AUTOMATIC FACIAL EXPRESSION RECOGNITION USING FACIAL ANIMATION PARAMETERS AND MULTI-STREAM HMMS

Petar S. Aleksic and Aggelos K. Katsaggelos

Department of Electrical and Computer Engineering
Northwestern University
2145 North Sheridan Road, Evanston, IL 60208
Email: {apetar, aggk} @ece.northwestern.edu

ABSTRACT

In this paper we present an automatic facial expression recognition system that utilizes multi-stream Hidden Markov Models (HMMs). The proposed system uses Facial Animation Parameters (FAPs), supported by the MPEG-4 standard, as features describing facial expressions. In particular, the FAPs controlling the movement of the outer-lips and eyebrows are used as visual features for classification. Experiments were performed under several different scenarios utilizing outer-lip and eyebrow FAPs individually and jointly. A new approach is proposed for introducing facial expression and FAP group dependent stream weights. The weights were chosen based on the facial expression recognition results obtained when FAP group streams are utilized individually. The proposed multi-stream HMM facial expression recognition system achieves relative reduction of the expression recognition error of 44%, compared to the single-stream HMM system.

1. INTRODUCTION

Automatic facial expression recognition [1] has many potential applications, such as emotion analysis, interactive video, indexing and retrieval of image and video databases, image understanding, and synthetic face animation. Most of the current human-computer interaction (HCI) techniques rely on modalities such as, key press, mouse movement, or speech input, and therefore do not provide natural human-to-human-like communication. The information about facial expressions contained in human faces is usually ignored. Human faces contain significant information about emotions and the mental state of a person that can be utilized in order to enable communication with computers in a natural way, similar to every day interaction between people.

Ekman and Friesen [2] defined six basic emotions (happiness, sadness, fear, disgust, surprise, and anger). Each of these six basic emotions corresponds to a unique facial expression (see Figure 1). They defined the Facial Action Coding System (FACS), a system developed in order to enable facial expression analysis through standardized coding of changes in facial motion. FACS consists of 46 Action Units (AU) which describe basic facial movements. It is based on muscle activity and describes in detail the effect of each AU on visual face features. Suwa *et al.* [3] and Mase and Petland [4] performed early work on automatic facial expression analysis. The important features used to describe facial expressions are the locations of facial actions, their intensity, and their dynamics. The intensity of facial feature changes is characterized by geometric deformations of certain regions in the face. Facial expression dynamics is contained in temporal changes of facial features. Sequences of facial images provide significant information about the dynamics of facial expressions.

Facial features used for automatic facial expression analysis can be obtained using two approaches. In the image-based approach, the whole face image, or images of parts of the face, are processed in order to obtain visual features. In the model-based approach, face models are used to describe movement of the visual features. Only the model parameters that change during facial expressions are used for expression recognition. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Discrete Cosine Transform (DCT), etc., are methods commonly used to decorrelate facial features and decrease their dimensionality. Such facial features provide more reliable training of classification systems and improve their performance. In order to perform person-independent automatic facial expression recognition it is important to normalize the values that correspond to facial feature changes using the facial features extracted from the person's neutral face. FACS have been used to describe visual features in automatic facial expression systems [5]. Low-level FAPs [6] (see Figure 2) have also been used as visual features in automatic facial expression recognition systems [7-9]. FAPs are normalized using Facial Animation Parameter Units (FAPUs) and, therefore, can be used for person-independent facial expression recognition. Facial expression analysis using FAPs has several advantages. One of these is that it secures compliance with the MPEG-4 standard. Another is that already existing FAP extraction systems or already available FAPs can be utilized to perform automatic facial expression recognition. The FAPs that contain significant information about facial expressions are those that control eyebrow (group 4) and mouth movements (group 8) (see Figure 2).

Facial feature classification approaches utilized in facial expression recognition systems can be divided into spatial and spatio-temporal. In spatial approaches visual features obtained from a single face image are used for classification [10]. Neural networks (NNs) are commonly used to perform spatial classification. Although spatial approaches can achieve good facial expression recognition in some cases, they do not model the dynamics of facial expressions and therefore do not utilize all of the information available in video sequences. Spatio-temporal approaches allow for such modeling by considering visual features extracted from each frame of a Facial expression video sequence. HMMs [11] are frequently used in the literature to perform spatio-temporal classification [8, 12].

In this paper we present a multi-stream HMM automatic facial expression recognition system that utilizes outer-lip and eyebrow FAPs. The proposed approach introduces facial expression and FAP group dependent stream weights.

