Automatic Facial Expression Recognition Using Facial Animation Parameters and Multistream HMMs

Petar S. Aleksic, *Member, IEEE,* and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—The performance of an automatic facial expression recognition system can be significantly improved by modeling the reliability of different streams of facial expression information utilizing multistream hidden Markov models (HMMs). In this paper, we present an automatic multistream HMM facial expression recognition system and analyze its performance. The proposed system utilizes facial animation parameters (FAPs), supported by the MPEG-4 standard, as features for facial expression classification. Specifically, the FAPs describing the movement of the outer-lip contours and eyebrows are used as observations. Experiments are first performed employing single-stream HMMs under several different scenarios, utilizing outer-lip and eyebrow FAPs individually and jointly. A multistream HMM approach is proposed for introducing facial expression and FAP group dependent stream reliability weights. The stream weights are determined based on the facial expression recognition results obtained when FAP streams are utilized individually. The proposed multistream HMM facial expression system, which utilizes stream reliability weights, achieves relative reduction of the facial expression recognition error of 44% compared to the single-stream HMM system.**

***Index Terms*—Facial expression recognition, multistream HMMs, facial animation parameters.**

I. Introduction

AUTOMATIC facial expression recognition has many potential applications in areas such as human-computer interaction (HCI), emotion analysis, interactive video, indexing and retrieval of image and video databases, image understanding, and synthetic face animation. Due to the increasing importance of computers in every day life, HCI has become very important in today's society. Most of the current 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 contained in facial expressions, eye movement, hand movement, etc., is usually ignored. Developing a system which could detect the presence of humans (using face detection), determine their identity (using face, voice, or audio-visual person recognition), and understand their behavior (using facial expression analysis, audio-visual speech recognition, etc.) in order to respond to their needs or requests, would significantly improve performance of HCI systems, and make them nonintrusive and natural to the user. Automatic facial expression analysis is an important part of such a system. Human faces contain significant information about emotions and the mental state of a person that can be utilized in order to enable nonverbal communication with computers [1], [2].

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 Fig. 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 face features. Suwa *et al.* [3] and Mase and Pentland [4] performed early work on automatic facial expression analysis. Most of the current automatic facial expression recognition systems analyze only the six basic facial expressions. However, there are also systems that define and analyze a number of possible facial expressions [5].

Robust face detection, and tracking are of great importance for automatic facial expression recognition. Face detection, in general, is a difficult problem, especially in cases where background, head pose, and lighting are varying [6]–[8]. After successful face detection, interesting facial features usually utilized for facial expression analysis, such as the mouth corners, eyebrows, eyes, nostrils, chin, etc., need to be located. Another important issue in the design and implementation of automatic facial expression systems is the choice of facial features and their robust extraction from a static face image or video. The various sets of facial features proposed in the literature can be grouped into two categories: a) image-based (appearance-based) features; and b) model-based (shape-based) features. The choice of facial features clearly mandates the tracking algorithms required for their extraction, but is also a function of image (video) data quality and resource constraints in the facial expression recognition application.

In the image-based facial feature extraction approaches, the pixel intensities of the whole face image or certain regions of the face image are processed in order to obtain facial features. This approach is usually very fast and simple, but also can provide facial features of high dimensionality, which can affect reliable training of a classification system. In order to improve performance of facial expression recognition systems, image compression techniques, such as principal component analysis (PCA) [6], linear discriminant analysis (LDA) [10], discrete cosine transform (DCT) etc. [9]–[11], are commonly used on image-based features to decrease their dimensionality. Image-based facial features have been regularly utilized for

Manuscript received July 2, 2004; revised March 9, 2005. An earlier version of this work appeared in the 6th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05) [40]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anil Jain.

The authors are with the Electrical Engineering and Computer Science Department, Northwestern University, Evanston, IL 60208 USA (e-mail: apetar@ece.northwestern.edu; aggk@ece.northwestern.edu).

Digital Object Identifier 10.1109/TIFS.2005.863510

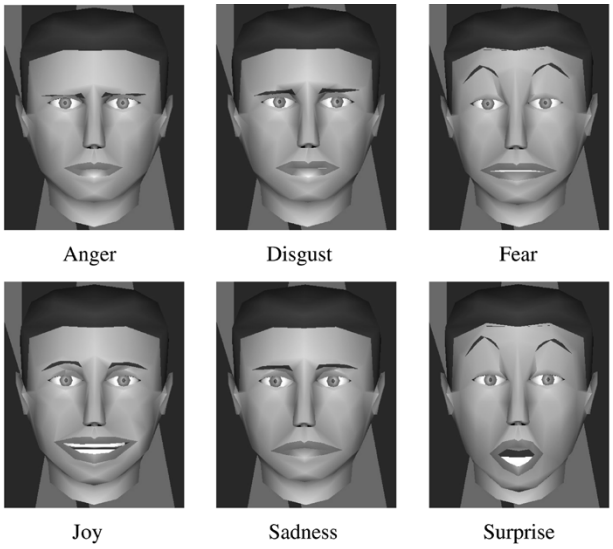

Fig. 1. Six synthesized basic facial expressions obtained using an MPEG-4 compliant avatar [21] and outer-lip and eyebrow FAPs.

automatic facial expression recognition [9]. In the model-based facial feature extraction approaches [5], [12]–[14], face models are used to describe facial features. Only the model parameters that change during facial expressions are utilized for classification of facial expressions. Such facial features are usually of low-dimensionality. However, in order to ensure good performance, the extraction of model-based features requires robust facial feature tracking, which can be difficult and computationally intensive. Significant information about facial expressions is contained not only in facial features, but also in dynamics of their temporal changes [9]. Facial expression recognition systems that utilize static images [15], [16] can provide good performance, however, video sequences contain more information about expressions, and it is expected that they provide improved recognition performance.

