EMOTION RECOGNITION BASED ON MPEG4 FACIAL ANIMATION PARAMETERS

Montse Pardàs, Antonio Bonafonte, José Luis Landabaso

Universitat Politècnica de Catalunya, Barcelona, Spain

ABSTRACT

In this paper a facial expression recognition system is presented. The system is based on the modelling of the expressions by means of Hidden Markov Models. The observations used to create the models are the MPEG-4 standardized Facial Animation Parameters (FAPs). The FAPs of a video sequence are first extracted and then analyzed using semi-continuous HMM. The system shows good performance for distinguishing isolated expressions and can also be used, with lower accuracy, to extract the expressions in long video sequences where speech is mixed with silence frames.

1. INTRODUCTION

The critical role that emotions play in rational decision-making, in perception and in human interaction has opened an interest in introducing the ability to recognize and reproduce emotions into the computers. In [13] many applications which could benefit from this ability are explored. The importance of introducing non verbal communication in automatic dialogue systems is highlighted in [2].

Psychological studies have indicated that at least six emotions are universally associated with distinct facial expressions: happiness, sadness, surprise, fear, anger and disgust [15].

Some research has been conducted on pictures that capture the subject's expression at its peak. For instance, in [9] and [3] classifiers for facial expressions were based on neural networks. Research has also been conducted on the extraction of facial expressions from video sequences. Most works in this area develop a video database from subjects making expressions on demand. Their aim has been to classify the six basic expressions that we have mentioned above, and they have not been tested in a dialogue environment.

Most approaches in this area rely on the Ekman and Friesen Facial Action Coding System [4]. The FACS is based on the enumeration of all Action Units of a face that cause facial movements. The combination of these actions units results in a large set of possible facial expressions. Some of the approaches presented in this context can be found in [1] and [14]. Other approaches employ physically-based models of heads including skin and musculature [5].

In this work we have developed an expression recognition technique consistent with the MPEG4 standardised parameters for facial definition and animation, FDP and FAP. Thus, the expression recognition process can be divided in two steps: facial parameter extraction and facial parameter analysis. The facial parameter extraction process is based on a feature point detection and tracking system described in [10] and [11]. These techniques allow to extract the facial features that are converted to MPEG-4 compliant parameters (FAP's and FDP's).

In this paper, we will describe the analysis of the Facial Animation Parameters that allows the expression recognition.

Using the MPEG4 parameters for the analysis of facial emotions has different advantages. In first place, the developed high level analysis can benefit from already existing low level analysis techniques for FDP and FAP extraction, as well as from any advances that are made in this area in the future years. Besides, the FDP and low-level FAP constitute a concise representation of the evolution of the expression of the face.

From the training database, and using the available FAP and FDP, spatio-temporal patterns for expressions are constructed.

Our approach for the interpretation of facial expressions uses Hidden Markov Models (HMM) to recognise different patterns of FAP evolution.

Besides, we will work on sequences that represent a conversation with an agent, in order to have data which reflects a real situation where the system could be applied. The system developed should be able to classify expressions in silence frames as well as detect important clues in speech frames. A first step has been to work on silence frames, but we have also performed experiments that allow separating different emotions in long sequences where silence frames are combined with speech frames.

2. SIX EMOTIONS RECOGNITION CONTEXT

The first approach to expression recognition that we describe takes as input the Low Level FAPs of a video sequence and extracts the predominant expression of this

0-7803-7402-9/02/$17.00 ©2002 IEEE

sequence. This system in based on Hidden Markov Models.

The Cohn-Kanade facial expression database [6] has been selected as basis for doing the training and recognition of the expressions. The whole database has been processed in order to extract a subet of the Low Level FAPs and perform the expression recognition experiments. Different experiments have been carried out to determine the best topology of the HMM to recognize expressions from the Low Level FAPs. The Low Level FAPs that have been extracted correspond to those obtained by tracking of the eyebrows and outer contour of the lips.

