Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Neural Networks 18 (2005) 423–435

Neural
Networks

www.elsevier.com/locate/neunet

2005 Special Issue

Emotion recognition through facial expression analysis based on a neurofuzzy network

Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis A. Tzouvaras,
Theofilos P. Mailis, Kostas C. Karpouzis, Stefanos D. Kollias*

Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering,
National Technical University of Athens, Zografou 15773, Greece

Received 24 March 2005; accepted 27 March 2005

Abstract

Extracting and validating emotional cues through analysis of users' facial expressions is of high importance for improving the level of interaction in man machine communication systems. Extraction of appropriate facial features and consequent recognition of the user's emotional state that can be robust to facial expression variations among different users is the topic of this paper. Facial animation parameters (FAPs) defined according to the ISO MPEG-4 standard are extracted by a robust facial analysis system, accompanied by appropriate confidence measures of the estimation accuracy. A novel neurofuzzy system is then created, based on rules that have been defined through analysis of FAP variations both at the discrete emotional space, as well as in the 2D continuous activation–evaluation one. The neurofuzzy system allows for further learning and adaptation to specific users' facial expression characteristics, measured though FAP estimation in real life application of the system, using analysis by clustering of the obtained FAP values. Experimental studies with emotionally expressive datasets, generated in the EC IST ERMIS project indicate the good performance and potential of the developed technologies.

Keywords: Facial expression analysis; MPEG-4 facial animation parameters; Activation evaluation emotion representation; Neurofuzzy network; Rule extraction; Adaptation

1. Introduction

Human–human interaction is an ideal model for designing affective interfaces, since the latter need to borrow many communication concepts to function intuitively. For example, in real situations, failure to receive or comprehend an input channel, such as prosody, is usually recovered by another channel, e.g. facial expression; this is an indication of the kind of robustness that is a major requirement of efficient multimodal HCI.

Facial features and expressions are critical to everyday communication. Besides speaker recognition, face assists a number of cognitive tasks: for example, the shape and motion of lips forming visemes can contribute greatly to speech comprehension in a noisy environment. While intuition may imply otherwise, social psychology research has shown that conveying messages in meaningful conversations can be dominated by facial expressions, and not spoken words. This result has led to renewed interest in detecting and analyzing facial expressions in not just extreme situations, but also in everyday human–human discourse.

A very important requirement for facial expression recognition is that all processes therein have to be performed without or with the least possible user intervention. This typically involves initial detection of face, extraction and tracking of relevant facial information, and facial expression classification. In this framework, actual implementation and integration details are enforced by the particular application. For example, if the application domain of the integrated system is behavioral science, real-time performance may not be an essential property of the system.

* Corresponding author. Address: Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 9, Heroon Politechniou str., Zografou GR-157 80, Greece. Tel.: +30 2107722488; fax: +30 2107722492.
E-mail address: stefanos@cs.ntua.gr (S.D. Kollias).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.03.004

As a general rule, an integrated recognition system should be able to classify all, or at least a wide range of visually distinguishable facial expressions; a robust and extensible face and facial deformation model is a vital requirement for this. Ideally, this would result in a particular face model setup uniquely describing a particular facial expression. A usual reference point is provided by the 44 facial actions defined in Facial Action Coding System (FACS) whose combinations form a complete set of facial expressions and facial expressions with a similar facial appearance. It has to be noted though that some of the facial action tokens included in FACS may not appear in meaningful facial expressions, since the purpose of FACS is to describe any visually distinguishable facial action and not to concentrate on emotional expressions.

Interpretation of the illustrated facial expression in terms of emotions is another important feature of a recognizer used in a multimodal HCI framework. Despite the domination of Ekman's theory of universal expressions, an ideal system should be able to adapt the classification mechanism according to the user's subjective interpretation of expressions, in order to cater for different emotion theories, such as Scherer's appraisal theory (Wehrle & Scherer, 2001) or the Ortony, Clore, and Collins (1988) (OCC) model. Besides this, discrete emotion models by definition cannot capture blended emotion displays, e.g. guilt. In order to provide accurate and purposeful results, an ideal recognizer should perform classification of facial expression into multiple emotion categories or utilize existing emotion modelling knowledge to create intermediate ones.

1.1. Review of facial expression recognition

The origins of facial expression analysis go back into the 19th century, when Darwin originally proposed the concept of universal facial expressions in man and animals. Since the early 1970s, Ekman and Friesen (1975) have performed extensive studies of human facial expressions, providing evidence to support this universality theory. These 'universal facial expressions' are those representing happiness, sadness, anger, fear, surprise, and disgust. To prove this, they provide results from studying facial expressions in different cultures, even primitive or isolated ones. These studies show that the processes of expression and recognition of emotions on the face are common enough, despite differences imposed by social rules. Ekman and Friesen used FACS to manually describe facial expressions, using still images of, usually extreme, facial expressions. This work inspired researchers to analyze facial expressions by tracking prominent facial features or measuring the amount of facial movement, usually relying on the 'universal expressions' or a defined subset of them. In the 1990s, automatic facial expression analysis research gained much interest, mainly thanks to progress, in the related fields such as image processing (face detection, tracking and recognition) and the increasing availability of relatively cheap computational power.

In one of the ground-breaking and most publicized works, Mase and Pentland (1990) used measurements of optical flow to recognize facial expressions. In the following, Lanitis et al. used a flexible shape and appearance model for face identification, pose recovery and facial expression recognition. Black and Yacoob (1997) proposed local parameterized models of image motion to recover non-rigid facial motion, which was used as input to a rule-based basic expression classifier; Yacoob and Davis (1996) also worked in the same framework, this time using optical flow as input to the rules. Local optical flow was also the basis of Rosenblum's work, utilizing a radial basis function network for expression classification. Otsuka and Ohya utilized the 2D Fourier transform coefficients of the optical flow as feature vectors for a hidden Markov model (HMM).

Regarding feature-based techniques, Donato, Bartlett, Hager, Ekman, and Sejnowski (1999) tested different features for recognizing facial AUs and inferring the facial expression in the frame. Oliver et al. tracked the lower face to extract mouth shape information and fed them to an HMM, recognizing again only universal expressions.

As shown above, most facial expression analysis systems focus on facial expressions to estimate emotion-related activities. Furthermore, the introduction and correlation of multiple channels may increase robustness, as well as improve interpretation disambiguation in real-life situations. Most attempts at channel fusion evaluate speech, in addition to facial expressions. Here, expressions may be conveyed by linguistic, as well as prosodic features, such as the fundamental frequency, intensity and pause timing. Cohn and Katz (1998) as well as Chen et al. focused on the fundamental frequency, as it is an important voice feature for emotion recognition and can be easily extracted. It has to be noted though, that introducing speech in the expression recognition picture has to be followed by separate provisions for aural and visual information synchronization. This is essential because, in the general case, events regarding these two channels do not occur simultaneously and may affect one another (e.g. visual information from the mouth area generally deteriorates when the subject is speaking).

The current paper introduces an expression recognition system which can be robust to facial expression variations among different users. This system evaluates facial expressions through the robust analysis of appropriate facial features. A novel neurofuzzy system is created, based on rules that have been defined through analysis of FAP variations both at the discrete emotional space, as well as in the 2D continuous activation–evaluation one. The neurofuzzy system allows for further learning and adaptation to specific users' facial expression characteristics, measured though FAP estimation in real life application of the system, using analysis by clustering of the obtained FAP values.

The rest of this paper is structured as follows: Section 2 presents the facial feature extraction process while Section 3 describes facial expression analysis and the derived rules for emotion recognition based on facial expression analysis. Section 4 describes a neurofuzzy platform for incorporating the emotion recognition rules, which also provides adaptation to specific user profiles, based on a learning procedure. Experimental results are presented in Section 5, while Section 6 presents the conclusions and future research.

