Facial Expression Classification Based on MPEG-4 FAPs: The Use of Evidence and Prior Knowledge for Uncertainty Removal

Manolis Wallace, Amaryllis Raouzaiou, Nicolas Tsapatsoulis, Stefanos Kollias
Image, Video and Multimedia Systems Laboratory (IVML)
Departmet of Computer Science, School of Electrical and Computer Engineering
National Technical University of Athens (NTUA)
15780 Zographou, Greece
E-mail: {wallace, araouz, ntsap}@image.ntua.gr, stefanos@cs.ntua.gr

***Abstract*— As low resolution shots, rotations of the head with respect to the camera, face deformation due to speech and so on inflict a great deal of uncertainty in FAP measurements, uncertainty is also inherent in the process of expression analysis. In this paper we tackle such uncertainty via the observation that user emotions do not typically alter rapidly very often. Thus, possibilistic evidence may be gathered from each frame about the user expression; evidence from the current and recent frames can be combined using evidence theory.**

I. Introduction

Facial expression recognition is expected to enhance interactivity and assist human-computer interaction issues, letting the system become accustomed to the current needs and feelings of the user. Prospective applications include educational environments, 3D video conferencing and collaborative workplaces, online shopping and gaming, virtual communities and interactive entertainment. Due to its importance and wide range of application, the field has received much attention in the MPEG-4 framework with the definition of explicit FDPs and FAPs.

In previous work we have defined expression vocabularies, i.e. the set of FAPs that each may be activated for each expression, and expression profiles, i.e. sets of FAP values, each profile representing a specific instance of the expression [1][4][5]. Each profile is easily transformed into a fuzzy rule, thus leading to the generation of a neurofuzzy classifier that, given the FAP values extracted from a still image as input, provides an estimation of the user expression as output.

However, estimation of FAPs is neither easy non error free. Low resolution shots, rotations of the head with respect to the camera, face deformation due to speech and so on inflict a great deal of uncertainty in FAP measurements, which in turn leads to uncertainty in the classifiers output. Moreover, some FAPs may not be extractable from a specific frame, while others may need to be ignored (e.g. mouth FAPs in video conferencing where mouth Feature Points' motion is based on phonemes rather than expressions), which may lead to the overlapping of rules. Due to all these, uncertainty is inherent in the overall task of facial expression estimation.

The importance of this effect is made obvious when applying techniques such as the one sketched above in consecutive frames of a single shot; it is common to estimate quite different expressions for the same sequence, even for frames that are, time wise, close to each other. In this paper we tackle such uncertainty via the observation that user emotions do not typically alter rapidly very often. Thus, possibilistic evidence may be gathered from each frame about the user expression; evidence from the current and recent frames can be combined using evidence theory.

The structure of the paper is the following: in section II we review the definition of FAPs and the extraction of FAP values and in section III we present the neurofuzzy classification scheme utilized. Section IV discusses the utilization of evidence theory for the limitation of uncertainty of the expression classification. Finally, section V lists some indicative experimental results and section VI lists our concluding remarks.

II. FAP Estimation in Still Images

In general, facial expressions and emotions are described by a set of measurements and transformations that can be considered atomic with respect to the MPEG-4 standard. Modeling facial expressions and underlying emotions through FAPs serves several purposes:

1) Archetypal expressions occur rather infrequently; in most cases, emotions are expressed through variation of a few discrete facial features which are directly related with particular FAPs.
2) FAPs do not correspond to specific models or topologies; the subject's model does not need to be specifically studied before facial expression can be estimated.

On the other hand, two basic issues should be addressed when modelling archetypal expression using FAPs:

1) estimation of FAPs that are involved in their formation.
2) definition of the FAP intensities.

We relate each archetypal expression to its own FAP vocabulary, each vocabulary containing all the FAPs that may

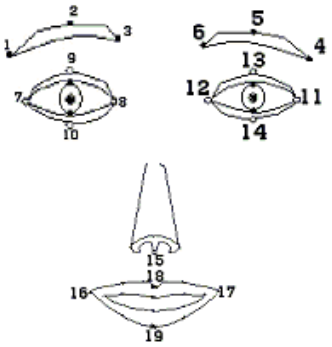

Fig. 1. FPs used for the definition of distances

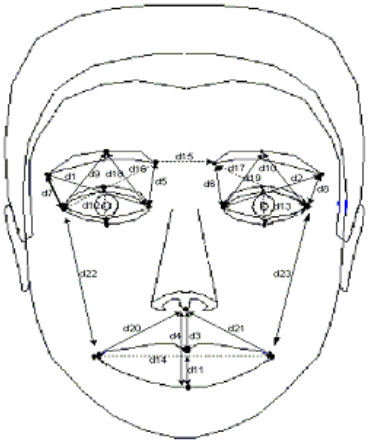

Fig. 2. Facial distances used for the definition of FAPs

be activated in some instance of the corresponding expression. A subset of this vocabulary that is activated during a specific expression, together with the corresponding FAP intensities, forms an expression profile.