MPEG-4 is an audio-visual object-based video representation standard supporting facial animation. MPEG-4 facial animation is controlled by the facial definition parameters (FDPs) and facial animation parameters (FAPs), which describe the face shape, and movement, respectively [1], [17], [18]. There are many applications that can use information contained in FAPs, including audio-visual automatic speech recognition (ASR) [19], audio-visual speaker recognition [20], facial expression recognition [5], [12]–[14], etc. FAPs can be used to concisely represent evolution of facial expressions. The FAPs that contain significant information about facial expressions are those that control eyebrow and mouth movements [14] (see Fig. 1). They are utilized in this work for classification of facial expressions.

Approaches for classification of facial expressions can be divided into spatial and spatio-temporal. In spatial approaches, facial features obtained from a single face image are utilized for classification of facial expressions. Artificial neural networks (ANNs) [22]–[24], support vector machines (SVMs) [10], LDA [10], etc., are commonly used to perform spatial classification in facial expression recognition systems. Although spatial approaches can achieve good facial expression recognition performance in some cases, they do not model the dynamics of facial expressions and therefore do not utilize all of the information about facial expressions available in video sequences. Spatio-temporal approaches allow for such modeling by considering facial features extracted from each frame of a facial expression video sequence. Hidden Markov models [25] are frequently used in the literature to perform spatio-temporal classification [14], [26], [27].

Littlewort *et al.* [10] utilize image-based facial features for classification. They automatically detect faces in images and rescale them to 48×48 pixels. The rescaled facial images are convolved with a bank of Gabor filters in order to obtain their Gabor magnitude representation. They define seven classes corresponding to the six basic facial expressions and the neutral facial expression. They utilize several classification approaches, including SVM, AdaBoost, and LDA. They obtain the best performance when they perform feature selection using AdaBoost and classification using SVMs. The best facial expression recognition rate that they achieve utilizing Cohn-Kanade database [28] is 93.33%. They also performed automatic recognition of FACS action units and achieved performance of 92.9%.

Lien *et al.* [29] developed a system that automatically recognizes individual facial action units or combinations of action units. They performed upper face facial expression recognition based on FACS, while utilizing facial feature point tracking, dense flow tracking with PCA, and high gradient component (furrow) detection, in order to extract facial expression features. They utilized HMMs for classification, and obtained upper face action unit recognition performance of 85%, 93%, and 85%, when feature point tracking, dense flow tracking, and furrow detection were utilized, respectively.

Several automatic facial expression and facial action recognition systems [29]–[33] utilize FACS as facial features. FAPs have also been used as facial features in automatic facial expression recognition systems [5], [12]–[14]. The recognition results obtained in [5], [12]–[14] showed that FAPs contain information important for expression recognition. It is very important to determine which FAP groups and classification approaches provide the best facial expression recognition performance.

In this paper we present a novel multistream HMM (MS-HMM) automatic facial expression recognition system that utilizes MPEG-4 compliant facial features. The outer-lip (Group 8) and eyebrow (Group 4) FAPs are utilized as observations (see Fig. 2). The proposed approach introduces facial expression and FAP group dependent stream reliability weights. The stream weights are determined based on the automatic facial expression recognition results obtained when FAP streams are utilized individually. The proposed MS-HMM facial expression system, which utilizes stream reliability weights, achieves high recognition performance and significantly outperforms the single-stream HMM (SS-HMM) facial expression recognition system.

The rest of the paper is organized as follows: Section II describes facial features, and their extraction. Section III describes the proposed automatic facial expression recognition system. In Section IV automatic facial expression recognition experiments are described, while Section V summarizes the results and draws conclusions.

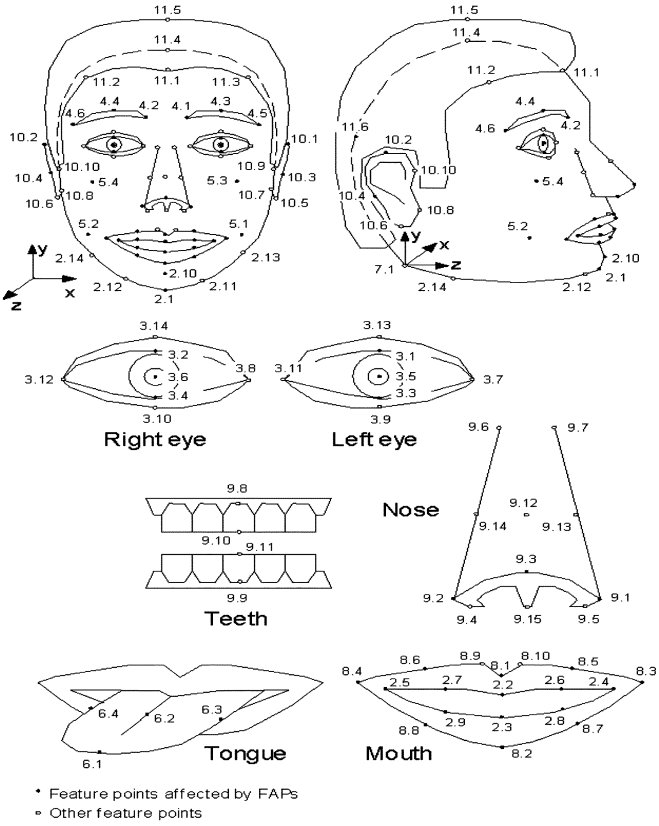

Fig. 2. Facial animation parameters [17].

TABLE I
NUMBER OF AVAILABLE VIDEO SEQUENCES FOR EACH OF THE SIX BASIC FACIAL EXPRESSIONS

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

II. FACIAL FEATURES

The Cohn-Kanade facial expression database [28] is utilized in this work. It consists of 284 recordings of 90 subjects. Each recording contains one of the six basic facial expressions (*anger, disgust, fear, joy, sadness*, and *surprise*) (see Fig. 1). The recordings corresponding to a particular subject represent different facial expressions. The video rate is 30 frames/s. 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. Image sequences were digitalized into either 640×490 or 640×480 pixel arrays. The number of available sequences in the database for each of the six basic expressions is shown in Table I.

