

AN INTELLIGENT SCHEME FOR FACIAL EXPRESSION RECOGNITION

A. Raouzaïou, S. Ioannou, K. Karpouzis, N. Tsapatsoulis, S. Kollias¹, R. Cowie²

¹Department of Electrical and Computer Engineering
National Technical University of Athens, Heron Polytechniou 9, 157 73 Zographou, Greece
Phone: +30-10-7722491, Fax: +30-10-7722492
email: {araouz, sivann, kkar pou, ntsap}@image.ntua.gr,
stefanos@cs.ntua.gr

²Department of Psychology
Queen's University of Belfast
Northern Ireland, United Kingdom
email: r.cowie@qub.ac.uk

Abstract: This paper addresses the problem of emotion recognition in faces through an intelligent neuro-fuzzy system, which is capable of analysing facial features extracted following the MPEG-4 standard, associating these features to symbolic fuzzy predicates, and reasoning on the latter, so as to classify facial images according to the underlying emotional states. Results are presented which illustrate the capability of the developed system to analyse and recognise facial expressions in human computer interaction applications.

1. Introduction

There has recently been high interest in affective computing, especially in interfaces which can analyse their users' emotional state and generate synthetic agents that possess convincing expression characteristics. Automatic emotion recognition in faces is a hard problem, requiring a number of pre-processing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, e.g., characteristic points in these regions, or model facial gestures using anatomic information about the face.

Most of the above techniques are based on a well-known system for describing "all visually distinguishable facial movements", called the Facial Action Coding System (FACS) [4], [6]. FACS is an anatomically oriented coding system, based on the definition of "action units" that cause facial movements. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard [7]. In particular, the Facial Definition Parameter (FDP) set and the Facial Animation Parameter (FAP) set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, as well as the animation of faces reproducing expressions, emotions and speech pronunciation. By monitoring facial gestures corresponding to FDP feature points (FP) and/or FAP movements over time, it is possible to derive cues about user's expressions/emotions [1], [3].

Research on facial expression analysis and synthesis has tended to concentrate on primary or archetypal emotions, i.e., sadness, anger, joy, fear, disgust and surprise. Very few studies which explore non-archetypal emotions have appeared in the computer science literature [2], [8]. However, psychological researchers working in different traditions [9] have investigated a broader variety of emotions. An extensive survey on emotion analysis can be found in [10]. According to these studies, emotions can be modelled as points in a space with a relatively small number of dimensions. Two dimensions, *activation* and *evaluation*, are sufficient for a first approximation. *Evaluation* summarises how positive or negative the user feels; *activation* indicates how energetically he or she is disposed to act.

In this work we present a methodology for analysing both primary and intermediate expressions, taking into account the above-mentioned results and particularly the *activation* parameter. This is performed through a neuro-fuzzy system which first translates FP movements into FAPs and reasons on the latter to recognize the underlying emotion in facial video sequences.

2. Modelling Facial Expressions Using FAPs

Two basic issues should be addressed when modelling archetypal expression: (i) estimation of FAPs that are involved in their formation, (ii) definition of the FAP intensities. Table 1 illustrates the description of “joy” and “sadness”, using MPEG-4 FAPs. Descriptions for all archetypal expressions can be found in [1].

Table 1:FAP vocabulary for description of “joy” and “sadness”

Joy	<i>open_jaw</i> (F_3), <i>lower_t_midlip</i> (F_4), <i>raise_b_midlip</i> (F_5), <i>stretch_l_cornerlip</i> (F_6), <i>stretch_r_cornerlip</i> (F_7), <i>raise_l_cornerlip</i> (F_{12}), <i>raise_r_cornerlip</i> (F_{13}), <i>close_t_l_eyelid</i> (F_{19}), <i>close_t_r_eyelid</i> (F_{20}), <i>close_b_l_eyelid</i> (F_{21}), <i>close_b_r_eyelid</i> (F_{22}), <i>raise_l_m_eyebrow</i> (F_{33}), <i>raise_r_m_eyebrow</i> (F_{34}), <i>lift_l_cheek</i> (F_{41}), <i>lift_r_cheek</i> (F_{42}), <i>stretch_l_cornerlip_o</i> (F_{53}), <i>stretch_r_cornerlip_o</i> (F_{54})
Sadness	<i>close_t_l_eyelid</i> (F_{19}), <i>close_t_r_eyelid</i> (F_{20}), <i>close_b_l_eyelid</i> (F_{21}), <i>close_b_r_eyelid</i> (F_{22}), <i>raise_l_i_eyebrow</i> (F_{31}), <i>raise_r_i_eyebrow</i> (F_{32}), <i>raise_l_m_eyebrow</i> (F_{33}), <i>raise_r_m_eyebrow</i> (F_{34}), <i>raise_l_o_eyebrow</i> (F_{35}), <i>raise_r_o_eyebrow</i> (F_{36})