Although FAPs are practical and very useful for animation purposes, they are somehow inadequate for analyzing facial expressions from video scenes or still images. The main reason for that is the absence of a clear quantitative definition of FAPs (at least of most of them). In order to be able to measure FAPs in real images and video sequences we define a way of describing them through the movement of some points that lie in the facial area and are able to be automatically detected (Feature Points, FPs), some of which are constant during expressions and are used as reference points.

Such a description can get advantage of the extended research made on automatic facial points detection [2][3]. In figure 1 we can see the FPs utilized, and in figure 2 the FP distances that are measured for the estimation of FAP values. More information on the definition of FAPs based on FPs and their utilization can be found in [4].

III. The Neurofuzzy Classification Scheme

Fuzzy systems are numerical model-free estimators. While neural networks encode sampled information in a parallel-distributed framework, fuzzy systems encode structured, empirical (heuristic) or linguistic knowledge in a similar numerical framework. Although they can describe the operation of the system in natural language with the aid of human-like if-then rules, they do not provide the highly desired characteristics of learning and adaptation. The use of neural networks in order to realize the key concepts of a fuzzy logic system enriches the system with the ability of learning and improves the sub symbolic to symbolic mapping. Neural network realization of basic operations of fuzzy logic, such as fuzzy complement, fuzzy intersection and fuzzy union, can be implemented in terms of the activation function of neurons to provide fuzzy logic inference.

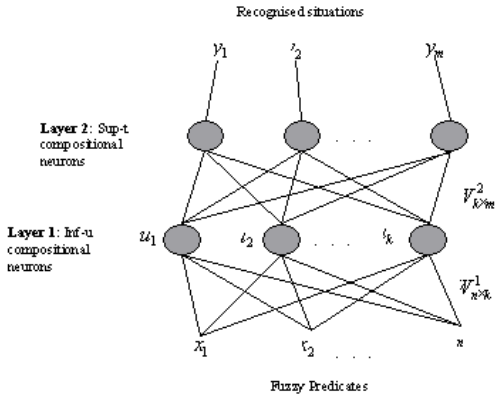

Fig. 3. The neurofuzzy architecture

The utilized architecture is a two-layer neural network of compositional neurons [7]. The first layer consists of the $inf-u$ neurons and the second layer consists of the $sup-t$ neurons. $W^1_{n\times k}$ is the weight matrix of the first layer and $W^2_{k\times m}$ is the weight matrix of the second layer (figure 3).

The network is loaded with prior knowledge through the utilization of "if - then" rules for the creation and parameter initialization of its nodes. Each rule defines as a predicate the activated FAPs together with the corresponding FAP values and provides as output the expression estimation. Operation of the network can be further enhanced in the presence of labelled data through credit assignment learning.

IV. Possibilistic Uncertainty Removal

Applying the above methodology on a still image, the neurofuzzy system provides in its output the fuzzy set of expressions estimated as activated; degrees in this fuzzy set indicate the activation level of the corresponding output node. Utilizing the principle of minimum uncertainty, the expression that is estimated as most activated is selected as the one appearing in the still image. Of course, numerous cases exist in which this is not satisfactory:

1) more than one expressions are activated to a high degree.
2) none of the expressions is activated to a high degree.
3) the difference in level of activation between the first (selected) expression and others in not great.

In all these cases, the still image is classified with a large probability of error. Moreover, in some special cases of video shots, such as shots of talking people, where mouth motion is governed by speech rather than by emotion, this type of uncertainty does not characterize just one frame but most of the frames in the sequence. Thus, applying the above technique in such a sequence, we get as output a sequence of highly insecure expression classifications.

On the other hand, the fact that the frames are placed close to each other in time allows as to assume frequent shifts between certain "incompatible" expressions as improbable. Based on this assumption, evidence theory may be utilized to combine uncertain information from distinct frames in order to refine the overall estimations.

