Intelligent Facial Analysis and Expression Recognition

S. Ioannou, M. Wallace, and S. Kollias

Abstract— **Since facial expressions are a key modality in human communication, the automated analysis of facial images and video for the estimation of the displayed expression is central in the design of intuitive and human friendly computer interaction systems. In this paper we present an intelligent feature extraction system which combines analysis from multiple channels based on their confidence, to result in better, error resilient facial feature boundary detection. Neural networks are a key component of the system. Issues such as uncertainty and lack of confidence in the process of feature extraction are considered during the expression analysis and recognition. Various results are presented which illustrate the performance of the method.**

I. Introduction

INTERPERSONAL communication is for the most part completed via the face. The face is the mean to identify a colleague or friend, to assist interpretation of what has been said via lip reading, and to understand someone's emotional state and intentions on the basis of the shown facial expression. Despite common belief, social psychology research has shown that conversations are usually dominated by facial expressions, and not spoken words, indicating the speaker's predisposition towards the listener. Mehrabian indicated that the linguistic part of a message, that is the actual wording, contributes only for seven percent to the effect of the message as a whole; the paralinguistic part, that is how the specific passage is vocalized, contributes for thirty eight percent, while facial expression of the speaker contributes for fifty five percent to the effect of the spoken message [1]. This implies that the facial expressions form the major modality in human communication.

In most real-life applications nearly all video media have reduced vertical and horizontal color resolution. A 4:2:0 video signal (eg. H-261, MPEG-2 where Cr and Cb are each sub-sampled by a factor of 2 both horizontally and vertically) is still considered to be a very good quality signal; moreover, the face usually occupies only a small percentage of the whole frame and illumination is far from perfect. When dealing with such input we have to accept that color quality and video resolution will be very poor.

While it is usually feasible to detect the presence and location of face and all facial features with high accuracy, it is very difficult in such conditions to find the exact boundary of each one (eye, eyebrow, mouth) in order to estimate its de-formation from a neutral-expression frame [2].

S.Ioannou, M.Wallace and S.Kollias are with the School of Electrical and Computer Engineering, National Technical University of Athens, Heroon Polytechniou 9, 157 80 Zographou, Greece. Corresponding Author: stefanos@image.ntua.gr, +302107722488 +302107722521. Work has been partially supported by the HUMAINE FP6 European IST Network of Excellence.

To accommodate for such problems, in this work we propose an intelligent facial feature extraction method which relies on the fusion of several facial feature masks derived from multiple feature extractors. Various neural network components are included in this approach, with a committee machine based fusion process, providing the facial feature values that are further exploited in a fuzzy and possibilistic rule based system for effective facial expression recognition.

II. Feature Extraction

A. Overview

An overview of the system is given in Fig. 1. At first face detection is performed using nonparametric discriminant analysis with a Support Vector Machine (SVM) [3], 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 [4] 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.

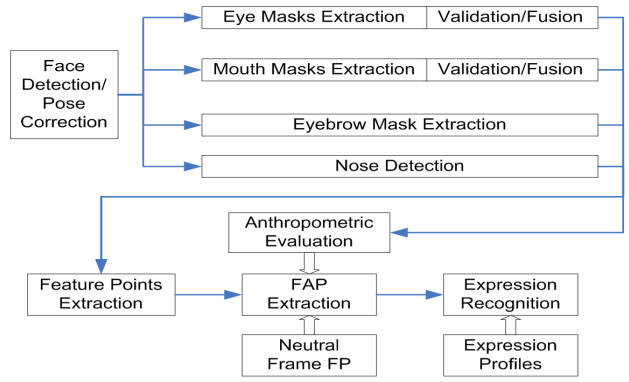


Fig. 1. Diagram of the proposed methodology

To estimate the head pose we first locate the left and right eyes in the detected corresponding eye candidate areas. After locating the eyes, we can 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

backpropagation neural network with a sigmoidal activation function. The multi-layer perceptron (MLP) we adopted employs Marquardt-Levenberg learning [5][6] while the optimal architecture obtained through pruning has two 20 node hidden layers and 13 inputs.

