EXPLORING THE TIME COURSE OF FACIAL EXPRESSIONS WITH A FUZZY SYSTEM

F. Piat and N. Tsapatsoulis
Dept of Electrical and Computer Engineering,
National Technical University of Athens, Zographou 15773, Greece
piat@image.ntua.gr , ntsap@image.ntua.gr

ABSTRACT

Recognizing facial expressions is one of the important challenges of current research in Human-Computer Interaction (HCI). Previous research show the limits of recognition based on a single static image, and analyzing video sequences seems more promising. We explore here three fuzzy systems for the classification of basic facial expressions and compare their performances with a template-correlation approach. We then use those to examine the time course dynamics of facial expressions. The system's inputs are the relative variations of distances defined by salient facial points from one frame to the next. For maximum compatibility, those facial points (eyebrows, eyes, mouth) were chosen from the set of points defined in the standard MPEG-4 specifications, and so that their automatic extraction is tractable. The first results suggest that some expressions can be recognized early after the onset. For other expressions, it is in general possible to reduce significantly the number of possibilities. Forming early hypotheses regarding the expression could be necessary for a system to work in real-time, since other steps may have to follow: prediction of user's action, choice of computer's action, etc... This also has implications for the recognition of milder expressions.

1. INTRODUCTION

One of the most important features of multimedia systems is the interactivity they offer to the user. We increasingly want systems able to adapt themselves to the user and to customize their behaviour accordingly. Computers may start to be truly intelligent when they understand the user's goals and expectations. A non-intrusive way to get some of this information is for the computer to decode the user's face expression, as this in general continuously provides a lot of information relating to his satisfaction. This information can be used by a system to get feed-back regarding the result or appropriateness of the last action, or to anticipate the user's next action, for instance. Some research [1] suggests that the meaning of an emotion may reside in the predispositions for possible actions it entails.

One of the numerous problems to recognize facial expressions is that they are ambiguous and we often need additional information in order to interpret them, like the speech accompanying the expression, or some contextual information. In particular, knowing the temporal position of a frame relative to the whole sequence is absolutely necessary. Without this information, we cannot distinguish a mild expression at its maximum of intensity (called apex) from the onset of a strong one: we would confuse a light smile with the beginning of a hearty laughter. This problem imposes severe restrictions on the possibility to recognize static expressions. Therefore it is necessary to explore the dynamics of the expressions, the way facial changes unfold in time over the course of the whole sequence. The speed of some movements and their relative temporal order may be a valuable information to improve recognition[2][3], even for off-line systems. Further, this should give us clues as to how early before the apex can an expression be recognized, and how to achieve early recognition, which is important for on-line systems. Recognizing expressions as early as possible could be crucial for many real-time applications, since recognition will most likely not be the ultimate goal of the application, but should be followed by some action, which choice and execution also require computational time. Even when early recognition cannot be achieved, it is likely that restricting the number of possible responses to just a few will be helpful, as this will play a role similar to 'orienting the attentional focus' of the system: fewer hypotheses can be more thoroughly tested for improved accuracy. Finally, understanding the development of expressions over time can have implications for the recognition of mild expressions, as mentioned above.

In the next sections we experiment with three fuzzy systems built to recognize expressions from salient facial points, and compare their performances to that of a simple template approach. For tractability and database availability reasons, we limit the recognition to the 5 basic expressions represented in the MIT facial expressions database: Anger, Smile, Disgust, Raise brows and Surprise. 'Raise brows' is not a 'universal' expression in itself but it is interesting to include it as it is a component of several other basic expressions (e.g., fear and surprise), and also used in isolation to punctuate speech. The database contains 20 sequences of different lengths, adding up to 212 frames. Section 3 details the results of the analysis of whole sequences with the fuzzy systems.

2. EMOTION RECOGNITION SYSTEMS

It is important for the generality of such a system to be compatible with the standards defined for video sequences, so we adopt the framework outlined in [4]. It defines sets of parameters that allow the description and animation of human bodies and faces for the MPEG-4 standard. Over 50 feature facial points (Facial Definition Parameters or FDPs) are used to define a given face, and as many basic actions (Facial Action Parameters or FAPs) that a face can perform are used to describe its movements and lead to the rendering of any expression. The FAPs express relations between FDPs, for instance the distance between two feature points. They are normalized according to some distances

independent of the expression, in order to get consistent values regardless of the scale of the picture, distance from the camera, etc… Specifically, those FAP units are the distance between the eyes (ESo) for horizontal distances and the distance from the middle of the eyes to the tip of the nose (ENSo) for vertical distances. Each FAP is divided by Eso or ENSo depending on whether it is a horizontal or vertical distance.

