Dealing with feature uncertainty in facial expression recognition

Manolis Wallace*

Department of Computer Science,
University of Indianapolis, Athens Campus,
9 Ipitou Str., 105 57, Syntagma, Athens, Greece
E-mail: wallace@uindy.gr
Website: http://www.uindy.gr/faculty/cv/wallace_manolis/
*Corresponding author

Spiros Ioannou, Amaryllis Raouzaiou,
Kostas Karpouzis and Stefanos Kollias

Department of Electrical and Computer Engineering,
National Technical University of Athens,
Heroon Polytechniou 9, 157 80 Zographou, Greece
E-mail: sivann@image.ntua.gr E-mail: araouz@image.ntua.gr
E-mail: kkarpou@image.ntua.gr E-mail: stefanos@image.ntua.gr
Website: http://www.image.ntua.gr/

Abstract: Since facial expressions are a key modality in human communication, the automated analysis of facial images for the estimation of the displayed expression is central in the design of intuitive and human friendly human–computer interaction systems. In existing approaches, over-formalised description of knowledge concerning the human face and human expressions, as well as failures of the image and video processing components, often lead to misclassification. In this paper, we propose the utilisation of extended fuzzy rules for the more flexible description of knowledge, and the consideration of uncertainty and lack of confidence in the process of feature extraction from image and video. The two are combined using a flexible possibilistic rule evaluation structure, leading to more robust overall operation. The proposed approach has been implemented as an extension to an existing expression analysis system and conclusions from comparative study have been drawn.

Keywords: facial expression recognition; facial feature extraction; information fusion.

Reference to this paper should be made as follows: Wallace, M., Ioannou, S., Raouzaiou, A., Karpouzis, K. and Kollias, S. (xxxx) 'Dealing with feature uncertainty in facial expression recognition', *Int. J. Intelligent Systems Technologies and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Manolis Wallace obtained his Diploma from NTUA in 2001 and his PhD from the Computer Science Division of NTUA in 2005. His main research interests include Handling of Uncertainty, Information Systems, Data Mining, Personalisation and Applications of Technology in Education. He has published more than 40 papers in the above fields, 10 of which were in international journals. Manolis Wallace is the author of a book

on Data Structures, the guest editor of 2 journal special issues, co-author of a book on Image Processing and the organising committee Chair of AIAI 2006. Since 2001 he has been with the University of Indianapolis, Athens Campus, where he serves now as an Assistant Professor and Chair of the Department of Computer Science.

Spiros Ioannou graduated from the Department of Electrical and Computer Engineering of the National Technical University of Athens in 2000 and is currently pursuing his PhD degree at the Image, Video, and Multimedia Systems Laboratory at the same University. His current research interests lie in the areas of Expression Analysis, Facial Feature Extraction and Machine Vision. He is a member of the Technical Chamber of Greece. He is with the team of NoE HUMAINE (IST-2002-2.3.1.6 Multimodal Interfaces.).

Amaryllis Raouzaiou graduated from the Department of Electrical and Computer Engineering, the National Technical University of Athens in 2000. She is currently pursuing her PhD degree at the Image, Video, and Multimedia Systems Laboratory at the same University. Her current research interests lie in the areas of Human-computer interaction, Synthetic-natural Hybrid Video Coding and Machine Vision. She is a member of the Technical Chamber of Greece. She is with the team of NoE HUMAINE (IST-2002-2.3.1.6 Multimodal Interfaces.).

Kostas Karpouzis graduated from the Department of Electrical and Computer Engineering, the National Technical University of Athens in 1998 and received his PhD degree in 2001 from the same University. His current research interests lie in the areas of Human-computer Interaction, Image and Video Processing, 3D Computer Animation and Virtual Reality. He is a member of the technical committee of the International Conference on Image Processing (ICIP). Since 1995 he has participated in seven research projects at the Greek and European levels.

Stefanos Kollias received his Diploma from NTUA in 1979, his MSc in Communication Engineering in 1980 from UMIST in England and his PhD in Signal Processing from the Computer Science Division of NTUA. He has been with the Electrical Engineering Department of NTUA since 1986 where he serves now as a Professor. Since 1990 he has been Director of the Image, Video and Multimedia Systems Laboratory of NTUA. He has published more than 120 papers in the above fields, 50 of which were in international journals. He has been a member of the Technical or Advisory Committee, and has been an invited speaker at 40 International Conferences. He is a reviewer of 10 IEEE Transactions and of 10 other journals. He and his team have been participating in 38 European and National projects.

1 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 7% to the effect of the message as a whole; the paralinguistic part, that is how the specific passage is vocalised, contributes for 38%, while facial expression of the speaker contributes for 55% to the effect of the spoken message (Mehrabian, 1968). This implies that the facial expressions form the major modality in human communication.

Facial expressions are generated by contractions of facial muscles; these result in temporally deformed prominent facial features such as eyelids, eyebrows and lips, often indicated by wrinkles. Hence, one can model a particular expression as a set of given concurrent deformations. In this framework, facial expression intensities may be measured by determining the geometric deformation of the particular facial features and examining their relation to the ones depicted in the priori represented expressions; barring situations of extreme or acted expressions, in most circumstances more than one of these representations may be close enough to the actual measurements. An overview of the methodologies used for automatic analysis of facial expression can be found in (Fasel and Luettin, 2003). A usual approach to measuring deformation, fortified by the fact that there are inter-personal variations of facial action amplitude, is to refer to the neutral-expression face of a given person.

