author.
E-mail addresses: gcari@image.ntua.gr (G. Caridakis), kkarpou@image.ece.ntua.gr (K. Karpouzis), stefanos@cs.ntua.gr (S. Kollias).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.11.043

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.

1.2. Related research

1.2.1. Affective states

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 [8,17]. These include happiness, anger, sadness, surprise, disgust, and fear. Data from both Western and traditional societies suggests that non-verbal communicative signals (especially facial and vocal expression) involved in these basic emotions are displayed and recognized cross-culturally. In opposition to this view, Russell [32] 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 [18,35,36]. 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 labeled. For example, Fridlund [11] argues that human facial expressions should not be labeled in terms of emotions but in terms of Behavioural Ecology interpretations, which explain the influence a certain expression has in a particular context. 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 [21,38]. 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, it is not only discrete emotional states like surprise or anger that are of importance for the realization of proactive human–machine interactive systems. Sensing and responding to behavioral cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are also essential [26]. 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 [25,26].

1.2.2. Recognition of emotions and context of interaction

Let us first focus on facial expression recognition. Facial expressions can vary in intensity. By intensity we mean the relative degree of change in facial expression as compared to a relaxed, neutral facial expression. It has been experimentally shown that the expression decoding accuracy and the perceived intensity of the underlying affective state vary linearly with the physical intensity of the facial display [13]. Hence, explicit analysis of expression intensity variation is very important for accurate expression interpretation, and is also essential to the ability to distinguish between spontaneous and posed facial behavior. While Facial Action Coding System (FACS) provides a 5-point intensity scale to describe AU intensity variation and enable manual quantification of AU intensity [9], fully automated methods that accomplish this task are yet to be developed. However, first steps toward this goal have been made. Automatic coding of intensity variation was explicitly compared to manual coding in Bartlett et al. [2]. They found that the distance to the separating hyperplane in their learned classifiers correlated significantly with the intensity scores provided by expert FACS coders.

Rapid facial signals do not usually convey exclusively one type of messages. For instance, squinted eyes may be interpreted as sensitivity of the eyes to bright light if this action is a reflex (a manipulator), as an expression of disliking if this action has been displayed when seeing someone passing by (affective cue), or as an illustrator of friendly anger on friendly teasing if this action has been posed (in contrast to being unintentionally displayed) during a chat with a friend, to mention just a few possibilities. As already mentioned earlier, to interpret an observed facial expression, it is important to know the context in which the observed expression has been displayed—where the expresser is (outside, inside, in the car, in the kitchen, etc.), what his or her current task is, are other people involved, and who the expresser is. Knowing the expresser is particularly important as individuals often have characteristic facial expressions and may differ in the way certain states (other than the basic emotions) are expressed. Since the problem of context sensing is extremely difficult to solve (if possible at all) for a general case, pragmatic approaches (e.g., activity/application- and user-centered approach) should be taken when learning the grammar of human facial behavior [25,26]. However, except for a few works on user-profiled interpretation of facial expressions like those of Fasel et al. [10] and Ioannou et al., [14], virtually all existing automated facial expression analyzers are context insensitive. Similar is the case with systems dealing with other modalities, such as speech and audio, hand and body gestures.

Regarding personalized expressivity, it is well known (see, for example, recent results, on emotional signs from signals, of the Humaine network of Excellence [22]) that in human–computer interaction (HCI), the emotional characteristics and signs of signals captured from a specific user, although adhering to some general descriptive theories and psychological models, differ, sometimes significantly, between different persons. Thus, emotion recognition is a research problem, the solution of which highly depends on individual human characteristics and way of behavior.

Emotion recognition systems are generally based on a rule base system, or on a system that has learnt to solve the problem through extensive training. In either case, if such a system is to be used in a real-life experiment, it further needs to take into account, i.e., to adapt its knowledge to the specific user characteristics as well as behavioral and environmental conditions, i.e., the context of interaction.

As richer the information provided by the interaction is, so more cues can be derived for extracting the interaction context and for achieving a better emotion recognition performance. The case of using multiple modalities, referred as multimodal emotion recognition, is, therefore, of crucial importance and research interest. Integrating, however, cues from different modalities, is not an easy task. Various types of problems, such as need for synchronization, temporal integration and semantic fusion, cause this difficulty.

In all cases, it is essential that systems are derived which are able to adapt their performance to environmental changes, by detecting deterioration of their performance, and refining it with data obtained by the specific environment and respective cues provided by the user or by cross-correlating different modalities. Neural networks fit well with this requirement, since adaptation is their main advantage when compared with knowledge-based systems, where updating of knowledge is a complex, generally off-line procedure. Both supervised, such as multilayered feed-forward networks, and unsupervised networks, such as SOM or k-NN-based approaches can be used for this purpose. In the rest of the paper an adaptive supervised feed-forward network is described and used for HCI enriched with emotion analysis capabilities, showing that it can provide an effective approach to handling of the above-described problems. The basic methodology can be extended to unsupervised, clustering techniques.

Section 2 refers to the problem of emotion recognition and the need for multimodal input fusion. Section 3.1 describes the adaptive network architecture, while its use in different contexts is presented in Section 3.2. An experimental study, with emotion datasets showing, not only extreme emotions, but also intermediate real-life ones, generated in the framework of the EC IST Humaine Network of Excellence, is given in Section 4, while conclusions and further work are discussed in Section 5.

2. Multimodal input fusion and emotion recognition

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) [15,23]. Jaimes and Sebe use a human-centered 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 [19]. 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 [39]. 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 (HMM) 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 labeled for training purposes.