2.1 Inputs

In order to limit the amount of data to be dealt with and computation time, we choose to focus on the most expressive parts of the face: eyes, eyebrows and mouth. This should still contain enough information for categorization, and those are salient points that can be detected automatically but at this stage the localization of the points is still manual. From these 19 facial points, we compute 14 of the FAPs defined by their distances. This is shown in Figure 1, along with the distances used to normalize the face in order to have scale invariance. The FAP values depend on the expression of the face, but also on the particular structural configuration of a given face. To feed the classification systems information that relates only to the facial movements, the relative FAP changes from a neutral reference frame are computed to constitute the input for each frame:

$$I_t^i = \left(FAP_t^i - FAP_0^i\right) \Big/ FAP_0^i$$

Assuming symmetry of the expression, we can reduce these 14 inputs to 8 only, by keeping the most significant input (highest absolute value) of a symmetric pair. This allows the greatest movement to be taken into account. The symmetry assumption holds for the basic expressions we are dealing with for now, but it will have to revised to process more complex, asymmetric expressions.

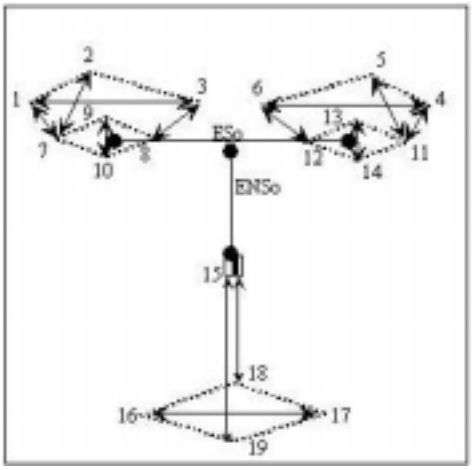

Figure 1. The 19 FDP points and the FAP distances they define, which time derivatives constitute the input to the fuzzy system. Eso and ENSo are the distances used for normalization.

Future improvements of the system will also include the addition of new FAPs and the automatic detection of the feature points. After inspection of emotional sequences, it became apparent that some attitudes (e.g., position of the head) are very revealing of the emotional state but not captured by any of these 14 FAPs, so 6 others were added, some being combinations of MPEG FAPs, some totally new. To locate the points, we first need to extract the eyes, eyebrows and mouth. Two methods are currently being explored. The first one uses a sequence of contour detection, erosion and dilatation, then uses constraints of symmetry and location within the face to select the appropriate features from the 'blobs' obtained. The second method seems very promising and less computationally intensive: it is based on the fact that the facial features induce a lot of horizontal contrast, and one vertical scan of the image line by line generally suffices to find their vertical position.

2.2 Template Approach

The input values of the frames are averaged for each category, to obtain five 8-dimensional template vectors. Then, for each frame, a correlation coefficient is computed with each template, giving a degree of belief that the frame belongs to each category. The frame is classified in the category which template correlates the most with its inputs. The classification rate for each expression is detailed in Table 1. This method classifies correctly 70% of the frames, but this drops to 60% when tested on generalization with the leave-one-(sequence-)out method.

2.3 Fuzzy Inference Systems

The continuity of the emotion space as well as the uncertainty involved in the feature estimation process, whether automatic or manual, make the use of fuzzy logic appropriate for the feature-to-emotion mapping. The input is the same as above, an 8-tuple for each frame, which components describe the increment (or decrement) of the corresponding FAP. The system is in fact made up of 5 subsystems, one for each category. Each subsystem outputs a value reflecting the degree of belief that the frame belongs to the corresponding category. Each subsystem has 8 Input Membership Functions (MF), which define a fuzzy linguistic partition on each input: it qualifies the input as being 'Low' or 'High' with a certain degree of confidence. The linguistic terms of the fuzzy partitions (for example medium open_jaw) are then connected with the aid of the IF-THEN rules of the Rule Base. The activation of the antecedents of a rule causes the activation of the consequences, i.e. the degree of belief that the emotion is X concluded from the degree of the increment (or decrement) of the FAPs after the stages of fuzzification and fuzzy inference. This is done for all 5 expressions and the expression with the highest degree of belief is considered the winner.

