IMPROVED FACE AND FEATURE FINDING FOR AUDIO-VISUAL SPEECH RECOGNITION IN VISUALLY CHALLENGING ENVIRONMENTS

Jintao Jiang[1], *Gerasimos Potamianos*[2], *Harriet Nock*[2], *Giridharan Iyengar*[2], *Chalapathy Neti*[2]

[1]Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA
[2]IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
jjt@icsl.ucla.edu, {gpotam, hnock, giyengar, cneti}@us.ibm.com

ABSTRACT

Visual information in a speaker's face is known to improve robustness of automatic speech recognizers. However, most studies in audio-visual ASR have focused on "visually clean" data to benefit ASR in noise. This paper is a follow up on a previous study that investigated audio-visual ASR in visually challenging environments. It focuses on visual speech front end processing, and it proposes an improved, appearance based face and feature detection algorithm that utilizes Gaussian mixture model classifiers. This method is shown to improve the accuracy of face and feature detection, and thus visual speech recognition, over our previously used baseline system. In turn, this translates to improved audio-visual ASR, resulting in a 10% relative reduction of the word-error-rate in noisy speech.

1. INTRODUCTION

Visual speech has been shown to improve ASR in noise [1][2]. These studies, which have used "visually clean" data, have contributed to establishing the rationale for audio-visual ASR (AV-ASR). However, recording high-quality visual speech is not always a feasible or affordable way to augment ASR in noise. Indeed, real applications often encounter visually challenging environments with variations in the speaker's head pose and characteristics, as well as in the environment lighting and background, and the video acquisition sensor's quality. However, very few studies [3][4] have investigated AV-ASR performance in realistic and non-ideal environments, where in addition to possibly noisy audio, the visual channel quality is poor. Accurate detection and acquisition of facial features is an indispensable first step for AV-ASR, and, as expected, it becomes an issue in visually challenging domains. Indeed, in a previous study [3], where face and feature detection was performed using Fisher's Linear Discriminant Analysis (LDA), although visual speech still benefited ASR in noise, the *visual-only* speech recognition error rate of connected-digit strings degraded significantly from a studio-like environment (29.5%) to more challenging office (46.1%) and moving-car (66.7%) domains. This degradation was attributed to poor face and facial feature detection performance in such domains. Improving their performance is the subject of this paper.

In the literature, most studies on face detection use appearance-based methods based on neural networks [5], LDA [3][4][6][7], support vector machines [8], eigenfaces [9], hidden Markov models (HMM) [10], or Gaussian mixture models (GMM) [11]. These studies have used a strong learner for face detection or recognition. Viola and Jones [12] proposed a quite different but very fast face detection algorithm based on AdaBoost and feature selection (weak learner).

Our baseline AV-ASR system uses appearance based face and facial feature detection, by means of an LDA projection and eigenfaces [3][4][6][7]. In this paper, and in order to improve its performance in visually challenging domains, we extend it to utilize GMM classifiers. In addition, the effect of various algorithm parameters, such as the face template size is also explored. To benchmark the visual front end improvement due to the GMM based face and feature detection algorithm, introduced in this paper, we report visual-only and audio-visual speech recognition results on a visually challenging dataset.

The paper is organized as follows: Section 2 describes the baseline visual front end. Section 3 reviews the main components of the AV-ASR system. Section 4 presents the GMM based face and facial feature detection algorithm. Section 5 briefly describes the database used in this work. The results are presented in Section 6 and a summary in Section 7.

2. BASELINE VISUAL FRONT END

The visual front end is based on a previous IBM face and facial feature detection system [3][4][6][7]. The method is a two-stage algorithm, and it is described in Figure 1. Given the video of a spoken utterance, face tracking is applied to find faces in video images. If faces are found, 26 facial features are then subsequently located. At the face detection stage, the images are scanned at different scales to find face candidates, since face size is unknown. At the feature detection stage, a detected face image is rotated and scaled based on the parameters of the detected face. Then, each feature is searched for over a specific region (based on statistics on training). At each stage, face or feature candidate vectors $\boldsymbol{x}$ (consisting of the grey-level pixel values, normalized in a rectangular template) are scored by a two-class Fisher discriminant as well as their "*distance from feature space*" (DFFS), defined as

