# HMM RECOGNITION OF EXPRESSIONS IN UNRESTRAINED VIDEO INTERVALS

*José Luis Landabaso, Montse Pardàs, Antonio Bonafonte*

Universitat Politècnica de Catalunya, Barcelona, Spain

## ABSTRACT

This paper discusses the application of a facial expression recognition system in unrestrained video intervals. The system is based on the modeling 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 basic recognizer, presented in [12], shows good performance in distinguishing entire expressions, previously marked inside video sequences. This paper describes the adaptation of the technique to deal with unrestrained expression intervals in video sequences, that is, expressions the boundaries of which have not been previously marked inside the scene. To design the new system, we have taken advantage of the symmetry of the expressions to extract a new topology of the HMMs and we have trained the models without requiring the recording and analysis of a new database. We have also used a discarding process with the aim to eliminate expressions not comprised within the models together with the intervals without any expressions at all. The combination of the presented techniques is suitable for temporal block sampling with later expression classification or discarding.

## 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 [14] 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 [17].

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.

* This work has been supported by the European project InterFace and TIC2001-0996 of the Spanish Government and it is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA

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 [16]. Other approaches employ physically-based models of heads including skin and musculature [5].

In [12], an expression recognition technique was presented which was consistent with MPEG-4 standardized parameters for facial definition and animation, FDP and FAP. These parameters constitute a concise representation of the evolution of the face expression and can be extracted using any low level analysis existing tool. In the technique, the recognition process was divided into two steps. The first one consisted of the facial parameters extraction using the techniques discussed in [10] and [11]. Following, a facial parameters analysis using Hidden Markov Models (HMM) was employed. The analysis allowed recognizing entire expression video sequences and could classify, with lower accuracy, concatenation of different expressions.

In this paper, the adaptation of a HMM based technique to work on unrestrained video intervals together with a discarding technique are described. The adaptation is intended to applications focused towards expression recognition in long video scenes by means of eventual sampling and subsequent processing. These kind of techniques are a possible answer to the low accuracy results reached with other techniques based on expression separation followed by a recognizer [12]. In essence, the analysis is based on new HMM topology which allows more freedom to temporal evolutions. The training is accomplished using different intervals of the same expression.

Concerning the discarding process, it is aimed to eliminate sequences not comprised within the models. It is based on a study of the probability of a sequence to belong to any of the existing models.

In section 2, the fundamentals of the HMM based recognition using FAPs and the general framework are explained. The basic

restrained recognition system is presented in 3. In 4 the new enhancements for unrestrained temporal interval recognition are described and, finally, the discarding process and conclusions are presented.

## 2. HMM BASED RECOGNITION USING FAPS

The expression recognition system that we describe takes as input the Low Level FAPs of a video sequence and extracts the predominant expression of this sequence. The systems are based on Hidden Markov Models.

The Cohn-Kanade facial expression database [6] has been used as basis for doing the training and recognition of the expressions. 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.

The whole database has been processed with a FAP extraction tool [11] in order to extract some of the Low Level FAPs and perform the expression recognition experiments. The extracted Low Level FAPs are divided into two subsets corresponding 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 modeling. Markov sources model temporal series assuming a (hidden) temporal structure. HMM 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.

The HMM [15] are defined by a number N of states connected by transitions. The output of these states are the observations, which belong to a set of symbols. The time variation is associated with transitions between states.

Each state of each emotion models the observed FAPs using a probability function. Having sparse data, as it is our case, better results are achieved if all the states of all the emotion models share the same 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. For each of the two FAPs 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 in a given state is computed as the product of the probability of each subset (independence assumption).

Different experiments have been carried out to determine the best parameters of the HMM to recognize expressions from the Low Level FAPs. The chosen HMM topology is different depending on whether we try to recognize entire marked expressions or isolated intervals that may contain any part of a expression. In the two recognition methods there are also different types of training sequences. The first recognition method is explained in section 3 while the new more general method that allows non restricted expression recognition is explained in 4.

## 3. MARKED EXPRESSION RECOGNITION

Each considered emotion (sad, anger, fear, joy, disgust and surprise) reflects a temporal structure: let's say start, middle and peak of expression. In case that we know in advance the sequence will evolve from a neutral to a peak state, we can model it using a 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. As presented in [13], we select the topology experimentally. The experiments show that using HMM with only 2 states the recognition rate is 4% lower. Using more than 3 states increases the complexity of the model without producing any improvement in the recognition results. The selected topology is represented in the following figure.

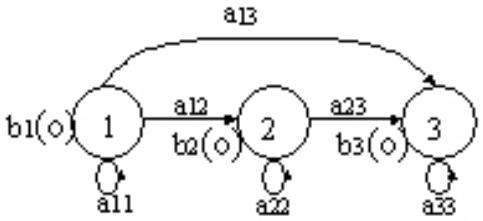


**Figure 1.** Topology of Hidden Markov Models for complete, neutral to peak intensity, expression recognition.

During the training, we use all those sequences from the Cohn-Kanade facial expression database manually classified as belonging to this emotion. 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.

All 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.

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).

Finally, 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 number of confusions.

## 4. UNRESTRAINED EXPRESSION INTERVALS RECOGNITION

The inconvenient of the recognition method presented in Section 3 is that it is assumed that the classifier is applied exactly to the same part of the emotion in the video sequence (from neutral to peak). This can be illustrated with a simple experiment which consists of employing the left-to-right models previously described to recognize expression sequences evolving from peak to neutral intensity instead of neutral to peak. The recognition rate decreases from 84% to 68,73%. This shows how bad the left-to-right topology performs with different temporal intervals of the expressions than the ones used during the training. However, emotions are quite symmetrical, therefore we can enrich the database using a priori knowledge and in this way accept unrestrained intervals during recognition.