2. Facial feature extraction

An overview of the facial analysis and feature extraction system is given in Fig. 1. At first face detection is performed using non-parametric discriminant analysis with a Support Vector Machine (SVM) (Fransens & De Prins, 2003), which classifies face and non-face areas by reducing the training problem dimension to a fraction of the original with negligible loss of classification performance. The face detection step provides us with a rectangle head boundary, which includes the whole face area. The latter is segmented roughly using static anthropometric rules (Young, 1993) into three overlapping rectangle regions of interest which include both facial features and facial background; these three feature-candidate areas include the left eye/eyebrow, the right eye/eyebrow and the mouth. Continuing, we utilize these areas to initialize the feature extraction process. Facial feature extraction performance depends on head pose, thus head pose needs to be detected and the head restored in the upright position; in this work we are mainly concerned with roll rotation, since it is the most frequent rotation encountered in real life video sequences. To estimate the head pose we first locate the left and right eyes in the corresponding eye candidate areas and estimate head roll rotation by calculating the angle between the horizontal plane and the line defined by the eye centers.

For eye localization, we propose an efficient technique using a feed-forward back propagation neural network with a sigmoidal activation function. The multi-layer perceptron (MLP) we adopted employs Marquardt–Levenberg learning (Kollias & Anastassiou, 1989), while the optimal architecture obtained through pruning has two 20 node hidden layers for 13 inputs. The network is applied separately on the left and right eye-candidate face regions; for each pixel in those regions the 13 NN inputs are the luminance Y, the Cr & Cb chrominance values and the 10 most important DCT coefficients (with zigzag selection) of the neighbouring 8×8 pixel area. Using additional input color spaces such as Lab, RGB or HSV to train the network, has not increased its distinction efficiency. The MLP has two outputs, one for each class, namely eye and non-eye, and it has been trained with more than 100 hand-made eye masks that depict eye and non-eye area in random frames from the ERMIS (FP5 IST ERMIS project) database, in images of diverse quality, resolution and lighting conditions. The network's output for facial images outside the training set is good for locating the eye, however, it cannot provide accurate information near the eye boundaries. The output of the aforementioned network is fed to other techniques in order to create facial feature masks, i.e. binary maps indicating the position and extent of each facial feature. The left, right, top and bottom-most coordinates of the eye and mouth masks, the left, right and top coordinates of the eyebrow masks as well as the nose coordinates, are used to define the considered feature points (FPs).

For the nose and each of the eyebrows, a single mask is created. On the other hand, since the detection of eyes and mouth can be problematic in low-quality images, a variety of methods is used each resulting in a different mask.

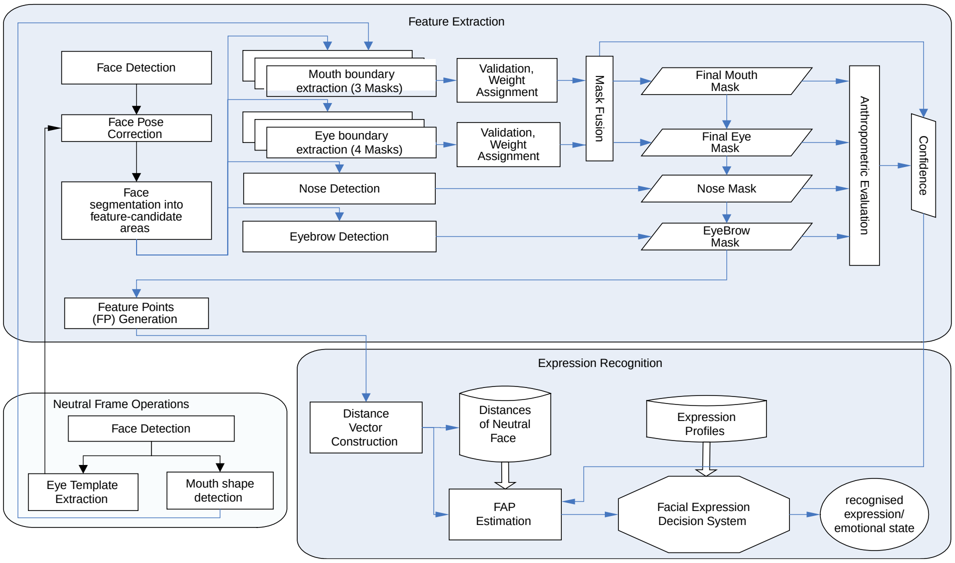
Feature Extraction
Face Detection
Face Pose Correction
Face segmentation into feature-candidate areas
Mouth boundary extraction (3 Masks)
Validation, Weight Assignment
Eye boundary extraction (4 Masks)
Validation, Weight Assignment
Nose Detection
Eyebrow Detection
Mask Fusion
Final Mouth Mask
Final Eye Mask
Nose Mask
EyeBrow Mask
Anthropometric Evaluation
Confidence
Feature Points (FP) Generation
Neutral Frame Operations
Face Detection
Eye Template Extraction
Mouth shape detection
Expression Recognition
Distance Vector Construction
Distances of Neutral Face
Expression Profiles
FAP Estimation
Facial Expression Decision System
recognised expression/ emotional state

Fig. 1. The facial feature extraction and facial expression analysis system.

In total, we have four masks for each eye and three for the mouth. These masks have to be calculated in near real-time thus we had to avoid utilizing complex or time-consuming feature extractors. The feature extractors developed for this work are briefly described in the following.

2.1. Facial feature mask extraction

Eyebrows are detected with a procedure involving morphological edge detection and feature selection using data from (Young, 1993). Nose detection is based on nostril localization. Nostrils are easy to detect due to their low intensity (Gorodnichy, 2002). Connected objects (i.e. nostril candidates) are labeled based on their vertical proximity to the left or right eye, and the best pair is selected according to its position, luminance and geometrical constraints from (Young, 1993).

For the eyes the following masks are constructed:

- A refined version of the original neural-network derived mask. The initial eye mask is extended by using an adaptive low-luminance threshold on an area defined from the neural network high-confidence output. This mask includes the top and bottom eyelids in their full extent that are usually missing from the initial mask (Fig. 2e).
- A mask expanding in the area between the upper and lower eyelids. Since the eye-center is almost always detected correctly from the neural network, the horizontal edges of the eyelids in the eye area are used to limit the eye mask in the vertical direction. A modified Canny edge operator is used due to its property of providing good localization. The operator is limited to ignore movements in the most vertical directions (Fig. 2b).
- A region-growing technique that takes advantage from the fact that texture complexity inside the eye is higher compared to the rest of the face. This process consists of thresholding the iteratively reduced grayscale eye image with its 3×3 standard deviation map, while the resulting binary eye mask center remains close to the original. This process is found to perform very well for images of very-low resolution and low color quality (Fig. 2c).

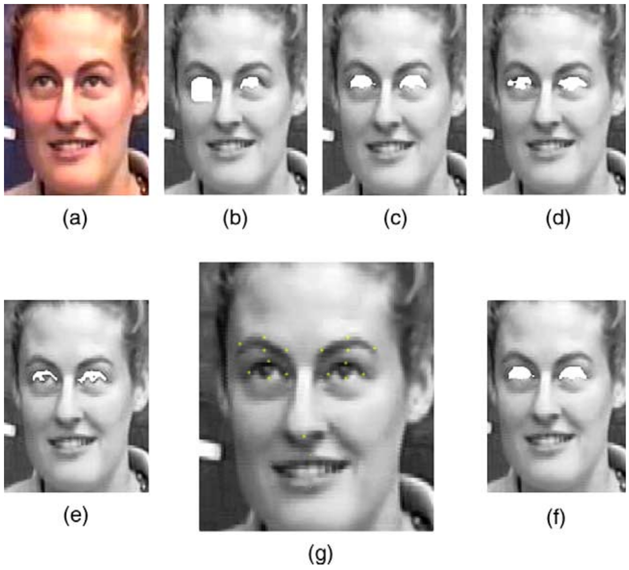

Fig. 2. (a) Original frame, (b)–(e) the four detected masks, (f) final mask for the eyes and (g) all detected feature points from the final mask.