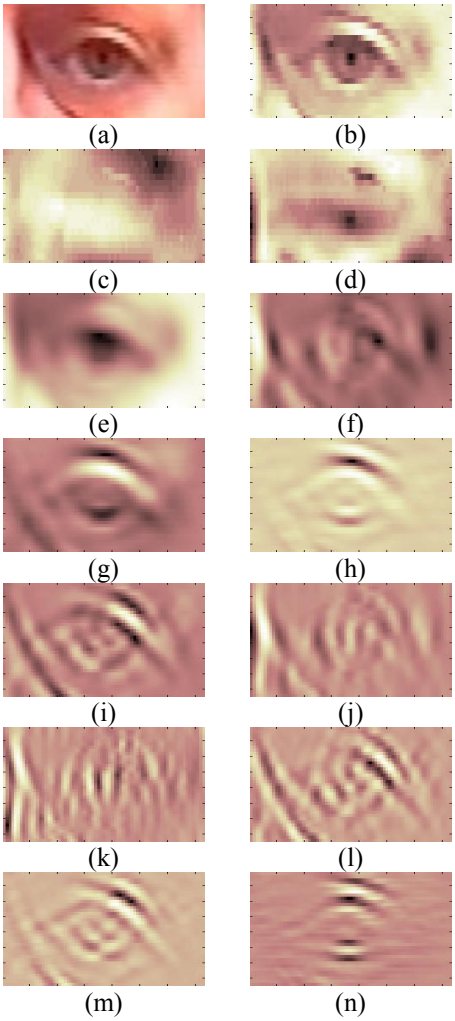


Fig. 2. (a) A typical 80x44 color eye image, (b)-(n) Neural Network inputs. (b)(c)(d) $Y/C_r/C_b$ channels, (e)-(n) 10 most important DCT coefficients calculated in 8x8 blocks and stitched together for illustration purposes.

We apply the network separately on the left and right eye-candidate face regions. For each pixel in these regions the 13 inputs to the neural network are the luminance Y, the Cr & Cb chrominance values and the 10 most important DCT coefficients (with zigzag selection) of the neighboring 8x8 pixel area, depicted in Fig. 2. Using alternative input color spaces such as Lab, RGB or HSV to train the network, has not changed 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 and HUMAINE [7],[8] databases, in images of diverse quality, resolution and lighting conditions.

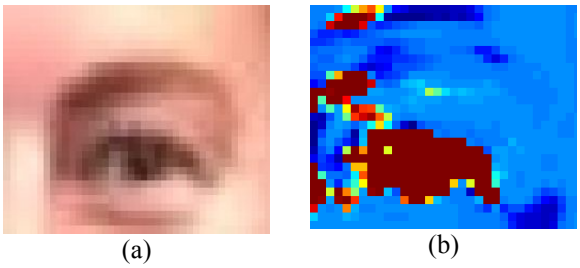


Fig. 3. (a) left eye input image (b) network output on left eye, darker pixels correspond to higher output

The output of the aforementioned network, depicted in Fig. 3 is used to locate the eyes and is also combined with other feature detectors in the fusion process described below, 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. 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 avoiding to utilize complex or time-consuming feature extractors. The use of the afore-mentioned neural network greatly serves this scope. The feature extractors developed for this work are briefly described in the following.

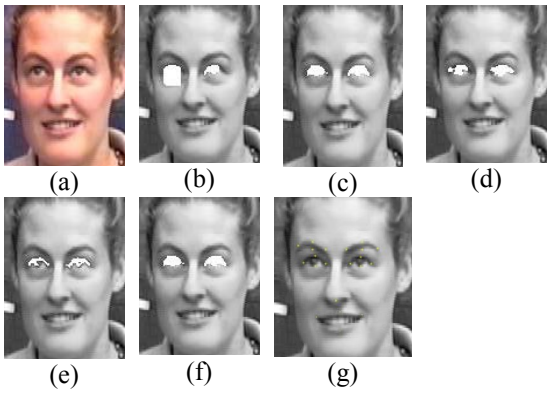
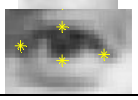


Fig. 4. (a):original frame, (b),(c),(d),(e): the four detected masks, (f):final mask for the eyes, (g):all detected feature points from the final masks

B. Mask Extraction

Eyebrows are detected with a procedure involving morphological edge detection and feature selection using data from [4]. Nose detection is based on nostril localization.

Table I Mask fusion examples on the left eye
with corresponding validation tags and detected feature points.