These inconveniences and restrictions can be overcome by creating expression models using several sequences belonging to different temporal intervals of the overall expression to model. That is, we can be given a neutral to peak but also a peak to neutral interval or any other combination as shown in figure 2. Regardless of the different period behaviors, the models created must still characterize the overall expression.

The structure of the problem leads us to adopt a topology as in figure 3, where it is not imposed the left-to-right restriction. In addition, the model is not forced to begin at the first state as it can start at any point of the expression sequence and, therefore the initial probability is equally distributed along the states.

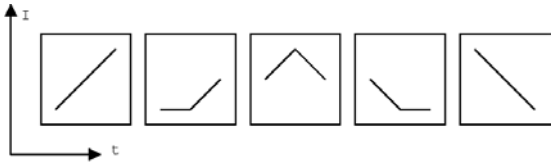


**Figure 2.** 5 temporal behaviours that we can find in unrestricted intervals of the video sequences. The y-axis shows expression intensity evolving in time (x-axis).

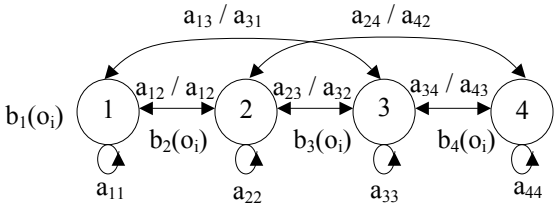


**Figure 3.** Topology of Hidden Markov Models for random temporal intervals of expression recognition.

The training of each expression model has been performed by using the 5 temporal intervals in figure 2 for all the subjects in the database in table 1 belonging to the emotion to model.

During the tests we use as input 5 intervals belonging to an expression and try match them against all the models. As in the experiments shown in the previous section, all 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.

When using the HMM topology for unrestrained expression recognition, the results of the experiment at the beginning of the section raised again to 82.12%.

In table 3 it is shown in detail the new models behavior when recognizing the 6 basic expressions in different temporal intervals. It can be clearly observed that the short recognition rate decay is only a minor drawback to the feature of identifying expressions at any possible interval.

| |  |  |  |  |  |
|---|---|---|---|---|---|
| Fear | 47,38% | 73,68% | 78,94% | 73,68% | 47,38% |
| Joy | 88,46% | 92,30% | 92,30% | 92,30% | 88,46% |
| Sad | 74,19% | 67,74% | 64,51% | 67,74% | 74,19% |
| Anger | 65,21% | 69,56% | 78,26% | 69,56% | 65,21% |
| Surprise | 72,41% | 100% | 100% | 100% | 72,41% |
| Disgust | 57,89% | 89,47% | 84,21% | 89,47% | 57,89% |

**Table 3.** Recognition rate for several temporal intervals and expression types.

The overall recognition rate obtained is 82,12%.

## 5. DISCARDING PROCESS

In the scenario described above, the expression recognizer is forced to choose a certain expression within the models for any input sequence whereas it may not belong to any of them. There are typical situations where it may not be interesting to decide anything. Situations such as when the subject is talking, in neutral expression and even when we are just not interested in detecting a particular expression.

For this reason, we introduce a validation process which will also reduce the error probability in sequences which correspond to emotions.

Taking the above into account, a reliability factor can be defined as in the following:

$$R = P_{chosen} - P_{below}$$

The $R_{threshold}$ threshold has been adjusted empirically, in order to achieve a low error probability discarding the minimum number or correct sequences. Expression sequences with an R factor below the threshold, will be discarded.

By means of two different experiments, a $R_{threshold}$ of 0.0133 was established. The first experiment consisted of preventing possible HMM errors due to sequences having similar probabilities. Without applying the validation process we obtained an 18% of error probability. Using the validation, the error probability decreases to 4.5%, while the correct detection decreases also to 78% (that is, 4% of the correct sequences were also discarded). In the second experiment, sequences not comprised within the models (in this case, sequences with people talking) were employed. The discarding process let us eliminate 60% of the sequences, and therefore a big bulk of wrong decisions were avoided.

## 6. CONCLUSIONS

We have presented a system which, using a standard database for facial expressions where video sequences comprise only the part of the expression evolving from neutral to peak, can be applied to random intervals of the expression. Taking advantage of the symmetry of the expressions, we have extracted a new topology of the HMMs adapted to any kind of interval. We have also trained the models whitout requiring the recording and analysis of a new database.

The system developed shows a high recognition rate when applied to random intervals of expressions in video sequences. The system can in this way be applied to long video expression detection and recognition by sampling and processing, as a validation process discards the parts of the video without useful expression information.

## 7. 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] M. Pardàs, A. Bonafonte, J.L. Landabaso, "Emotion recognition based on MPEG4 facial animation parameters", Proc. of IEEE Acoustics, Speech, and Signal Processing, 2002, Volume: 4 , pp. 3624 –3627, 2002.

[13] M. Pardàs, A. Bonafonte, "Facial Animation Parameters extraction and Expression detection using HMM", Special Issue on Image processing techniques for Virtual Environments and 3D Imaging of the Signal Processing: Image Communication Journal, 2002.

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

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

[16] 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.

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