Taylor and Fragopanagos describe a neural network architecture (see [33,34]) 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 feed-forward 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 an emotional dictionary), prosodic features (as determined by a prosodic speech feature extraction), and facial features (facial animation parameters (FAPs)). The obtained results are different for different viewers who helped to annotate the datasets. 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

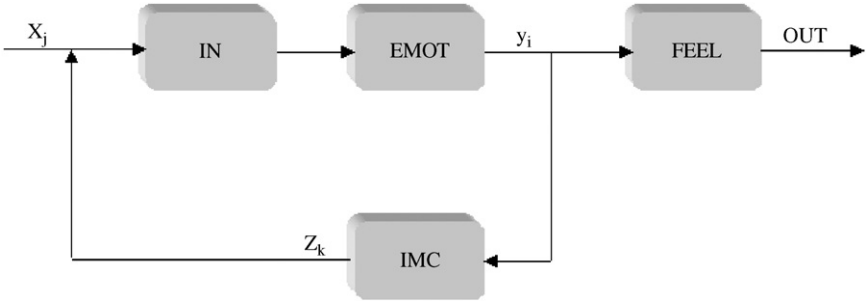


Fig. 1. Information flow in the system: IMC = inverse model controller; EMOT = hidden layer emotional state; and FEEL = output state emotion classifier.

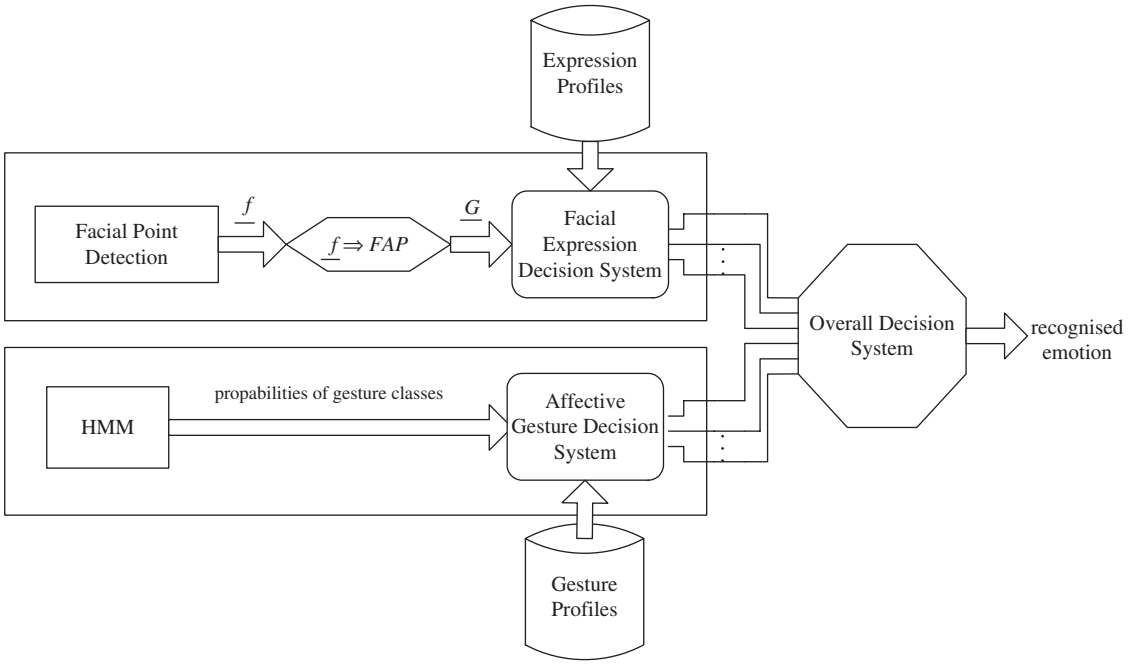


Fig. 2. Overall architecture of the multimodal emotion recognition process.

Table 1
Recognized gestures and mapping to emotion labels

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

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, majority vote, etc. As a general rule, semantic fusion builds on individual recognizers, followed by an integration process; individual recognizers can be trained using unimodal data, which are easier to collect.

We have developed such a system for emotion recognition, based on decision-level fusion, when dealing with two visual modalities, i.e., facial expressions and hand and body gestures [1]. In this case, a fuzzy-logic-based system was derived, based on the formulation shown in Fig. 2.

In Fig. 2, the 'facial expression decision system' is trained to recognize facial expressions based on extraction of facial points and FAPs according to the MPEG-4 ISO standard. The 'affective gesture decision system' is trained to recognize specific gestures, which are mapped to relevant emotion labels, using HMM to extract probabilities of gesture classes, as shown in Table 1. The 'overall decision system' takes into account the outputs of the two aforementioned systems, using a fuzzy-logic rule-based approach. While the overall system outperforms both unimodal ones, from these experiments it has become clear that the ability of the system to adapt to the specific characteristics and user/situation contexts of the interaction is crucial.

The mapping of an image to an input vectors is based on feature extraction process both on facial and gesture domains. Concerning the detection of prominent facial points which leads to the calculation of FAPs the estimated location of candidate facial regions follows the detection of head position and pose (roll rotation). Several techniques, presented in [14], including both computer vision and artificial intelligence techniques produce a set of masks which in turn are fused and a final mask per feature family (eyes, eyebrows, and mouth) is produced. Hand coordinates are extracted using a region extraction method using a color skin model, morphological operations, and motion information from the input image. Acoustic processing aims to quantify the prosodic variations in speech based on a feature set including pitch, intensity, duration, spectrum, and stability measures.

