INNER LIP FEATURE EXTRACTION FOR MPEG-4 FACIAL ANIMATION

Zhilin Wu, and Petar S. Aleksic

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

ABSTRACT

It is very important to accurately track the mouth of a talking person for many applications, such as face recognition, audio-visual speech recognition and human computer interaction. This is in general a difficult problem due to the complexity of shapes, colors, textures, and changing lighting conditions. In this paper we develop techniques for inner lip feature extraction using a matching function based on a color module and a gradient module. Our numerical results show that the extraction using both modules outperforms that with color module only. From the extracted continuous lip contours, FAPs are extracted which are used to drive an MPEG-4 decoder. FAPs are also applied in our audio-visual automatic speech recognition (AV-ASR) system to improve the recognition rate.

1. INTRODUCTION

MPEG-4 is an emerging multimedia compression standard expected to have an important impact on a number of future consumer electronic products. It is the first audiovisual object-based representation standard as opposed to most existing frame-based standards for video representation. One of the prominent features of MPEG-4 is facial animation. By controlling the Facial Definition Parameters (FDPs) and the Facial Animation Parameters (FAPs), a face can be animated with different shapes, textures, and expressions. This kind of animation can be used in a number of applications, such as web-based customer service with talking heads or games. It can also be of great help to hearing impaired people by providing visual information. In addition, for video conferencing MPEG-4 facial animation objects could be a rather cost-efficient alternative. The animation objects can imitate a real person and animate the talking head satisfactorily as long as the parameters are extracted accurately.

Both outer and inner lip movements are important in facial animation and both convey information in lip reading and multimodal applications. However, relatively few results have been reported on the extraction and usefulness of inner lip information.

A reason for this is that inner lip extraction imposes more challenges than outer lip extraction. These challenges arise primarily due to the fact that the area inside the mouth has a similar color, texture, and luminance as that of the inner lip. In addition, teeth appear and disappear in typical conversations, which further complicates matters. Markers on the inner lip contours are still being used by researchers [1] in inner lip extraction. This is still a difficult problem, since for instance, when a speaker extrudes his lips, the inner lip boundaries are physically inside the lips. Other tracking methods include deformable templates [2], edge scanning [3] and color segmentation [4]. These methods work well when the teeth are visible. However, when the teeth are barely visible, these techniques are less accurate.

We developed an algorithm for outer lip extraction in [5, 6]; yet it is considerably harder to extract the inner lips than the outer lips using the same approach. Therefore, we propose a matching function which is based on color histogram comparison and intensity gradient searching to determine the inner lip boundaries. The inner lip boundaries can be achieved by maximizing the matching function.

The paper is organized as follows. The inner lip model and the procedure for extraction are described in Sec. 2. The inner lip extraction with a color module combined with a gradient module is described in Sec. 3, followed by the numerical evaluation in Sec. 4. The generation of FAPs is described in Sec. 5. In Sec. 6, the FAPs are applied in an AV-ASR system. Conclusions are given in Sec. 7.

2. INNER LIP MODEL AND EXTRACTION PROCEDURE

In this inner lip extraction study, two parabolas are used as inner lip templates, as was done in [2] when using deformable templates for lip tracking. Figure 1 shows the mouth model, where the outer curve is the continuous contour for the outer lip. The inner lip model consists of two parabolas, which share two inner corners, defining a line at angle θ with respect to the horizontal axis. The two curves modeling the outer and inner lips define two areas, one between them (denoted by x_1) and one inside the inner lips (denoted by x_2).

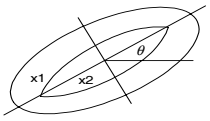

Figure 1: Outer and inner lip model

For this study, we utilize the Bernstein database [7] in which speaker rarely tilts her head. Therefore, for this set of experiments, θ is set equal to zero for all frames. The goal is to model the position of the inner lips with these two parabolas.