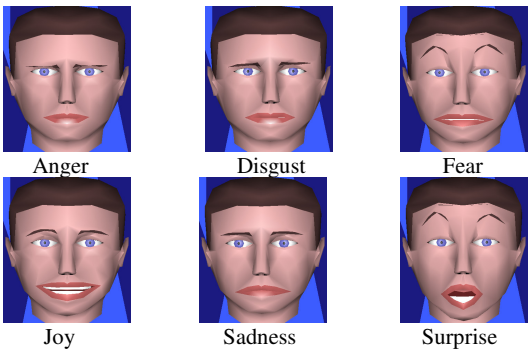

Figure 1: Six basic facial expressions represented by an MPEG-4 compliant player, using outer-lip and eyebrow FAPs.

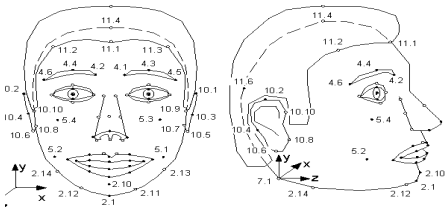

Figure 2: Outer-lip and eyebrow FAPs [6].

The remainder of the paper is organized as follows: Section 2 describes the database used. Section 3 describes the proposed facial expression recognition system. In Section 4 facial expression recognition experiments are described, while Section 5 summarizes the results and draws conclusions.

2. COHN-KANADE FACIAL EXPRESSION DATABASE

The Cohn-Kanade facial expression database [13] is used for training and testing of the developed automatic facial expression recognition system. It consists of recordings of 90 subjects. Each recording contains one of the basic six expressions (*anger, disgust, fear, joy, sadness,* and *surprise*) (see Figure 1). The video rate is 30 frames per second. Only full-face frontal views with constant illumination are captured. The subjects were not previously trained in displaying facial expressions, however, they practiced the expressions with an expert prior to video recording. Each expression recording starts at neutral expression and ends at the peak of the expression. The number of available sequences in the database for each of the six basic expressions is shown in Table 1. The FAPs describing outer-lip and eyebrow positions, extracted from each of the video sequences in the database, are obtained from [9] and used as features in the classification experiments. There are ten outer-lip and eight eyebrow FAPs.

3. FACIAL EXPRESSION RECOGNITION SYSTEM

The block diagram of the facial expression recognition system that we developed is shown in Figure 3. This approach exploits temporal facial feature changes, which contain significant information about facial expressions. We modeled each of the six basic expressions with a left-to-right HMM. HMMs had three states and continuous probability density functions. The means and variances of all the states of all six models were tied in order to perform more reliable training and overcome the lack

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Num. of Seq.	34	37	34	62	53	64

Table 1: Number of available sequences for each of the six basic facial expressions.

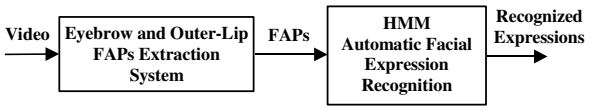

Figure 3: Block diagram of the automatic facial expression recognition system.

of training data. Iterative mixture splitting, and re-estimation were performed during the training in order to obtain the final HMMs. The six HMMs were first trained with all the subjects except one ("leave-one-out" training approach) using the extracted eyebrow and outer-lip FAPs as observation vectors. The testing of the trained automatic facial expression recognition system was then performed on the expression sequences of the subject that had not participated in the training. This process was repeated for all the subjects in the database. Experiments were performed under three different scenarios. In the first scenario, only the eyebrow (*e*) FAP vectors ($\boldsymbol{o}_t^e$) were used as observations. In the second scenario, only the outer-lip (*ol*) FAP vectors ($\boldsymbol{o}_t^{ol}$) were utilized, while in the third scenario, eyebrow FAP vectors were appended to the outer-lip FAP vectors in order to form joint observation vectors ($\boldsymbol{o}_t^j$). The observation vectors (outer-lip, eyebrow, and joint) were used to train a single-stream HMM model.

Since outer-lip and eyebrow FAPs provide different amounts of information about facial expressions, we should rely more on the FAPs that contain more information. In addition, the amount of the information also depends on the particular facial expression. Hence, it is desirable to develop a system that could model the reliability of the facial expression information contained in the outer-lip and eyebrow FAP streams with respect to different facial expressions.