Decision-level fusion still fails to model the interplay between different modalities, a fact which one can exploit to fortify the results obtained from an individual modality (e.g. correlation between visemes, the visual equivalent of phonemes, 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 resulting approach is termed Dominant Modality Recoding Model. Nevertheless, identification 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.

In the rest of this paper, we examine the confidence produced by each classifier, such as a feed-forward multilayer neural network, handling a single modality-focusing on facial expressions—and we derive an efficient methodology for adapting the classifier's performance, when detecting such a need, by collecting

data from its specific environment. Thus, in the framework presented here, facial expression is considered as the dominant modality; this means that most of the time classification is performed using the facial features as input. In cases where the network trained with the facial data does not perform well (hence, the need to adapt arises), speech prosody or gestures can be utilized as "fall-back" solutions, possibly providing the expected output for the adaptation process.

3. Adaptation procedure

3.1. 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 $\underline{x}_i$ containing the features extracted from the input signal. A neural network produces a p-dimensional output vector $\overline{y}(\overline{x}_i)$

$$\overline{y}(\overline{x}_i) = [\, p^i_{\omega_1} \;\; p^i_{\omega_2} \ldots p^i_{\omega_p}]^{\mathrm{T}} \tag{1}$$

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, $S_{\mathrm{b}} = \{(\overline{x}'_1, \; \overline{d}'_1), \cdots, (\overline{x}'_{m_{\mathrm{b}}}, \; \overline{d}'_{m_{\mathrm{b}}})\}$, where vectors $\overline{x}'_i$ and $\overline{d}'_i$ with $i = 1,2,\cdots, m_{\mathrm{b}}$ denote the ith input training vector and the corresponding desired output vector consisting of p elements.

Then, let $\overline{y}(\overline{x}_i)$ 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 $\overline{w}_{\mathrm{b}}$ include all weights of the network before adaptation, and $\overline{w}_{\mathrm{a}}$ the new weight vector which is obtained after adaptation is performed. To perform the adaptation, a training set S_{c} has to be extracted from the current operational situation composed of say, m_{c} inputs; $S_{\mathrm{c}} = \{(\overline{x}_1, \; \overline{d}_1), \cdots, (\overline{x}_{m_{\mathrm{c}}}, \; \overline{d}_{m_{\mathrm{c}}})\}$, where $\overline{x}_i$ and $\overline{d}_i$ with $i = 1,2,\cdots,m_{\mathrm{c}}$ similarly correspond to the ith input and desired output data used for adaptation. The adaptation algorithm that is activated, whenever such a need is detected, computes the new network weights $\underline{w}_{\mathrm{a}}$, minimizing the following error criteria with respect to weights

$$\begin{aligned} E_{\mathrm{a}} &= E_{\mathrm{c,a}} + \eta E_{\mathrm{f,a}} \\ E_{\mathrm{c,a}} &= \tfrac{1}{2}\sum_{i=1}^{m_{\mathrm{c}}} \left\| \underline{z}_{\mathrm{a}}(\underline{x}_i) - \underline{d}_i \right\|_2 \\ E_{\mathrm{f,a}} &= \tfrac{1}{2}\sum_{i=1}^{m_{\mathrm{b}}} \left\| \underline{z}_{\mathrm{a}}(\underline{x}'_i) - \underline{d}'_i \right\|_2 \end{aligned} \tag{2}$$

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

The goal of the training procedure is to minimize (2) and estimate the new network weights $\overline{w}_{\mathrm{a}}$. The adopted algorithm has been proposed by the authors in [6]. Let us first assume that a small perturbation of the network weights (before adaptation) $\overline{w}_{\mathrm{b}}$ is enough to achieve good classification performance. Then

$$\overline{w}_{\mathrm{a}} = \overline{w}_{\mathrm{b}} + \Delta\overline{w}$$

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

Eq. (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 in (2), one can replace the first term by the constraint that the actual network outputs are equal to the desired ones, that is

$$z_{\mathrm{a}}(\overline{x}_i) = d_i \quad i = 1, \ldots, m_{\mathrm{c}}, \;\; \text{for all data in } \; S_{\mathrm{c}} \tag{3}$$

Through linearization, solution of (3) with respect to the weight increments is equivalent to a set of linear equations

$$\overline{c} = \mathbf{A} \cdot \Delta\overline{w} \tag{4}$$

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

$$\overline{c} = [d_1 \cdots d_{m_{\mathrm{c}}}]^{\mathrm{T}} - [z_{\mathrm{b}}(\overline{x}_1) \cdots z_{\mathrm{b}}(\overline{x}_{m_{\mathrm{c}}})]^{\mathrm{T}} \tag{5}$$

Moreover, minimization of the second term of (2), which expresses the effect of the new network weights over data set S_{b}, can be considered as minimization of the absolute difference of the error over data in S_{b} with respect to the previous and the current network weights. This means that the weight increments are minimally modified, with respect to the following error criterion

$$E_{\mathrm{S}} = \left\| E_{\mathrm{f,a}} - E_{\mathrm{f,b}} \right\|_2 \tag{6}$$

with $E_{\mathrm{f,b}}$ defined similarly to $E_{\mathrm{f,a}}$, with $\overline{z}_{\mathrm{a}}$ replaced by $\overline{z}_{\mathrm{b}}$ in (2).

It can be shown [27] that (6) takes the form of

$$E_{\mathrm{S}} = \tfrac{1}{2}(\Delta\overline{w})^{\mathrm{T}} \cdot \mathbf{K}^{\mathrm{T}} \cdot \mathbf{K} \cdot \Delta\overline{w} \tag{7}$$

where the elements of matrix $\mathbf{K}$ are expressed in terms of the previous network weights $\overline{w}_{\mathrm{b}}$ and the training data in S_{b}. The error function defined by (7) is convex since it is of squared form. Thus, the weight increments can be estimated through solution of (7). The gradient projection method has been used in [6] to estimate the weight increments.