$$DFFS = |\boldsymbol{x}|^2 - \sum_{i=1}^{12} (\boldsymbol{x} \cdot \boldsymbol{v}_i)^2 \,, \qquad (1)$$

where $\boldsymbol{v}_i$ are the eigenvectors of a training set of vectors $\{\boldsymbol{x}\}$. In this work, 26 facial features (see Table 1) are tracked and searched hierarchically. The top-level features are searched first using a rectangular template of size 14x11 pixels, with respect to a normalized eye separation of 15 pixels (see the large box in

Figure 1). The bottom-level features are subsequently detected, over search regions defined relative to the top-level features, using a template of 14x11 pixels, with respect to a normalized eye separation of 45 pixels (see the small box in Figure 1).

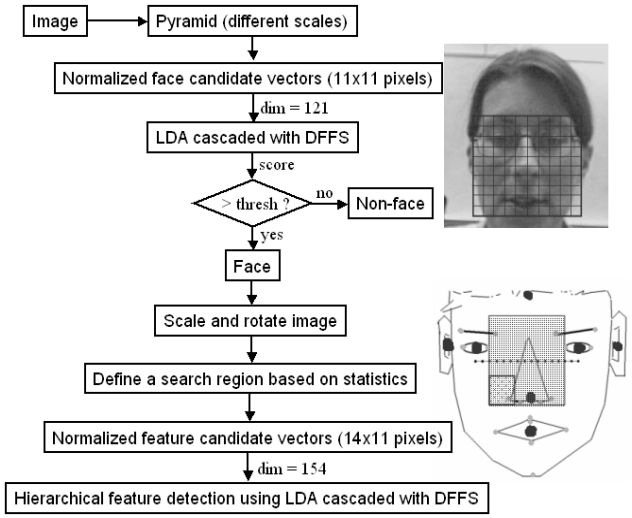

Figure 1. Baseline face and feature detection system. The templates for face detection, as well as top and bottom level feature finding are also depicted, relative to eye-separation.

Table 1. 26 facial features used in this study (L.:Left, R.: Right, I.: Inner, O: Outer). Top-level features are depicted in boldface.

Hairline	L.I.Eye	L.Nostril	L.O.Eyebrow
R.I.Eyebrow	**L.Eye**	L.Nose	Nose bridge
L.I.Eyebrow	L.O.Eye	R.Mouth	T.O.Lip
R.Ear	**L.Ear**	**Mouth**	T.I.Lip
R.O.Eye	R.Nose	L.Mouth	L.O.Lip
R.Eye	R.Nostril	Chin	
R.I.Eye	**Nose**	R.O.Eyebrow	

3. AUDIO-VISUAL ASR SYSTEM COMPONENTS

There are three main components of an AV-ASR system: The visual front end design, the audio-visual integration strategy, and the speech recognition method. This work focuses on the visual front end design. Section 2 described how face and facial features are located. The next step is to decide the type of visual speech features to be fed into the AV-ASR system. As summarized in [3], there are three possibilities in this regard: Appearance-based features that typically seek a suitable transformation of the pixel values within a visual region-of-interest (ROI), shape-based features that consist of a geometric or statistical representation of the lip contour, and combination of the two strategies.

In this study, an appearance-based method is used. The procedure to obtain visual speech features is the same as in [3] and is shown in Figure 2. Based on the face and feature detection results, the mouth location, size, and orientation are tracked and then smoothed over a temporal window to improve robustness. Based on the resulting estimates, a 64x64 pixel ROI is obtained for every video frame. This ROI is further normalized to alleviate differences in rotation, size, and lighting. Subsequently, a 2-D DCT, LDA, and a maximum likelihood linear transformation (MLLT) [2] are applied, and finally, visual features of dimension 41 are derived.