In addition to issues related to expression representation, an important parameter of this approach is the effectiveness of the image processing procedures. In actual situations, such as processing visual data from talk shows, many kinds of noise may hinder feature extraction: subjects turning their heads or moving their hands may lead to feature occlusion, or bad and uneven lighting may hamper edge- or colour-based feature extraction algorithms. As a result, the appearance and deformation of one or more features may not be available for a given frame of a video sequence; worse yet, an erroneous deformation estimate may be unknowingly fed into the knowledge representation infrastructure.

In these circumstances, the easiest (and safest) way for an expression recogniser to get around would be to provide no label for the given sequence. However, the lack of evidence for a particular feature being deformed, when this feature is used in the representation of an expression, should not always be considered as absence of this feature: it may be attributed to a mistake of the image processing algorithms or to the fact that the feature may not be essential for the representation of the particular expression. A flexible recogniser should be able to handle the absence of information or evidence and incorporate it into the final estimate.

In this paper, we quantify the uncertainty generated during the image processing for feature extraction phase through validation of the results against a set of anthropometric criteria and propose a methodology based on which fuzzy rules containing knowledge on expression analysis and estimation can be evaluated in an uncertain environment. The structure of the paper is as follows: In Section 2, we briefly review expression representation as proposed by psychologists and explain how these are ported to expression analysis practice by computer scientists. Continuing, in Section 3, we explain how information required to evaluate rule antecedents can be extracted from still facial images, and how uncertainty in the image processing steps can be both minimised and measured. Section 4 discusses the evaluation of the fuzzy rules representing the mapping between measure features and estimated expression, given the uncertainty contained in the input provided by the image processing steps of Section 3. Section 5 lists results from the application of the proposed approach to an annotated database of static and moving facial images. A more conventional approach with rule evaluation that disregards input

uncertainty is also applied on the same data and conclusions are drawn through comparisons. Finally, Section 6 lists our concluding results.

2 Preliminaries

In the 1990s, automatic facial expression analysis research gained much interest – thanks mainly to progress in the related fields such as image processing (face detection, tracking and recognition) and the increasing availability of relatively cheap computational power (Fasel and Luettin, 2003). In one of the groundbreaking and most publicised works, Mase and Pentland (1991) used measurements of optical flow to recognise facial expressions. In the following, Lanitis et al. (1997) used a flexible shape and appearance model for face identification, pose recovery and facial expression recognition. Black and Yacoob (1997) proposed local parameterised models of image motion to recover non-rigid facial motion, which was used as input to a rule-based basic expression classifier. Local optical flow was the basis of Rosenblum's (1996) work, utilising a radial basis function network for expression classification. Regarding feature-based techniques, Donato et al. (1999) tested different features for recognising facial AUs and inferring the facial expression in the frame. Oliver et al. (1997) tracked the lower face to extract mouth shape information and fed them to an HMM, recognising again only universal expressions.

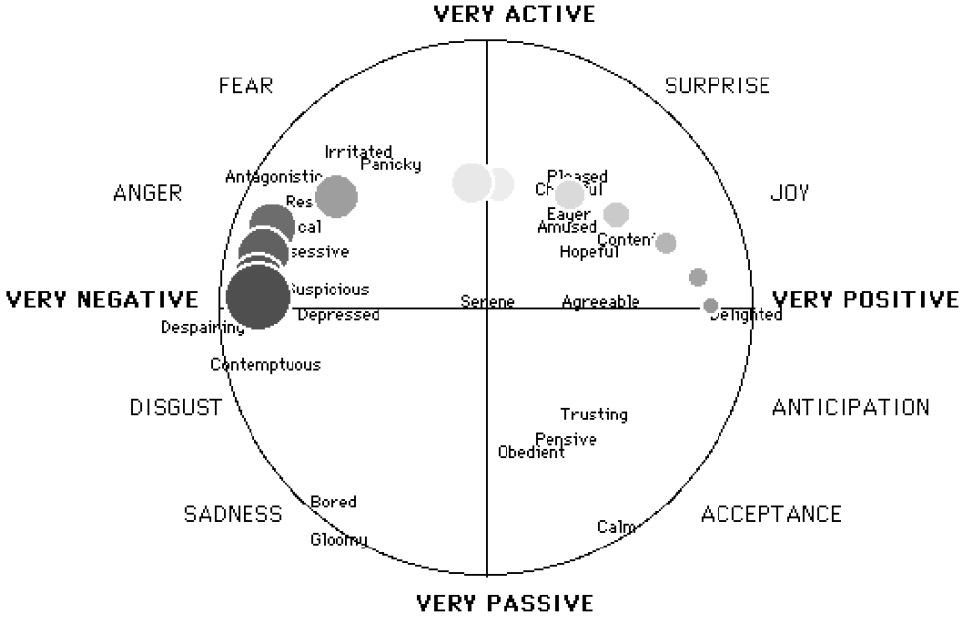
The obvious goal for expression analysis applications is to assign category labels that identify expressional states. However, labels as such are very poor descriptions, especially, since humans use a daunting number of labels to describe expression. Therefore, we need to incorporate a more transparent, as well as continuous representation, that matches closely our conception of what expression are or, at least, how they are displayed and perceived. Activation–emotion space (Cowie et al., 2001) is a representation that is both simple and capable of capturing a wide range of significant issues in expression. It rests on a simplified treatment of two key themes.

- *Valence*. The clearest common element of emotional and expressional states is that the person is materially influenced by feelings that are 'valenced', i.e., they are centrally concerned with positive or negative evaluations of people, or things or events; the link between emotion, expression and valencing is widely agreed.
- *Activation level*. Research has recognised that emotional and expressional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e., the strength of the person's disposition to take some action rather than none.