Each time the decision mechanism ascertains that adaptation is required, a new training set S_{c} is created, which represents the current condition. Then, new network weights are estimated taking into account both the current information (data in S_{c}) and the former knowledge (data in S_{b}). Since the set S_{c} has been optimized only for 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 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 S_{c} is created and the previous one is discarded, while new weights are estimated based on the current set S_{c} and the old one S_{b}, which remains constant throughout network operation.

3.2. 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 first assume that a network adaptation has taken place and let us focus on visual inputs. Let $\bar{x}(k)$ denote the feature vector of the kth image or image frame, following the time at which adaptation occurred. Index k is therefore reset each time adaptation takes place, with $\bar{x}(0)$ corresponding to the feature vector of the image where the adaptation of the network was accomplished. At this input, the network performance had deteriorated, i.e., the network output deviated from the desired one. Let us recall that vector $\bar{c}$ in (5) expresses the difference between the desired and the actual network outputs based on weights $\bar{w}_b$ and applied to the current data set. As a result, if the norm of vector $\bar{c}$ increases, network performance deviates from the desired one and adaptation should be applied. On the contrary, if vector $\bar{c}$ takes small values, then no adaptation is required. In the following we use the difference between the output of the adapted network and of that produced by the initially trained classifier to approximate the value of $\bar{c}$. Moreover, we assume that the difference computed when processing input $\bar{x}(0)$ constitutes a good estimate of the level of improvement that can be achieved by the adaptation procedure. Let us denote by $e(0)$ this difference and let $e(k)$ denote the difference between the corresponding classifiers' outputs, when the two networks are applied to $\bar{x}(k)$. It is anticipated that the level of improvement expressed by $e(k)$ will be close to that of $e(0)$ 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 $e(k)$, which is quite different from $e(0)$, is generally due to a change of the environment. Thus, the quantity $a(k) = |e(k) - e(0)|$ 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:

$$a(k) < T \tag{8}$$

where T is a threshold which expresses the max tolerance, beyond which adaptation is required for improving the network performance.

Such an approach detects with high accuracy the adaptation time instances both in cases of abrupt and gradual changes of the operational environment since the comparison is performed between the current error difference $e(k)$ and the one obtained right after adaptation, i.e., $e(0)$. In an abrupt operational change, error $e(k)$ will not be close to $e(0)$; consequently, $a(k)$ exceeds threshold T and adaptation is activated. In case of a gradual change, error $e(k)$ will gradually deviate from $e(0)$ so that the quantity $a(k)$ gradually increases and adaptation is activated at the frame where $a(k) > T$.

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

4. Experimental study

4.1. Corpus

In this section results of extensive experimentation, based on the above-described adaptive neural network classifier are provided.

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 [22]. 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 [16]. The "Wizard of Oz" methodology is an experimental simulation, in which experimental participants are given the impression that they are interacting with a program that understands English as well as another human would. The experimenter, acting as "Wizard", surreptitiously intercepts communications between participant and program, supplying answers and new inputs as needed.

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 [4]. This is a descriptive tool that has been developed at Queen's University Belfast using dimensional representations, which provides time-sensitive dimensional

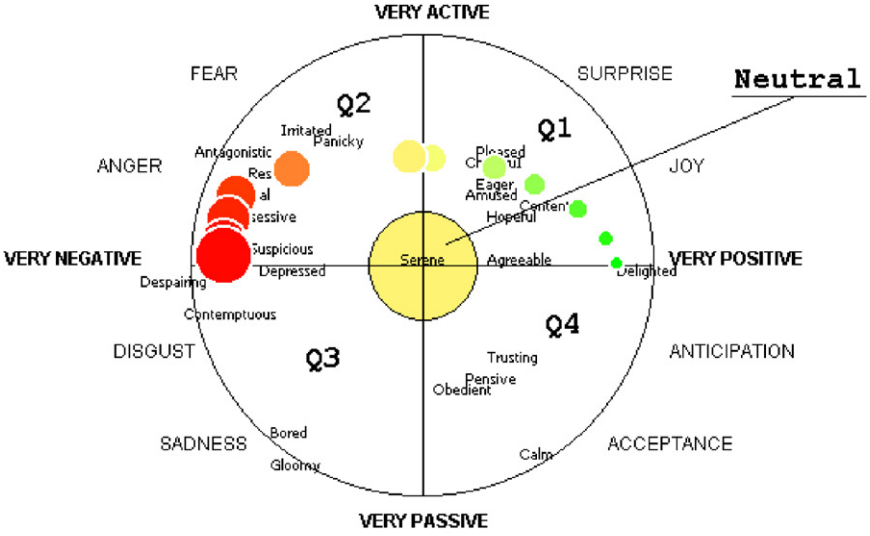


Fig. 3. The Whissel's wheel activation/valence dimensional representation [37].

representations. It lets observers track the emotional content of a time-varying stimulus as they perceive it. Fig. 3 illustrates the kind of display that FeelTrace users see as well as a particular trace across a tune. The tune is defined loosely by either a meaningful sentence or a temporal segment during which the subject has a steady emotional state. We observed that in the majority of the cases from a subjective point of view, a tune defined using criteria such as audio pauses to be a good temporal segmentation. This segmentation refers only to the temporal axis and is not dependent on extracted features from any modality.