In order to find the inner lip boundary, we utilize four displacement variables and denote the mouth region R, as shown in Figure 2. The positions of the two parabolas can be set by these four displacement variables. If we can find a matching function f which has the maximum value when the four displacement variables are corresponding to the two parabola

inner lip boundaries, we may define the mouth region as the result of the maximization of:

$$(D_1, D_2, J_1, J_2) = \arg \max_{d_1, d_2, j_1, j_2} f(d_1, d_2, j_1, j_2) \quad (1)$$

over all possible combinations (d_1, d_2, j_1, j_2). The optimal (D_1, D_2, J_1, J_2) is then used to define two parabolas for the upper and lower inner lips.

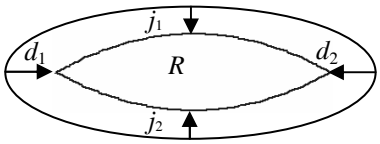

Figure 2: Inner lip extraction procedure

3. COLOR MODULE WITH GRADIENT MODULE

In the mouth area, teeth and tongue may be visible individually or simultaneously. Simple edge detection may get a number of detailed structures, where the true inner lip boundaries very likely emerge. Color histogram is well suited to describe the entire mouth area with color distribution of all the detailed structures while overlooking their physical shapes. In S. Birchfield's head tracking [8], histogram intersection was used and good tracking results were achieved. Histogram intersection was experimented in our inner lip tracking as well, yet without satisfying results because this method only compares the overlapping portion of two histograms, which may not be able to differentiate two similar histograms. In our color module, we compare the histograms for every pixel within a region, which is similar to F. Huang's face tracking [9].

Assuming that the histograms of the lip and the mouth regions, h_1 and h_2, respectively, are known, a region R can be classified as inside mouth region if

$$f_c(R) = \sum_{q \in R} \log \frac{h_2(q)}{h_1(q)} \quad (2)$$

is maximized over all possible shapes of R. Practically, both the two histograms are offset by a small number to ensure non-zero values in the whole color space. A detailed discussion can be found in [5].

Inner lip extraction only with color module sometimes fails when the inside-mouth area and the lip area have very similar color distributions. Nevertheless, the gradient along the inner lip boundaries still exists and is helpful in inner lip extraction. In this study, the gradient along a perimeter is found as in [9], which is the gradient module.

We will use the notation s for the inner lip parabolas corresponding to the inside mouth area R. The gradient is measured as the averaged gradient magnitude perpendicular to the inner lip perimeter (parabolas) for parabola s

$$f_g(s) = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{n}_s(i) \cdot \mathbf{g}_s(i)|, \quad (3)$$

where $\mathbf{n}_s(i)$ is the unit vector normal to the inner lip parabolas at pixel i, $\mathbf{g}_s(i)$ is the intensity gradient at pixel i, N is the number of pixels on the parabolas, and $(\cdot)$ is the dot product. The maximization of $f_g(s)$ is sought over s.

When both color and gradient modules are taken into account, the matching function is written as

$$f = f_c + w_g f_g, \quad (4)$$

where w_g is the relative weight for the gradient module, which is set experimentally. Since the color module provides better accuracy in inner lip extraction than the gradient module, the weight w_g is chosen such that the maximum value of $w_g f_g$ is smaller than the maximum value of f_c. The four displacement parameters (d_1, d_2, j_1, j_2) are obtained by maximizing the function in Eq. (1), where f is given by Eq. (4). Shown in Figure 3 are some inner lip extraction results with both the color and gradient modules.

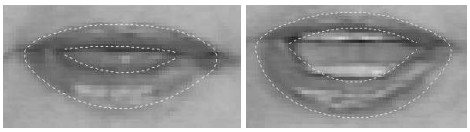

Figure 3: Lip extraction results

In order to preserve some form of temporal continuity, the values of the displacement parameters resulting from the maximization in Eq. (1) are then smoothed based on the maximum value of f [6].

4. NUMERICAL EVALUATION

To evaluate the inner lip extraction results, we manually labeled all inner lips for certain image sequences. If a pixel does not lie in both the extracted and the hand labeled inner lip region, it is treated as an error pixel. The extraction error e_t for frame t is defined as the ratio of all error pixels divided by the number of pixels inside the hand-labeled mouth area. To evaluate the extraction results for an entire sequence, we calculate the mean of the error ratios of all the frames within the sequence, denoted as e_m and

$$e_m = \frac{1}{N_f} \sum_{t=1}^{N_f} e_t \quad (5)$$

where N_f is the total number of frames within the sequence.