The axes of the activation–evaluation space reflect those themes. The vertical axis shows activation level, the horizontal axis evaluation. A basic attraction of that arrangement is that it provides a way of describing emotional and expressional states which is more tractable than using words, but which can be translated into and out of verbal descriptions. Translation is possible because emotion-related words can be understood, at least to a first approximation, as referring to positions in activation–emotion space. Various techniques lead to that conclusion, including factor analysis, direct scaling and others (Whissel, 1989).

A surprising amount of emotional discourse can be captured in terms of activation–emotion space. Perceived full-blown emotions are not evenly distributed in activation–emotion space; instead they tend to form a roughly circular pattern. From that and related evidence, work presented in Plutchik (1980) shows that there is a circular structure inherent in emotionality. In this framework, identifying the centre as a natural origin has several implications. Emotional strength can be measured as the distance from the origin to a given point in activation–evaluation space. The concept of a full-blown expression can then be translated roughly as a state where emotional and expressional strength has passed a certain limit. An interesting implication is that strong expressions are more sharply distinct from each other than weaker expressions with the same emotional orientation. A related extension is to think of primary or basic expressions as cardinal points on the periphery of an expression circle. Plutchik has offered a useful formulation of that idea, the 'emotion wheel'; the emotion wheel is presented in Figure 1.

Figure 1 The activation–emotion space



In the framework of MPEG-4 standard, parameters have been specified for Face and Body Animation (FBA) by defining specific Face and Body nodes in the scene graph; the initial goal of FBA definition is the animation of both realistic and cartoonist characters. Thus, MPEG-4 has defined a large set of parameters and the user can select subsets of these parameters according to the application. MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for Facial Animation Parameter (FAP) definition; these feature points are presented in Tekalp and Ostermann (2000). FAPs are defined through the comparison of distances between pairs of feature points on the observed and the neutral face. Most of the techniques for facial animation are based on the well-known system for describing "all visually distinguishable facial movements", FACS. FACS is an anatomically oriented coding system, based on the definition of 'Action Units' (AU) of a face that cause facial movements. An Action Unit could

combine the movement of two muscles or work in the reverse way, i.e., split into several muscle movements. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard (Tekalp and Ostermann, 2000). In particular, the Facial Definition Parameter (FDP) and the Facial Animation Parameter set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture through FDPs, thus eliminating the need for specifying the topology of the underlying geometry, and the animation of faces through FAPs, thus reproducing expressions, emotions and speech pronunciation (see Table 1).

3 Feature extraction

Besides expression representation, an important parameter of the expression analysis process is the effectiveness of the image processing procedures. Automatic analysis systems usually require good input to avoid misclassification or errors, which is often ensured by the use of specific environment conditions such as in Pantic and Rothkrantz (2000a, 2000b). In actual situations, such as processing visual data from talk shows, many kinds of noise may hinder feature extraction: subjects turning their heads, or moving their hands may lead to feature occlusion or bad and uneven lighting may hamper edge- or colour-based feature extraction algorithms. As a result, the appearance and deformation of one or more features may not be available for a given frame of a video sequence; worse yet, an erroneous deformation estimate may be unknowingly provided as input to the subsequent expression analysis and classification procedures.

In this work, we utilise our recent work in feature extraction methods described in Ioannou et al. (2005). Precise facial feature extraction is performed resulting in a set of 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 feature points. 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, three for the mouth and one for each one of the eyebrows. The methodologies applied in the extraction of these masks include.

- A feed-forward back propagation neural network trained to identify eye and non-eye facial area. The network has 13 inputs; for each pixel on the facial region the NN inputs are luminance Y, chrominance values Cr and Cb and the ten most important DCT coefficients (with zigzag selection) of the neighbouring 8×8 pixel area.
- A second neural network, with similar architecture to the first one, trained to identify mouth regions.
- Luminance based masks, which identify eyelid and sclera regions.
- Edge-based masks.
- A region growing approach based on standard deviation.

Since, as we 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; of course, different criteria are applied to masks of different facial features. 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; in this manner, a validity confidence degree is generated for each one of the initial feature masks. For example, two criteria that can be used for the validation of the eye masks are the following:

$$M_{\text{eye}}^{1c} = 1 - \left| 1 - \frac{d_2 / d_6}{0.49} \right| \tag{1}$$

and

$$M_{\text{eye}}^{2c} = 1 - \frac{|d_4|}{d_5} \tag{2}$$

where M_{eye}^{1c} and M_{eye}^{2c} are the confidence degrees acquired through the application of each validation criterion on an eye mask. The former of the two criteria is based on Young (1993), where the ration of eye width over bipupil breadth is reported as constant and equal to 0.49. In almost all cases these validation criteria, as well as the other criteria utilised in mask validation, produce confidence values in the [0,1] range. In the rare cases that the estimated value exceeds the limits, it is set to the closest extreme value, 0 for negative values and one for values exceeding one. The features measured for the application of the two example criteria are explained in Table 2.