Table 2
Emotion classes

Label	Location in FeelTrace [5] diagram
Q1	Positive activation, positive evaluation (+/+)
Q2	Positive activation, negative evaluation (+/−)
Q3	Negative activation, negative evaluation (−/−)
Q4	Negative activation, positive evaluation (−/+)
Neutral	Close to the center

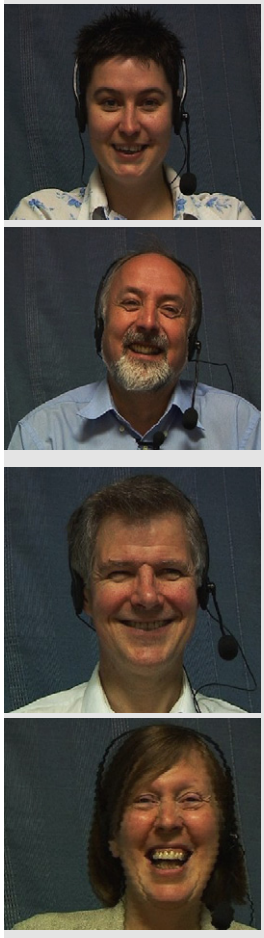
Table 3
Class distribution in the SAL dataset

	Neutral	Q1	Q2	Q3	Q4	Total
Tunes	47	205	90	63	72	477
Percentages	9.85	42.98	18.87	13.21	15.09	100.00

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 center 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 coordinates of the cursor at any time.

The x–y coordinates of the mouse movements on the two-dimensional user interface are mapped to the five emotional categories presented in Table 2. 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 up to 174 frames. A bias towards Q1 exists in the database, as 42.98% of the tunes are classified to Q1, as

Table 4
Samples from the different subjects displaying emotions

1st quadrant (+,+)	2nd quadrant (−,+)	3rd quadrant (−,−)	4th quadrant (+,−)
			

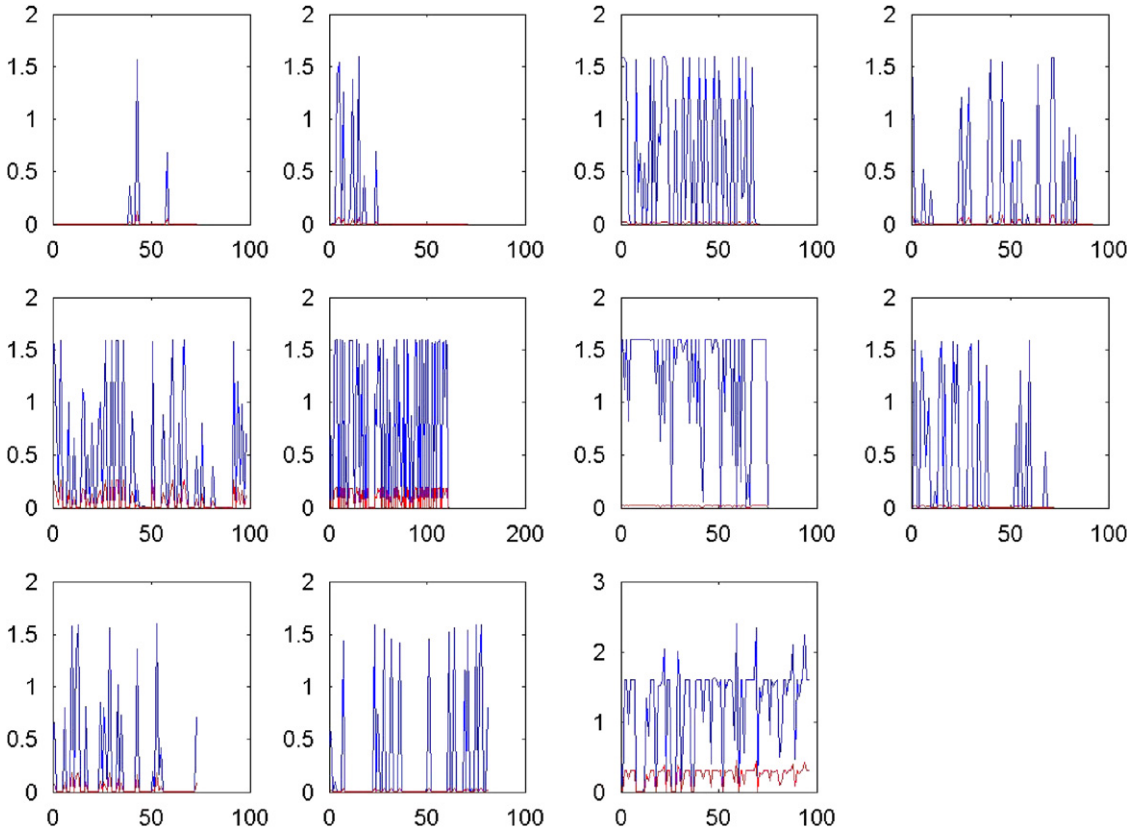


Fig. 4. MSE of NetProm (blue) and Net$_i$ (red).

shown in Table 3. Table 4 shows the four different subjects displaying emotions from a variety of tunes and quadrants.

4.2. Experiments

Our experiments aimed at investigating the practical stand of the proposed adaptation procedure. The main idea of the experimental study was to explore the performance of the adapted networks over inputs belonging to the same tune, but not used for adaptation, as well as to tunes of the same emotional quadrant as the one used for adaptation purposes.

Out of approximately 35,000 frames, belonging to 477 tunes of the SAL database, we selected a merely 500 frames—from all four subjects—for training a feed-forward back-propagation network referred from now as NetProm. The architecture details for NetProm are three layers consisting of 10 and 5 neurons on the first and second hidden layers, respectively, and 5 neurons of the output layer. The targets were formatted as a 5×1 vector for every frame so as to only one, of the 5 candidate classes, was equal to 1. So for example if the frame used for training belonged to the first quadrant the output vector would be $[1\ -1\ -1\ -1\ -1]$. The fifth class of the classification problem corresponds to the neutral emotional state and the other four to the four quadrants of the Whissel's wheel.