Figure 4 shows the mean error rates for the first 20 sequences in the Bernstein database. The dash line shows the mean error ratios of inner lip extraction results with the color module only and the solid line corresponds to both color module and gradient module extraction results. It is clearly seen that the combined module outperforms the single color module, especially in those sequences with higher extraction errors.

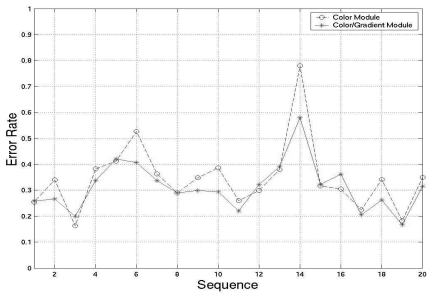

Figure 4: Mean errors e_m for color module and combined color/gradient module

5. FAP GENERATION AND EVALUATION

FAPs defined in the MPEG-4 standard are the minimum facial animation parameters responsible for describing the movements

of the face. They manipulate key feature points on a mesh model of a head to animate all kinds of facial movements and expressions. These parameters are either low level (i.e., displacement of a specific single point of the face) or high level (i.e., production of a facial expression) [10]. There are totally 68 FAPs, divided into 10 groups. We are interested in the 10 inner lip FAPs which are in group 2. Figure 5 shows the outer and inner lip position FAPs.

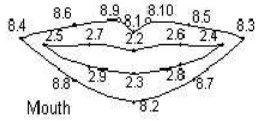

Figure 5: Outer and inner lip position FAPs

All FAPs are expressed in terms of Facial Animation Parameter Units (FAPU). These units are normalized by certain essential facial feature distances in order to give an accurate and consistent representation. Two FAPUs are involved in mouth-related FAPs: mouth width and separation between the horizontal line of nostrils and the horizontal line of neutral mouth corners. Each distance is normalized to 1024.

In the Bernstein database, the first frame in each sequence is a neutral face. Therefore we can get the two FAPUs from the first frame and apply them to all the remaining frames. The extracted outer lip [5, 6] of the first frame is the neutral outer lip. Since a neutral mouth is a closed mouth, the line connecting the two outer lip corners is the neutral inner lip. In each following frame, the positions of the 10 inner lip FAP points are compared to the neutral inner lip FAP positions and are normalized by FAPUs. This difference represents the inner lip movement.

Our system automatically reads all sequences from the Bernstein audio-video database and generates all inner lip FAPs. These parameters with our previously extracted outer lip FAPs are then input to an MPEG-4 facial animation player [11] to generate MPEG-4 sequences. Based on the visual evaluation of the synthesized video, mouth movements are very realistic and close to the original video. Some results are shown in Figure 6. Images (a) and (c) are the two original frames with extracted outer and inner lips. Images (b) and (d) are corresponding frames generated by the MPEG-4 decoder with FAPs extracted from the original frames. With the Facial Animation Engine [12] all animated MPEG-4 face sequences were well synchronized with acoustic signals.

To further objectively evaluate our inner lip extraction and FAP generation results, we compare the FAPs generated from the hand labeled inner lip contours and from our automatic extraction algorithm. The mean squared error (MSE) is used to compare two FAP sequences, A and B, and is defined by

$$MSE_{AB} = \frac{\sum_{i=1}^{f_{num}} \sum_{j=1}^{FAP_{num}} (FAP_A^{ij} - FAP_B^{ij})^2}{f_{num} \times FAP_{num}}, \quad (6)$$

(a) (b) (c) (d)

Figure 6: (a) (c) Original images with tracked outer and inner lips, (b) (d) MPEG-4 facial animations

where f_{num} denotes the number of frames in a sequence, and FAP_{num} denotes number of FAPs compared. For inner lip FAPs, FAP_{num} equals to ten. The percentage normalized mean error (PNME) is calculated from MSE by

$$PNME_{AB} = \frac{\sqrt{MSE_{AB}}}{1024} \times 100 \quad (7)$$

and represents the percentage error. Figure 7 shows the PNME calculated for the first 20 sequences of the Bernstein database. The average percentage error of the first 20 sequences is 5.59%, which is considered to be relatively low.