Table 1 FAPs vocabulary for archetypal expression description

Joy	F_3, F_4, F_5, F_6, F_7, F_{12}, F_{13}, F_{19}, F_{20}, F_{21}, F_{22}, F_{33}, F_{34}, F_{41}, F_{42}, F_{53}, F_{54}
Sadness	F_{19}, F_{20}, F_{21}, F_{22}, F_{31}, F_{32}, F_{33}, F_{34}, F_{35}, F_{36}
Anger	F_4, F_5, F_{16}, F_{18}, F_{19}, F_{20}, F_{21}, F_{22}, F_{31}, F_{32}, F_{33}, F_{34} , F_{35}, F_{36}, F_{37}, F_{38}
Fear	F_3, F_4, F_5, F_8, F_9, F_{10}, F_{11}, F_{19}, F_{20}, F_{21}, F_{22}, F_{31}, F_{32}, F_{33}, F_{34}, F_{35}, F_{36}, F_{37}, F_{38}
Disgust	F_3, F_4, F_5, F_8, F_9, F_{10}, F_{11}, F_{19}, F_{20}, F_{21}, F_{22}, F_{33}, F_{34}, F_{55}, F_{56}, F_{57}, F_{58}, F_{59}, F_{60}
Surprise	F_3, F_5, F_6, F_7, F_{10}, F_{11}, F_{19}, F_{20}, F_{21}, F_{22}, F_{31}, F_{32}, F_{33}, F_{34}, F_{35}, F_{36}, F_{37}, F_{38}, F_{53}, F_{54}

Table 2 Eye mask features used in the process of mask validation

d_6	Bipupil breadth
d_2	Eye width
d_4	Distance of eye's middle vertical coordinate and eyebrow's middle vertical coordinate
d_5	Eyebrow width

For the features for which more than one mask has been detected using different methodologies, the multiple masks are 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, they provide independent information on the location on the mask; combining information from independent sources has the property of alleviating a portion of the uncertainty present in the individual information components. In other words, the final masks that are acquired via mask fusion are accompanied by lesser uncertainty than each one of the initial masks.

The fusion algorithm is based on a Dynamic Committee Machine structure that combines the masks based on their validity confidence, thus producing a final mask together with the corresponding estimated confidence (Krog and Vedelsby, 1995; Dietterich, 2000). As already explained, this confidence degree is always higher than the degree of any of the considered initial masks. A final, more refined, confidence value can be acquired when also taking into account the temporal information from the video sequence. The final confidence for each feature mask is based on three parameters: absolute anthropometric measurements based on Young (1993), face symmetry exploitation and examination of the facial feature size constancy over a period of ten frames. The outcome of this procedure is a set of final masks along with the final confidence of their validity.

A way to evaluate our feature extraction performance is Williams' Index (*WI*) (Williams, 1976), which compares the agreement of an observer with the joint agreement of other observers. An extended version of WI, which deals with multivariate data, can be found in Chalana and Kim (1997). The modified Williams' Index *I*' 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:

$$WI = \frac{\frac{1}{n}\sum_{j=1}^{n}(1/D_{0,j})}{\frac{2}{n(n-1)}\sum_{j}\sum_{j':j'>j}(1/D_{j,j'})} \tag{3}$$

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

$$D_{j,j'} = \frac{1}{D_{bp}} \| M_j^x \veebar M_{j'}^x \| \tag{4}$$

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

From a dataset of about 50,000 frames, 250 frames were selected at random and the 19 FPs were manually selected from two observers. *WI* was calculated using equation (3) for each feature and for each frame separately. Distribution of the average *WI* calculated over the two eyes and mouth for each frame is shown in Figure 2, while Figure 3 depicts the average *WI* calculated on the two eyebrows.

Figure 2 Williams index distribution (average on eyes and mouth)

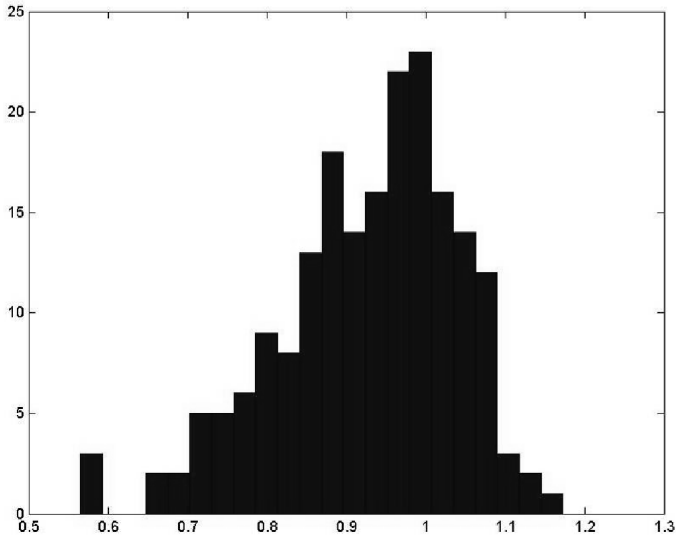
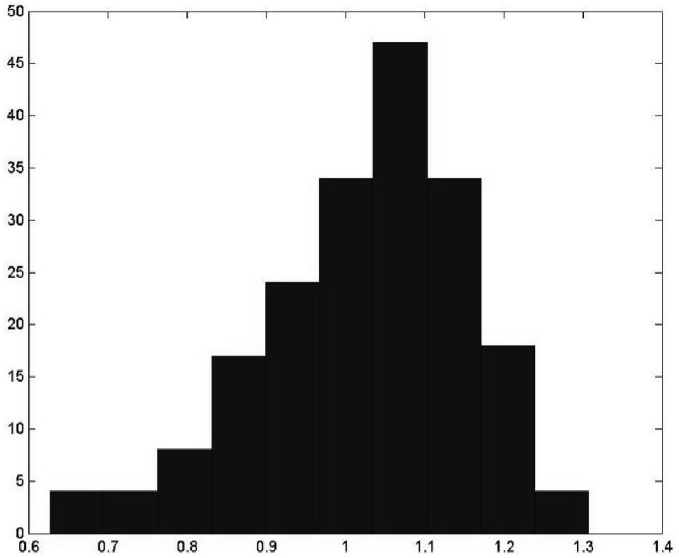