In this work we exploit relative timing and temporal evolution of facial feature changes by utilizing FAPs as facial features. The MPEG-4 standard defines 68 FAPs. They are divided into ten groups, which describe the movement of the face [1], [17], [18]. These parameters are either high level parameter (Group 1), that is, parameters that describe visemes and facial expressions, or low-level parameters (Groups 2–10; see Fig. 2), that is, parameters describing displacement of the specific single point of the face. FAPs control the key features of the model of a head, shown in Fig. 1, and can be used to animate facial movements and expressions [1]. 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. In addition, FAPs are expressed in terms of facial animation parameter units (FAPUs). These units are normalized by important facial feature distances, such as mouth width, mouth-nose, eye-nose, or eye separation, in order to give an accurate and consistent representation. This is particularly useful for facial expression recognition, since normalizing facial features corresponding to different subjects enables better modeling of facial expressions. In this work, the parameters which describe outer-lip (Group 8) and eyebrow movement (Group 4) are utilized as features for classification. Group 8 and 4 parameters are expected to be the FAPs that contain significant information for facial expression recognition [9], [14]. There are ten FAPs describing the outer-lips position and eight FAPs describing the eyebrow position. The effectiveness of information contained in the outer-lip and eyebrow FAPs for facial expression recognition is demonstrated in this work. There exist different approaches for extraction of FAPs, using active contour [14], [19], geometric templates [19], [34]–[35] or combination algorithms [19].

A number of researchers have developed facial expression recognition systems, which vary in the choice of database, facial features, information fusion, and classification techniques. However, in order to enable fair comparison of different systems, standard evaluation procedures and experiments should be defined. Furthermore, the effect of different facial features or classification techniques on the overall facial expression recognition performance can be evaluated separately, only if a controlled experiment is constructed, i.e., all other elements of the system remain fixed. Therefore, since the main goal of this paper is to determine the advantages and improvements related to the proposed MS-HMM system, we wanted to be able to compare our system in terms of recognition performance with a system utilizing the same database and the same facial features but a different information fusion approach. Hence, we report facial expression recognition results utilizing FAP sequences that we obtained from [14] which were extracted from the Cohn–Kanade database utilizing an active contour algorithm, and compare them to the SS-HMM system facial expression recognition results obtained in [14].

The active contour method is commonly used for extraction of visual features. It is very useful in cases when it is hard to present the shape of an object with a simple template. An active contour (snake) is an elastic curve defined by a set of control points [36]. It is used for finding important visual features, such as lines, edges, or contours. Snake deformation is controlled by the iterative search for a local minimum of the energy function, which depends on the external and internal force fields. The internal snake forces are defined based on the required tension and rigidity of the snake, while the external forces are derived from the image data usually using gradient operator on original or blurred images. An example of outer-lip tracking process is shown in Fig. 3, where a gray level mouth image is shown in

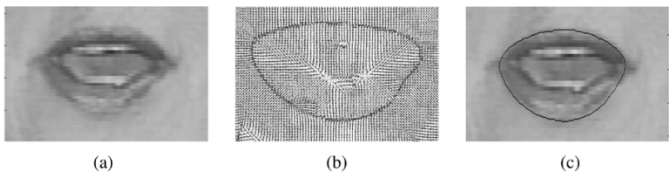

Fig. 3. An example of (a) an original mouth image; (b) corresponding external force field; and (c) the final snake position.

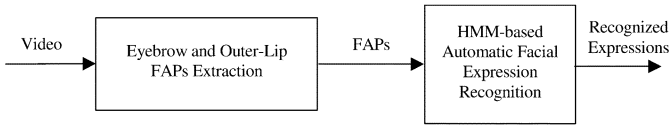

Fig. 4. Block diagram of the automatic HMM-based facial expression recognition system.

Fig. 3(a), the corresponding external force field in Fig. 3(b), and the resulting snake position in Fig. 3(c). In the FAP extraction process, after the outer-lip and eyebrow contours are tracked for each sequence frame, they are compared to the corresponding contours of the neutral (first) frame of the sequence in order to calculate FAPs in terms of FAPUs. They are calculated by aligning the current frame contours with the neutral reference contours, calculating the movement, and normalizing the distances between corresponding locations (see Fig. 2).

III. Automatic Facial Expression Recognition System

The block diagram of the automatic facial expression recognition system is shown in Fig. 4. HMMs, commonly used tool for automatic speech recognition, are utilized in this work as a classification approach. In general, an HMM is characterized by the number of states in the model, state transition probability matrix, type of state observations, observation probability distribution, and initial state distribution. In this work, eyebrow FAPs ($\boldsymbol{o}_t^e$), outer-lip FAPs ($\boldsymbol{o}_t^{ol}$), or joint FAP vectors ($\boldsymbol{o}_t^j$), extracted at time t from a face image, were used as observations. The joint FAP vectors were obtained by appending the eyebrow to the outer-lip FAP vectors. The single-stream HMMs [25] model generation of a single observation sequence, and can be used to model temporal changes of facial features which occur during facial expressions. The continuous observation probability distribution corresponding to an HMM state is usually modeled by Gaussian mixture densities, given by

$$\boldsymbol{b}_i\left(\boldsymbol{o}_t^k|F\right) = \sum_{m=1}^{M} c_{im}\boldsymbol{N}\left(\boldsymbol{o}_t^k; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}\right), \quad k \in \{e, ol, j\}. \quad (1)$$

In (1) subscript i denotes a state of the HMM corresponding to the facial expression $F (F \in \{\text{anger}, \text{disgust}, \text{fear}, \text{joy}, \text{sadness}, \text{surprise}\})$, while k denotes a type of observations used. M denotes the number of mixtures, c_{im} the weight of the mth mixture component, and $\boldsymbol{N}$ a multivariate Gaussian with mean observation $\boldsymbol{\mu}_{im}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{im}$. The mixture weights c_{im} are positive and their sum is equal to 1.