HMM are one of the basic probabilistic tools used for time series modelling. They are definitely the most used model in speech recognition, and they are beginning to be used in vision, mainly because they can be learned from data and they implicitly handle time-varying signals by dynamic time warping. They have already been successfully used to classify the mouth shape in video sequences [8] and to combine different information sources (optical flow, point tracking and furrow detection) for expression intensity estimation [8]. We will extend their use creating the feature vectors from the available low-level FAP.

This first approach tries to estimate isolated expressions. That is, we take a sequence which contains the transition from a neutral face to a given expression and we have to decide which is this expression. All the sequences from the Cohn-Kanade facial expression database belong to this type.

3. EXTRACTION OF THE FAPS

The facial parameter extraction process used is described in [11]. It is based on a feature point detection and tracking system. These techniques allow to extract the facial features that are converted to MPEG-4 compliant Facial Animation Parameters (FAPs).

The algorithm is applied to initialize and track the eyebrows and mouth, thus producing the Low Level FAPs for the eyebrows 31 32 33 34 35 36 37 38 and for the mouth 51 52 53 54 55 56 57 58 59 60.

Both the initialization and tracking are based on active contours or snakes. Conventional snake approaches find the position of the snake by finding a minimum of its energy, composed of internal and external forces. The external forces pull the contours toward features such as lines and edges. Our approach, for initialization, introduces higher level information by an statistical characterization of the snaxels that should represent the contour. For tracking, a measurement of the local correlation with the previous image is introduced both in the energy of the active contour and in the search strategy.

Two examples of initialization and tracking results are shown in Figure 1.

Figure 1. Initialization and tracking of contours

4. SELECTION OF THE HMM TOPOLOGY

Markov sources model temporal series assuming a (hidden) temporal structure. HMM [13] can be used to create models of the expressions to be recognized. Once these models are trained we will be able to compute, for a given sequence of observations (the FAPs of the video sequence in our case), the probability that this sequence was produced by each of the models. Thus, this sequence of observations will be assigned to the expression that has a higher probability of generating it.

Each considered emotion (sad, anger, fear, joy, disgust and surprise) reflects a temporal structure: let's say start, middle and end of the emotion. This structure can be modeled using left-to-right HMM. This topology is appropriated for signals whose properties change over time in a successive manner. As the time increases, the observable symbols in each sequence either stay at the same state or increase in a successive manner. In our case, we have defined for each emotion a four-states HMM, and we only allow to skip one state. This topology has been selected after testing different configurations. It is described in the following figure.

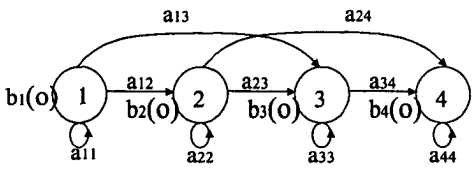

Figure 2. Topology of Hidden Markov Models

5. ESTIMATION OF THE OBSERVATION PROBABILITIES

Each state of each emotion models the observed FAPs using a probability function, which can be either a discrete or a continuous one. In the first case, each set of FAPs for a given image has to be discretized using vector quantification. In the continuous case, a parametric pdf is defined for each state. Typically, multivariate Gaussian or Laplacian pdf's are used. In the case of having sparse data for training, as it is our case, better results can be achieved if all the states of all the emotion models share the Gaussian mixtures (mean and variance). Then, the estimation of the pdf's for each state is reduced to the estimation of the contribution of each mixture to the pdf.

Our training sequences are all those sequences that we select as representatives of the given emotion. In our case, we use all those sequences from the Cohn-Kanade facial expression database manually classified as belonging to this emotion. Each set of FAPs has been divided in two subsets corresponding to eyebrows and mouth, following the MPEG4 classification. For each of these subsets, a set of Gaussian mixtures have been estimated (mean and variance) applying a clustering algorithm on the FAPs of the training database. At each frame, the probability of the whole set of FAPs is computed as the product of the probability of each subset (independence assumption).