Figure 3 Williams index distribution (average on left and right eyebrows)



These 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} is estimated as:

$$F_{37} = \| FP^n_{4.5} - FP^n_{3.11} \| - \| FP_{4.5} - FP_{3.11} \| \tag{5}$$

where FP^n_i, FP_i are the locations of feature point i on the neutral and the observed face, respectively, and $\|FP_i - FP_j\|$ 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

$$F^c_{37} = \min(FP^c_{4.5}, FP^c_{3.11}) \tag{6}$$

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

$$F^1_{31} = \| FP^n_{3.1} - FP^n_{3.3} \| - \| FP_{3.1} - FP_{3.3} \| \tag{7}$$

or as

$$F^2_{31} = \| FP^n_{3.1} - FP^n_{9.1} \| - \| FP_{3.1} - FP_{9.1} \| . \tag{8}$$

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 an FAP, the final value and confidence for the FAP are as follows:

$$F_i = \frac{F^1_i + F^2_i}{2}. \tag{9}$$

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

$$E^1_i = 1 - F^{1c}_i \tag{10}$$

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:

$$E_i = t(E^1_i, E^2_i) \tag{11}$$

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

$$E_i = 1 - \min(1, ((1 - E^1_i)^w + (1 - E^2_i)^w)^w). \tag{12}$$

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

$$F^c_i = 1 - E_i. \tag{13}$$

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 the fuzzy numbers defined for each FAP, and confidence degrees x^c_j are inherited from the FAP:

$$x_j^c = F_i^c \tag{14}$$

where F_i is the FAP based on which antecedent x_j is defined.

4 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 analyse 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.

A conventional approach to the evaluation of fuzzy rules of the form

$$\text{IF } x_1, x_2, ..., x_n \text{ THEN } y \tag{15}$$

is as follows (Klir and Yuan, 1995):

$$y = t(x_1, x_2, ..., x_n) \tag{16}$$

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

$$t(x_1, x_2, ..., x_n) = \min(x_1, x_2, ..., x_n) \tag{17}$$

the algebraic product

$$t(x_1, x_2, ..., x_n) = x_1 \cdot x_2 \cdot \cdots \cdot x_n \tag{18}$$

the bounded sum

$$t(x_1, x_2, ..., x_n) = x_1 + x_2 + \cdots + x_n + 1 - n \tag{19}$$

and so on. Another well-known approach in rule evaluation is described in Lee and Takagi (1993) and utilises a weighted sum instead of a t-norm in order to combine information from different rule antecedents:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n. \tag{20}$$

Both approaches are well studied and established in the field of fuzzy automatic control. 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, as was explained in Section 3, FAPs may well be estimated with a very low confidence, or not estimated at all, owing to low video quality, speech interference, occlusion, noise and so on. Thus, a more flexible rule evaluation scheme is required, which is able to incorporate such uncertainty as well.

Moreover, the second one of the conventional approaches, owing 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 an 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 generalisation of the *t*-norm based conventional approach.

In the *t*-norm operation described in equation (16), antecedents with lower values affect most the resulting value of y, while antecedents with values close to 1 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 $c(k(x))$ 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:

$$c(k(x)) = t(x^c, c(x)) \tag{21}$$

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

$$k(x) = u(c(x^c), x) \tag{22}$$

where u is a fuzzy *s*-norm. It is easy to see that equation (22) 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,
- $x^c \to 0$, 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:

$$y = t(k(x_1), k(x_2), ..., k(x_n)). \tag{23}$$

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 utilised inputs, i.e., by the confidence values of the rule antecedents, as follows:

$$y^c = \frac{x_1^c + x_2^c + \cdots + x_n^c}{n}. \tag{24}$$

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.

In order to have a complete possibilistic representation of the rule evaluation process, together with the plausibility of the expression profile we need to estimate the corresponding belief, i.e., the degree to which available evidence suggests that the expression profile is present in the considered input.

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:

$$\text{Bel} = t(y, y^c) \tag{25}$$

$$\text{Pl} = y. \tag{26}$$

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
- 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, owing 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 (25), will be low.

An additional flexible approach, to dealing with situations in which the output of the rule evaluation process does not provide a clear and confident output, is the combination of the output of the application of facial expression analysis on multiple (almost) contiguous frames (Wallace et al., 2004). Once more, the reasoning of the approach is that combining information from multiple sources alleviates a portion of the uncertainty related to each independent bit of information.

5 Experimental results

The goal of IST project ERMIS is the development of a prototype system for human–computer interaction than can interpret its users' attitude or emotional state, e.g., interest, boredom, anger, etc., in terms of their speech and their facial gestures and expressions (IST Project: Emotionally Rich Man-Machine Interaction Systems (ERMIS), 2001–2003). In this framework, a software prototype of the expert system has been developed that is able to automatically categorise facial expressions observed on real faces. As far as the knowledge of the system is concerned, facial expression information is coded using MPEG-4 FAPs (Raouzaiou et al., 2002) and expressed through conventional fuzzy rules (Ioannou et al., 2004). The evaluation of the fuzzy rules is also performed in the conventional manner.