Although FAPs are practical and very useful for animation purposes, they are inadequate for analysing facial expressions from video scenes or still images. In order to measure FAPs in real images and video sequences, it is necessary to define a way of describing them through the movement of points that lie in the facial area and that can be automatically detected. Such a description could gain advantage from the extended research on automatic facial point detection [11].

Quantitative modelling of FAPs can be implemented using the features labelled as f_i ($i=1...15$) in the third column of Table 2 [12]. The feature set employs FDP feature points that lie in the facial area. It consists of distances (noted as $s(x,y)$, where x and y correspond to FDP feature points ranked in terms of their belonging to specific facial areas [14]), some of which are constant during expressions and are used as reference points. It should be noted that not all FAPs can be modelled by distances between

facial protuberant points (e.g. *raise_b_lip_lm_o*, *lower_t_lip_lm_o*). In such cases, the corresponding FAPs are retained in the vocabulary and their ranges of variation are experimentally defined based on facial animations. Moreover, some features serve for the estimation of the range of variation of more than one FAP (e.g. features f_{12} - f_{15}).

Table 2: Quantitative FAP modelling: (1) $s(x,y)$ is the Euclidean distance between FPs x and y , (2) $D_{i-NEUTRAL}$ refers to distance D_i with the face in neutral position

FAP name	Main Feature for description	Utilized Main Feature
squeeze_l_eyebrow (F ₃₇)	$D_1=s(4.6,3.8)$	$f_1= D_{1-NEUTRAL} - D_1$
squeeze_r_eyebrow (F ₃₈)	$D_2=s(4.5,3.11)$	$f_2= D_{2-NEUTRAL} - D_2$
lower_t_midlip (F ₄)	$D_3=s(9.3,8.1)$	$f_3= D_3 - D_{3-NEUTRAL}$
raise_b_midlip (F ₅)	$D_4=s(9.3,8.2)$	$f_4= D_{4-NEUTRAL} - D_4$
raise_l_i_eyebrow (F ₃₁)	$D_5=s(4.2,3.8)$	$f_5= D_5 - D_{5-NEUTRAL}$
raise_r_i_eyebrow (F ₃₂)	$D_6=s(4.1,3.11)$	$f_6= D_6 - D_{6-NEUTRAL}$
raise_l_o_eyebrow (F ₃₅)	$D_7=s(4.6,3.12)$	$f_7= D_7 - D_{7-NEUTRAL}$
raise_r_o_eyebrow (F ₃₆)	$D_8=s(4.5,3.7)$	$f_8= D_8 - D_{8-NEUTRAL}$
raise_l_m_eyebrow (F ₃₃)	$D_9=s(4.4,3.12)$	$f_9= D_9 - D_{9-NEUTRAL}$
raise_r_m_eyebrow (F ₃₄)	$D_{10}=s(4.3,3.7)$	$f_{10}= D_{10} - D_{10-NEUTRAL}$
open_jaw (F ₃)	$D_{11}=s(8.1,8.2)$	$f_{11}= D_{11} - D_{11-NEUTRAL}$
close_t_l_eyelid (F ₁₉) –close_b_l_eyelid (F ₂₁)	$D_{12}=s(3.2,3.4)$	$f_{12}= D_{12} - D_{12-NEUTRAL}$
close_t_r_eyelid (F ₂₀) –close_b_r_eyelid (F ₂₂)	$D_{13}=s(3.1,3.3)$	$f_{13}= D_{13} - D_{13-NEUTRAL}$
stretch_l_cornerlip (F ₆) (stretch_l_cornerlip_o)(F ₅₃) – stretch_r_cornerlip (F ₇) (stretch_r_cornerlip_o) (F ₅₄)	$D_{14}=s(8.4,8.3)$	$f_{14}= D_{14} - D_{14-NEUTRAL}$
squeeze_l_eyebrow (F ₃₇) AND squeeze_r_eyebrow (F ₃₈)	$D_{15}=s(4.6,4.5)$	$f_{15}= D_{15-NEUTRAL} - D_{15}$