The selection of the 500 frames used for training the NetProm network was made following a prominence criterion. More specifically, for every frame, a metric was assigned denoting the distance of the values of the FAPs for that specific frame with reference to the mean values of the FAPs of the other frames belonging to the same class. This metric of FAP variance was the sorting parameter for the frames. Under the constraint that each class should be represented as equally as possible we selected the 500 most prominent frames and used it as input for training the NetProm network.

With regard to the adaptation phase we selected 11 tunes consisting of the largest number of frames. This selection was based on the idea that it would not make much sense selecting very short tunes, because the adaptation data would be very sparse as will be explained later. Also we made sure that no frame belonging to these 11 tunes was used for training NetProm. Each of the 11 tunes was divided into two groups of frames, the adaptation group and the testing group containing 30% and 70% of the total frames of the original tune, sorted by the prominence criterion, respectively.

NetProm was adapted using the adaptation group of the eleven tunes and produced 11 new networks Net$_i$, $i = 1,\ldots,11$. Each Net$_i$ was then tested on the testing group of the respective tune and the results can be seen in Fig. 4. It is clear that the adaptation procedure has been beneficial and greatly reduced the MSE for every tune it was applied.

Furthermore, we tested the procedure proposed in Section 4 for detecting when adaptation is necessary. In particular, we used the above-derived Net$_i$ and compared their performance with that of NetProm through criterion (8) in 11 synthetic experiments, shown in Fig. 5. The reader is prompted to notice the different scaling of y-axis of this figure because this way the point that this figure wants to point out becomes more evident. On the other hand, the x-axis represents the number of frames which is also different across tunes. In the first 6 experiments and in the 9th, there was no change of the subject showing the expression. It can be verified that the value of $e(k)$, for all values of k shown in the horizontal axis, are close to the $e(0)$ value, so no need for adaptation was detected. On the contrary, the 7th, 8th, 10th, and 11th experiments contained one or more frames where a different subject (the first) showed a similar expression. In most of these

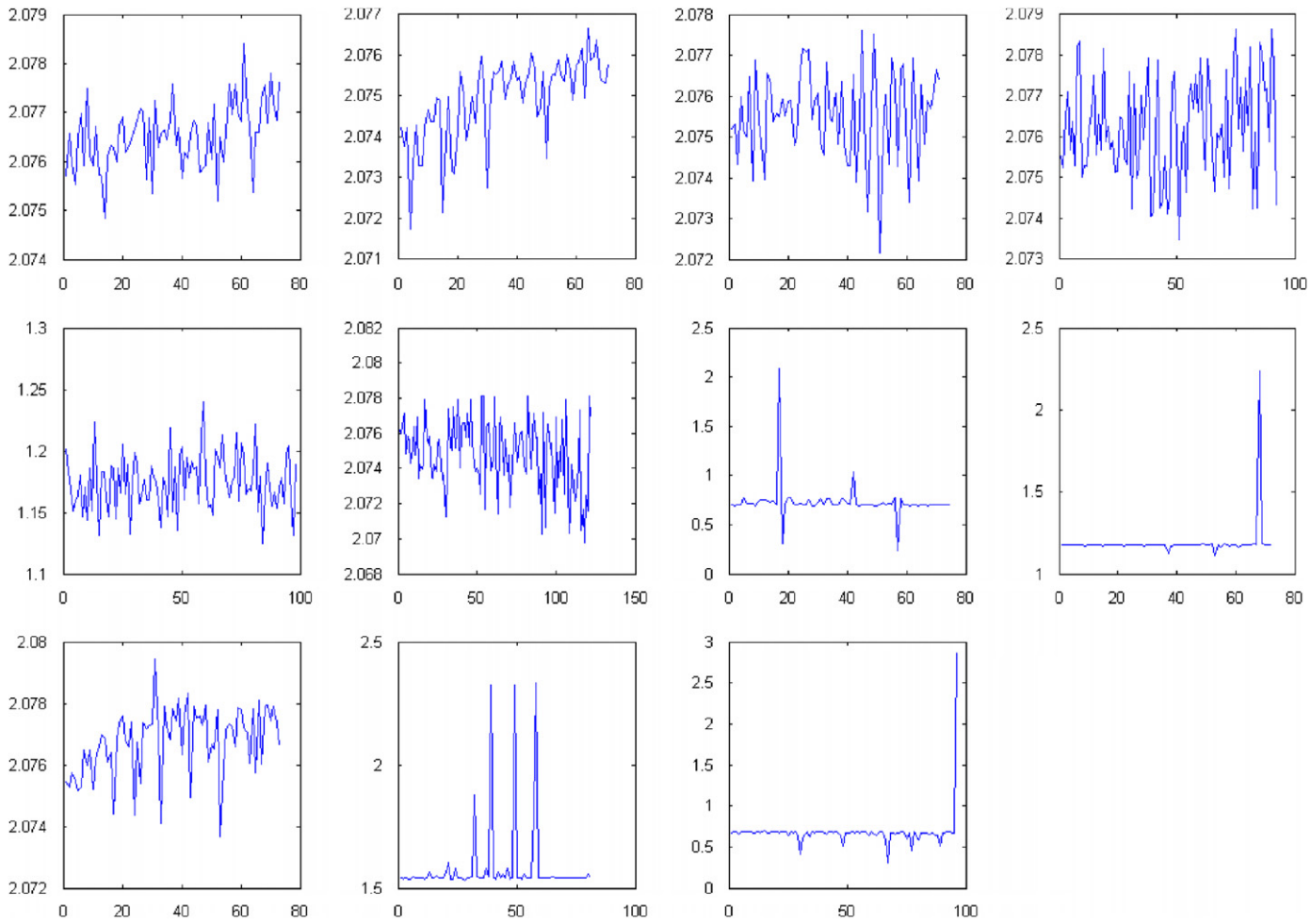


Fig. 5. Detecting the need for network adaptation using the criterion of Eq. (8).