Based on this structure, we can build many different systems by choosing different MFs or different rules; we experiment here with three of them. The first one uses trapezoidal MFs, which in fact behave like Boolean gate functions. This makes the fuzzy system a special case where it degenerates into a Boolean system. The value associated to an input is maximal when it is within the min-Max limits of this input over all the frames of a particular expressions, otherwise 0. The rule giving a high degree of belief to the expression is a conjunction of all the MFs: a frame is given the highest degree of belief if all its inputs are within the acceptable range for that expression. By construction, this system accepts all the frames in their correct category. Only, when a frame is compatible with several expressions, the same high degree of belief is given to these, and we have many

ambiguous predictions, with 2, 3 or even 4 categories at the same time regarded as being exactly as possible. The row 'Ties' in Table 1 shows the proportion of the frames that could not yield a unique and clear prediction.
The second fuzzy system uses the full range of fuzzy values through the use of triangular MFs, bounded by the same min and Max values as above but peaking only for an input value equal to the template (average) value of this input. The rules are the same as above, and classify 77% of the frames correctly.
The third fuzzy system was designed to test whether we can predict expressions based on only one feature, the one that distinguishes best between the given expression and all the others (e.g., width of the mouth for 'Smile'). The MF used are also triangular: using trapezoidal MFs on only one input would give too many possible responses but using the whole range of fuzzy values disambiguates the predictions. Only 48% of the frames are classified correctly, which suggests that considering only the most significant input of each expression does not provide enough information to classify the frames accurately.

3. RESULTS

3.1 Frame Classification

Table 1 shows the performance of the different systems for each expression. The classifications of the systems were recorded for all frames after the first (neutral) one for each of the 20 video sequences, totaling 192. The number of frames varied with each sequence. Besides the performance of the one-feature-only-based fuzzy system, the average classification rate was acceptable, around 70-75% for the template approach and properly fuzzy system. Most of the errors come from confusions between anger and disgust. Even human judgment is ambiguous on some of these sequences. On the other hand, the 'Boolean' fuzzy system achieves 100% classification by construction, but gives many ambiguous predictions, unlike the other systems. This can be seen in Figure 2 by comparing the graphs in the right column. However this is enough to classify the sequences with perfect accuracy, as explained below. Further, we see from these graphs that the number of correct unique classifications rises steadily as the sequence unfolds (as the expression becomes more pronounced, getting close to the apex), whereas the number of ambiguous classifications generally decreases.

3.2 From Frame to Sequence classification

Once all the frames of a sequence have been classified, we need one resulting global prediction for the whole sequence. Three methods have been explored here: 1) Summing the degrees of belief for each expression over the whole sequence, and choosing the highest total. 2)Considering only the winning category for each frame, and choosing the category with the most wins over the sequence. 3) Replacing the degrees of belief by their rank in decreasing order, and choosing the category that minimizes the rank summed over the sequence.

The results obtained by these three methods are in general very consistent, they all classify the same number of sequences, plus or minus one. The last column of Table 1 gives the range of performance obtained with these 3 methods. Only the 'Boolean' fuzzy system reaches perfect classification, with only one disgust sequence classified ambiguously as disgust or anger. The graphs on the left of Figure 2 show for each expression when the sequence is correctly classified, after discretising the sequences into 4 temporal quarters. We can see that the properly fuzzy system needs only half the sequence to classify, whereas the other approaches need up to 3/4 of the sequence. The template method identified surprise and raise brows early in the sequence, whereas those are the latest to be recognized by the Boolean/fuzzy system. This points to the complementarity of those 2 approaches: the former depends on positive clues for classification, the latter works more by elimination until all candidate expressions are reduced to only one. Thus the optimal method suggested by those results is that we should use the fuzzy system with triangular MF for early prediction, then confirm that choice with the trapezoidal MF fuzzy system.

4. REFERENCES

[1] Review of existing techniques for human emotion understanding and their applications in human-computer interaction, Technical Report, Research contract FMRX-CT97-0098 (DG12-BDNC), October 1998. http://www.image.ntua.gr/physta/

[2] Yaccob, Y. and Davis, L.S. Recognizing human facial expressions. *The Second Workshop on Visual Form*, Capri, 1994, pages 584-593.

[3] Yaccob, Y. and Davis, L.S. Computing spatio-temporal representations of human faces. Proc. of the Computer Vision and Pattern Recognition Conference, 1994, pages 70-75.

[4] ISO/IEC JTC1/SC29/WG11 MPEG96, "MPEG4 SNHC: Face and body definition and animation parameters ", 1996. http://drogo.cselt.stet.it/mpeg/chicago/animation.htm

	Anger (40 frames)	Smile (60 frames)	Disgust (26 frames)	Raise Brows (18 frames)	Surprise (48 frames)	Range of Sequence correct
Template	.55	.75	.62	.89	.73	75 – 85 %
Fuzzy (one_input)	.48	.62	.50	.28	.35	40 – 70 %
Fuzzy (MF=tri)	.85	.88	.73	.22	.81	70 – 90%
Fuzzy (MF=trap)	1.00	1.00	1.00	1.00	1.00	95 – 100 %
Ties	0.40	0.40	0.50	0.67	0.17	

Table 1: Compared performance of the template approach and the 3 fuzzy systems, for frames and sequence classifications.