In order to experimentally validate the approach proposed in this paper, we have used the software prototype of ERMIS as a test bed. Specifically, we have altered the rule evaluation component to the more flexible possibilistic evaluation methodology described in Section 4. Of course, the feature extraction module was also edited, as to allow for the estimation of the confidence that accompanies the results it produces, as described in Section 3.

Figure 4 presents frame A, one of the frames that lead the original prototype to failure. As we can see in Figure 5, where the masks for the eyes detected using the various implemented approaches are presented, the utilised methodologies do not provide reliable eye region detection. As a consequence, the FAP specifications acquired using any of these approaches are unreliable and lead to poor performance of the expression classification component. When, on the other hand, we combine these masks, as described in Section 3, considering at the same time the confidence in their validity, we acquire the greatly improved result presented in Figure 6(a), which, as expected, allows the following expression classification process to operate without problems. The most important feature points detected on frame A, using the feature masks that resulted from the process of the fusion of multiple masks, are presented in Figure 6(b). The original prototype did not provide any output owing to the asymmetry in the detection of the eye related points, whereas the proposed methodology activates (to a high degree) the following three rules, all corresponding to profiles of the archetypal expression of joy.

Figure 4 Original frame A

Figure 5 Masks for the eyes in frame A detected using different methodologies

Figure 6 (a) Final mask for the eyes in frame A and (b) detected feature points on frame A

(a) (b)

IF close_left_eye IS close_left_eye_low
AND close_right_eye IS close_right_eye_low
AND raise_left_inner_eyebrow IS raise_left_inner_eyebrow_high
AND raise_right_inner_eyebrow IS raise_right_inner_eyebrow_high
AND raise_left_medium_eyebrow IS raise_left_medium_eyebrow_high
AND raise_right_medium_eyebrow IS raise_right_medium_eyebrow_high
AND raise_left_outer_eyebrow IS raise_left_outer_eyebrow_high
AND raise_right_outer_eyebrow IS raise_right_outer_eyebrow_high
AND squeeze_left_eyebrow IS squeeze_left_eyebrow_low
AND squeeze_right_eyebrow IS squeeze_right_eyebrow_low
THEN output IS quadrant_1
with $y = 0{,}3015$ and $y^0 = 0{,}714883$

IF close_left_eye IS close_left_eye_low
AND close_right_eye IS close_right_eye_low
AND raise_left_inner_eyebrow IS raise_left_inner_eyebrow_high
AND raise_right_inner_eyebrow IS raise_right_inner_eyebrow_high
AND raise_left_medium_eyebrow IS raise_left_medium_eyebrow_high
AND raise_right_medium_eyebrow IS raise_right_medium_eyebrow_high
AND raise_left_outer_eyebrow IS raise_left_outer_eyebrow_high
AND raise_right_outer_eyebrow IS raise_right_outer_eyebrow_high
THEN output IS quadrant_1
with $y = 0{,}3015$ and $y^0 = 0{,}716673$

IF close_left_eye IS close_left_eye_low
AND close_right_eye IS close_right_eye_low
AND raise_left_inner_eyebrow IS raise_left_inner_eyebrow_high
AND raise_right_inner_eyebrow IS raise_right_inner_eyebrow_high
AND raise_left_medium_eyebrow IS raise_left_medium_eyebrow_high
AND raise_right_medium_eyebrow IS raise_right_medium_eyebrow_high
AND raise_left_outer_eyebrow IS raise_left_outer_eyebrow_high
AND raise_right_outer_eyebrow IS raise_right_outer_eyebrow_high
AND squeeze_left_eyebrow IS squeeze_left_eyebrow_low
AND squeeze_right_eyebrow IS squeeze_right_eyebrow_low
AND wrinkles_between_eyebrows IS wrinkles_between_eyebrows_low
THEN output IS quadrant_1
with $y = 0{,}3015$ and $y^0 = 0{,}69938$

The overall results of the two evaluation approaches (conventional and possibilistic) are summarised in Table 3. We can see that although the conventional approach totally fails to provide any output, the proposed possibilistic approach both identifies quadrant one as the correct output and incorporates the inputs' uncertainty in the output.

Table 3 Summary of results

Quadrant	*Ground truth*	*Conventional*	*Belief*	*Plausibility*
1	1	0	0,21608	0,3015
2	0	0	0,06160	0,09135
3	0	0	0,00238	0,00352
Neutral	0	0	<0,00001	<0,00001

As a different example, let us consider frame B, presented in Figure 7(a). The original prototype fails to activate any of the rules in the rule base for this frame as well. As can be seen from Figure 7(b) and (c), where the final masks and feature points for the eyes and the mouth are presented, this is not a case that can be handled successfully by simply considering multiple masks; the resulting masks are again poor estimators of the real feature positions. Anthropometric validation of these masks yields a low confidence degree (0.6) for the left eye (right eye as we observe the picture) and much lower confidence degrees for the mouth (0.4) and right eye (0). Since most FAPs considered in the rules of the expert system are defined considering at least one of the eyes and/or the mouth, no rule is activated with a high confidence. Still, as the right eye is totally ignored and the left eye and mouth are only partially considered, a number of expression profiles are indicated as having high plausibility. The gain, when considered to the output of the original prototype, is that the system now provides the information that, most probably, the observed expression is not a surprise, as the rules corresponding to expression profiles of the surprise archetypal expression have very low plausibility values, whereas the original prototype did not provide any information as output; in general, the original prototype, owing to the lack of optional rule components and the utilisation of a 'hard' approach in rule evaluation, does not provide any output in cases where asymmetries are detected on the face, as in frame B where one eye is estimated to be open and the other closed.