3. The Facial Expression Recognition System

In general, six general categories are used, each one characterized by an archetypal emotion. Within each category, intermediate expressions are described by different emotional and optical intensities, as well as minor variations in expression details.

A hybrid intelligent emotion recognition system is presented next, consisting of a connectionist (subsymbolic) association part and a symbolic processing part as shown in Figure 1. In this modular architecture the *Connectionist Association Module* (CAM) provides the system with the ability to ground the symbolic predicates (associating them with the input features), while the *Adaptive Resource Allocating Neuro Fuzzy Inference System* (ARANFIS) [15] implements the semantic reasoning process.

The system takes as input a feature vector \underline{f} that corresponds to the features f_i shown in the third column of Table 2. The particular values of \underline{f} are associated to the symbolic predicates – i.e., FAP values shown in the first column of the same table – through the CAM subsystem. The CAM's outputs form the input vector \underline{G} to the fuzzy inference subsystem, with the elements of \underline{G} expressing the observed value of a corresponding FAP. The CAM consists of a neural network that dynamically forms the above association, providing the emotion analysis system with the capability to adapt to peculiarities of the specific user. In the training phase, the CAM learns to analyse the feature space and provide estimates of the FAP intensities (e.g. low, high, medium). This step requires: (a) Using an appropriate set of training inputs f , (b) Collecting a representative set T_f of pairs (f, s) to be used for network training, and (c) Estimating a parameter set \mathbf{W}_f , which maps the input space F to the symbolic predicate space S .

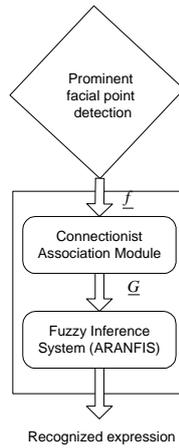


Fig. 1: The emotion analysis system

ARANFIS evaluates the symbolic predicates provided by the CAM subsystem and performs the conceptual reasoning process that finally results to the degree at which the output situations – expressions- are recognised. ARANFIS [15] is a variation of the SuPFuNIS system [5] that enables structured learning. ARANFIS embeds fuzzy rules of the form “If s_1 is LOW and s_2 is HIGH then y is [expression - e.g. anger], where LOW, and HIGH are fuzzy sets defined, respectively, on input universes of discourse (UODs) and the output is a fuzzified expression.

Input nodes represent the domain variables-predicates and output nodes represent the target variables or classes. Each hidden node represents a rule, and input-hidden node connections represent fuzzy rules antecedents. Each hidden-output node connection represents a fuzzy-rule consequent. Fuzzy sets corresponding to linguistic labels of fuzzy *if-then* rules (such as LOW and HIGH) are defined on input and output UODs and are represented by symmetric Gaussian membership functions specified by

a center and spread. Fuzzy weights w_{ij} from input nodes i to rule nodes j are thus modeled by the center w_{ij}^c and spread w_{ij}^s of a Gaussian fuzzy set and denoted by $w_{ij} = (w_{ij}^c, w_{ij}^s)$. In a similar fashion, consequent fuzzy weights from rule nodes j to output nodes k are denoted by $v_{jk} = (v_{jk}^c, v_{jk}^s)$. The spread of the i -th fuzzified input element is denoted as s_i^s while s_i^c is obtained as the crisp value of the i -th input element. Knowledge in the form of *if-then* rules can be either derived through clustering of input data or be embedded directly as a-priori knowledge.

It should be noted that in the previously described emotion analysis system, no hypothesis has been made about the type of recognizable emotions, that can be either archetypal or non-archetypal ones.