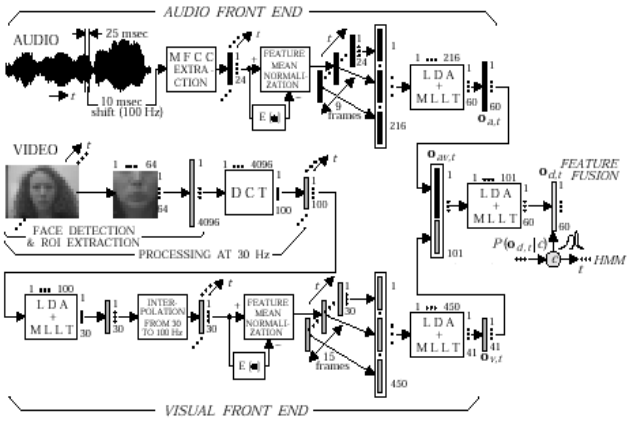

Figure 2. Block diagram of the AV-ASR system employed in this study. Time-synchronous, 60-dimensional audio feature vectors and 41-dimensional visual feature vectors are extracted both at a 100 Hz rate. A simple feature fusion method is used.

In addition to visual features, time-synchronous audio features are extracted at a frame rate of 100 Hz. First, 24 mel-frequency cepstral coefficients of the speech signal are computed and mean normalized to provide static features. Then, an LDA/MLLT cascade [2] is applied to produce 60-dimensional audio features (see also Figure 2).

Following feature extraction, a feature fusion strategy is used. The 101-dimensional concatenated audio-visual vectors are then projected onto a 60-dimensional space by an LDA/MLLT. The speech recognition module employs a hidden Markov model (HMM) with Gaussian mixture emission probabilities [3][4]. The HMM parameters are obtained by the traditional maximum likelihood approach, based on available training data.

4. GMM BASED VISUAL FRONT END

To apply the GMM based face and feature detection algorithm, first a 2-D DCT transform is performed to de-correlate and "compress" the input vectors. Thus, a reduced number of DCT components are used for GMM based modeling without significant performance degradation. This section describes the 2-D DCT transform, GMM based face detection, GMM based feature detection, face template size adjustment, and training feature sample generation.

4.1. 2-D DCT transform

A separable 2-D DCT transform is applied. Let $\boldsymbol{I}$ denote the normalized face or feature candidate of size MxN. Then the 2-D DCT transform $\boldsymbol{D}$ is computed as:

$$D_{i,j} = C_{i,j} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left[I_{m,n} \cdot \cos\left(i\pi \frac{2m+1}{2M}\right) \cdot \cos\left(j\pi \frac{2n+1}{2M}\right) \right] \qquad (2)$$

After the DCT transform, the $\boldsymbol{D}$ matrix is organized into a vector using a zig-zag scan, with only the first K = 32 or 50 coefficients used in classification.

4.2. GMM based face detection

Face detection is a two-class problem. To account for the variations in face size, pose, lighting, and speaker characteristics, a GMM based classification method is employed. The face samples are extracted from a limited number of video images manually annotated. To minimize the mismatch between training and test data, the face selection box is slightly translated and rotated to produce additional samples (a total of 10 variations for each annotated face). The non-face samples are randomly chosen from the video images (but located away from the annotated faces). Again, 10 random non-faces are generated for each annotated face. For both face and non-face GMMs, up to 50 mixture components are trained. Note that our implementation of EM training automatically selects the actual number of mixture components. The DCT vector dimensionality is 50.

4.3. Face template size

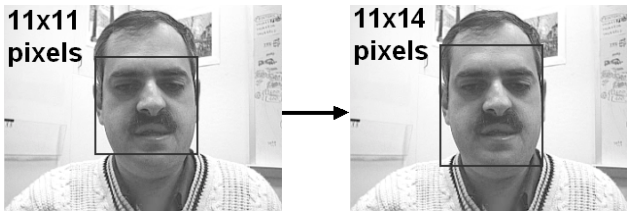

Figure 3. Adjustment of the face template.

In [3][4][6][7], the face template is chosen to be an 11x11 pixel square. The ratio of face height and face width is different from person to person. However, for most people, the faces look closer to rectangles than squares. Therefore, in this study, the face template is adjusted from 11x11 to 11x14 pixels, as shown in Figure 3.