cases the $\alpha(k)$ value was raised due to the inappropriateness of the adapted (to the fourth subject) network to cope well with the specific characteristics of the first subject. Consequently the need for (new) adaptation was detected through usage of criterion (8).

These results are very promising indicating that the proposed process can form an effective adaptation tool in expression/emotion recognition.

5. Conclusions

Recognition of facial expressions and hand gestures is a very important part of adapting HCI to the needs and feedback from the users, especially since psychological research has shown that the face is vital ingredient of human expressivity. However, in everyday HCI, emotions are usually subtle, hence difficult to pick out using a small set of universal labels; to tackle this, one needs to consider multiple modalities as a "fall-back" or reinforcement solution. In addition to this, personalized expressivity and context dependence make generalization of learning techniques a daunting task.

In this paper we proposed an extension of a neural network adaptation procedure, which caters for training from different modalities. After training and testing on a particular subject, the best-performing network is adapted using prominent samples from discourse with another subject, so as to adapt and improve its ability to generalize. Results shown here indicate that the performance of the network is improved using this approach, without the need to train a specific network for each subject, which would wipe out the nice generalization attribute of the network. Future work includes the extension of this work to include speech-related modalities, deployment on different naturalistic contexts and introduction of mechanisms to handle uncertainty in the various modalities and decide which of them would be the more robust to depend upon for co-training [3,20].

References

[1] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, S. Kollias, Emotion analysis in man–machine interaction systems, in: Samy Bengio, Hervé Bourlard (Eds.), Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, vol. 3361, Springer, New York, 2004, pp. 318–328.

[2] M. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behaviour, Proc. IEEE Conf. Automat. Face Gesture Recogn. (2006) 223–230.

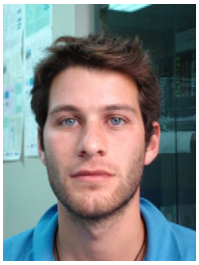
[3] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.

[4] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M. Schroeder, 'Feeltrace': an instrument for recording perceived emotion in real time, in: Proceedings of the ISCA Workshop on Speech and Emotion, 2000, pp. 19–24.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Processing Magazine, January 2001.

[6] N. Doulamis, A. Doulamis, S. Kollias, On-line retrainable neural networks: improving performance of neural networks in image analysis problems, IEEE Trans. Neural Networks 11 (1) (2000) 1–20.