Seq. / Mask	kk-1002	Validity (0..1)	kk-1998	Validity (0..1)	rd-12259	Validity (0..1)	al-27	Validity (0..1)
$\mathbf{M}^e_{nn}$								
$\mathbf{M}^e_1$		0.825		0.813		0.823		0.839
$\mathbf{M}^e_2$		0.782		0.581		0.763		0.810
$\mathbf{M}^e_3$		0.866		0.733		0.716		0.787
$\mathbf{M}^e_4$		0.883		0.917		0.826		0.872
$\mathbf{M}^e_f$								
FPs								
	D_{bp}: 58 px		D_{bp}: 58 px		D_{bp}: 96 px		D_{bp}: 36 px	

D_{bp} denotes bipupil breadth in pixels and is quoted as an image resolution indicative

$\mathbf{M}^e_i$ denotes eye mask i $\mathbf{M}^e_{nn}$ denotes the neural network output, $\mathbf{M}^e_f$ denotes the final mask

Nostrils are easy to detect due to their low intensity [9]. 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 [4].

For the eyes the following masks are constructed:

• A refined version of the original neural-network derived mask. The initial eye mask provided by the neural network 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. 4e)

• A mask expanding in the area between the upper and lower eyelids. Since the eye-center is almost always detected correctly by 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. 4b)

• 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 3x3 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. 4c)

• 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 include 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. 4d)

Finding the extent of a closed mouth in a still image is a relatively easy accomplished task [10]. In case of an open mouth, several methods have been proposed which make use of intensity [11] or color information [12]. 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 training data are acquired from the neutral image (where the mouth is closed) as follows: the mouth-candidate region of interest (ROI) is first filtered with Alternating Sequential Filtering by Reconstruction 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.

• An horizontal morphological gradient is calculated in the mouth area and the longest connected object which comply with constraints from [4] 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 labeled. 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 [4] 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 we already mentioned, the detection of a mask using the 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 consists 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.

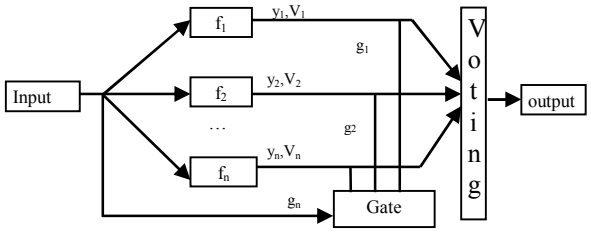


Fig. 5. Dynamic Committee Machine Architecture

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 the ones which were found to have the best performance taking also into account the lack of specific rules for extracting the facial areas, and the ability to use training data from the current environment so as to adapt the method to the local testbed characteristics. The fusion algorithm is based on a Dynamic Committee Machine (DCM) structure (depicted in Fig. 5 that combines the masks based on their validity confidence, producing a final mask together with the corresponding estimated confidence [13] 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 [14] that the combination error $y_{comb} - t$ from different machines f_i is guaranteed to be lower than the average error:

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

In a Static CM, the voting weight for a component is proportional to its error on a validation set. In DCMs, (Fig. 5) input is directly involved in the combining mechanism through a Gating Network (GN), which is used to modify those weights dynamically.

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:

$$m_f^x = \frac{1}{n}\sum_{i=1}^{n} m_i^x M_f^{cx_i} h^i g^i \quad (2)$$

$$h^k = \begin{cases} 1, M_f^{c,x_k} \geq \left(\mathrm{t_{vd}} \cdot \left\langle M_q^{c,x_k} \right\rangle_q\right) \\ 0, M_f^{c,x_k} < \left(\mathrm{t_{vd}} \cdot \left\langle M_q^{c,x_k} \right\rangle_q\right) \end{cases} \quad (3)$$

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 $M_f^{c,x_k} < \left(\mathrm{t_{vd}} \cdot \left\langle M_q^{c,x_k} \right\rangle_q\right)$ to contribute to the final mask. A sufficient value for $\mathrm{t_{vd}}$ is 0.8.