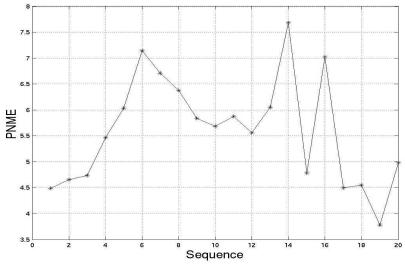

Figure 7: PNME calculated for the inner lip FAPs of hand labeled and automatic inner lip contours

6. AUDIO-VISUAL AUTOMATIC SPEECH RECOGNITION

Another objective metric we propose in evaluating the performance of the described lip-tracking algorithm, or the quality of the automatically extracted FAP sequences is in terms of the increase in speech recognition accuracy of an AV-ASR system over an audio only ASR (A-ASR) system. It provides a means to quantifying the amount of speechreading information contained in the FAP sequences, of interest in most, if not all, applications.

The use of visual information in addition to audio, improves speech understanding especially in noisy environments. Improving ASR performance, by exploiting the visual information of the speaker's mouth region is the main objective of AV-ASR [6, 13]. In this work we utilize the extracted FAPs as visual features, in the AV-ASR system proposed in [6], which is show in Fig. 8.

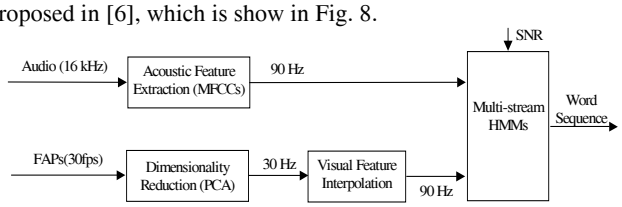

Figure 8: Audio-visual system for ASR.

The audio and visual speech information can be combined in AV-ASR systems using several techniques. A late integration approach (used in this work) is to model audio and visual features as separate feature streams, by use of multi-stream HMMs. The advantage of this approach is that stream log-likelihoods can be combined using the weights that capture the reliability of each particular stream. Therefore, we can choose larger values for the acoustic stream weights in order to rely more on acoustic data when there is not much acoustic noise in

the environment, and smaller values for the acoustic weights in the presence of significant acoustic noise.

In order to decrease the dimensionality of the visual feature vector, Principal Component Analysis (PCA) was performed on the 10-dimensional FAP vectors. The first six, two and one eigenvectors represent 99.5%, 99%, and 97% of the total statistical variance, respectively. When choosing the dimensionality of the visual features to be used for AV-ASR one should have in mind the trade-off between the number of HMM parameters that have to be estimated and the amount of the speechreading information contained in the visual features. Therefore we decided to use two-dimensional projection weights as visual features. These features were used in all AV-ASR experiments.

The Mel-Frequency Cepstral Coefficients (MFCC), signal energy and first and second derivatives, widely used in speech processing, were used as audio features. Since MFCCs were obtained at a rate of 90Hz, while FAPs at a rate of 30Hz, FAPs were interpolated in order to obtain synchronized data.