Figure 7 (a) Original frame B; (b) mouth and eyes mask for frame B and (c) detected feature points on frame B

(a)

(b) (c)

As a last example, let us consider frame C, presented in Figure 8(a). The original prototype fails to provide any output in this frame as well, owing to the poor performance of the mouth detection algorithms. As can be seen in Figure 8(b) and (d) none of the utilised methodologies can lead to the successful estimation of the mouth region in frame C. Moreover, even fusion of the masks cannot overcome the problem, as is made evident in Figure 9(a), where the final mask for the mouth is presented. Main feature points detected in frame C are presented in Figure 9(b). Owing to the fact that in the considered frame sequence the observed person is speaking throughout the recording, all rules have been edited as to make all FAPs that are defined using the mouth as optional. Thus, even if the detected mask had proper size, shape and location in order to be validated against anthropometric criteria (which is not the case of the mask in Figure 9(a)), the unreliable FAP estimations it would provide would not be allowed to characterise a profile as definitely not present; during speech, all FAPs that are defined using the mouth are considered as unreliable owing to the fact that mouth feature point positions are determined by phonemes rather than disposition. Rules activated in this case correspond to profiles of the disgust, fear and anger archetypal expressions.

Figure 8 (a) Original frame and (b) and (d) mouth masks detected in frame C

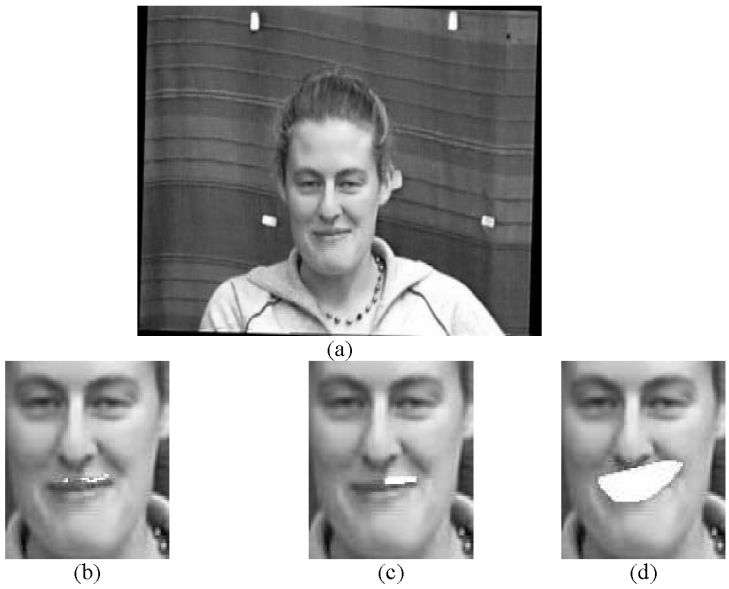


Figure 9 (a) Final mask for the mouth in frame C and (b) main feature points detected in frame C

Overall, through these sample frames, we can see that by the proposed approach, where imprecision as well as failure of the image processing process are considered, quantified and incorporated as information in the evaluation process, and optional antecedents are permitted in the rule base, a number of situations can be dealt with; these situations were not tractable by an otherwise successful system that did not have these characteristics. As a result, even in cases where insufficient information is available for the determination of the observed expression, the system is able to provide useful information by at least ruling out improbable cases.

6 Conclusions

Conventional facial expression analysis and classification systems often employ fuzzy rules for the representation of the knowledge utilised by the expert system. On the other hand, fuzzy rules and fuzzy expert systems are designed for problems where the input is provided in a constant and accurate manner by a set of sensors. In the case of facial expression analysis, where fuzzy inputs are the output of the imperfect process of feature extraction via image processing, conventional fuzzy rules and conventional rule evaluation methodologies are often inadequate and lead to extremely poor performance.

In this paper, we have chosen to independently apply multiple image processing methodologies and fuse their results, thus minimising the uncertainty that is inherent in this process. Moreover, we have utilised validation of feature masks against a set of anthropometric criteria in order to evaluate the quality of the information provided as input to the rule system by the image processing component, thus quantifying the related uncertainty; flexible rule evaluation has been proposed as the way to incorporate this information in the process of rule evaluation, thus tackling situations in which the traditional rule-based approach to facial expression recognition would have failed.

The final output of the proposed system is possibilistic rather than probabilistic. The activation level of a rule corresponds to the plausibility of the rule, thus indicating the degree to which available evidence does not contradict the rule. A combination of rule activation and confidence corresponds to the belief, thus indicating the degree to which available knowledge supports the rule. This is a reasonable feature of a system that aims to incorporate uncertainty and lack of confidence in its operation; probabilistic systems cannot provide meaningful or even reliable output in the case where insufficient input information is available.

Experimental application of the proposed methodology has indicated, as expected, that extended fuzzy rules, consideration of confidence in the process of feature extraction and flexible rule evaluation provide for more robust operation in an uncertain environment. Thus, the resulting system outperforms its conventional predecessor in cases where the image-processing component fails or the observed facial expression does not strictly comply to the specified rules by missing some optional characteristic.

As further extension to this work, we intent to examine the way the analysis of different modalities, such as speech, posture and gestures can be combined with facial expression analysis towards more accurate estimation of the human disposition. This will be pursued, among other ways, in the framework of the HUMAINE Network of Excellence (NoE 2004–2007).

References

Black, M. and Yacoob, Y. (1997) 'Recognizing facial expressions in image sequences using local parameterized models of image motion', *International Journal of Computer Vision*, Vol. 25, No. 1, pp.23–48.