4.4. GMM based facial feature detection

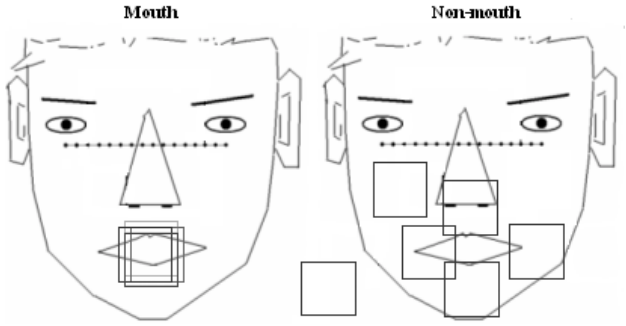

Figure 4. An example of mouth (left) and non-mouth (right) sample generation.

Facial feature detection is also a two-class problem. Similarly to the face detection, manually annotated feature samples are also slightly translated to produce additional samples (totally 5 variations for each annotated feature), whereas non-feature samples are chosen randomly from the video images with constraints: 4 random samples are generated near the annotated feature; one random sample is generated in medium distance from the annotated feature; and one random sample is located far away from the annotated feature (see Figure 4). Note that for feature detection, the top-level feature detection is performed in a low-resolution image, and thus a small translation would result in large deviation from the annotated position. Therefore, for top-level features, there is only one sample for each annotated feature. For both feature and non-feature GMMs, up to 50 mixture components are trained. The DCT vector dimensionality is 32.

5. DATABASE

Most AV-ASR research has concentrated on databases collected in ideal visual conditions. In this paper, we examine more realistic and challenging data from [3]. The corpus was captured using a laptop-based audio-visual data collection prototype. Wideband audio was recorded using the built-in laptop microphone and uncompressed video by means of an inexpensive web-cam, utilizing the USB 2.0 interface. The video (at 30 frames/sec) was obtained with automatic gain control present and at a 320x280 pixel size. The database subjects were recorded in their own offices without the use of a teleprompter, and thus, lighting, background, and head-pose vary greatly. A total of 109 subjects uttering connected digit strings are available.

The database is divided into a training set (4591 utterances; about 6 hours) and test set (537 utterances; about 40 minutes). For face and feature detection purposes, 1368 face images from the training set are manually annotated for building face and feature models (see Figure 5). In addition, 253 face images from the test set are also annotated for evaluating visual front end performance.

Figure 5. An example of manually annotated facial features.

6. RESULTS

We now proceed to report results on face detection, facial feature detection, and ASR experiments on the database.

6.1. AV-ASR paradigm

In this study, a multi-speaker scenario is considered, where separate data from all subjects are used for both training and testing. In addition to the original database acoustic signal (SNR ≈ 15 dB), audio-only and AV-ASR are also considered on artificially corrupted audio by additive "speech babble" (two cases: SNR ≈ 8 dB and SNR ≈ 4 dB). A two-stage stack decoding algorithm is employed for recognition, with unknown digit-string length. HMMs with 159 context dependent states and approximately 3.2k Gaussian mixture components are used.

6.2. Visual front end performance

The face detection accuracy is expressed as a percentage of detected faces within 20% of their manually annotated location, orientation, and scale. Table 2 lists the face detection accuracy using different methods. Clearly, GMM based face detection outperforms a simple LDA algorithm. Changing the face template size from 11x11 to 11x14 pixels also improves face detection performance. Hereafter, the face template size is 11x14, if not specified.

Table 2. Face detection accuracy.

Algorithm	LDA	GMM	
Face template size	11x11	11x11	11x14
Accuracy, %	91	97	98

Figure 6 shows the facial feature detection error rate. A feature is not considered detected if the location error is larger than 10% of the annotated eye separation. The figure shows that the improvement from GMM based face detection translates to improvement in LDA based feature detection, and that GMM based feature detection clearly outperforms LDA based feature detection.