We have chosen to utilize Dempster's rule [8] for the combination of the evidence from different frames; more sophisticated methodologies for the combination of evidence can be acquired from [9][10]. In order to apply the rule, we need to have focal elements than are clearly defined as far as their intersection / compatibility is concerned. Therefore, following the principle of maximum uncertainty, we do not apply the rule directly on expressions, but rather on classes of expressions; although some expression shifts are possible, shifting between particular expressions is not probable (e.g. surprise might shift to joy, but joy will not shift to fear within a very short time period).

The four expression classes considered in this work are defined using Whissel's wheel [6] (figure 4). Whissel has suggested that emotions are points in a space with a relatively small number of dimensions, which with a first approximation, seem to occupy two dimensions: activation and evaluation. Out of the four quarters of the considered space, only three are populated with actual expressions. Thus, we group expressions into four classes:

1) Quarter 1 (+/+ area of the circle, labeled A)
2) Quarter 2 (-/+ area of the circle, labeled B)
3) Quarter 3 (-/- area of the circle, labeled C)
4) Neutral expression (center of the circle, labelled N)

The application of the technique based above on frame i of a sequence produces activation levels a_i, b_i, c_i, n_i, each one corresponding to its respective class. Having selected the classes as to assure incompatibility, the only focal elements of our body of evidence are A, B, C, N, X, with

$$X = A \cup B \cup C \cup N$$

$$A \cap B = A \cap C = A \cap N = B \cap C = B \cap N = C \cap N = \emptyset$$

High activation of one of a class's expressions at the output of the neurofuzzy system can be considered as evidence that the corresponding class actually characterizes the specific frame. Thus basic probability assignments can be acquired as:

$$m_i(A) = a_i \tag{1}$$

$$m_i(B) = b_i \tag{2}$$

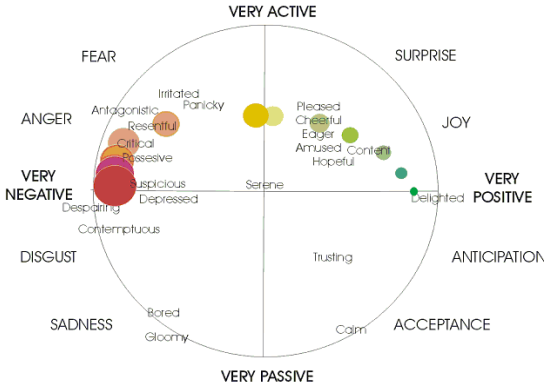

Fig. 4. The activation-evaluation space

$$m_i(C) = c_i \tag{3}$$

$$m_i(N) = n_i \tag{4}$$

The value of $m_i(X)$ is set as:

$$m_i(X) = max[0, 1 - \sum_{K \neq X} m_i(K)] \tag{5}$$

Dempster's rule states that

$$m_{i,j}(L) = \frac{\sum_{K_1 \cap K_2 = L} m_i(K_1) \cdot m_j(K_2)}{1 - \sum_{K_1 \cap K_2 = \emptyset} m_i(K_1) \cdot m_j(K_2)} \tag{6}$$

Taking into account equations 1, 2, 3, 7 and 5, it is easy to acquire from 6 the values of $m_{i,j}(L)$, for $L = A, B, C, N, X$, using the outputs of the neurofuzzy.

In our body of evidence, where focal elements are not nested, it is easy to see that

$$Bel(L) = \sum_{K \subset L} m_{i,j}(K) = m_{i,j}(L) \tag{7}$$

In the cases that the same class L is indicated by both frames, the above methodology indicates again the same frame, but with a greater degree of belief. In cases, on the other hand, that different classes are indicated, Dempster's rule will indicate the one for which the greatest belief is available, based on the available information.

Of course, in a manner quite similar the the one presented above, information from more that two frames can be combined for the extraction of a more secure conclusion.

V. Experimental Results

The methodology presented above has been applied on three sequences of test data. The test sequences display a woman who is talking; speech interferes with the values of FAPs thus disturbing the output of the neurofuzzy classification system. The classification of the output of the neurofuzzy classifier in the four general expression classes is presented in table I. Each frame has been classified to the class for which the neurofuzzy has had the greatest activation; question marks correspond to

TABLE I

OUTPUT OF THE NEUROFUZZY SYSTEM

sequence	classification
1st	??BABAA
2nd	A?A?A?A?
3rd	CCBAAAN?ACAACAA

TABLE II

FRAME CLASSIFICATION WITH THE PROPOSED METHODOLOGY

sequence	classification
1st	??BAAAA
2nd	AAAAAAAA
3rd	CCBAAANNAAAAAAA

frames for which no expression has a considerable activation (checked using a small threshold). Degrees of activation are omitted for the sake of space.