Chalana, V. and Kim, Y. (1997) 'A methodology for evaluation of boundary detection algorithms on medical images', *IEEE Transactions on Medical Imaging*, October, Vol. 16, No. 5, pp.642–652.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. (2001) 'Emotion recognition in human-computer interaction', *IEEE Signal Processing Magazine*.

Dietterich, T.G. (2000) 'Ensemble methods in machine learning', in Kittler, J. and Roli, F. (Eds.): *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, Springer-Verlag, New York, pp.1–15.

Donato, G., Bartlett, S., Hager, C., Ekman, P. and Sejnowski, J. (1999) 'Classifying facial actions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No, 10, pp.974–989.

Fasel, B. and Luettin, J. (2003) 'Automatic facial expression analysis: a survey', *Pattern Recognition*, Pergamon–Elsevier, Vol. 36, pp.259–275.

Ioannou, S., Raouzaiou, A., Karpouzis, K., Pertselakis, M., Tsapatsoulis, N. and Kollias, S. (2004) 'Adaptive rule-based facial expression recognition', in Vouros, G. and Panayiotopoulos, T. (Eds.): *Lecture Notes in Artificial Intelligence*, Springer-Verlag, SETN 2004, Samos, Greece, Vol. 3025, pp.466–475.

Ioannou, S., Wallace, M., Karpouzis, K., Raouzaiou, A. and Kollias, S. (2005) 'Combination of multiple extraction algorithms in the detection of facial features', *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, September, Genova, Italy.

IST Project: (2001) Emotionally Rich Man-Machine Interaction Systems (ERMIS) http://www.image.ntua.gr/ermis/.

IST Project: (2002) Emotionally Rich Man-Machine Interaction Systems (ERMIS) http://www.image.ntua.gr/ermis/.

IST Project: (2003) Emotionally Rich Man-Machine Interaction Systems (ERMIS) http://www.image.ntua.gr/ermis/.

Klir, G. and Yuan, B. (1995) *Fuzzy Sets and Fuzzy Logic, Theory and Applications*, Prentice-Hall, New Jersey.

Krog, A. and Vedelsby, J. (1995) 'Neural network ensembles, cross validation and active learning', in Tesauro, G., Touretzky, D. and Leen, T. (Eds.): *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, Vol. 7, pp.231–238.

Lanitis, A., Taylor, C. and Cootes, T. (1997) 'Automatic interpretation and coding of face images using flexible models', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp.743–756.

Lee, M.A. and Takagi, H. (1993) 'Integrating design stages of fuzzy systems using genetic algorithms', *Proceedings of IEEE International Conference on Fuzzy Systems*, San Francisco, CA, Vol. 1, pp.612–617.

Mase, K. and Pentland, A. (1991) 'Recognition of facial expression from optical flow', *IEICE Transactions*, Vol. E74, No. 10, pp.3474–3483.

Mehrabian, A. (1968) 'Communication without words', *Psychology Today*, Vol. 2, No. 4, pp.53–56.

NoE: (2004) Human-Machine Interaction Network on Emotion (HUMAINE) http://emotion-research.net/.

NoE: (2005) Human-Machine Interaction Network on Emotion (HUMAINE) http://emotion-research.net/.

NoE: (2006) Human-Machine Interaction Network on Emotion (HUMAINE) http://emotion-research.net/.

NoE: (2007) Human-Machine Interaction Network on Emotion (HUMAINE) http://emotion-research.net/.

Oliver, N., Pentland, A.P. and Berard, F. (1997) 'LAFTER: lips and face real time tracker', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June, San Juan, pp.123–129.

Pantic, M. and Rothkrantz, L.J.M. (2000a) 'Expert system for automatic analysis of facial expressions', *Image and Vision Computing*, Vol. 18, pp.881–905.

Plutchik, R. (1980) *Emotion: A Psychoevolutionary Synthesis*, Harper and Row, NY, USA.

Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K. and Kollias, S. (2002) 'Parameterized facial expression synthesis based on MPEG-4', *Eurasip Journal on Applied Signal Processing*, Vol. 2002, No. 10, pp.1021–1038.

Rosenblum, M., Yacoob, Y. and Davis, L. (1996) 'Human expression recognition from motion using a radial basis function network architecture', *IEEE Transactions on Neural Networks*, Vol. 7, No. 5, pp.1121–1138.

Tekalp, A.M. and Ostermann, J. (2000) 'Face and 2-D mesh animation in MPEG-4', *Signal Processing: Image Communication*, Vol. 15, pp.387–421.

Wallace, M., Raouzaiou, A., Tsapatsoulis, N. and 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)*, July, Budapest, Hungary, Vol. 1, pp.51–54.

Whissel, C.M. (1989) 'The dictionary of affect in language', in Plutchnik, R. and Kellerman, H. (Eds.): *Emotion: Theory, Research and Experience: The Measurement of Emotions*, Academic Press, New York, Vol. 4, pp.113–131.

Williams, G.W. (1976) 'Comparing the joint agreement of several raters with another rater', *Biometrics*, Vol. 32, pp.619–627.

Pantic, M. and Rothkrantz, L.J.M. (2000b) 'Automatic analysis of facial expressions: the state of the art', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp.1424–1445.

Young, J.W. (1993) *Head and Face Anthropometry of Adult US Civilians*, *Office of Aviation Medicine*, Federal Aviation Administration, July, Tech. Report No. R0221201, DOT/FAA/AM-93/10, p.40.