The HMMs used in this work were three-state continuous left-to-right HMMs (see Fig. 5). The left-to-right HMM topology is commonly used for modeling signals for which

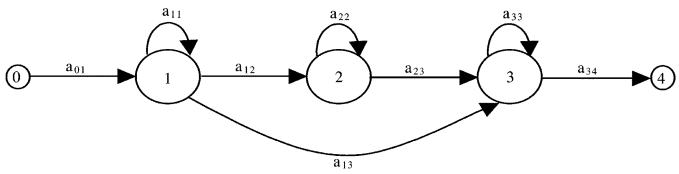

Fig. 5. Model topology for the HHMs used to model the six basic facial expressions.

properties change over time. The topology shown in Fig. 5 provided the best recognition results. Each of the six basic facial expressions is modeled with an HMM. We developed three systems by training SS-HMMs under three different scenarios. In the first and second scenarios, the observation sequences consisted of eyebrow, and outer-lip FAPs, respectively. In the third scenario, the joint FAP vectors were used as observations. In the HMM training process, training observation sequences are used to estimate and adjust the model parameters. The expectation-maximization (EM) [25] algorithm is commonly used to obtain maximum likelihood estimates of the HMM parameters, such as means, covariance matrices, mixture weights, and state transition probabilities, from the training observations. The means and variances of all the states of the six HMMs were tied in order to overcome the lack of training data and perform more reliable training. Hence, the observation probability distributions for each state differed only in their mixture weights. Iterative mixture splitting (increasing the number of Gaussian mixtures), and re-estimation were performed during the training process in order to obtain the final set of trained HMMs. The "leave-one-out" strategy was used for training and testing. The six HMMs were first trained with observation sequences corresponding to all the subjects except one. Subsequently, the testing of the facial expression recognition system was performed on the observation sequences corresponding to the remaining subject. This procedure was repeated for all the subjects in the database.

Since the outer-lip and eyebrow FAPs provide different amounts of information about facial expressions, this information should be weighted appropriately. Furthermore, the amount of the information contained in a FAP stream 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. Multistream HMMs can be used to model such reliability through the use of facial expression dependent stream weights. Their usage in the developed automatic facial expression recognition system is described next.

A. Multistream HMM System

Multistream HMMs allow for modeling reliability of different streams of information [37]–[39]. MS-HMMs can model generation of multiple observation sequences. In general, the observation probability distribution of the MS-HMM is the product of the observation likelihoods of its single-stream components, which are raised to appropriate stream exponents that capture the reliability of each stream. MS-HMMs have been commonly used for multiband audio-only ASR [37], [38]

and audio-visual ASR [19], [39]. For the problem at hand, we consider two streams of information, eyebrow, and outer-lip FAPs, and utilize two-stream HMMs. The two-stream HMM state observation probability distribution corresponding to the facial expression F, is given by

$$\boldsymbol{b}_{Fi}\left(\boldsymbol{o}_t^{ol},\boldsymbol{o}_t^{e}|F\right) = \prod_{s\in\{ol,e\}}\left[\sum_{m=1}^{M_S} c_{Fism}\boldsymbol{N}\left(\boldsymbol{o}_t^{s};\boldsymbol{\mu}_{Fism},\boldsymbol{\Sigma}_{Fism}\right)\right]^{\gamma_{Fs}} \quad (2)$$

where s denotes the stream ($s \in \{ol,e\}$), while subscript i denotes an HMM state of the HMM corresponding to the facial expression F. M_s denotes the number of mixtures in a stream, c_{Fism} the weight of the mth mixture of the stream s for the facial expression F, and $\boldsymbol{N}$ is a multivariate Gaussian with mean observation $\boldsymbol{\mu}_{Fism}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{Fism}$. The nonnegative stream weights are denoted by γ_{Fs}. In general, they can depend on the stream s, facial expression F, state i, and the time t. However, in this work, the stream weights depend only on the stream and the facial expression. They do not change over time or for different states of an HMM model. It is assumed that the outer-lip (γ_{Fol}) and eyebrow (γ_{Fe}) FAP stream weights corresponding to the facial expression F satisfy

$$\gamma_{Fol}+\gamma_{Fe}=1, \quad (3)$$

for all six facial expressions.

In order to train the MS-HMM system, we need to estimate HMM parameters, such as mixture weights, means, covariance matrices, state transition probabilities, and stream weights, for each of the streams. Maximum likelihood estimation by means of the EM algorithm can also be used to estimate MS-HMM parameters. We trained again six HMMs, one for each of the six basic facial expressions. The MS-HMMs utilized in this work were state-synchronous, that is the synchronicity between the streams is enforced at the state level. Therefore, the six MS-HMMs had the topology shown in Fig. 5.

Training of an MS-HMM system can be performed using two approaches. With the first approach, the parameters corresponding to each of the stream components are trained separately based on single-stream observation sequences. The EM algorithm is used to train two SS-HMMs. Subsequently, MS-HMMs can easily be obtained by combining observation probability densities corresponding to the SS-HMMs, as in (2). The transition matrices can be obtained by linear combination of the SS-HMM transition matrices, which can be weighted by the stream weights. The main disadvantage of this approach is that the SS-HMMs are trained asynchronously, while the (2) assumes that the stream components are state-synchronous. In the second approach, which is used in this work, the MS-HMMs are trained using both eyebrow and outer-lip FAP observation sequences at the same time in order to enforce state synchrony. The EM algorithm can also be utilized for training in this approach, due to the linear combination of stream log-likelihoods by means of (2). The training procedure was similar to the procedure described in the previous section.