3.1. Multi-Stream HMM System

Multi-stream HMMs allow for modeling the reliability of different streams of information. Eyebrow and outer-lip FAP stream weights can be defined for each of the HMMs modeling the six basic facial expressions. The multi-stream HMM observation vector probability distribution is given by

$$\boldsymbol{b}_{Fi}(\boldsymbol{o}_t^{ol}, \boldsymbol{o}_t^{e} \mid F) = \prod_{s\in\{ol,e\}} \left[\sum_{m=1}^{M_s} \boldsymbol{c}_{Fism}\, N(\boldsymbol{o}_t^{s}; \boldsymbol{\mu}_{Fism}, \boldsymbol{\Sigma}_{Fism}) \right]^{\gamma_{Fs}} . \quad (1)$$

In Eq. (1) *s* denotes the stream, while subscript *i* denotes an HMM state used to model facial expression *F*. M_s denotes the number of mixtures in a stream, c_{Fism} denotes the weight of the m^{th} mixture of the stream *S* for the facial expression *F*, and *N* a multivariate Gaussian with mean vector $\boldsymbol{\mu}_{Fism}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{Fism}$. The non-negative stream weights are denoted by γ_{Fs}, and they depend on the modality *s* and the facial expression *F*. It is assumed that the outer-lip (γ_{Fol}) and eyebrow FAP (γ_{Fe}) stream weights for the facial expression *F* satisfy

$$\gamma_{Fol} + \gamma_{Fe} = 1 , \quad (2)$$

for all six facial expressions. The stream weights can be determined based on the confidence of the streams and amount

Eyebrow FAPs							
	Anger	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Rec. [%]**
Anger	18	9	2	3	2	0	52.9
Disgust	6	28	1	2	0	0	75.7
Fear	4	5	2	5	13	5	5.9
Joy	0	0	3	52	6	1	84.1
Sadness	4	3	9	18	9	10	17.0
Surprise	0	0	2	0	4	58	90.6
Total							58.80

Table 2: Confusion matrix for the system that utilizes eyebrow FAPs.

Outer-Lip FAPs							
	Anger	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Rec [%]**
Anger	22	4	0	0	8	0	64.7
Disgust	2	34	0	0	0	1	91.9
Fear	0	0	28	5	1	0	82.4
Joy	0	0	3	59	0	0	95.2
Sadness	8	2	0	0	43	0	81.1
Surprise	0	2	0	0	0	62	96.9
Total							87.32

Table 3: Confusion matrix for the system that utilizes outer-lip FAPs.

of information contained in them. They can also be chosen by minimizing the facial expression recognition error on a held-out (development) data set, utilizing optimization techniques. Here, due to the training approach utilized, we determined the facial expression dependent stream weights based on the recognition results obtained when single-stream HMMs were utilized with eyebrow and outer-lip FAP vectors individually employed as observations. The stream weights were computed as

$$\gamma_{Fol} = \frac{R_F^e}{R_F^{ol} + R_F^e} \qquad \gamma_{Fe} = \frac{R_F^{ol}}{R_F^{ol} + R_F^e} \quad , \tag{3}$$

where R_F^e and R_F^{ol} denote recognition error rates for expression F, obtained when eyebrow, and outer-lip observations were utilized, respectively. This approach provides stream weights proportional to the amount of information contained in corresponding streams. It will be shown in the next section, the performance of the multi-stream HMM system is not very sensitive to the choice of stream weights, as long as they are proportional to the corresponding stream information.

4. FACIAL EXPRESSION RECOGNITION EXPERIMENTS

Single-stream HMM facial expression experiments were performed under three different scenarios, using eyebrow and outer-lip FAPs individually and jointly. The confusion matrices for the three scenarios are shown in Tables 2-4. Tables 2 and 3 reveal that the system that utilized outer-lip FAPs outperformed the system that utilized eyebrow FAPs. The overall expression recognition performance of the system that used outer-lip FAPs was 87.32%, while the performance of the system that used eyebrow FAPs was only 58.80%. It can be concluded from these results that outer-lip FAPs provide more information for classifying facial expressions. The recognition rates for *fear* and *sadness* facial expressions were particularly low when only eyebrow FAPs were used (see Table 2), due to the fact that the changes in eyebrow FAPs that occur during these expressions are very similar. The facial expression recognition rates for the remaining four facial expressions were high, even when only eyebrow FAPs were used. It is important to note that the recognition rates for the facial expressions *fear* and *sadness* decreased when both FAP groups were used, as compared to the

Outer-Lip and Eyebrow FAPs (Single-stream)							
	Anger	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Rec. [%]**
Anger	22	6	0	0	6	0	64.7
Disgust	1	36	0	0	0	0	97.3
Fear	0	0	27	3	1	3	79.4
Joy	0	0	0	61	0	1	98.4
Sadness	7	0	1	0	42	3	79.2
Surprise	0	0	0	0	0	64	100
Total							88.73

Table 4: Confusion matrix for the system that utilizes eyebrow and outer-lip FAPs and single-stream HMMs.