The role of the gating variable g^i is to favor the color-based feature extraction methods ($\mathbf{M}_1^e, \mathbf{M}_1^m$) 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 [15].Table I illustrates mask fusion examples for the left eye where some of the masks are problematic. Validity values refer to the corresponding mask anthropometric validation value while D_{bp} is quoted as an indication of the sequence resolution. For illustration purposes the feature points extracted from the final masks are presented verifying the precise extraction of the features and feature points, based on the mask fusion process

III. EXPRESSION ANALYSIS

The feature masks are used to extract the Feature Points (FPs) considered in the definition of the FAPs, used in this work. Each FP inherits the confidence level of the final mask from which it derives; for example, the four FPs (top, bottom, left and right) of the left eye share the same confidence as the left eye final mask. Continuing, FAPs can be estimated via the comparison of the FPs of the examined frame to the FPs of a frame that is known to be neutral, i.e. a frame which is accepted by default as one displaying no facial deformations. For example, FAP F_{37} *(squeeze_l_eyebrow)* is estimated as:

$$F_{37} = \left\| FP^n_{4.5} - FP^n_{3.11} \right\| - \left\| FP_{4.5} - FP_{3.11} \right\| \quad (4)$$

where FP^n_i, FP_i are the locations of feature point i on the neutral and the observed face, respectively, and $\left\| FP_i - FP_j \right\|$ is the measured distance between feature points i and j.

Obviously, the uncertainty in the detection of the feature points propagates in the estimation of the value of the FAP as well. Thus, the confidence in the value of the FAP, in the above example, is estimated as

$$F^c_{37} = \min(FP^c_{4.5}, FP^c_{3.11}) \quad (5)$$

On the other hand, some FAPs may be estimated in different ways. For example, FAP F_{31} is estimated as:

$$F^1_{31} = \left\| FP^n_{3.1} - FP^n_{3.3} \right\| - \left\| FP_{3.1} - FP_{3.3} \right\| \quad (6)$$

or as

$$F^2_{31} = \left\| FP^n_{3.1} - FP^n_{9.1} \right\| - \left\| FP_{3.1} - FP_{9.1} \right\| \quad (7)$$

As argued above, considering both sources of information for the estimation of the value of the FAP alleviates some of the initial uncertainty in the output. Thus, for cases in which two distinct definitions exist for a FAP, the final value and confidence for the FAP are as follows:

$$F_i = \frac{F^1_i + F^2_i}{2} \quad (8)$$

The amount of uncertainty contained in each one of the distinct initial FAP calculations can be estimated by

$$E^1_i = 1 - F^{1c}_i \quad (9)$$

for the first FAP and similarly for the other. The uncertainty present after combining the two can be given by some t-norm operation on the two:

$$E_i = t(E^1_i, E^2_i) \quad (10)$$

The Yager t-norm with parameter $w=5$ gives reasonable results for this operation:

$$E_i = 1 - \min\left(1, \left((1-E^1_i)^w + (1-E^2_i)^w\right)^w\right) \quad (11)$$

The overall confidence value for the final estimation of the FAP is then acquired as

$$F^c_i = 1 - E_i \quad (12)$$

Using statistical analysis of the FAP values over well known facial expression datasets, a set of expression profiles has been derived that can be used for facial expression recognition [14]. A fuzzy rule based expression recognition system has been derived thereafter [15].

While evaluating the expression profiles, FAPs with greater uncertainty must influence less the profile evaluation outcome, thus each FAP must include a confidence value. This confidence value is computed from the corresponding FPs which participate in the estimation of each FAP.

Finally, FAP measurements are transformed to antecedent values x_j for the fuzzy rules using fuzzy numbers defined for each FAP, and confidence degrees x^c_j are inherited from the FAP:

$$x^c_j = F^c_i \quad (13)$$

where F_i is the FAP based on which antecedent x_j is defined. More information about the expression profile based fuzzy recognition system can be found in [15],[16].

IV. POSSIBILISTIC RULE EVALUATION

In the process of exploiting the knowledge contained in the fuzzy rule base and the information extracted from each frame in the form of FAP measurements, with the aim to analyze and classify facial expressions, a series of issues has to be tackled:

FAP degrees need to be considered in the estimation of the overall result.

The case of FAPs that cannot be estimated, or equivalently are estimated with a low degree of confidence, needs to be considered.

The activation of contradicting rules needs to be considered.