4. Application Study

Let us examine the situation where a PC camera captures its user's image. In the pre-processing stage, skin color segmentation is performed and the face is extracted. A snake is then used to smooth the face mask computed at the segmentation subsystem output. Then, the facial points are extracted, and point distances are calculated. Assuming that the above procedure is first performed for the user's neutral image, storing the corresponding facial points, the differences between them and the FPs of the current facial image of the user are estimated.

An emotion analysis system is created in [13]. In the system interface shown in Figure 3, one can observe an example of the calculated FP distances, the rules activated by the neurofuzzy system and the recognised emotion ('surprise').



Fig. 3: System Interface

To train the CAM system, we used the PHYSTA database in [2] as training set and the EKMAN database [4], [10] as evaluation test. The coordinates of the points have been marked by hand for 300 images in the training set and 110 images in the test set. The CAM consisted of 17 neural networks, each of which associated less than 10 FP input distances (from the list of 23 distances defined as in Table 1 and mentioned in

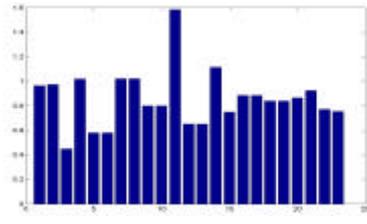
Table 3) to the states (high, medium, low, very low) of a corresponding FAP, and was trained using a variant of backpropagation learning algorithm [16]. Moreover, 41 rules were appropriately defined, half of them taken from the associated literature and half of them derived through training [14], and inserted in the ARAFNIS subsystem.

Let us provide an example of the performance of the overall system. Input is *image001* of the Ekman database, showing a happy expression. Figure 4a shows the feature vector provided at the input of the CAM subsystem; Figure 4b shows the output of the CAM, which is translated as: *open_jaw*-> Medium, *lower_t_midlip*-> Low, *raise_b_midlip*->VeryLow, ..., *raise_r_cornerlip_o*->High. Figure 4c presents the activation level of each of the 41 rules inserted in ARANFIS, while Figure 4d the final system output, in terms of the archetypal emotional states.

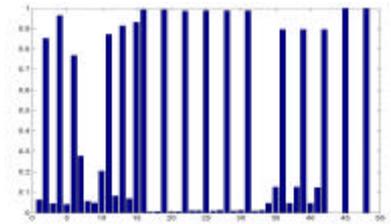
Table 4 illustrates the confusion matrix of the mean degree of beliefs (*not the classification rates*), for each of the archetypal emotions *anger*, *joy*, *disgust*, *surprise* and the *neutral* condition, computed over the EKMAN dataset, which verifies the good system performance, while Table 5 shows the more often activated rule for each of the above expressions.

Table3: Training the CAM module

<i>FAP name</i>	<i>Primary distance</i>	<i>Other distances</i>	<i>States (VL-VeryLow, L-Low, M-Medium, H-High)</i>
Squeeze_l_eyebrow (F ₃₇)	d ₂	d ₆ , d ₈ , d ₁₀ , d ₁₇ , d ₁₉ , d ₁₅	L, M, H
Squeeze_r_eyebrow (F ₃₈)	d ₁	d ₅ , d ₇ , d ₉ , d ₁₆ , d ₁₈ , d ₁₅	L, M, H
Lower_t_midlip (F ₄)	d ₃	d ₁₁ , d ₂₀ , d ₂₁	L, M
Raise_b_midlip (F ₅)	d ₄	d ₁₁ , d ₂₀ , d ₂₁	VL, L, H
Raise_l_l_eyebrow (F ₃₁)	d ₆	d ₂ , d ₈ , d ₁₀ , d ₁₇ , d ₁₉ , d ₁₅	L, M, H
Raise_r_l_eyebrow (F ₃₂)	d ₅	d ₁ , d ₇ , d ₉ , d ₁₆ , d ₁₈ , d ₁₅	L, M, H
Raise_l_o_eyebrow (F ₃₅)	d ₈	d ₂ , d ₆ , d ₁₀ , d ₁₇ , d ₁₉ , d ₁₅	L, M, H
Raise_r_o_eyebrow (F ₃₆)	d ₇	d ₁ , d ₅ , d ₉ , d ₁₆ , d ₁₈ , d ₁₅	L, M, H
Raise_l_m_eyebrow (F ₃₃)	d ₁₀	d ₂ , d ₆ , d ₈ , d ₁₇ , d ₁₉ , d ₁₅	L, M, H
Raise_r_m_eyebrow (F ₃₄)	d ₉	d ₁ , d ₅ , d ₇ , d ₁₆ , d ₁₈ , d ₁₅	L, M, H
Open_jaw (F ₃)	d ₁₁	d ₄	L, M, H
close_left_eye (F ₁₉ , F ₂₁)	d ₁₃	-	L, H
close_right_eye (F ₂₀ , F ₂₂)	d ₁₂	-	L, H
Wrinkles_between_eyebrows (F ₃₇ , F ₃₈)	d ₁₅	d ₁ , d ₂ , d ₅ , d ₆ , d ₇ , d ₈ , d ₉ , d ₁₆ , d ₁₇ , d ₁₈ , d ₁₉	L, M, H
Raise_l_cornerlip_o (F ₅₃)	d ₂₃	d ₃ , d ₄ , d ₁₁ , d ₂₀ , d ₂₁ , d ₂₂	L, M, H
Raise_r_cornerlip_o (F ₅₄)	d ₂₂	d ₃ , d ₄ , d ₁₁ , d ₂₀ , d ₂₁ , d ₂₃	L, M, H
widening_mouth (F ₆ , F ₇)	d ₁₁	d ₃ , d ₄ , d ₁₄	L, M, H