Both approaches require *a-priori* choice of stream weights. The stream weights corresponding to eyebrow and outer-lip FAPs can be determined based on the confidence of these streams and amount 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 ("leave-one-out"), we determined the facial expression dependent stream weights based on the recognition results obtained when SS-HMMs were utilized with eyebrow and outer-lip FAP vectors individually employed as observations. The stream weights were computed as

$$\gamma_{Fol}=\frac{R_F^e}{R_F^{ol}+R_F^e} \quad \gamma_{Fe}=\frac{R_F^{ol}}{R_F^{ol}+R_F^e} \quad (4)$$

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. Similarly, these weights can be determined utilizing SS-HMM recognition rates obtained on a development data set. However, as it will be shown in the next section, the performance of the MS-HMM system is not very sensitive to the choice of stream weights, provided that they are proportional to the corresponding stream information.

As a result of the MS-HMM training process, we obtained a set of six state-synchronous MS-HMMs corresponding to the six basic facial expressions. Subsequently, we performed automatic facial expression recognition experiments in order to determine the effect of the introduction of stream reliability control to the recognition performance.

IV. Automatic Facial Expression Recognition Experiments

Automatic facial expression recognition experiments were performed utilizing both single- and multistream HMM systems. In the testing process, we calculate likelihoods that an observations sequence is generated by each of the HMMs corresponding to the six basic facial expressions. Subsequently, the observation sequence is assigned to the facial expression that correspond to the HMM for which the likelihood of generating the observation sequence is the largest. The performance of an HMM-based facial expression recognition system strongly depends on the choice of facial expression features (observations), their dimensionality, and the amount of available training data. There is a trade-off between the number of HMM parameters that need to be estimated and the dimensionality of the observations (and the amount of the information contained in them). The higher the dimensionality of the observations, the more classifying information is available; however, the number of HMM parameters also increases and more training data is required for their accurate estimation. In order to investigate this , we performed PCA on the eyebrow, outer-lip, and joint FAP observations, with the purpose of decreasing their dimensionality and decorrelating them. The FAP PCA training sets consisted of L FAP observations, which were obtained utilizing all observation sequences in the database in order to determine whether

TABLE II
FACIAL EXPRESSION RECOGNITION PERFORMANCE FOR DIFFERENT CLASSIFICATION FEATURES

Facial Expression Recognition Performance [%]							
	Classification Features (FAPs)						
	Original	PCA	Original + D	Original + D + A	Only D	Only A	Only D + A
Eyebrow	58.80	42.62	57.54	59.67	56.84	53.87	56.14
Outer-lip	87.32	71.13	85.21	86.97	85.86	84.86	84.72
E and OL	88.73	67.96	88.38	88.03	88.03	78.52	85.92

PCA can improve the performance of the system. The covariance matrices $\boldsymbol{C}^k$ were computed as

$$\boldsymbol{C}^k = \frac{1}{L}\sum_{l=1}^{L}\left(\boldsymbol{o}_l^k - \bar{\boldsymbol{o}}^k\right)\left(\boldsymbol{o}_l^k - \bar{\boldsymbol{o}}^k\right)^T, \quad k \in \{e, ol, j\} \tag{5}$$

where $\boldsymbol{o}_l^k$ and $\bar{\boldsymbol{o}}^k$ denote the lth training FAP observation and the mean FAP observation, respectively, for the type of observations k. After the covariance matrices were obtained and their eigenvectors and eigenvalues determined, the FAP observations, $\boldsymbol{o}_t^k$, were projected onto the eigenspace defined by the first P eigenvectors

$$\boldsymbol{o}_t^k = \bar{\boldsymbol{o}}^k + \boldsymbol{E}\cdot\hat{\boldsymbol{o}}_t^k. \tag{6}$$

In (6), $\boldsymbol{E} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \ldots\ \boldsymbol{e}_P]$ is the matrix of P eigenvectors, which corresponds to the P largest eigenvalues, while $\hat{\boldsymbol{o}}_t^k$ denotes the $P \times 1$ vector of corresponding projection weights. The PCA of the eyebrow observations resulted in two-dimensional $(P = 2)$ eyebrow FAP PCA projection weights. The first two principal components accounted to 98.81% of the total statistical variance. Similarly, we obtained three-dimensional outer-lip and four-dimensional joint FAP PCA projection weights. The first three and four principal components accounted for 99.15% and 99.01% of the total variance in outer-lip and joint FAP PCA, respectively. The SS-HMM system was used to perform facial expression experiments utilizing original eyebrow, outer-lip, or joint FAP vectors, as well as their projection weights as observations, with the purpose of determining the effect of the PCA on the recognition performance. The observations obtained as a result of the PCA produced poorer facial expression recognition performance, than the original FAP observations (see Table II). We concluded that the amount of training data was not sufficient to provide reliable PCA and produce recognition performance improvement.