[7] P. Ekman, T.S. Huang, T.J. Sejnowski, J.C. Hager, (Eds.), NSF understanding the face, A Human Face eStore, Salt Lake City, USA (see Library), 1993.
[8] P. Ekman, W.F. Friesen, The repertoire of nonverbal behavioural categories—origins, usage, and coding, Semiotica 1 (1969) 49–98.
[9] P. Ekman, W.F. Friesen, J.C. Hager, Facial action coding system, A Human Face, Salt Lake City, USA, 2002.
[10] B. Fasel, F. Monay, D. Gatica-Perez, Latent semantic analysis of facial action codes for automatic facial expression recognition, in: Proceedings of the ACM International Workshop on Multimedia Information Retrieval, 2004, pp. 181–188.
[11] A. Fridlund, The new ethology of human facial expression, in: J.A. Russell, J.M. Fernandez-Dols (Eds.), The Psychology of Facial Expression, Cambridge University Press, Cambridge, MA, USA, 1997, pp. 103–129.
[12] D. Goleman, Emotional Intelligence, Bantam Books, New York, NY, USA, 1995.
[13] U. Hess, S. Blairy, R.E. Kleck, The intensity of emotional facial expressions and decoding accuracy, J. Nonverbal Behav. 21 (4) (1997) 241–257.
[14] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, S. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy network, Special Issue on Emotion: Understanding & Recognition, Neural Networks, Elsevier, vol. 18(4), May 2005, pp. 423–435.
[15] A. Jaimes, N. Sebe, Multimodal Human Computer Interaction: A Survey, Computer Vision and Image Understanding, 2007 (available online).
[16] J. Kelley, Natural language and computers: six empirical steps for writing an easy-to-use computer application, unpublished Doctoral Dissertation, The Johns Hopkins University, 1983.
[17] D. Keltner, P. Ekman, Facial expression of emotion, in: M. Lewis, J.M. Haviland-Jones (Eds.), Handbook of Emotions, Guilford Press, New York, NY, USA, 2000, pp. 236–249.
[18] Promises and problems with the circumplex model of emotion, in: M.S. Clark (Ed.), Review of Personality and Social Psychology, Sage Publications, Newbury Park, USA, 1992.
[19] A. Legin, A. Rudnitskaya, B. Seleznev, Yu. Vlasov, Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie, Anal. Chim. Acta 534 (2005) 129–135.
[20] C. Mario Christoudias, Kate Saenko, Louis-Philippe Morency, Trevor Darrell, Co-adaptation of audio-visual speech and gesture classifiers, in: Proceedings of the Eighth International Conference on Multimodal Interfaces, Banff, Alberta, Canada, 2006, pp. 84–91.
[21] D. Matsumoto, Cultural similarities and differences in display rules, Motiv. Emotion 14 (1990) 195–214.
[22] Humaine Network of Excellence on Emotions, 〈www.emotion-research.net〉.
[23] S. Oviatt, Ten myths of multimodal interaction, Commun. ACM 42 (11) (1999) 74–81.
[24] M. Pantic, Affective computing, in: M. Pagani, (Ed.), Encyclopedia of Multimedia Technology and Networking, vol. 1, Idea Group Reference, Hershy, PA, USA, 2005, pp. 8–14.
[25] M. Pantic, N. Sebe, J.F. Cohn, T.S. Huang, Affective multimodal human-computer interaction, in: Proceedings of the ACM International Conference on Multimedia, 2005, pp. 669–676.
[26] M. Pantic, A. Pentland, A. Nijholt, T.S. Huang, Human computing and machine understanding of human behaviour: a survey, in: Proceedings of the ACM International Conference on Multimodal Interfaces, 2006, pp. 239–248.
[27] D. Park, M.A. EL-Sharkawi, R.J. Marks II, An adaptively trained neural network, IEEE Trans. Neural Networks 2 (1991) 334–345.
[28] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, I. Poggi, Embodied contextual agent in information delivering application, in: Proceedings of the International Conference on Autonomous Agents & Multi-Agent Systems, 2002.
[29] A. Pentland, Socially aware computation and communication, IEEE Comput. 38 (3) (2005) 33–40.
[30] R. Picard, Affective Computing, The MIT Press, Cambridge, MA, USA, 1997.
[31] R. Picard, Affective computing: challenges, Int. J. Hum.-Comput. Stud. 59 (1–2) (2003) 55–64.
[32] J.A. Russell, Is there universal recognition of emotion from facial expression?, Psychol. Bull. 115 (1) (1994) 102–141.
[33] J. Taylor, N. Fragopanagos, The interaction of attention and emotion, Neural Networks 18 (4) (2005) 353–369.
[34] J. Taylor, N. Fragopanagos, Modelling human attention and emotions, in: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, 2004, pp. 501–506.
[35] D. Watson, L.A. Clark, K. Weber, J. Smith-Assenheimer, M.E. Strauss, R.A. McCormick, Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples, J. Abnorm. Psychol. 104 (1995) 15–25.
[36] D. Watson, K. Weber, J.S. Assenheimer, L.A. Clark, M.E. Strauss, R.A. McCormick, Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales, J. Abnorm. Psychol. 104 (1995) 3–14.
[37] C. Whissel, The dictionary of affect in language, in: R. Plutchnik, H. Kellerman (Eds.), Emotion: Theory, Research and Experience: The Measurement of Emotions, vol. 4, Academic Press, New York, 1989, pp. 113–131.
[38] A. Wierzbicka, Reading human faces, Pragmatics Cognit. 1 (1) (1993) 1–23.
[39] L. Wu, S. Oviatt, P. Cohen, Multimodal integration—a statistical view, IEEE Trans. Multimedia 1 (4) (1999) 334–341.



George Caridakis graduated from the Department of Informatics and Telecommunications from the School of Sciences of the National and Kapodistrian University of Athens in 2004 and is now a Ph.D. candidate at the Department of Electrical and Computer Engineering, of the National Technical University of Athens. His current research interests lie in the areas of human computer interaction, artificial intelligence, neural networks, image and video processing, 3D computer animation, gesture and sign language recognition, sign language synthesis and virtual reality. He has published more than 15 papers in international journals and proceedings of international conferences. Since 2004 he has participated in 6 research projects at the Greek and European levels.

Kostas Karpouzis graduated from the Department of Electrical and Computer Engineering from the National Technical University of Athens in 1998 and received his Ph.D. degree in 2001 from the same University. His current research interests lie in the areas of human computer interaction, artificial intelligence, neural networks, image and video processing, 3D computer animation, sign language synthesis and virtual reality. Dr. Karpouzis has published more than 70 papers in international journals and in proceedings of international conferences. He is a member of the technical committee of the International Conference on Image Processing (ICIP) and a reviewer in many international journals. Since 1995 he has participated in more than 10 research projects at the Greek and European levels. Dr. Karpouzis is an associate researcher at the Institute of Communication and Computer Systems (ICCS) and holds an adjunct lecturer position in the University of Piraeus, teaching Medical Informatics and Image Processing. He is also a national representative in IFIPWorking Groups 12.5 'Artificial Intelligence Applications' and 3.2 'Informatics and ICT in Higher Education'.

Stefanos Kollias (S'81–M'85) was born in Athens, Greece, in 1956. He received his Diploma in electrical and computer engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1979, his M.Sc. degree in communication engineering in 1980 from the University of Manchester Institute of Science and Technology, Manchester, UK, and his Ph.D. degree in signal processing from the Computer Science Division, NTUA. He has been with the Electrical Engineering Department, NTUA, since 1986, where he currently serves as a Professor. Since 1990, he has been the Director of the Image, Video, and Multimedia Systems Laboratory, NTUA. His current research interests lie in the areas of image and video processing, analysis, coding, storage, retrieval, multimedia systems, computer graphics and virtual reality, artificial intelligence, neural networks, human computer interaction, and medical imaging. He has published more than 120 papers, 50 of which in international journals. He has been a member of the technical or advisory committee or invited speaker in 40 international conferences. He is a reviewer of ten IEEE Transactions and of ten other journals. Ten graduate students have completed their doctorate under his supervision, while another ten are currently performing their Ph.D. thesis. He and his team have been participating in 38 European and national projects.