The conventional approach to the evaluation of fuzzy rules of the form IF $x_1, x_2, ..., x_n$ THEN y is as follows [17]:

$$y = t(x_1, x_2, ..., x_n) \quad (14)$$

where t is a fuzzy t-norm, such as the minimum

$$t(x_1, x_2, ..., x_n) = \min(x_1, x_2, ..., x_n) \quad (15)$$

the algebraic product

$$t(x_1, x_2, ..., x_n) = x_1 \cdot x_2 \cdot ... \cdot x_n \quad (16)$$

the bounded sum

$$t(x_1, x_2, ..., x_n) = x_1 + x_2 + ... + x_n + 1 - n \quad (17)$$

and so on. Another well known approach in rule evaluation is described [18] and utilizes a weighted sum instead of a t-norm in order to combine information from different rule antecedents:

$$y = w_1 x_1 + w_2 x_2 + ... + w_n x_n \quad (18)$$

Both approaches are well studied and established in the field of fuzzy system analysis. Still, they are not adequate for the case of facial expression estimation: their main disadvantage is that they assume that all antecedents are known, i.e. that all features are measured successfully and precisely. In the case of facial expression estimation, FAPs may well be estimated with a very low confidence, or not estimated at all, due to low video quality, speech interference, occlusion, noise and so on. Thus, a more flexible rule evaluation scheme is required, that is able to incorporate such uncertainty as well.

Moreover, the second one of the conventional approaches, due to the summation form, has the disadvantage of possibly providing a highly activated output even in the case that an important antecedent is known to be missing; obviously it is not suitable for the case examined in this paper, where the non activation of a FAP automatically implies that the expression profiles that require it are not activated either. Therefore, the flexible rule evaluation scheme that we propose is in fact a generalization of the t-norm based conventional approach.

In the t-norm operation described in equation (14), antecedents with lower values affect most the resulting value of y, while antecedents with values close to one have trivial and negligible affect on the value of y. Having that in mind, we can demand that only antecedents that are known with a high confidence will be allowed to have low values in that operation. More formally, we demand that the degree $k(x)$ to which antecedent x is considered in the operation is low, i.e. its complement $c\left(k(x)\right)$ is high, only when the confidence x^c with which the value of x is known is high and the value of x is low. This can be expressed as:

$$c\left(k(x)\right) = x^c \cap c(x) \quad (19)$$

where c is a fuzzy complement. Applying de Morgan's law we have that the degree to which antecedent x is considered is:

$$k(x) = c(x^c) \cup x \quad (20)$$

It is easy to see that equation (20) satisfies the desired marginal conditions: when $x^c \to 1$, then $c(x^c) \to 0$ and $k(x) \to x$, i.e. the antecedent is considered normally, while when $x^c \to 0$, then $c(x^c) \to 1$ and $k(x) \to 1$, i.e. the antecedent is not allowed to affect the overall evaluation of the rule; the formula that provides the overall evaluation assumed in this discussion is the one followed by the conventional approach, with the exception that antecedents participate with their considered values:

$$y = t\left(k(x_1), k(x_2), ..., k(x_n)\right) \quad (21)$$

It is easy to see that in the case that all antecedents are known with a confidence of one the rule will be evaluated in the same way as in the conventional methodology. When, on the other hand, all antecedents are known with a confidence of zero, i.e. when no information is available, the rule will be evaluated with a degree of one. Thus, the activation level of a rule with this approach can be interpreted in a possibilistic manner, i.e. it can be interpreted as the degree to which the corresponding output is possible, according to the available information; in the literature, this possibilistic degree is referred to as plausibility.

As far as the confidence in the calculated output is concerned, the conventional approach always displays a total confidence in the output, which originates from the assumption that all inputs are precisely known. In the extended approach followed herein, where we accept that one or more of the rule antecedents may be unknown or known with a confidence other than one, it does not make sense to always have total confidence in the calculated output. Quite the contrary, the calculated output is only complete in information when associated with a corresponding degree of confidence.

The confidence is determined by the confidence values of the utilized inputs, i.e. by the confidence values of the rule antecedents, as follows:

$$y^c = \frac{1}{n} x_1^c + x_2^c + ... + x_n^c \quad (22)$$

The definition of y^c in this manner has the desired effect that $y^c = 0$ is equivalent to the complete lack of information, as it can only happen when all inputs are known with confidence zero; this property is essential in possibilistic reasoning.