- A mask computed using the normal probability of luminance using a simple adaptive threshold on the eye area. This mask includes the darkest areas of the eye area, which usually includes the sclera and eyelashes but can extend outside the eye area when illumination is not uniform, thus it is cut vertically at its thinnest points from both sides of the eye centre and the convex hull of the result is used (Fig. 2d).

Finding the extent of a closed mouth in a still image is a relatively easy accomplished task (Hagan & Menhaj, 1994). In case of an open mouth, several methods have been proposed which make use of intensity (Yin, 2001) or color information (Leung, Wang, & Lau, 2004). In this work, we propose three different approaches that are then fused in order to produce the final mask:

- An MLP neural network is trained to identify the mouth region using the neutral image. The network has similar architecture as the one used for the eyes. The train data are acquired from the neutral image (where the mouth is closed) as follows: the mouth-candidate ROI is first filtered with Alternating Sequential Filtering by Reconstruction (ASFR) to simplify and create connected areas of similar luminance. Simple but effective luminance thresholding is then used to find the area between the lips in the neutral image where the mouth is closed. This area is dilated vertically and the data depicted by this area are used to train the network.
- A horizontal morphological gradient is calculated in the mouth area and the longest connected object which comply with constrains from (Fransens & De Prins, 2003) and the nose position is selected as a possible mouth mask.
- This final approach takes advantage of the relative low luminance of the lip corners and contributes to the correct identification of horizontal mouth extent which is not always detected by the previous methods in cases of smiling and apparent teeth. A short summary of the procedure is as follows: the image is simplified and thresholded and connected objects are labelled. Two cases are examined separately: either we have no apparent teeth and the mouth area is denoted by a cohesive dark area or there are teeth and thus two dark areas appear at both sides of the teeth. In the first case mouth extend is straightforward to detect; in the latter mouth centre proximity of each object is assessed through (Fransens & De Prins, 2003) and the appropriate objects are selected. The convex hull of the result is then merged through morphological reconstruction with an

horizontal edge map to include the upper and bottom lips. The result is the third mouth mask.

Since, as was already mentioned, the detection of a mask using any of these applied methods can be problematic, all detected masks have to be validated against a set of criteria. Each one of the criteria examines the masks in order to decide whether they have acceptable size and position for the feature they represent. This set of criteria consist of relative anthropometric measurements, such as the relation of the eye and eyebrow vertical positions, which when applied to the corresponding masks produce a value in the range [0,1] with zero denoting a totally invalid mask.

For the features for which more than one masks have been detected using different methodologies, the multiple masks have then to be fused together to produce a final mask. The choice for mask fusion, rather than simple selection of the mask with the greatest validity confidence, is based on the observation that the methodologies applied in the initial masks' generation produce different error patterns from each other, since they rely on different image information or exploit the same information in fundamentally different ways. Thus, combining information from independent sources has the property of alleviating a portion of the uncertainty present in the individual information components.

The mask fusion approach described in the following is not bound to specific feature extractors; more and different extractors than those described above can be developed for each feature, as long as they provide better results in difficult situations where other extractors fail. The feature extractors briefly described above are merely the ones developed for this specific work. The fusion algorithm is inspired from the structure of Dynamic Committee Machines (DCM) combining the masks based on their validity confidence and producing a final mask together with the corresponding estimated confidence (Dietterich, 2000) for each facial feature. Each of those masks represents the best-effort result of the corresponding mask-extraction method used. The most common problems, especially encountered in low quality input images, are connection with other feature boundaries or mask dislocation due to noise. If y_{comb} is the combined machine output and t the desired output it has been proven in the committee machine (CM) theory (Krog & Vedelsby, 1995) that the combination error y_{comb}-t from different machines f_i is guaranteed to be lower than the average error:

$$(y_{\mathrm{comb}}-t)^2 = \frac{1}{M}\sum_i (y_i-t)^2 - \frac{1}{M}\sum_i (y_i-y_{\mathrm{comb}})^2 \qquad (2.1)$$

In a Static CM, the voting weight for a component is proportional to its error on a validation set. In DCMs (Fig. 3), input is directly involved in the combining mechanism through a Gating Network (GN), which is used to modify those weights dynamically.

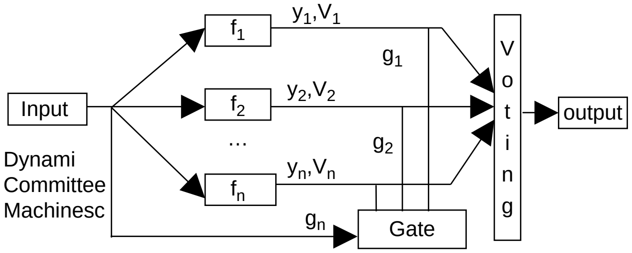

Fig. 3. Dynamic committee machine architecture.

In our case, the final masks for the left eye, right eye and mouth, $\mathbf{M}_f^{e_L}$, $\mathbf{M}_f^{e_R}$, $\mathbf{M}_f^{m}$ are considered as the machine output and the final confidence values of each mask for feature x $M_x^{c_f}$ are considered as the confidence of each machine. Therefore, for feature x, each element m_f^x of the final mask $\mathbf{M}_f^x$ is calculated from the n masks as

$$m_f^x = \frac{1}{n}\sum_{i=1}^{n} m_i^x M_f^{cx_i} h^i g^i \qquad (2.2)$$

$$h^k = \begin{cases} 1, & M_f^{c,x_k} \geq (t_{\mathrm{vd}} \cdot \langle M_q^{c,x_k} \rangle_q) \\ 0, & M_f^{c,x_k} < (t_{\mathrm{vd}} \cdot \langle M_q^{c,x_k} \rangle_q) \end{cases} \qquad (2.3)$$

where m_i^x is the element of mask M_i^x, M_f^{c,x_i} the validation value of mask i and h^i is used to prevent the masks with $M_f^{c,x_k} < (t_{\mathrm{vd}} \cdot \langle M_q^{c,x_k} \rangle_q)$ to contribute to the final mask. A sufficient value for t_{vd} is 0.8.

The role of the gating variable g^i is to favor the color-based feature extraction methods ($\mathbf{M}_1^{\mathrm{e}}$, $\mathbf{M}_1^{\mathrm{m}}$) in images of high color and resolution. In this stage, two variables are taken into account: image resolution and color quality. More information about the used expression profiles can be found in (Raouzaiou, Tsapatsoulis, Karpouzis, & Kollias, 2002).

3. Facial expression recognition

In our former research on emotion recognition, a rule-based system was created, characterizing a user's emotional state in terms of the six universal, or archetypal, expressions (joy, surprise, fear, anger, disgust, sadness).

We have created rules in terms of the MPEG-4 FAPs for each of these expressions, by analysing the FAPS extracted from the facial expressions of the Ekman dataset (Raouzaiou et al., 2002). This dataset contains several images for every one of the six archetypal expressions, which, however, are rather exaggerated. A result of this fact is that the rules extracted from this dataset if used in real data, cannot have accurate results, especially if the subject is not very expressive.

3.1. Quadrants

Newer psychological studies claim that the use of quadrants of emotion's wheel (see Fig. 4) (Whissel, 1989) instead of the six archetypal expressions is more accurate.

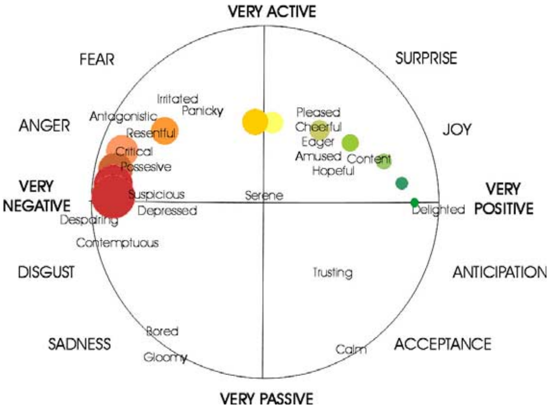

Fig. 4. The activation–emotion space.