(a)



(b)

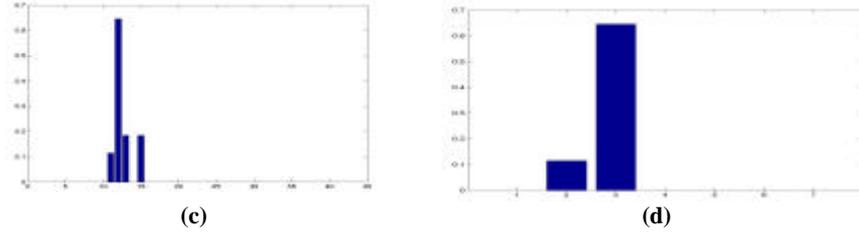


Fig. 4: (a) Example of a feature vector feeding the CAM, (b) An instance of CAM's output, (c) Activation of each of the 41 rules, (d) The final system output

Table 4: Results in images of different expressions

	Anger	Joy	Disgust	Surprise	Neutral
Anger	0.611	0.01	0.068	0	0
Joy	0.006	0.757	0.009	0	0.024
Disgust	0.061	0.007	0.635	0	0
Surprise	0	0.004	0	0.605	0.001
Neutral	0	0.123	0	0	0.83

Table 5: Activated rules

Expressions	Rule more often activated (% of examined photos)
Anger	[<i>open_jaw_low, lower_top_midlip_medium, raise_bottom_midlip_high, raise_left_inner_eyebrow_low, raise_right_inner_eyebrow_low, raise_left_medium_eyebrow_low, raise_right_medium_eyebrow_low, squeeze_left_eyebrow_high, squeeze_right_eyebrow_high, wrinkles_between_eyebrows_high, raise_left_outer_cornerlip_medium, raise_right_outer_cornerlip_medium</i>] (47%)
Joy	[<i>open_jaw_high, lower_top_midlip_low, raise_bottom_midlip_verylow, widening_mouth_high, close_left_eye_high, close_right_eye_high</i>] (39%)
Disgust	[<i>open_jaw_low, lower_top_midlip_low, raise_bottom_midlip_high, widening_mouth_low, close_left_eye_high, close_right_eye_high, raise_left_inner_eyebrow_medium, raise_right_inner_eyebrow_medium, raise_left_medium_eyebrow_medium, raise_right_medium_eyebrow_medium, wrinkles_between_eyebrows_medium</i>] (33%)
Surprise	[<i>open_jaw_high, raise_bottom_midlip_verylow, widening_mouth_low, close_left_eye_low, close_right_eye_low, raise_left_inner_eyebrow_high, raise_right_inner_eyebrow_high, raise_left_medium_eyebrow_high, raise_right_medium_eyebrow_high, raise_left_outer_eyebrow_high, raise_right_outer_eyebrow_high, squeeze_left_eyebrow_low, squeeze_right_eyebrow_low, wrinkles_between_eyebrows_low</i>] (71%)
Neutral	[<i>open_jaw_low, lower_top_midlip_medium, raise_left_inner_eyebrow_medium, raise_right_inner_eyebrow_medium, raise_left_medium_eyebrow_medium, raise_right_medium_eyebrow_medium, raise_left_outer_eyebrow_medium, raise_right_outer_eyebrow_medium, squeeze_left_eyebrow_medium, squeeze_right_eyebrow_medium, wrinkles_between_eyebrows_medium, raise_left_outer_cornerlip_medium, raise_right_outer_cornerlip_medium</i>] (70%)