In addition, in order to determine the best choice of classification features, we performed experiments using first (delta-D) and second (acceleration-A) order derivatives as observations, by appending them to the original FAP vectors or using them individually. The recognition results obtained in these experiments are also shown in Table II. Since the original FAPs produced the best results they were used as observations in the remaining experiments. The confusion matrices and the facial expression recognition rates obtained when original FAP vectors were used as observations are shown in Tables III–V. The best recognition performance was achieved when ten Gaussian mixtures were used. Tables III and IV reveal that the facial expression recognition system that utilized outer-lip FAP vectors as observations outperformed the system that utilized eyebrow

TABLE III
CONFUSION MATRIX AND AUTOMATIC FACIAL EXPRESSION RECOGNITION PERFORMANCE FOR THE SS-HMM SYSTEM THAT UTILIZES EYEBROW FAP VECTORS AS OBSERVATIONS

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 IV
CONFUSION MATRIX AND AUTOMATIC FACIAL EXPRESSION RECOGNITION PERFORMANCE FOR THE SS-HMM SYSTEM THAT UTILIZES OUTER-LIP FAP VECTORS AS OBSERVATIONS

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 V
CONFUSION MATRIX AND AUTOMATIC FACIAL EXPRESSION RECOGNITION PERFORMANCE FOR THE SS-HMM SYSTEM THAT UTILIZES JOINT FAP VECTORS AS OBSERVATIONS

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

FAP vectors. The overall expression recognition performance of the systems that utilized individually outer-lip, and eyebrow FAPs was 87.32%, and 58.80%, respectively. It can be concluded from these results that outer-lip FAPs provide more information for classifying a facial expression sequence as one of the basic six facial expressions. The recognition rates for facial expressions *fear* and *sadness* were particularly low when only eyebrow FAPs were used. This was expected, due to the fact that the changes in eyebrow FAPs that occur during these expressions are very similar to the changes that occur during several other facial expressions. This caused many classification errors, as can be seen in the confusion matrix shown in Table III. The facial expression recognition rates for the remaining four facial expressions were relatively high, even when only eyebrow FAPs were used, due to the clear motion of the outer-lips and eyebrows corresponding to these expressions. It is important to note that the recognition rates for the facial expressions *fear* and *sadness* decreased when joint observations were used, as compared to the recognition rates achieved when only outer-lip FAPs were used (see Table V). Utilizing the joint observations increased the dimensionality of the features used for classification and affected reliable training. The amount of additional information

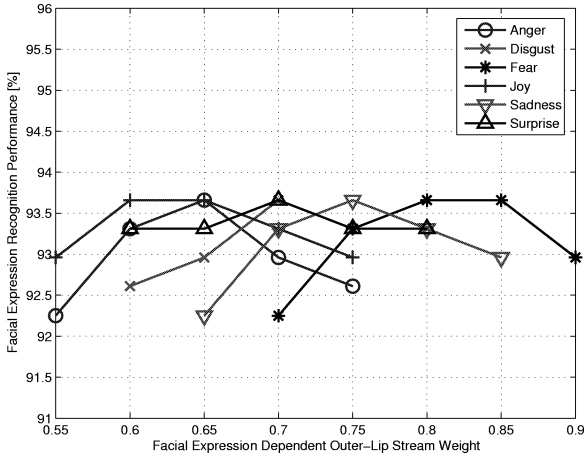

Fig. 6. Facial expression recognition performancetion obtained when each of the outer-lip weights corresponding to the six basic facial expression MS-HMMs was varied around the value that provided the best performance, while the remaining five weights were fixed.

about expressions *fear* and *sadness* contained in the eyebrow FAPs was insufficient to overcome the effect of the increased dimensionality of the observations on reliable training. Therefore, it is desirable to rely more on information contained in outer-lip FAPs when describing expressions *fear* and *sadness* and less on the information contained in the eyebrow FAPs. In order to achieve that, we utilized MS-HMMs and facial expression dependent stream weights to model the reliability of the information contained in the outer-lip and eyebrow FAP streams.

A. Multistream HMM Facial Expression Recognition Experiments

Outer-lip FAPs provided more information and produced better expression recognition results. 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. Stream weights were determined according to (4), utilizing the facial expression recognition results obtained when eyebrow and outer-lip FAPs were used individually. After the stream weights were chosen, MS-HMMs, described in the previous section, were utilized to perform facial expression recognition experiments and determine the effect of the introduction of stream reliability control to the recognition performance. The training and testing procedures were similar to procedures described in the previous section. The recognition performance obtained for the stream weights chosen based on (4) was 93.31%. In general, stream weights should be estimated on a development data set, either using (4) and the SS-HMM facial expression recognition results or by maximizing the MS-HMM recognition performance utilizing optimization techniques. Nevertheless, the MS-HMM recognition results are not very sensitive to the choice of the stream weights, provided that the larger weight is assigned to the information stream that provides 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 (4). We show in Fig. 6 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. It can be concluded from Fig. 6 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 MS-HMMs are utilized. In addition, the best recognition performance (93.66%) was obtained for several different sets of stream weights. The confusion matrix for the set of stream weights that provide the best recognition rates is shown in Table VI. The overall recognition rates for all the systems tested are shown in Table VII together with the MS-HMM facial expression dependent stream weights.

TABLE VI
CONFUSION MATRIX AND AUTOMATIC FACIAL EXPRESSION RECOGNITION PERFORMANCE FOR THE MS-HMM SYSTEM THAT UTILIZES EYEBROW AND OUTER-LIP FAP VECTORS AS OBSERVATIONS

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 VII
AUTOMATIC FACIAL EXPRESSION RECOGNITION PERFORMANCE FOR DIFFERENT OBSERVATIONS AND DIFFERENT HMM SYSTEMS

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**	0.65
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	0.65
Sadness	17.0	81.1	79.2	**96.2**	0.75
Surprise	90.6	96.9	100	100	0.7
Total	58.80	87.32	88.73	**93.66**	

The experiments were performed for different number of Gaussian mixtures used for eyebrow and outer-lip FAPs. The number of mixtures for each of the streams varied from two to ten. The recognition results obtained are shown in Fig. 7. The best recognition performance was obtained when four mixtures were used for each of the FAP groups. In general, the optimal number of Gaussian mixtures for eyebrow and outer-lip observation streams depends on the dimensionality of the observations and the amount of available training data, and should be determined on a development data set.

It is important to note that the overall facial expression recognition performance increased by approximately 5% when the MS-HMM system was used compared to the joint feature SS-HMM system. The relative reduction of the expression recognition error achieved, compared to the joint feature SS-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. The proposed system outperforms the system described in [14] which achieves the recognition performance of 84% utilizing SS-HMMs, the same facial features and the training procedure on the same database.