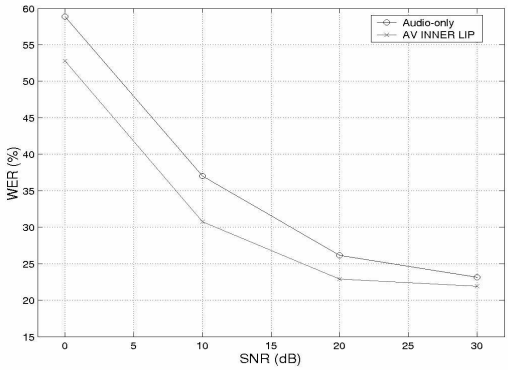

Figure 9. Audio-only, and audio-visual system WERs vs. SNR

6.1 AV-ASR Experiments

All ASR systems were developed using the HTK toolkit version 3.2 [14]. The experiments used the portion of the audio-visual Bernstein database with the female speaker. Context dependent phoneme models, biphones, were used as speech units. HMMs used for state-synchronous multi-state systems were left-to-right, with 3 states. Approximately 80% of the data was used for training, 18% for testing, and 2% as a development set for obtaining roughly optimized stream weights, word insertion penalty and the grammar scale factor. The bi-gram language model, used for decoding, was created based on the transcriptions of the training data set, and its perplexity was approximately 40. The same training and testing procedures were used for both audio-only and audio-visual experiments. To test the algorithm over a wide range of SNRs (0, 10, 20, and 30 dB), white Gaussian noise was added to the audio signals. The ASR results are summarized in Figure 9, where WER represents the word error rate. It can be observed that the A-ASR performance is severely affected by additive noise.

The AV-ASR system was trained using the extracted FAP sequences as visual features. The testing was performed for FAP sequences describing inner lip movement and the results were compared to the A-ASR results. As can be clearly seen, the AV-ASR system performs better than the A-ASR system, for all SNR values. The relative reduction in WER achieved by the AV-ASR system, when FAPs describing inner lip movement were used for testing, compared to the audio-only WER, ranges from 5% for the noisy audio with SNR of 30 dB to 10% for SNR of 0 dB. The improvement in the ASR performance achieved when FAP sequences describing inner lip movement were used was not as large as in the case of FAPs describing outer lip movement [6]. It is important to point out that the considerable performance improvement was achieved with the use of only two-dimensional visual features vectors. The AV-ASR performance improvement also confirms the usefulness of the extracted FAPs in speechreading applications and therefore the "accuracy" of the proposed lip-tracking algorithm.

7. CONCLUSIONS

Based on our earlier results of the outer lip extraction, the inner lip contours are extracted using a matching function with a color module and a gradient module with a temporal smoothing constraint. Numerical evaluation of the inner lip tracking results indicates that the merged module has better results than that of the color model only. The continuous lip contours were then used to estimate MPEG-4 FAPs. With MPEG-4 facial animation decoder, a realistic talking head was achieved. The inner lip FAPs were also used by our AV-ASR system in order to improve ASR performance over a wide range of SNRs.

8. REFERENCES

[1] P. Daubias and P. Deleglise, "Statistical lip-appearance models trained automatically using audio information," *EURASIP J. on Applied Signal Processing*, vol. 2002, no. 11, pp. 1202-1212, 2002.

[2] A.L. Yuille, P.W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int J. of Computer Vision*, vol. 8(2), pp. 99-111, 1992.

[3] T. Goto, S. Kshirsagar, and N. Magnenat-Thalmann, "Real time facial feature tracking and speech acquisition for cloned head," *IEEE Signal Processing Magazine*, Vo. 18, No. 3, pp 17-25., May, 2001

[4] E. Cosatto and H.P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. On Multimedia*, pp. 152-163, vol. 2, no. 3, Sept. 2000.

[5] Z. Wu, A.S. Petar, and A.K. Katsaggelos, "Lip tracking for MPEG-4 facial animation," *Int. Conf. on Multimodal Interfaces (ICMI'02)*, Pittsburgh, PA, Oct. 2002.

[6] P.S. Aleksic, J.J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP Journal on Applied Signal Processing, special Issue on Audio-Visual Speech Processing*, pp. 1213-1227, November 2002.

[7] L. E. Bernstein, Lipreading Corpus V-VI: Disc 3., Gallaudet University, Washington, D.C., 1991.

[8] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," *IEEE Conf. on Computer Vision and Pattern Recog*., Santa Barbara, CA, June, 1998.

[9] F.J. Huang and T. Chen, "Tracking of multiple faces for human-computer interfaces and virtual environments," *IEEE Int. Conf. On Multimedia and Expo*., New York, July 2000.

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

[11] F. Lavagetto, R. Pockaj, "The Facial Animation Engine: Toward a High-Level Interface for the Design of MPEG-4 Compliant Animated Faces", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 2, pp. 277-289, March, 1999.

[12] http://www-dsp.com.dist.unige.it/~pok/RESEARCH/MPEG/fae.htm

[13] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech" To appear: *Proc. IEEE*, 2003.

[14] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Entropic Ltd., Cambridge, 2002.