The belief should be high when plenty of information is available during the evaluation of the rule, and that information suggests that the rule should be activated. The amount of information that was available during the evaluation of the rule is provided by the calculated confidence value, while the degree to which this information suggests that the specific rule should be activated is provided by the activation level. Thus, the complete possibilistic representation of the calculated output is provided as:

$$Bel = t(y, y^c) \quad (23)$$

$$Pl = y \quad (24)$$

The extreme cases are $Bel = Pl = 1$, which occurs when $y = y^c = 1$ and implies absolute confidence that the specific profile is the one perfectly matching the observed face, $Bel = Pl = 0$, which occurs when $y = 0$ and implies absolute confidence that the specific profile is not one matching the observed face and $Bel = 0$, $Pl = 1$ which occurs when $y = 1$, $y^c = 0$ and implies absolute ignorance. The case of activation of multiple and incompatible rules of the rule base is not an issue for our approach. In that case it is expected that confidence values will be low, which can be interpreted as the case in which, due to poor performance of the image processing module, more than one possible outputs cannot be ruled out. Still, the belief that they are indeed the ones matching the observed face, as reported by equation (23), will be low.

V. Experimental Results

A. Feature Extraction

A way to evaluate our feature extraction performance is the modified Williams' Index (WI) [19] which compares the agreement of an observer with the joint agreement of other observers and also deals with multivariate data. The modified WI divides the average number of agreements (inverse disagreements, $D_{j,j'}$) between the computer (observer 0) and *n-1* human observers (*j*) by the average number of agreements between human observers:

$$WI = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{0,j}} \Bigg/ \frac{2}{n(n-1)}\sum_{j}\sum_{j':j'>j}\frac{1}{D_{j,j'}} \qquad (25)$$

and in our case we define the average disagreement between two observers *j,j'* as:

$$D_{j,j'} = D_{bp}^{-1}\left\| M_j^x \veebar M_{j'}^x \right\| \qquad (26)$$

where $\veebar$ denotes the pixel-wise xor operator, $\left\| M_j^x \right\|$ denotes the cardinality of feature mask *x* constructed by observer *j*, and D_{bp} is the bipupil breadth, used as a normalization factor to compensate for camera zoom on video sequences.

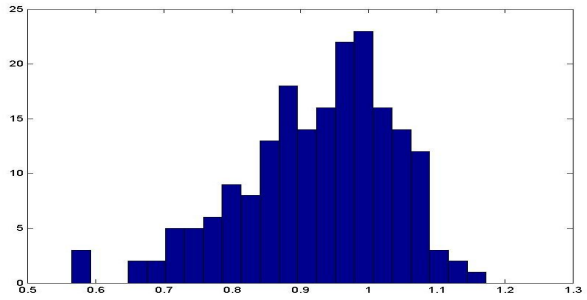


Fig. 6. Williams Index distribution (average on eyes and mouth)

From a dataset of about 50000 frames, 250 frames were selected at random and the 19 FPs were manually selected from two observers. WI was calculated using (25) for each feature and for each frame separately. Distribution of the average WI calculated over the two eyes and mouth is shown in. Fig. 6.

B. Expression Recognition

In order to experimentally validate the approach proposed in this paper, we have used the software prototypes and datasets of ERMIS and HUMAINE as a test bed. Since the corresponding datasets were created by engaging participants to emotional dialogue, facial expressions in these sets were not acted but are mostly naturalistic. In this case, even manual expression categorization into the four quadrants is often difficult, when relying only on the visual cue. A way to enhance the recognition performance is to have a measure of a frame's expressiveness, before the actual expression recognition is performed. The latter was realized by first extracting audio tunes (a tune being the portion of the pitch contour that lies between two audio pause boundaries [20], and then performing expression analysis on the most expressive frame in each tune. Expressiveness is measured as a deviation from the neutral state. If we denote as F_j^n a vector containing *j*-ith FAP value of the neutral frame and as F_j^i the *j*-ith FAP value from frame *i*, then from each tune we select the frame that satisfies equation (27):

$$i = \arg\max_i \sum_{j=1}^{n}\left|F_j^i - F_j^n\right| \qquad (27)$$

We evaluated our data in both the full data set, and the automatically chosen "expressive" data subset from a total of about 30000 frames. Expression analysis results were tested against manual multimodal annotation and the results are presented in Table II.

TABLE II
RECOGNITION RESULTS

Full set		Expressive subset	
Probabilistic	Possibilistic	Probabilistic	Possibilistic
27.8%	38.5%	65.1%	78.4%
Annotator Dissagreement			
20.01%			

As shown in the last column of Table III, even the human experts classifying the frames in order to generate the ground truth make different evaluations once every five frames, which is clearly indicative of the ambiguity of the procedure. From the comparative study of the conventional and proposed systems' performance on the remaining frames, as shown in the first four columns, it is also clear that the proposed approach outperforms its predecessor. Finally, it is worth underlining that this system achieves a 78% classification rate while operating based solely on expert knowledge provided by humans in the form of fuzzy rules, without weights for the rule antecedents. Allowing for the specification of antecedence importance as well as for rule

optimization through machine learning is expected to provide for even further enhancement of the achieved results.