Thus, for every defined facial expression (sad, anger, fear, joy, disgust and surprise) a HMM is trained with the extracted FAPs. In the test phase, the HMM system computes the maximum likelihood of the input sequence with respect to all the models and assigns to the sequence the expression with the highest probability.

6. TESTS

The Cohn-Kanade facial expression database [6] has been used for training the models. It consists of the recording from 90 subjects, each one with several basic expressions. The recording was done with a standard VHS video camera, with a video rate of 30 frames per second, with constant illumination and only full-face frontal views were captured. Although the subjects were not previously trained in displaying facial expressions, they practiced the expressions with an expert prior to video recording. Each posed expression begins from neutral and ends at peak expression. The number of sequences for each expression is the following:

	F	Sa	Su	J	A	D
# Seq	33	52	60	61	33	37

Table 1. Number of available sequences for each expression (F: Fear, Sa: Sad, Su: Surprise, J: Joy, A: Anger, D: Disgust).

The FAP's of these sequences were first extracted using our FAP extraction tool [11].

The tests have been performed by first training the six models with all the subjects except one, and then testing the recognition with this subject which had not participated in the training. This process has been repeated for all the subjects, obtaining the recognition results shown in Table 2. The overall recognition rate obtained is 84%.

	F	Sa	Su	J	A	D	%Cor
F	26	3	0	3	0	1	78,7%
Sa	8	31	2	2	6	3	59,6%
Su	0	0	60	0	0	0	100%
J	4	0	0	56	0	0	93,3%
A	0	2	0	1	24	6	72,7%
D	0	0	0	0	1	36	97,3%

Table 2. Number of sequences correctly detected and nimber of confusions.

7. INTRODUCTION OF THE "TALK" MODEL

An important objective of our work is to use sequences that represent a conversation with an agent, in order to have data which reflects a real situation where the system could be applied. The system developed should be able to classify expressions in silence frames as well as to detect important clues in speech frames. A first step has been to work on silence frames. However, we also want to separate the motion due to the expressions from the motion due to speech. One possibility that will be exploited in the future is to combine the audio analysis with the video analysis. The other possibility that we have developed up to now consists on the generation of a new model, in addition to those of the six emotions, to classify the speech sequences.

	F	J	Sa	A	Su	D	T	%Cor
F	27	3	2	0	1	1	0	79,4%
Sa	6	2	29	7	3	1	5	54,7%
Su	0	0	0	0	64	0	0	100%
J	4	57	0	0	1	0	0	91,9%
A	0	2	1	22	0	7	2	64,7%
D	1	0	0	2	0	34	0	91,9%
T	0	0	3	4	0	0	21	75,0%

Table 3. Number of sequences correctly detected and number of confusions adding the "talk" (T) sequence.

In order to train this additional model, that we have designated as "talk", we have selected 28 sequences that contain speech, from available data sequences. These sequences have been processed with the FAP extraction tool, in the same way that the emotion models had been previously analysed. Then, the recognition tests were repeated including these sequences. The performance of

the recognition algorithm in this case is shown in table 3. The overall recognition rate decreases to 81%.

8. EXPERIMENTS WITH CONNECTED RECOGNITION

As discussed before, the first experiments performed are oriented to the extraction of an expression from the input sequence. However, in real situations, we will have a continuous video sequence where the person is talking or just looking at the camera and, at a given point, the expression is performed by the person. So, we need to define another recognition experiment that is able to separate, from a video sequence, different parts where different emotions occur.

In this case the emotions can be decoded using the same HMM by means of a one-stage dynamic programming algorithm. This algorithm is widely used in connected and continuous speech recognition. The basic idea is to allow during the decoding a transition from the final state of each HMM (associated to each emotion) to the first state of the other HMMs. This can be interpreted as a large HMM composed from the emotion HMMs. The transitions along this large HMM indicate both the decoded emotions and the frames associated to each emotion. In order to recover the transitions, the algorithm needs to save some backtracking information, as it is usually the case in dynamic programming.