Outer-Lip and Eyebrow FAPs (Multi-stream)							
	Anger	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Rec. [%]**
Anger	24	4	0	0	6	0	70.6
Disgust	0	36	1	0	0	0	97.3
Fear	0	0	30	2	1	1	88.2
Joy	0	0	0	61	0	1	98.4
Sadness	2	0	0	0	51	0	96.2
Surprise	0	0	0	0	0	64	100
Total							93.66

Table 5: Confusion matrix for the system that utilizes eyebrow and outer-lip FAPs and multi-stream HMMs.

FAPs / Exp	**Eyebrow (E) [%]**	**Outer-Lip (OL) [%]**	**E and OL [%]**	**E and OL [%] (Multi-stream)**	**OL stream weight**
Anger	52.9	64.7	64.7	70.6	0.7
Disgust	75.7	91.9	97.3	97.3	0.7
Fear	5.9	82.4	79.4	88.2	0.8
Joy	84.1	95.2	98.4	98.4	0.6
Sadness	17.0	81.1	79.2	96.2	0.7
Surprise	90.6	96.9	100	100	0.6
Total	58.80	87.32	88.73	93.66	

Table 6: Automatic facial expression recognition results for different FAP groups and different HMM systems.

the rates achieved when only outer-lip FAPs were utilized. Generation of the joint features increased the dimensionality of the observations used for classification and affected reliable training. The amount of additional information about expressions *fear* and *sadness* contained in the eyebrow FAPs was insufficient to overcome the effect of the increased dimensionality. Therefore, it is desirable to rely more on outer-lip FAPs information when describing expressions *fear* and *sadness* and less on the eyebrow FAPs information.

Multi-stream HMMs and facial expression dependent stream weights were used to model the reliability of the information contained in outer-lip and eyebrow FAP streams. Stream weights for eyebrow and outer-lip FAP streams were determined based on the facial expression recognition results obtained when only eyebrow or outer-lip FAPs were utilized (see Eq. (3)). Hence, the outer-lip stream weights were set in the experiments to be larger for the facial expressions for which eyebrow FAPs did not contain sufficient information. The resulting confusion matrix is shown in Table 5. The facial expression recognition rates for all systems tested are shown in Table 6 together with the multi-stream HMM facial expression dependent stream weights for which the best recognition rates were obtained. The recognition performance obtained for the stream weights chosen based on Eq. (3) was 93.31%. In general, stream weights should be estimated on a development data set, either using Eq. (3) or by maximizing the multi-stream HMM recognition performance utilizing optimization techniques. Nevertheless, the multi-stream HMM recognition results are not very sensitive to the choice of the stream weights, as long as the larger weight is assigned to the information stream that provides

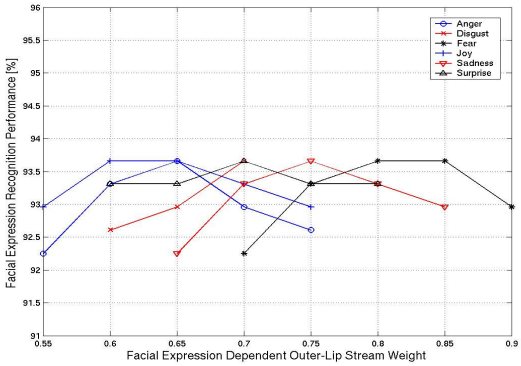

Figure 4: Facial expression recognition performance obtained when each of the weights corresponding to the six basic facial expressions was varied around the value that provided the best performance, while the remaining five weights were fixed.

better recognition results. In order to demonstrate that, we performed recognition experiments utilizing a number of different stream weight sets. The best recognition performance achieved was 93.66%, which was slightly better than the performance obtained utilizing stream weights obtained from Eq. (3). We show in Figure 4 the recognition performance obtained when the weights for five of the MS-HMMs corresponding to five facial expressions were fixed and one of them is varied around the value that provided the best performance (see [14] for details). It can be concluded from Figure 4 that the proposed system is not very sensitive to the choice of stream weights. Furthermore, significant recognition performance improvement is achieved for a large set of stream weights when multi-stream HMMs are utilized. In addition, the best recognition performance (93.66%) was obtained for several different sets of stream weights.