So the creation of rules for every one of the first-three quadrants—no emotion is lying in the fourth quadrant—was necessary. A newer statistical analysis, taking into account the results of Whissel's study and in particular the *activation* parameter, was realised.

In order to do this, we translate facial muscle movements into FAPs. FAPs of expressions of every quadrant are also experimentally verified through analysis of prototype datasets. In order to make comparisons with real expression sequences, we model FAPs employed in the facial expression formation through the movement of particular Feature Points (FPs)—the selected FPs can be automatically detected from real images or video sequences. In the next step, we estimate the range of variation of each FAP. This is achieved by analyzing real images and video sequences as well as by animating synthesized examples. Table 1 illustrates three examples of rules which were created based on the developed methodology.

In order to use these rules in a system dealing with the continuous activation–emotion space and fuzzy representation, we transformed the rules replacing the range of variation with the terms *high*, *medium*, *low* after having normalized the corresponding partitions. The full set of rules is illustrated in Table 2.

4. Network platform: learning/adaptation to specific users

In the following, a neurofuzzy learning platform is described, which can both incorporate the above 41 rules and adapt to the specific expression characteristics of each individual user Table 2.

Any conditional (If–Then) fuzzy proposition can be expressed in terms of a fuzzy relation R between the two variables involved. One way to determine R is using the fuzzy implication, which operates on fuzzy sets involved in the fuzzy proposition. However, the problem of determining R for a given conditional fuzzy proposition can be detached from fuzzy implications and determine R using fuzzy relational equations.

The equation to be solved for fuzzy modus ponens has the form

$$B = A \circ^t R, \tag{4.1}$$

where A and B are given fuzzy sets that represent, respectively, the IF-and the THEN-part in the conditional fuzzy proposition involved and t is a t-norm. Eq. (4.1) is solvable for R if $A \circ^{\omega_t} B$ is a solution, where

$$\omega_t(a, b) = \sup\{x \in [0,1] | t(a,x) \leq b\} \tag{4.2}$$

for every a, $b \in [0,1]$ and a continuous t-norm t.

In the following section, we present a complete algorithm for solving fuzzy relational equations for the interpretation of inference rules in the respective fuzzy extension of propositional logics. The proposed interpretation algorithm is realized using a hybrid neurofuzzy architecture shown in Fig. 5.

4.1. Neurofuzzy network

Let $y = [y_1, y_2, \ldots, y_m]$ denote a fuzzy set defined on the set of output predicates, the truth of which will be examined. Actually, each y_i represents the degree in which the ith output fuzzy predicate is satisfied. The input of the proposed neurofuzzy network is a fuzzy set $x = [x_1, x_2, \ldots, x_n]$ defined on the set of the input predicates, with each x_i representing the degree in which the ith input predicate is detected. The proposed network represents the association $f: X \rightarrow Y$ which is the knowledge of the system, in a neurofuzzy structure. After the evaluation of the input predicates, some output predicates represented in the knowledge of the system can be recognized with the aid of fuzzy systems' reasoning (Klir & Yuan, 1995). One of the widely used ways of constructing fuzzy inference systems is the method of approximate reasoning which can be implemented on the basis of

Table 1
Rules with FAP range of variation in MPEG-4 units

$F_6 \in [160,240]$, $F_7 \in [160,240]$, $F_{12} \in [260,340]$, $F_{13} \in [260,340]$, $F_{19} \in [-449,-325]$, $F_{20} \in [-426,-302]$, $F_{21} \in [325,449]$, $F_{22} \in [302,426]$, $F_{33} \in [70,130]$, $F_{34} \in [70,130]$, $F_{41} \in [130,170]$, $F_{42} \in [130,170]$, $F_{53} \in [160,240]$, $F_{54} \in [160,240]$	$(+,+)$
$F_{16} \in [45,155]$, $F_{18} \in [45,155]$, $F_{19} \in [-330,-200]$, $F_{20} \in [-330,-200]$, $F_{31} \in [-200,-80]$, $F_{32} \in [-194,-74]$, $F_{33} \in [-190,-70]$, $F_{34} \in [-190,-70]$, $F_{37} \in [65,135]$, $F_{38} \in [65,135]$	$(-,+)$
$F_3 \in [400,560]$, $F_5 \in [-240,-160]$, $F_{19} \in [-630,-570]$, $F_{20} \in [-630,-570]$, $F_{21} \in [-630,-570]$, $F_{22} \in [-630,-570]$, $F_{31} \in [460,540]$, $F_{32} \in [460,540]$, $F_{33} \in [360,440]$, $F_{34} \in [360,440]$, $F_{35} \in [260,340]$, $F_{36} \in [260,340]$, $F_{37} \in [60,140]$, $F_{38} \in [60,140]$	$(-,+)$