VI. CONCLUSIONS

An intelligent approach to facial feature extraction for expression recognition purposes has been presented in this paper. Neural networks have been used in the feature extraction procedure, providing the approach with adaptation to specific data regarding the subject or environmental conditions. A committee machine approach has been used to fuse the results with different masks, while an anthropometric measure has been used to provide confidence values to the obtained facial features. The involved uncertainty is handled in the following through the use of a fuzzy rule based system and a possibilistic rule evaluation procedure that is proposed and used in the paper. The whole approach indicates the ability of neural networks and machine learning techniques to provide efficient solutions to real life multimedia analysis and human computer interaction applications.

REFERENCES

[1] A. Mehrabian, Communication without Words, Psychology Today, vol. 2, no. 4, pp. 53-56, 1968.

[2] M. Pantic, L.J.M Rothkrantz, Automatic Analysis of Facial Expressions: The State of the Art, IEEE Transactions on PAMI, Vol.22, No.12, December 2000

[3] R. Fransens, Jan De Prins, SVM-based Nonparametric Discriminant Analysis, An Application to Face Detection, Ninth IEEE International Conference on Computer Vision Volume 2, October 13 - 16, 2003

[4] J.W. Young, Head and face anthropometry of adult U.S. civilians, FAA Civil Aeromedical Institute, 1993.

[5] S. Kollias and D. Anastassiou, An adaptive least squares algorithm for the efficient training of artificial neural networks, IEEE Transactions on Circuits and Systems, Volume: 36 , Issue: 8 , Aug. 1989 pp.1092-1101

[6] M. T. Hagan, and M. Menhaj, Training feedforward networks with the Marquardt algorithm, IEEE Transactions on Neural Networks, vol. 5, no. 6, 1994, pp. 989-993.

[7] European FP5 IST ERMIS project, (Emotionally Rich Man-machine Intelligent System) http://www.image.ntua.gr/ermis

[8] European FP6 IST HUMAINE Network of Excellence, 2004-2007.

[9] D. Gorodnichy. On Importance of Nose for Face Tracking, Proc. Intern. Conf. on Automatic Face and Gesture Recognition (FG'2002), Washington DC, May 20-21, 2002.

[10] Hagan, M. T., and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, vol. 5, no. 6, pp. 989-993, 1994.

[11] Lijun Yin, Generating Realistic Facial Expressions with Wrinkles for Model-Based Coding, Computer Vision and Image Understanding 84, 201-240 (2001)

[12] Leung et al: Lip image segmentation using fuzzy clustering incorporating an alliptic shape function, IEEE Trans. on image processing, vol.13, No.1, January 2004

[13] T.G. Dietterich, Ensemble methods in machine learning, Proceedings of First International Conference on Multiple Classifier Systems, 2000.

[14] A. Krog, J. Vedelsby, Neural network ensembles, cross validation and active learning, in Tesauro G., Touretzky D., Leen T. (Eds) Advances in neural information processing systems 7, pp. 231-238, Cambridge, MA. MIT Press, 1995.

[15] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4", EURASIP Journal on Applied Sig-nal Processing, Vol. 2002, No. 10, pp. 1021-1038, Hin-dawi Publishing Corporation, October 2002.

[16] K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis and S. Kollias. "Facial expression and gesture analysis for emotionally-rich man-machine interaction" N. Sarris,

[17] G. Klir, B.Yuan, Fuzzy Sets and Fuzzy Logic, Theory and Applications, New Jersey, Prentice Hall, 1995.

[18] M.A. Lee, H. Takagi, Integrating design stages of fuzzy systems using genetic algorithms, proceedings of IEEE International conference on fuzzy systems, 1993.

[19] Vikram Chalana and Yongmin Kim, A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images, IEEE Transactions on Medical Imaging, Vol.16, No.5 October 1997

[20] R. Cowie, E. Douglas-Cowie, Automatic statistical analysis of the signal and prosodic signs of emotion in speech. Proceedings of the 4th International Conference of Spoken Language Processing (pp. 1989–1992). 1996, Philadelphia, USA.