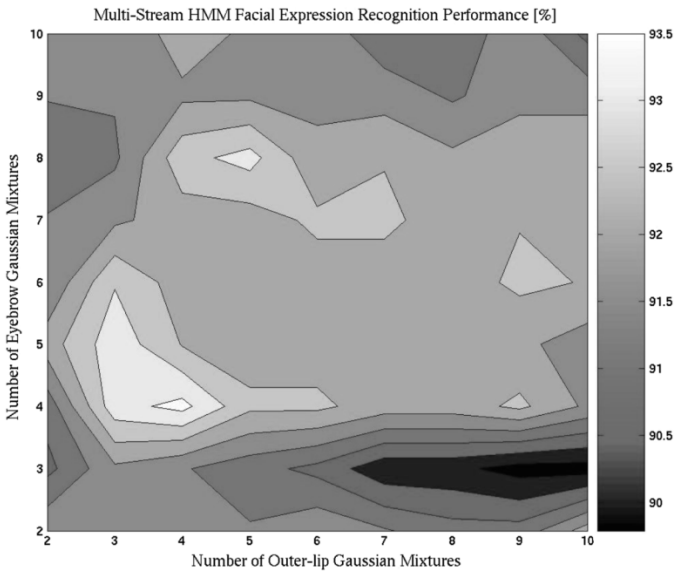

Fig. 7. Facial expression recognition performance obtained for different number of Gaussian mixtures used for outer-lip and eyebrow FAP streams.

The best recognition performance obtained (93.66%) was very similar to the performance obtained in [10] (93.33%), where image-based features are utilized on the same database.

V. Conclusions and Future Work

The automatic facial expression recognition systems developed in this work utilize eyebrow and outer-lip FAPs for classification. Using FAPs as classification features secures compliance with the MPEG-4 standard and enables utilization of already available FAP extraction tools and FAP streams for the purpose of automatic facial expression recognition. In addition, the achieved high automatic facial expression recognition rates show that significant amount of information about facial expressions is contained in FAPs. The relative reduction of the expression recognition error achieved when the MS-HMM system was used, compared to the SS-HMM system, was 44%. The introduction of expression dependent stream weights enabled modeling of the reliability of information contained in FAP streams. The stream weights were chosen based on the effectiveness with which the FAP streams describe a particular facial expression.

It is important to point out that the proposed MS-HMM facial expression recognition approach can be utilized with any kind of facial features, be it model- or image-based. For example, in order to utilize image-based facial features, an image of the face can be divided into several regions, which describe movement of the mouth, eyes, nose, cheeks, etc. The facial features extracted from such facial regions can be utilized as observation streams. The system can be easily extended to utilize more than two streams of information employing three-, four-, or five-stream HMMs. Furthermore, model-based and image-based facial features can be combined as two separate streams of information employing two-stream HMMs. As a result, both shape and intensity information can be exploited, and their contribution appropriately weighted, in order to achieve improved recognition performance. Training procedure described in previous sections can be used with such systems. The number of Gaussian mixtures and the stream weights can be determined on a development data set, based on the dimensionality of the observations, the amount of training data available, and the SS-HMM recognition performance.

It is expected that the use of additional information about facial expressions contained in the video sequences would further improve recognition performance. Sufficient amount of training data could also be used to perform PCA on outer-lip and eyebrow FAPs in order to decorrelate them and decrease their dimensionality, which should result in more reliable training of the MS-HMMs and ultimately better recognition results. It would also be very important to develop an MS-HMM system that would utilize useful acoustic information. Since facial expressions usually occur during verbal communications, audio information about facial expressions could be combined with the information contained in the person's face in order to provide improvement in automatic facial expression recognition performance. Introducing weights that depend on reliability of the information contained in a particular stream would provide better control of the information integration process and better recognition performance. The reliability of audio information could be determined based on acoustic noise and amount of information contained in them.

Investigating possible applications of expression FAPs (high-level FAPs) for facial expression analysis would also provide directions for future research.

References

[1] I. S. Pandzic and R. Forchheimer, Eds., *MPEG-4 Facial Animation*. New York: Wiley, 2002.
[2] P. Ekman and W. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press, 1978.
[3] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *Proc. 4th Int. Joint Conf. on Pattern Recognition*, 1978, pp. 408–410.
[4] K. Mase and A. Pentland, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E 74, no. 10, pp. 3474–3483, 1991.
[5] N. 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 and R. Forchheimer, Eds, U.K.: Wiley, 2002.
[6] A. W. Senior, "Face and feature finding for a face recognition system," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication*, Washington, DC, 1999, pp. 154–159.
[7] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.
[8] K.-K. Sung and T. Poggio, "Example-based learning for view based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 1, pp. 39–51, 1998.
[9] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, 2003.
[10] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *Proc. CVPR*, 2004.
[11] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and expression recognition: Development and application to human-computer interaction," in *Proc. CVPR*, 2003.
[12] J. L. Landabaso, M. Pardàs, and A. Bonafonte, "HMM recognition of expressions in unrestrained video intervals," in *Proc. ICASSP*, Hong Kong, China, 2003, pp. 197–200.
[13] M. Pardàs, A. Bonafonte, and J. L. Landabaso, "Emotion recognition based on MPEG-4 facial animation parameters," in *Proc. ICASSP*, vol. 4, Orlando, FL, 2002, pp. 3624–3627.