Table 2
Rules used in the emotion recognition system

Rule	FAP	Quadrant
1	F3_H+F4_L+F5_VL+[F53+F54]_H+[F19+F21]_H+[F20+F22]_H	(+,+)
2	F3_M+F4_L+F5_L+[F53+F54]_H+[F19+F21]_H+[F20+F22]_H	(+,+)
3	F3_M+F4_L+F5_H+[F53+F54]_H+[F19+F21]_H+[F20+F22]_H	(+,+)
4	F3_H+F4_L+F5_L+[F53+F54]_H+[F19+F21]_H+[F20+F22]_H	(+,+)
5	F3_L+F4_M+F5_H+[F53+F54]_M+[F19+F21]_H+[F20+F22]_H+F31_M+F32_M+F33_M+F34_M+F37_M+F38_M+F59_H+F60_H	(+,+)
6	F3_H+F4_L+F5_VL+[F53+F54]_M+[F19+F21]_H+[F20+F22]_H	(+,+)
7	F3_L+F4_L+F5_H+[F53+F54]_H+[F19+F21]_H+[F20+F22]_H+[F37+F38]_M+F59_H+F60_H	(+,+)
8	F3_H+F5_VL+[F53+F54]_L+[F19+F21]_L+[F20+F22]_L+F31_H+F32_H+F33_H+F34_H+F35_H+F36_H+F37_L+F38_L+[F37+F38]_L	(+,+)
9	F3_H+F5_VL+[F53+F54]_M+[F19+F21]_L+[F20+F22]_L+F31_H+F32_H+F33_H+F34_H+F35_H+F36_H+F37_L+F38_L	(+,+)
10	F3_M+F5_L+[F53+F54]_L+[F19+F21]_L+[F20+F22]_L+F31_H+F32_H+F33_H+F34_H+F35_H+F36_H	(+,+)
11	F3_H+F5_VL+[F53+F54]_L+[F19+F21]_L+[F20+F22]_L+F31_H+F32_H+F33_H+F34_H+F35_M+F36_M	(+,+)
12	F3_H+F5_VL+[F53+F54]_L+[F19+F21]_L+[F20+F22]_L+F31_M+F32_M+F33_H+F34_H+F35_M+F36_M	(+,+)
13	F3_L+F4_M+F5_H+F31_L+F32_L+F33_L+F34_L+F37_H+F38_H+[F37+F38]_H+F59_M+F60_M	(−,+)
14	F3_L+F4_M+F5_L+F31_L+F32_L+F33_L+F34_L+F37_M+F38_M+[F37+F38]_H	(−,+)
15	F3_L+F4_M+F5_H+F31_M+F32_M+F33_L+F34_L+F37_H+F38_H+[F37+F38]_H	(−,+)
16	F3_L+F4_M+F5_L+F31_L+F32_L+F33_L+F34_L+F37_H+F38_H+[F37+F38]_H+F59_M+F60_M	(−,+)
17	F3_H+F4_L+F5_VL+[F53+F54]_L+F31_M+F32_M+F33_L+F34_L+F35_L+F36_L+F37_H+F38_H+[F37+F38]_H+F59_L+F60_L	(−,+)
18	F3_H+F4_M+F5_VL+[F53+F54]_L+F31_L+F32_L+F33_L+F34_L+F35_L+F36_L+F37_H+F38_H+[F37+F38]_H	(−,+)
19	F3_H+F4_M+F5_VL+[F53+F54]_L+F31_L+F32_L+F33_L+F34_L+[F37+F38]_H+F59_L+F60_L	(−,+)
20	F3_H+F4_L+F5_VL+[F53+F54]_L+F31_M+F32_M+F33_M+F34_M+F37_H+F38_H+[F37+F38]_H+F59_M+F60_M	(−,+)
21	F3_M+F4_L+F5_L+[F53+F54]_L+F31_L+F32_L+F33_L+F34_L+F37_H+F38_H+[F37+F38]_H+F59_M+F60_M	(−,+)
22	F3_L+F4_M+F5_H+F31_M+F32_M+F33_L+F34_L+F37_H+F38_H+[F37+F38]_M	(−,+)
23	F3_L+[F53+F54]_M+F5_H+[F19+F21]_L+[F20+F22]_L+F31_M+F32_M+F33_M+F34_M	(−,+)
24	F3_M+[F19+F21]_L+[F20+F22]_L+F31_H+F32_H+F37_M+F38_M	(−,+)
25	F3_M+F4_M+F5_H+[F19+F21]_L+[F20+F22]_L+F33_M+F34_M+F35_H+F36_H	(−,+)
26	F3_M+F5_L+[F19+F21]_H+[F20+F22]_H+F31_H+F32_H+F33_M+F34_M+F35_M+F36_M+[F37+F38]_H	(−,−)
27	F3_M+F4_L+F5_L+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+[F37+F38]_M	(−,−)
28	F3_M+F4_L+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+F35_L+F36_L+F37_M+F38_M+[F37+F38]_L	(−,−)
29	F3_L+F4_L+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_M+F32_M+F33_M+F34_M+[F37+F38]_M	(−,−)
30	F3_L+F4_L+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+[F37+F38]_H	(−,−)
31	F3_L+F4_L+F5_H+[F53+F54]_M+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+[F37+F38]_M	(−,−)
32	F3_L+F4_M+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+[F37+F38]_H+F59_H	(−,−)
33	F3_M+F4_L+F5_H+[F53+F54]_M+[F19+F21]_H+[F20+F22]_H+F31_M+F32_M+F33_M+F34_M+F35_M+F36_M+[F37+F38]_H+F60_H	(−,−)
34	F3_L+F4_L+[F53+F54]_L+F31_M+F32_M+F33_M+F34_M+F35_M+F36_M+F37_M+F38_M+[F37+F38]_M	(−,−)
35	F3_H+F4_L+F5_VL+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+F35_L+F36_L+F37_H+F38_H+[F37+F38]_H+F59_H	(−,−)
36	F3_L+F4_M+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+[F37+F38]_H+F60_H	(−,−)
37	F3_M+F4_L+F5_H+[F53+F54]_M+[F19+F21]_H+[F20+F22]_H+F31_M+F32_M+F33_M+F34_M+F35_M+F36_M+[F37+F38]_H+F59_H	(−,−)
38	F3_L+F4_L+F5_H+[F53+F54]_M+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+[F37+F38]_H+F59_H+F60_H	(−,−)

(*continued on next page*)

Table 2 (*continued*)

Rule	FAP	Quadrant
39	F3_L+F4_L+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_M+F32_M+F33_M+F34_M+[F37+F38]_H+F59_H+F60_H	(−,−)
40	F3_H+F4_L+F5_VL+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+F35_L+F36_L+F37_H+F38_H+[F37+F38]_H+F60_H	(−,−)
41	F3_L+F4_M+F31_M+F32_M+F33_M+F34_M+F35_M+F36_M+F37_M+F38_M+[F37+F38]_M+F59_M+F60_M	Neutral

compositional rule of inference (Klir & Yuan, 1995). The need for results with theoretical soundness lead to the representation of fuzzy inference systems on the basis of generalized sup-*t*-norm compositions (Jenei, 1998; Kosko, 1992).

The class of *t*-norms has been studied by many researchers (Hirota & Pedrycz, 1996; Jenei, 1998; Lin and Lee, 1995). Using the definition ω_t in Eq. (4.2) two additional operators $\hat{\omega}_t, \breve{\omega}_t : [0,1]\times[0,1]\rightarrow[0,1]$ are defined by the following relations

$$\hat{\omega}_t(a,b)=\begin{cases}1 & a<b\\ a^{\hat{t}}b & a\geq b\end{cases}, \qquad \breve{\omega}_t(a,b)=\begin{cases}0 & a<b\\ a^{\check{t}}b & a\geq b\end{cases}$$

where $a^{\hat{t}}b=\sup\{x\in[0,1]:t(a,x)=b\}$, $a^{\check{t}}b=\inf\{x\in[0,1]:t(a,x)=b\}$.

With the aid of the above operators, compositions of fuzzy relations can be defined. These compositions are used in order to construct fuzzy relational equations and represent the rule-based symbolic knowledge with the aid of fuzzy inference (Stamou & Tzafestas, 2000).

Let X, Z, Y be three discrete crisp sets with cardinalities n, l and m, respectively, and $A(X,Z)$, $B(Z,Y)$ be two binary fuzzy relations. The definitions of sup-t and inf-$\hat{\omega}_t$ compositions are given by:

$$(A\circ^{t}B)(i,j)=\sup_{k\in N_l} t\{A(i,k),B(k,j)\},\quad i\in N_n,\ j\in N_m \tag{4.3}$$

$$(A\circ^{\hat{\omega}_t}B)(i,j)=\inf_{k\in N_l} \hat{\omega}_t\{A(i,k),B(k,j)\},\quad i\in N_n,\ j\in N_m \tag{4.4}$$

Let us now proceed to a more detailed description of the proposed neurofuzzy architecture shown in Fig. 7. It consists of two layers of compositional neurons which are extensions of the conventional neurons (Fig. 6) (Stamou & Tzafestas, 2000). While the operation of the conventional neuron is described by the equation

$$y=\alpha\left(\sum_{i=1}^{n} w_i x_i+\vartheta\right), \tag{4.5}$$

where α is non-linearity, ϑ is threshold and w_i are the weights, the operation of the sup-t compositional neuron is described by the equation

$$y=a'\left\{\sup_{j\in N_n} t(x_i,w_i)\right\}, \tag{4.6}$$

where t is a t-norm and α is the following activation function:

$$a'(z)=\begin{cases}0, & x\in(-\infty,0)\\ x, & x\in[0,1]\\ 1, & x\in(0,+\infty)\end{cases}. \tag{4.7}$$

A second type of compositional neuron is constructed using the $\hat{\omega}_t$ operation. The neuron equation is given by:

$$y=a'\left\{\inf_{j\in N_n} \hat{\omega}_t(x_i,w_i)\right\} \tag{4.8}$$

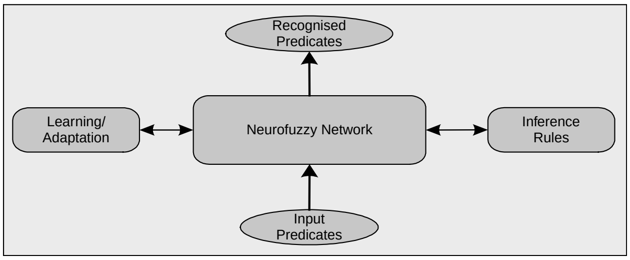

Fig. 5. The neurofuzzy architecture.

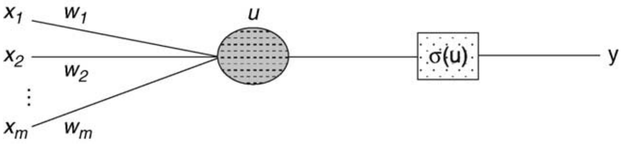

Fig. 6. The structure of a compositional neuron.