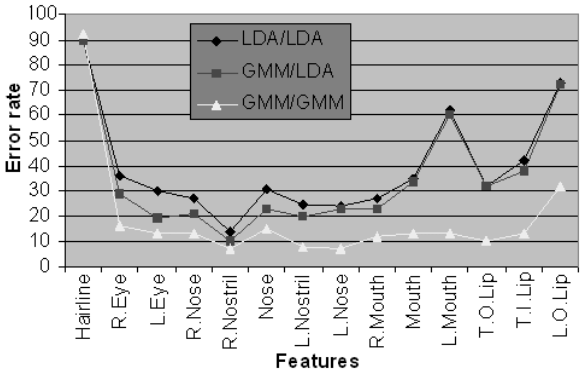

Figure 6. Facial feature detection error rate using LDA/LDA, GMM/LDA, and GMM/GMM for face/feature detection.

As mentioned in the introduction, face and feature detection is a crucial step in AV-ASR. Improvements in facial feature detection should result in more accurate ROI extraction, and thus better visual-only ASR. This is the case (see Table 3), the GMM based face and feature detection algorithm contributes to a visual-only WER reduction of about 8% (46.51% → 42.96%).

6.3. AV-ASR

Potamianos and Neti [3] report that the visual modality remains of benefit to ASR even in visually challenging environments. In Table 3, the results show that the improvement in visual-only ASR translates to improved AV-ASR. The GMM based face and feature detection improves visual-only ASR by about 8% (46.51% → 42.96%) and AV-ASR in noise by about 10% (13.89% → 12.51%) in WER, compared to the LDA based algorithm. Note that in clean audio environments, the good AV-ASR performance (WER = 2.24%), the small test set, and the nature of EM based HMM training may have contributed to the higher WER (2.38%) for the GMM based feature detection than that (2.24%) for an LDA based method.

Table 3. WER for visual-only (VI), audio-only (AU), and audio-visual ASR with different visual front ends (face/feature detection: LDA/LDA; GMM/LDA; GMM/GMM). Three acoustic conditions are considered: clean (A0, SNR ≈ 15 dB) and two with additive babble noise (A1, SNR ≈ 8 dB; A2; SNR ≈ 4 dB).

	AU	Face and feature detection algorithm		
		LDA/LDA	GMM/LDA	GMM/GMM
VI		46.51	45.42	42.96
A0+V	2.51	2.53	2.24	2.38
A1+V	12.64	7.44	7.31	6.89
A2+V	24.91	13.89	13.22	12.51

7. DISCUSSION

We investigated a new visual front end for AV-ASR in a "challenging" environment that presents difficulties for accurate visual processing. The new visual front end includes GMM based face and feature detection, a face template size of 11x14, and random non-feature generation with constraints. The results show that this visual front end improves AV-ASR over a previously used baseline system in visually challenging environments.

8. REFERENCES

[1] E.D. Petajan, *Automatic lipreading to enhance speech recognition*, Ph.D. Thesis, U. Ill. Urbana-Champaign, 1984.

[2] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual speech recognition," *Proc. ICASSP*, Salt Lake City, pp. 165-168, 2001.

[3] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," *Proc. Eurospeech*, Geneva, pp. 1293-1296, 2003.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, 91: 1306-1326, 2003.

[5] H. Rowley, S. Baluja, and T. Kanade. "Neural network-based face detection," *IEEE Trans. Patt. Anal. Mach. Intell.*, 20: 23-38, 1998.

[6] A.W. Senior, "Face and feature finding for a face recognition system," *Proc. Int. Conf. Audio- Video-based Biometric Person Authent.*, Washington, pp. 154-159, 1999.

[7] A.W. Senior, "Recognizing faces in broadcast video," *IEEE Work. Real-Time Analysis and Tracking of Faces and Gestures in Real-Time Systems*, Kerkyra, 1999.

[8] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 1997.

[9] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Patt. Anal. Mach. Intell.*, 19: 711-720, 1997.

[10] A.V. Nefian and M.H. Hayes III, "Face detection and recognition using hidden Markov models," *Proc. ICIP*, Chicago, pp. 141-145, 1998.

[11] K. Sung and T. Poggio, "Example-based learning for view-based face detection," *IEEE Trans. Patt. Anal. Mach. Intell.*, 20: 39-51, 1998.

[12] P. Viola and M. Jones, "Robust real-time object detection," *Proc. Int. Work. on Statistical and Computational Theories of Vision*, Vancouver, 2001.