[14] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression detection using HMM," in *Signal Process.: Image Commun.*, vol. 17, 2002, pp. 675–688.
[15] G. W. Cottrell and J. Metcalfe, "EMPATH: Face, emotion, and gender recognition using holons," in *Neural Inform. Process. Syst.*, vol. 3, 1991, pp. 564–571.
[16] C. Padgett and G. Cottrell, "Identifying emotion in static face images," in *Proc. 2nd Joint Symp. Neural Computation*, vol. 5, Univ. California, San Diego, pp. 91–101.
[17] Text for ISO/IEC FDIS Visual, ISO/IEC JTC1/SC29/WG11 N2502, Nov. 1998.
[18] Text for ISO/IEC FDIS 14496-1 Systems, ISO/IEC JTC1/SC29/WG11 N2502, Nov. 1998.
[19] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," in *EURASIP J. Appl. Signal Process.*, vol. 2002, Nov. 2002, pp. 1213–1227.
[20] P. S. Aleksic and A. K. Katsaggelos, "An audio-visual person identification and verification system using FAPs as visual features," in *Proc. Works. Multimedia User Authentication*, Santa Barbara, CA, Dec. 11–12, 2003, pp. 80–84.
[21] Facial Animation Engine (FAE), Tecnologia Automazione Uomo S.c.r.l., Genova, Italy.
[22] C. Padgett and G. W. Cottrell, "Representing face image for emotion classification," in *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 894–900.
[23] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. 2nd IEEE Int. Conf. on Auto. Face and Gesture Recognition (FG'98)*, 1998, pp. 454–459.
[24] J. Zhao and G. Kearney, "Classifying facial emotions by back propagation neural networks with fuzzy inputs," in *Proc. Int. Conf. on Neural Information Proc.*, vol. 1, 1996, pp. 454–457.
[25] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[26] N. Oliver, A. Pentland, and F. Berard, "LAFTER: A real-time lips and face tracker with facial expression recognition," in *Proc. CVPR97*, S. Juan and P. Rico, Eds., 1997.
[27] T. Otsuka and J. Ohya, "Extracting facial motion parameters by tracking feature points," in *Proc. First Int. Conf. Advanced Multimedia Content Proc.*, 1998, pp. 442–453.
[28] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Con. on Automatic Face and Gestures Reco*, France, 2000.
[29] J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Automated facial expression recognition based on FACS action units," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Reco. (FG'98)*, Nara, Japan, 1998.
[30] J. Cohn, A. Zlochower, J. J. Lien, Y. T. Wu, and T. Kanade, "Automated face coding: A computer-vision based method of facial expression analysis," in *7th Eur. Conf. Facial Expression Measurement and Meaning*, July 1997, pp. 329–333.
[31] G. Donato, S. Bartlett, C. J. Hager, P. Ekman, and J. T. Sejnowski, "Classifying facial actions," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Oct. 1999.
[32] A. Kapoor, Y. Qi, and R. W. Picard, "Fully automatic upper facial action recognition," in *IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, 2003.
[33] M. Pantic and J. M. Rothcrantz, "Automatic analysis of facial expressions: State of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[34] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in *Proc. 2nd Workshop on Multimedia Sig. Proc.*, 1998, pp. 65–70.
[35] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Comput. Vis.*, vol. 8, no. 2, pp. 99–111, 1992.
[36] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.
[37] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP*, vol. 1, 1996, pp. 426–429.
[38] S. Okawa, T. Nakajima, and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," in *Proc. Eur. Conf. Speech Communication and Technology (EUROSPEECH)*, vol. 2, Budapest, Hungary, 1999, pp. 603–606.
[39] "Final Workshop 2000 Report Center for Language and Speech Processing," The Johns Hopkins Univ., Baltimore, MD, 2000.
[40] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multi-stream HMMs," in *6th E. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05)*, Montreux, Switzerland, Apr. 2005.

Petar S. Aleksic (M'02) received the B.S. degree in electrical engineering from University of Belgrade, Serbia, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Northwestern University, Evanston, IL, in 2001 and 2004, respectively.

He has been a Member of the Image and Video Processing Lab at Northwestern University, where he is currently a postdoctoral fellow, since 1999. His primary research interests include visual feature extraction and analysis, audio-visual speech recognition, audio–visual biometrics, multimedia communications, computer vision, and pattern recognition.

Aggelos K. Katsaggelos (F'98) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Greece, in 1979 and the M.S. and Ph.D. degrees both in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical and Computer Engineering at Northwestern University, Evanston, IL, where he is currently Professor. He was the holder of the Ameritech Chair of Information Technology (1997–2003). He is also the Director of the Motorola Center for Communications and a member of the Academic Affiliate Staff, Department of Medicine, at Evanston Hospital.

Dr. Katsaggelos is a member of the Publication Board of the *Proceedings of the IEEE*, the IEEE Technical Committees on Visual Signal Processing and Communications, and Multimedia Signal Processing, the Editorial Board of Academic Press, Marcel Dekker: Signal Processing Series, *Applied Signal Processing*, and *Computer Journal*. He has served as editor-in-chief of the *IEEE Signal Processing Magazine* (1997–2002), a member of the Publication Boards of the IEEE Signal Processing Society, the IEEE TAB Magazine Committee, an Associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990–1992), an area editor for the journal *Graphical Models and Image Processing* (1992–1995), a member of the Steering Committees of the IEEE TRANSACTIONS ON IMAGE PROCESSING (1992–1997) and the IEEE TRANSACTIONS ON MEDICAL IMAGIN (1990–1999), a member of the IEEE Technical Committee on Image and Multi-Dimensional Signal Processing (1992–1998), and a member of the Board of Governors of the IEEE Signal Processing Society (1999–2001). He is the editor of *Digital Image Restoration* (Springer-Verlag 1991), co-author of *Rate-Distortion Based Video Compression* (Kluwer, 1997), and co-editor of *Recovery Techniques for Image and Video Compression and Transmission*, (Kluwer 1998). He is the co-inventor of ten international patents, and recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), and an IEEE Signal Processing Society Best Paper Award (2001).