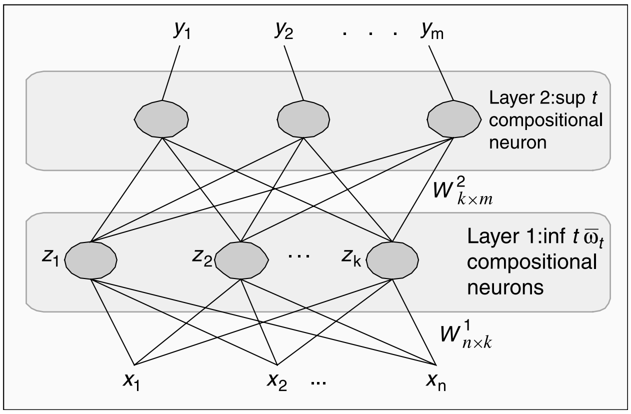
y_1
y_2
. . .
y_m
Layer 2:sup t compositional neuron
$W^2_{k\times m}$
z_1
z_2
…
z_k
Layer 1:inf $t\,\bar{\omega}_t$ compositional neurons
$W^1_{n\times k}$
x_1
x_2
…
x_n

Fig. 7. The neurofuzzy network.

The proposed architecture is a two-layer neural network of compositional neurons as shown in Fig. 7. The first layer consists of the inf-$\hat{\omega}_t$ neurons and the second layer consists of the sup-t neurons. The system takes as input, predicates, and gives to the output the recognized output predicates. The first layer computes the antecedents of the mapping rules, while the second implements the fuzzy reasoning using the fuzzy modus ponens schema.

The rules are used to initialize the neurofuzzy network (giving its initial structure and weights). During the learning process the number of neurons in the hidden layer and the weights of the two layers may change with the aid of a learning with the objective of the error minimization. The learning algorithm that supports the above network is applied in each layer independently. During the learning process, the weight matrices are adapted in order to approximate the solution of the fuzzy relational equation describing the association of the input with the output. Using a traditional minimization algorithm (for example the steepest descent), we cannot take advantage of the specific character of the problem. The algorithm that we use is based on a more sophisticated credit assignment that 'blames' the neurons of the network using the knowledge about the topographic structure of the solution of the fuzzy relation equation (Stamou & Tzafestas, 2000). After the learning process, the network keeps its transparent structure and the new knowledge represented in it can be extracted in the form of mapping If–Then rules.

4.2. Learning operation

In the process of knowledge adaptation, the If–Then rules are inserted into the proposed neurofuzzy system. This refers to automatically transforming the structured knowledge provided by the knowledge base in order to perform the followings:

(a) Define the required input predicates as *input predicate(1), input predicate(2),…, input predicate(n)*. The input predicates will define the set $X=\{x_1,x_2,\ldots,x_n\}$.

(b) Define the required output predicates as *output predicate(1), output predicate(2),…, output predicate(n)*. The output predicates will define the set $Y=\{y_1,y_2,\ldots,y_m\}$.

(c) Insert the a priori knowledge given in If–Then rules of the form 'if *input predicate(1)* and *input predicate(2)* and … then *output predicate(5)*' into the neurofuzzy structural elements (the weights of the neurofuzzy system). The number of different antecedents (If parts of the rules) defines the set $Z=\{z_1,z_2,\ldots,z_l\}$. The predicates could be associated with confidence levels in order to produce the antecedents; this means that the antecedents could have the form (*input predicate(1), input predicate(2), 0.7, 0.9*), with the 0.7 and 0.9 values corresponding to confidence levels. The above degrees are used in order to define the weights $\mathbf{W}^1_{ij}$, $i\in N_n, j\in N_l$ of the first layer. Furthermore, the consequences could also be associated with confidence levels, i.e. 'if *input predicate(1)* and *input predicate(2)* and … then *output predicate(5) with confidence 0.7*'. These values are used in order to define the weights $\mathbf{W}^2_{ij}$, $i\in N_l, j\in N_m$ of the second layer.

The knowledge refinement provided by the proposed neurofuzzy system will be now described. Let $X=\{x_1,x_2,\ldots,x_n\}$ and $Y=\{y_1,y_2,\ldots,y_m\}$ be the input and output, respectively, predicate sets and let also $R=\{r_1,r_2,\ldots,r_p\}$ be the set of rules describing the knowledge of the system. The set of antecedents of the rules is denoted by $Z=\{z_1,z_2,\ldots,z_l\}$ (see the structure of the neurofuzzy system given in Fig. 7). Suppose now that a set of input-output data $D=\{(A_i,B_i,i\in N_q)\}$, where $A_i\in \mathsf{F}(X)$ and $B_i\in\mathsf{F}(Y)$ ($\mathsf{F}(*)$ is the set of fuzzy sets defined on *), is given sequentially and randomly to the system (some of them are allowed to reiterate before the first appearance of some others). The data sequence is described as $(A^{(q)}, B^{(q)})$, $q\in N$, where $(A^{(i)}, B^{(i)})\in N$. The problem that arises is finding of the new weight matrices and $\mathbf{W}^2_{ij}$, $i\in N_l$, $j\in N_m$ for which the following error is minimised:

$$\varepsilon\sum_{i\in N_q}\|B_i-y_i\| \tag{4.9}$$

where y^i, $i\in N_q$ is the output of the network when the input A_i is given. The process of the minimization of the above error is based on the resolution of the following fuzzy relational equations

$$\mathbf{W}^1\circ^{\hat{\omega}_t}\mathbf{A}=\mathbf{Z} \tag{4.10}$$

$$\mathbf{Z}\circ^t\mathbf{W}^2=\mathbf{B}, \tag{4.11}$$

where t is a continuous t-norm and $\mathbf{Z}$ is the set of antecedents fired when the input $\mathbf{A}$ is given to the network.

For the resolution of the above problem the adaptation process changes the weight matrices $\mathbf{W}^1$ and $\mathbf{W}^2$ in order to approximate a solution of the above fuzzy relational equations. During its operation the proposed network can generalize in a way that is inspired from the theory of fuzzy systems and the generalized modus ponens. Let us here describe the adaptation of the weights of the second layer (the adaptation of the first layer is similar). The proposed algorithm converges independently for each neuron. For simplicity and without loss of generality, let us consider only the single neuron case. The response of the neuron $f^{(k)}$ at time k is given by

$$f^{(k)}=\sup_{i\in N_l} t(z_i^{(k)},w_i^{(k)}), \tag{4.12}$$

where $w_i^{(k)}$ are the weights of the neuron and $z_i^{(k)}$ the input, at time k. The desired output at time k is $B_i^{(k)}$. The algorithm has as following:

Initialize the weights as $w_i^{(0)}$, $i\in N_l$.

Process the input $\mathbf{z}^{(k)}$ and the desired output $B^{(k)}$, compute the response of the network $f^{(k)}$ and update the weight accordingly (on-line variant of learning):

$$w_i^{(k+1)} = w_i^{(k)} + \Delta w_i^{(k)}$$

$$\Delta w_i^{(k)} = \eta l_s$$

$$l_s = \begin{cases} \eta_1(\breve{\omega}_t(z_i^{(k)}, B^{(k)}) - w_i^{(k)}), & \text{if } w_i^{(k)} < \breve{\omega}_t(z_i^{(k)}, B^{(k)}) \\ \eta_2(w_i^{(k)} - \hat{\omega}_t(z_i^{(k)}, b^{(k)})), & \text{if } w_i^{(k)} > \hat{\omega}_t(z_i^{(k)}, B^{(k)}) \end{cases}$$

where η, η_1, η_2 are the learning rates. The adaptation is activated only if $|\varepsilon(B^{(k)}, y^{(k)})| > \varepsilon_c$, where ε_c is an error constant.

If the *t*-norm is Archimedean, then the learning signal is computed as:

$l_s = (\hat{\omega}_t(z_i^{(k)}, b^{(k)}) - w_i^{(k)})$, if $z_i^{(k)} \geq b^{(k)}$ and $z_i^{(k)} \neq 0$, else $l_s = 0$.