The first experiments that we have performed in this direction are simplified due to the lack of an extensive database where this situation occurs. However, it was possible to test the connected recognition algorithms by concatenating the FAP files extracted from the "emotions" and "talk" sequences. Results are given in next table. The recognition rate for these sequences is 64%.

	F	J	Sa	A	Su	D	T	%Cor
F	25	3	4	0	1	1	0	73,5%
Sa	5	2	25	8	4	2	7	47,1%
Su	2	0	2	6	37	4	9	57,8%
J	2	56	2	0	1	0	0	91,8%
A	0	2	2	16	1	4	8	47,0%
D	2	0	0	3	9	19	3	51,3%
T	1	3	8	6	4	1	56	70,9%

Table 4. Number of sequences correctly detected and number of confusions in the connected sequences experiment.

9. CONCLUSIONS

The system developed shows a high recognition rate when applied to isolated sequences with expressions. This rate decreases when the sequences are connected. Our future work will be devoted to increase the recognition rate by using the speech information.

Our systems uses only the MPEG-4 standardized Facial Animation Parameters (FAPs). The results can also be improved by introducing additional visual information that can be extracted from the video sequence.

10. REFERENCES

[1] M. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, Int. Journal of Computer Vision, 25 (1), 1997, 23-48.

[2]F. Cassell, K.R. Thorisson, "The power of a nod and a glance: Envelope vs. Emotional Feedback in Animated Conversational Agents", Applied Artificial Intelligence 13: 519-538, 1999.

[3] G.W. Cottrell and J. Metcalfe, "EMPATH: Face, emotion, and gender recognition using holons", in Neural Information Processing Systems, vol.3, pp. 564-571, 1991.

[4] P. Ekman and W. Friesen, Facial Action Coding System. Consulting Psychologists Press Inc., 577 College Avenue, Palo Alto, California 94306, 1978.

[5] I. Essa and A. Pentland, Facial Expression Recognition using a Dynamic Model and Motion Energy, Proc. of the Int. Conf. on Computer Vision 1995, Cambridge, MA, May 1995.

[6] Kanade, T., Cohn, J.F., Tian. Y., Comprehensive Database for Facial Expression Analysis, Proceedings of the Fourth IEEE International Conference on Automatic Face and Gestures Recognition, Grenoble, France, 2000.

[7] J. Lien, T.Kanade, J. Cohn, C. Li, Subtly different Facial Expression Recognition And Expression Intensity Estimation, in Proc. Of the IEEE Int. Conference on Computer Vision and Pattern Recognition, pp. 853-859, Santa Barbara, Ca, June 1998.

[8] N. Oliver, A. Pentland, F. Berard, "LAFTER: A Real-time Lips and Face Tracker with Facial Expression Recognition", in Proc. of IEEE Conf. on Computer Vision, Puerto Rico, 1997.

[9] C. Padgett, G. Cottrell, Identifying emotion in static face images, in Proc. Of the 2nd Joint Symp. on Neural Computation, Vol.5, pp.91-101, La Jolla, CA, Uni. of California, San Diego.

[10] M. Pardàs, E. Sayrol, "A new approach to active contours for tracking" in Proceedings of Int. Conf on Image Processing, ICIP 2000, Vancouver, Canada, September 2000.

[11] M. Pardàs, Marcos Losada, Facial Parameters Extraction System based on Active Contours", in Proc. of Int. Conf on Image Processing, ICIP01, Thesaloniki, Greece, October 2001.

[12] R.W. Picard, "Affective Computing", MIT Press, Cambridge 1997.

[13] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proceedings IEEE, pp. 257-284, February 1989.

[14] M. Rosenblum, Y. Yacoob, L.S. Davis, Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture, IEEE Trans. On Neural Networks, 7 (5), 1996, 1121-1138.

[15] A. Young and H. Ellis (eds.), Handbook of Research on Face Processing, Elsevier Science Publishers 1989.