5. Conclusions

Facial expression recognition has been investigated in this paper, based on neuro-fuzzy analysis of facial features extracted from a user's image following the MPEG-4 standard. A hybrid intelligent system has been described that performs extraction of fuzzy predicates and inference, providing an estimate of the user's emotional state. Work is currently being done, extending and validating the above developments in the framework of the IST ERMIS project [13].

References:

1. N. Tsapatsoulis, A. Raouzaïou, S. Kollias, R. Cowie and E. Douglas-Cowie, "Emotion Recognition and Synthesis based on MPEG-4 FAPs," in *MPEG-4 Facial Animation*, Igor Pandzic, R. Forchheimer (eds), John Wiley & Sons, UK, 2002.
2. EC TMR Project "PHYSTA: Principled Hybrid Systems: Theory and Applications," <http://www.image.ece.ntua.gr/physta>.
3. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, "Emotion Recognition in Human-Computer Interaction", *IEEE Signal Processing Magazine*, 18 (1), p. 32-80, January 2001.
4. P. Ekman and W. Friesen, *The Facial Action Coding System*, Consulting Psychologists Press, San Francisco, CA, 1978 (<http://www.paulekman.com>)
5. S. Paul and S. Kumar, "Subsethood-Product Fuzzy Neural Inference System (SuPFuNIS)," *IEEE Trans. on Neural Networks*, vol. 13, no 3, pp. 578-599, May 2002.
6. ISO/IEC JTC1/SC29/WG11 N3205, "Multi-users technology (Requirements and Applications)", December 1999, Maui
7. A.M. Tekalp, J. Ostermann, "Face and 2-D mesh animation in MPEG-4", *Signal Processing: Image Communication*, Vol. 15, No. 4-5 (Tutorial Issue on the MPEG-4 Standard), pp. 387-421, January 2000.
8. EC TMR Project PHYSTA Report, "Development of Feature Representation from Facial Signals and Speech," January 1999.
9. P. Ekman, "Facial expression and Emotion," *Am. Psychologist*, vol. 48 pp.384-392, 1993
10. P. Ekman and W.V. Friesen, "Pictures of Facial Affect", Palo Alto, CA: Consulting Psychologists Press, 1978.
11. P. Chellapa, C. Wilson and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proceedings of IEEE*, vol.83, no. 5, pp. 705-740, 1995.
12. K.Karpouzis, N. Tsapatsoulis and S. Kollias, "Moving to Continuous Facial Expression Space using the MPEG-4 Facial Definition Parameter (FDP) Set," in *Proc. of the Electronic Imaging 2000 Conference of SPIE*, San Jose, CA, USA, January 2000.
13. IST Project: Emotionally Rich Man-Machine Interaction Systems (ERMIS), 2001-2003
14. A. Raouzaïou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4", *EURASIP Journal on Applied Signal Processing*, Vol. 2002, No. 10, pp. 1021-1038, Hindawi Publishing Corporation, October 2002.
15. M Pertselakis, N. Tsapatsoulis, S. Kollias and A. Stafylopatis, "An Adaptive Resource Allocating Neural Fuzzy Inference System," submitted for possible presentation to the "IEEE Intelligent Systems Application to Power Systems" (ISAP'03), Lemnos, Greece, 2003.
16. S. Haykin, "Neural Networks: a Comprehensive Foundation", Macmillan College Publishing Company, Inc., New York, 1994