The experiments were performed for different number of Gaussian mixtures used for eyebrow and outer-lip FAPs. The number of mixtures varied from two to ten. The best recognition performance was obtained when four mixtures were used for each of the FAP groups (see Figure 5). It is important to note that the overall facial expression recognition performance increased by approximately 5% when the multi-stream HMM system was used compared to the single-stream HMM system. The relative reduction of the expression recognition error achieved, compared to the single-stream HMM facial expression recognition system, was 44%. In addition, recognition rates for facial expressions *fear* and *sadness* increased by addition of the eyebrow information due to the stream reliability modeling by appropriately chosen stream weights. The proposed system outperforms the system described in [8] which achieves the recognition performance of 84% on the same database.

5. CONCLUSIONS

The automatic facial expression recognition systems developed in this work utilize eyebrow and outer-lip FAPs. The introduction of expression dependent stream weights enabled modeling of the reliability of information contained in FAP streams, when multi-stream HMMs were used.

It is expected that the use of additional visual information about facial expressions (cheek, eye FAPs) would further improve recognition performance. It is important to point out

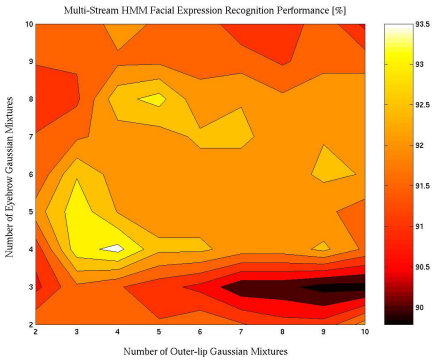

Figure 5: Facial expression recognition performance obtained for different number of Gaussian mixtures used for outer-lip and eyebrow FAP streams.

that the proposed multi-stream HMM approach can be utilized with any kind of facial features, be it model- or image-based. Furthermore, model-based and image-based facial features can be combined as two separate streams of information.

6. REFERENCES

[1] B. Fasel and J. Luettin, "Automatic facial expression analysis: A Survey," *Pattern Recognition*, vol. 36(1), pp. 259-275, 2003.

[2] P. Ekman and W. Friesen, Facial action coding system, Consulting Psychologists Press Inc., Palo Alto, California 94306, 1978.

[3] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," *Proc. of the 4th Int. Joint Conf. on Pattern Recognition*, pp. 408-410, 1978.

[4] K. Mase and A. Pentland, "Recognition of Facial Expression from Optical Flow," *Trans. of IEICE*, E 74(10), pp. 3474-3483, Oct. 1991.

[5] J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Automated facial expression recognition based on FACS action units," *IEEE Proc. of the Second Int. Conf. on Automatic Face and Gesture Recognition (FG'98)*, pp. 390-395, Japan, April 14-16 1998.

[6] Text for ISO/IEC FDIS 14496-2 Visual, ISO/IEC JTC1/SC29/WG11 N2502, Nov. 1998.

[7] Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie and E. Douglas-Cowie, "Emotion recognition and synthesis based on MPEG-4 FAPs," in MPEG-4 Facial Animation, I. Pandzic, R. Forchheimer, Eds., John Wiley & Sons, UK, 2002.

[8] L. Landabaso, M. Pardàs, and A. Bonafonte, "HMM recognition of expressions in unrestrained video intervals," *Proc. of ICASSP*, pp. 197-200, China, April 6-10, 2003.

[9] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression detection using HMM," *Sig. Proc.: Image Communication,* vol.17, pp. 675-688, 2002.

[10] Padgett and G. W. Cottrell, "Representing face image for emotion classification," in M. Mozer, M. Jordan, and T. Petsche, Eds., *Advances in Neural Information Proc. Systems*, vol. 9, pp. 894-900, Cambridge, MA, MIT Press, 1997.

[11] R Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[12] N. Oliver, A. Pentland, and F. Berard, "LAFTER: A real-time lips and face tracker with facial expression recognition," *Proc. of the CVPR*, pp. 123-129, Puerto Rico, 1997.

[13] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proc. of the 4th IEEE Int. Conf. on Automatic Face and Gestures Recognition*, pp. 46-53, France, 2000.

[14] P. S. Aleksic and A. K. Katsaggelos, "Automatic Facial Expression Recognition Using Facial Animation Parameters and Multi-Stream HMMs," to appear in *IEEE Trans. on Sig. Proc. Supplement on Secure Media,* October 2005.