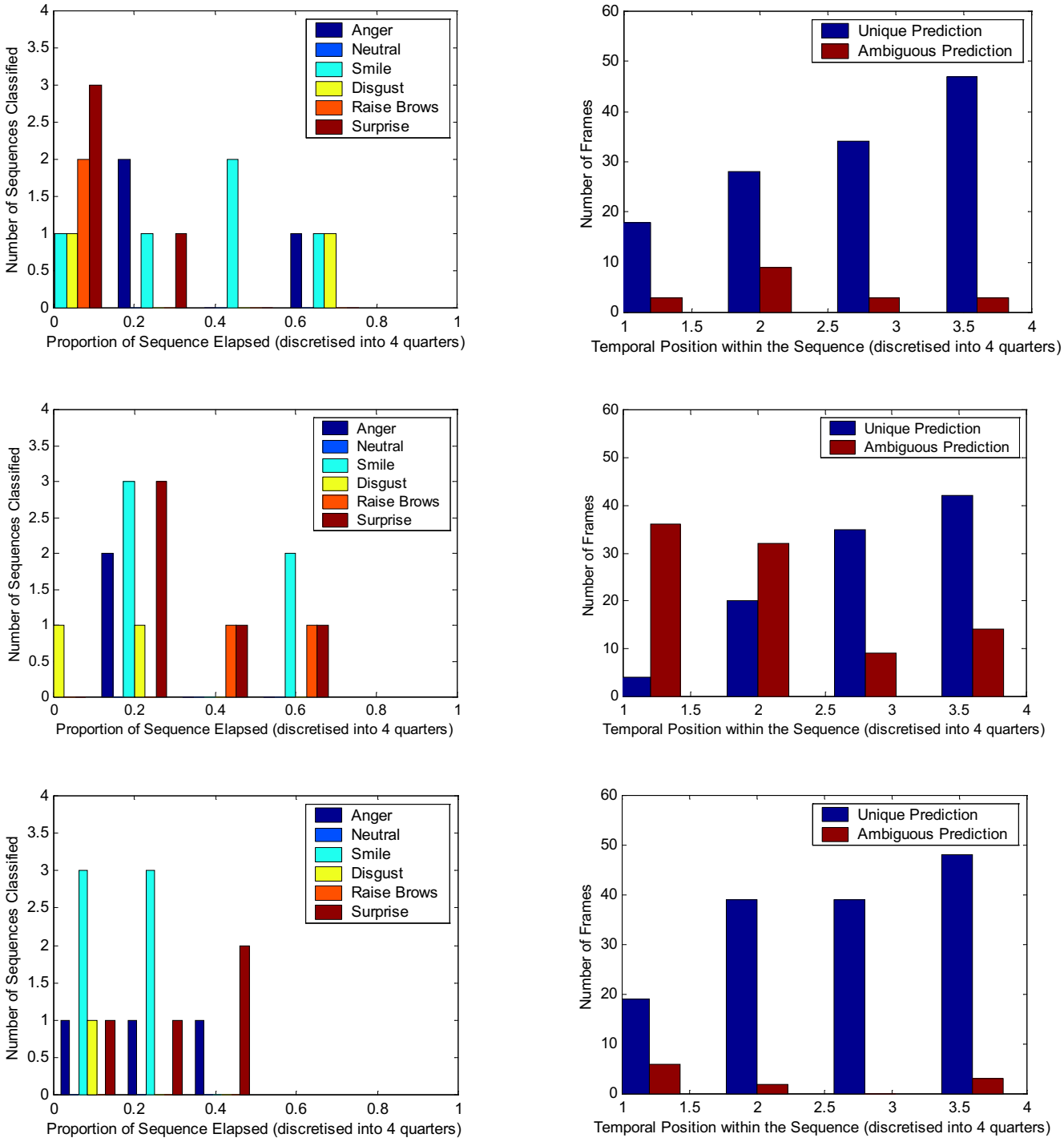

Figure 2. The correct classification of sequences (left) and of frames (right) as a function of the time elapsed relative the sequence length, for the template approach (top), the fuzzy system with trapezoidal MFs (middle) and the fuzzy system with triangular MFs (bottom).