With the aid of the above learning process (and similar for the first layer, since the operator $\hat{\omega}_t$ is also used in order to solve the fuzzy relational equation of the first layer (Klir & Yuan, 1995)), the network approximates the solutions of the fuzzy relational equations given above and thus minimize the error.

5. Experimental study

The data used to investigate and illustrate the techniques presented in this paper were obtained from the naturalistic database generated by Queen's University of Belfast (QUB) in the EC FP5 IST ERMIS project (FP5 IST ERMIS project) and further extended into the EC FP6 IST HUMAINE Network of Excellence (FP6 NOE HUMAINE project, Human–machine interaction network on emotion).

The experimental study first tested the performance of the developed emotion recognition system on some single user's facial expressions which had not been used for the creation of the rule based system. Then, the focus was on adapting the neurofuzzy system knowledge, taking into account annotated facial expressions of the specific user. In a real-life HCI environment, this procedure corresponds to a retraining/adaptation procedure performed in a semi-automatic way with the aid of the user-customer who buys the neurofuzzy emotion recognition system and uses it for the first time. In this case the system asks the user to react in specific expressive ways and records (through a PC camera) its facial expressions. An ELIZA-based environment that has been created by QUB which can be used in this framework to trigger user's reactions towards specific attitudes.

In the following, we concentrated on the analysis of one user's (Roddy) dataset, so as to illustrate the adaptation capability of the proposed expression analysis and emotion recognition system to each specific user under examination. Since the database is audiovisual, the data set has been separated in time intervals corresponding to tunes identified by the analysis of the respective speech recordings. The frame, which was the most 'facially expressive' (with large FAP values and a high confidence level of estimation) in each tune was then selected, thus generating a set of 100 frames. These frames were visually annotated with respect to the four 2D emotion representation quadrants, and 54 of them were selected, evenly covering three quadrants (the positive-passive did not contain data and was neglected). This set was mainly used for training/adaptation purposes. We also created and visually annotated a second data set of about 930 frames, which were used for testing purposes.

First, we applied the generated neurofuzzy system to the above datasets. Its performance was around 58% on these datasets, which was well beyond chance level, but did not take into account the specific user's ways of expression.

Fig. 8 shows three frames for which the initial rules failed to recognize the expressed emotion correctly. For example, Fig. 8(a) was feeltraced in the first quadrant (+,+) but the calculated FAPs fired a rule which classified it in the second quadrant (−,+). This can be attributed to the unique way the subject communicates his emotions, which requires adaptation of the initial rules. The procedure followed to generate additional rules to adapt the neurofuzzy system's knowledge was the following.

First, we performed clustering of the 54 FAP sets, each consisting of 17 FAP values. Based on the extracted cluster canters and standard deviations, we generated a user specific correspondence of his/her FAP values with a low/medium/high rating, and provided the neurofuzzy system with the generated specific user's training dataset of predicates and desired responses to adapt its performance.

Of significant interest is usage of unsupervised hierarchical clustering (we developed a methodology for clustering with high-dimensional extensions and probabilistic refinement (Wallace & Kollias, 2003)), since this can form a basis for future merging of different emotional representations (i.e. different hierarchical levels), and categorization in either coarser or more detailed classes (half-plane, quadrants, discrete emotions).

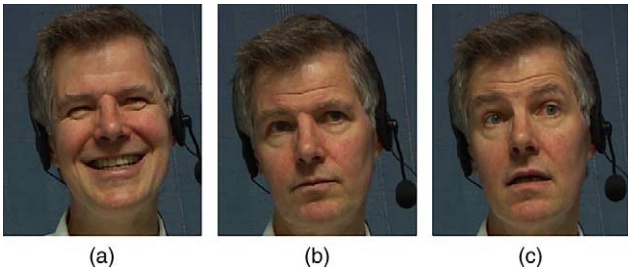

Fig. 8. Example frames which the expression analysis system failed to recognize without adaptation.

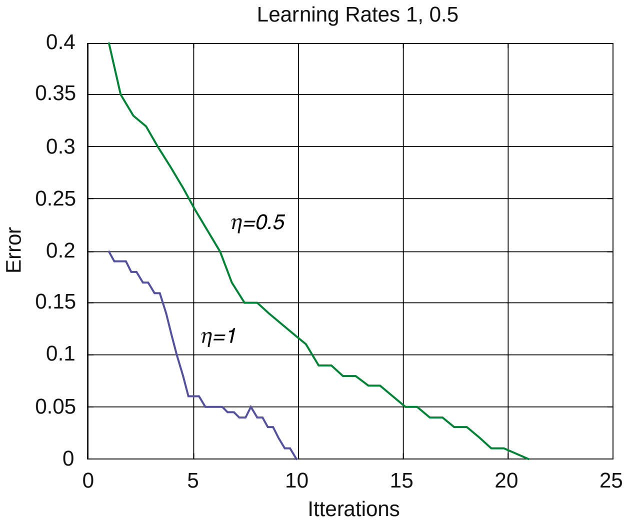

Fig. 9. The error of the neural network during the adaptation process.

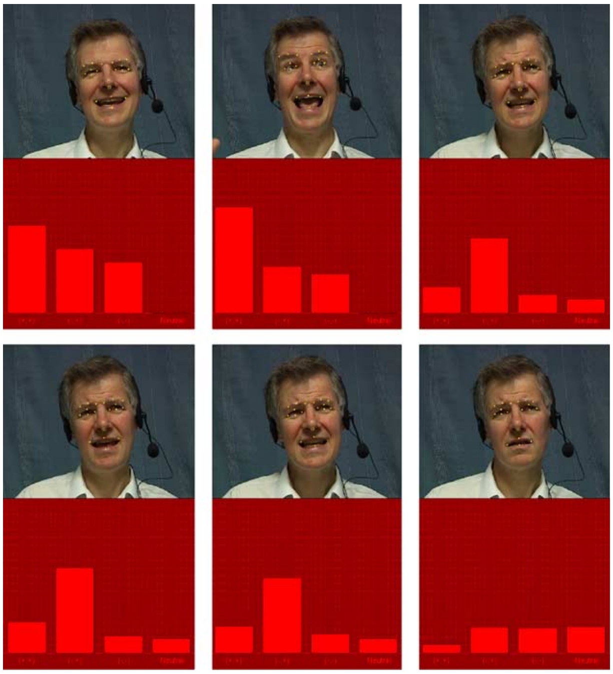

Fig. 10. Example results showing the detected FPs along with network output for the three quadrants (+,+),(−,+),(−,−) and the neutral case shown in the four bars.

We used an hierarchical agglomerative clustering (HAC) approach which performs an on-line scaling and soft selection of features to consider in the agglomeration process. In clustering we have not taken into account the FAP corresponding to vertical mouth opening (No. 1), so as to ignore the effect of speaking onto the recognition process. Moreover, a naive Bayes Classification approach has been adopted in order to allow for the refinement of the initial partitioning results (Wallace, Mylonas, & Kollias, 2003). This classifier is automatically built from the output of the agglomeration process and is able to re-evaluate cluster assignments for all considered elements. Through recursive application, an optimization of the count of detected clusters as well as of their cardinalities, centers and spreads in each direction is achieved in a fully unsupervised manner, thus overcoming the susceptibility to errors in the initial steps of the agglomeration process. Hierarchical clustering was applied to the generated data set to produce clusters of similar data samples. An aggregating Mahalanobis distance function was used to identify the underlying patterns and produce clusters of similar data samples. 10 clusters have been obtained in this way, and the resulting centers were used to define the rules by which the neurofuzzy system would then adapt its knowledge and performance to the specific user.