Applying the proposed methodology on the these data, we acquire the classification presented in table II. For the extraction of expression classification information for each frame we used the same frame together with the one before it as sources of evidence. In order to apply the methodology to the first frame of each sequence, following the principle of maximum uncertainty, we initialized with

$$m_0(A) = m_0(B) = m_0(C) = m_0(N) = 0$$

$$m_0(X) = 1$$

which indicates complete lack of information.

We can observe that

- some of the frames that were not classified by the neurofuzzy system are now classified to some quarter of Whissel's wheel.
- Some shifts between classes of expressions have been eliminated.
- Degrees of belief for the classification of most frames have been augmented.

Fig. 5. One of the frames. The interference of speech with the facial expression is evident

VI. CONCLUSIONS AND FUTURE WORK

We started this paper by briefly reviewing the definition of FAP values based on FP distance measurements, as well as the utilization of a neurofuzzy system, loaded with a priori knowledge, for the automated classification of facial expressions. As we have explained, this process in not free of uncertainty.

Based on the justified assumption that emotions cannot change often within a small period of time, we can utilize evidence theory to combine information from consecutive frames in order to restrict this uncertainty. Indicative experimental results have been presented that demonstrate the potential of the proposed methodology.

As future work, attention can be given to the utilization of more sophisticated evidence combination methodologies, so that information from more than two frames is utilized, and sources of information have a weighted credibility based on their temporal proximity to the examined frame.

ACKNOWLEDGMENTS

The authors wish to extend their gratitude to researhers Spiros Ioannou and Vassilis Tzouvaras for their contribution in the implementation of the FAP extraction and neurofuzzy modules.

REFERENCES

[1] Karpouzis, K., Raouzaiou, A., Drosopoulos, A., Ioannou, S., Balomenos, T., Tsapatsoulis N. and Kollias, S., "Facial expression and gesture analysis for emotionally-rich man-machine interaction", Sarris, N and Strintzis, M. (eds.), 3D Modeling and Animation: Synthesis and Analysis Techniques, Idea Group Publ., 2003.

[2] Chellapa, P., Wilson, C. and Sirohey, S., "Human and Machine Recognition of Faces: A Survey", Proceedings of IEEE, vol. 83, no. 5, pp. 705-740, 1995.

[3] Rowley, H.A., Baluja, S. and Kanade, T., "Neural Network-Based Face Detection", IEEE Trans. on PAMI, vol. 20, no. 1, pp. 23-28, 1998.

[4] Raouzaiou, A., Tsapatsoulis, N. Karpouzis, K. and Kollias, S., "Parameterized facial expression synthesis based on MPEG-4", Eurasip Journal on Applied Signal Processing vol. 2002, no. 10, pp. 1021-1038, 2002.

[5] Tsapatsoulis, N., Raouzaiou, A., Kollias, S., Cowie, R. and Douglas-Cowie, E., "Emotion Recognition and Synthesis based on MPEG-4 FAPs", Pandzic, I. and Forchheimer, R.(eds), MPEG-4 Facial Animation, John Wiley & Sons, UK, 2002.

[6] Whissel, C.M., "The dictionary of affect in language", Plutchnik, R. and Kellerman, H. (eds) "Emotion: Theory, research and experience: vol. 4, The measurement of emotions", Academic Press, New York, 1989

[7] Raouzaiou, A., Tsapatsoulis, N., Tzouvaras, V., Stamou, G. and Kollias, S., "A hybrid intelligence system for facial expression recognition", EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems, 2002.

[8] Dempster, A.P., "Upper and lower probabilities induced by a multivalued mapping", Annals of mathematical statistics vol. 38, pp. 325-339, 1967.

[9] Guan, J.W. and Bell, D.,A., "Evidence theory and its applications" vol. 1, North-Holland, New York, 1991.

[10] Guan, J.W. and Bell, D.,A., "Evidence theory and its applications" vol. 2, North-Holland, New York, 1992.

[11] Klir, G., Yuan, B., "Fuzzy Sets and Fuzzy Logic: Theory and Applications", Prentice Hall, 1995.