The adaptation procedure described in the former section was then applied to this training data set. More specifically, an aggregating Mahalanobis distance function was used to identify the underlying patterns and produce clusters of similar data samples. 10 clusters have been obtained in this way, and the resulting centres were used to define the rules. The aim of the learning algorithm of the neurofuzzy network is to adapt these rules. The error of the learning algorithm, as shown in Fig. 9 becomes zero. The error performance is illustrated, using learning rates, 0.5 and 1. If we use high learning rate, the learning algorithm converges in 10 iterations. However, using high learning rate, in some applications, the algorithm might not converge. Using medium learning rate the algorithm converges in 20 iterations. It must be noticed that the structure of the rules is not changed. The algorithm only adapts the activated predicates and not inserting new predicated in the rules, which may result in altering the knowledge provided by the experts Fig. 10.

Table 3
New rules obtained through adaptation

Rule	FAPs
5	F3_H+F4_M+F5_L+[F53+F54]_M+[F19+F21]_L+[F20+F22]_L+F31_H+F32_H+F33_H+ F34_H+F37_M+F38_M+F59_M+F60_M(−,+)
29	F3_L+F4_L+F5_H+[F53+F54]_M+[F19+F21]_M+[F20+F22]_M+F31_M+F32_M+F33_L+ F34_L+[F37+F38]_M(−,−)
39	F3_L+F4_L+F5_H+[F53+F54]_L+[F19+F21]_H+[F20+F22]_H+F31_L+F32_L+F33_L+F34_L+ [F37+F38]_H+F59_H+F60_H(+,+)

The adaptation procedure created a set of new rules, three of which are shown in Table 1. A comparison of the rules shown in Tables 1 and 3 shows the specific changes made to fit the characteristics of the specific user.

Then we tested the performance of the adapted system to the testing data set. The performance of this was increased to 78%, which is very satisfactory for the non-extreme emotion recognition problem. Results from the application of this network to the dataset are shown below.

6. Conclusions

This paper describes an emotion recognition system, which combines psychological findings about emotion representation with analysis and evaluation of facial expressions. The performance of the proposed system has been investigated with experimental real data. More specifically, a neurofuzzy rule based system has been first created and used to classify facial expressions using a continuous 2D emotion space, obtaining high rates in classification and clustering of data to quadrants of the emotion representation space. To improve these rates for a specific user, the initial set of rules that captured the a-priori knowledge was then adapted via a learning procedure of the neurofuzzy system (Tzouvaras, Stamou, & Kollias, 2004; Wallace, Raouzaiou, Tsapatsoulis, & Kollias, 2004), so as to capture unique expressivity instances. Future extensions will include emotion recognition based on combined facial and gesture analysis (Karpouzis et al., 2004; Balomenos et al., 2004). These can provide the means to create systems that combine analysis and synthesis of facial expressions, for providing more expressive and friendly interactions (Raouzaiou et al., 2004). Moreover, development of rule-based emotion recognition provides the possibility to combine the results obtained within the framework of the ERMIS project with current knowledge technologies, e.g. in implementing an MPEG-4 visual ontology for emotion recognition (Tzouvaras et al., 2004).

References

Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., & Kollias, S. (2004). *Emotion analysis in man–machine interaction systems Workshop on multimodal interaction and related machine learning algorithms, Switzerland.*

Black, M., & Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, *25*(1), 23–48.

Cohn, J. F., & Katz, G. S. (1998). *Bimodal expressions of emotion by face and voice Workshop on face/gesture recognition and their applications, the sixth ACM international multimedia conference, Bristol, England.*

Dietterich, T. G. (2000). *Ensemble methods in machine learning Proceedings of first international conference on multiple classifier systems.*

Donato, G., Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(10).

Ekman, P., & Friesen, W. (1975). *Unmasking the face*. Prentice-Hall.

FP5 IST ERMIS project. *Emotionally rich man–machine intelligent system*, http://www.image.ntua.gr/ermis.

FP6 NOE HUMAINE project. *Human–machine interaction network on emotion*, http://www.emotion-research.net.

Fransens, R. De Prins, J. (2003). SVM-based non-parametric discriminant analysis, an application to face detection. *Ninth IEEE international conference on computer vision* (Vol. 2).

Gorodnichy, D. (2002). On importance of nose for face tracking. *Proceedings of the international conference on automatic face and gesture recognition (FG'2002), Washington DC*, May 20–21.

Hagan, M. T., & Menhaj, M. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, *5*(6), 989–993.

Hirota, K., & Pedrycz, W. (1996). Solving fuzzy relational equations through logical filtering. *Fuzzy Sets and Systems*, *81*, 355–363.

Jenei, S. (1998). On Archimedean triangular norms. *Fuzzy Sets and Systems*, *99*, 179–186.

Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. New Jersey: Prentice Hall.

Kollias, S., & Anastassiou, D. (1989). An adaptive least squares algorithm for the efficient training of artificial neural networks. *IEEE Transactions on Circuits and Systems*, *36*(8), 1092–1101.

Kosko, B. (1992). *Neural networks and fuzzy systems: A dynamical approach to machine intelligence*. Englewood Cliffs, NJ: Prentice Hall.

Krog, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. In G. Tesauro, D. Touretzky, & T. Leen, *Advances in neural information processing systems* (Vol. 7) (pp. 231–238). Cambridge, MA: MIT Press.

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 Publ.

Leung, S. H., Wang, S. L., & Lau, W. H. (2004). Lip image segmentation using fuzzy clustering incorporating an alliptic shape function. *IEEE Transactions on Image Processing*, *13*(1).

Lin, C.-T., & Lee, C. S. (1995). *Neural fuzzy Systems: A neuro-fuzzy synergism to intelligent systems*. Englewood Cliffs, NJ: Prentice-Hall.

Mase, K., Pentland, A. (1990). Lip reading by optical flow. *IEICE of Japan, J73-D-II, 6*, 796–803.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge, MA: Cambridge University Press.

Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., & Kollias, S. (2002). Parameterized facial expression synthesis based on MPEG-4. *EURASIP Journal on Applied Signal Processing*, *2002*(10), 1021–1038. Hindawi Publishing Corporation.

Raouzaiou, A., Karpouzis, K., & Kollias, S. (2004). *Emotion synthesis in the MPEG-4 framework IEEE international workshop on multimedia signal processing (MMSP), Siena, Italy..*

Stamou, G. B., & Tzafestas, S. G. (2000). Neural fuzzy relational systems with a new learning algorithm. *Mathematics and Computers in Simulation* , 301–304.

Tzouvaras, V., Stamou, G., & Kollias, S. (2004). A fuzzy knowledge based system for multimedia applications. In G. Stamou, & S. Kollias (Eds.), *Multimedia content and semantic web: Methods, standards and tools*. New York: Wiley.

Wallace, M., & Kollias, S. (2003). *Soft attribute selection for hierarchical clustering in high dimensions Proceedings of the international fuzzy systems association world congress (IFSA), Istanbul, Turkey.*

Wallace, M., Mylonas, P., & Kollias, S. (2003). *Detecting and verifying dissimilar patterns in unlabelled data Eighth online world conference on soft computing in industrial applications (WSC8), September–October*.

Wallace, M., Raouzaiou, A., Tsapatsoulis, N., & Kollias, S. (2004). *Facial expression classification based on MPEG-4 FAPs: The use of evidence and prior knowledge for uncertainty removal Proceedings of the IEEE international conference on fuzzy systems (FUZZ-IEEE), Budapest, Hungary*.

Wehrle, T., & Scherer, K. (2001). Toward computational modeling of appraisal theories, in appraisal processes in emotion: Theory, methods, research. In Scherer, Schorr, & Johnstone (Eds.), *Series in affective science* (pp. 350–365). New York: Oxford University Press.

Whissel, C. M. (1989). The dictionary of affect in language. In R. Plutchnik, & H. Kellerman, *Emotion: Theory, research and experience*. *The measurement of emotions* (Vol. 4). New York: Academic Press.

Yacoob, Y., & Davis, L. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(6), 636–642.

Yin, L. (2001). Generating realistic facial expressions with wrinkles for model-based coding. *Computer Vision and Image Understanding*, *84*, 201–240.

Young, J. W. (1993). *Head and face anthropometry of adult US civilians*. FAA Civil Aeromedical Institute.