Lipreading Using Profile Versus Frontal Views

Patrick Lucey*
Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
Brisbane, QLD 4001, Australia
Email: p.lucey@qut.edu.au

Gerasimos Potamianos
Human Language Technology Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
Email: gpotam@us.ibm.com

Abstract— **Visual information from a speaker's mouth region is known to improve automatic speech recognition robustness. However, the vast majority of audio-visual automatic speech recognition (AVASR) studies assume frontal images of the speaker's face. In contrast, this paper investigates extracting visual speech information from the speaker's profile view, and, to our knowledge, constitutes the first real attempt to attack this problem. As with any AVASR system, the overall recognition performance depends heavily on the visual front end. This is especially the case with profile-view data, as the facial features are heavily compacted compared to the frontal scenario. In this paper, we particularly describe our visual front end approach, and report experiments on a multi-subject, small-vocabulary, bimodal, multisensory database that contains synchronously captured audio with frontal and profile face video. Our experiments demonstrate that AVASR is possible from profile views, however the visual modality benefit is decreased compared to frontal video data.**

I. Introduction

Over the past two decades, considerable research activity has concentrated on utilizing visual speech extracted from a speaker's face in conjunction with the acoustic signal, in order to improve robustness of automatic speech recognition (ASR) systems [1]. Even though a great deal of progress has been achieved in audio-visual ASR (AVASR), so far the vast majority of works in the field have focussed on using video of a speaker's fully frontal face. This is mainly due to the lack of any large corpora which can accommodate poses other than frontal. But as more work is being concentrated within the confines of a "meeting room" [2] or "smart room" [3] environment, data is becoming available that allows visual speech recognition from multiple views to become a viable research avenue.

In the literature, only three studies were found to be related to visual speech from side views. In the first paper, Yoshinaga et al. [4] extracted lip information from the horizontal and vertical variances from the optical flow of the mouth image. In this paper, no mouth detection or tracking was performed. Yoshinaga et al. [5], refined their system by incorporating a mouth tracker which utilizes Sobel edge detection and binary images, and uses the lip angle and its derivative for the visual feature on a limited data set. The improvement sought from these primitive features was minimal as expected, essentially due to the fact that only two visual features were used, compared to most other frontal-view systems that utilize significantly more features [1]. The third study was a comprehensive psychological study conducted by Jordan and Thomas [6]. Their findings were rather intuitive, as the authors determined that human identification of visual speech became more difficult as the angle (from frontal to profile view) increased. To the best of our knowledge, no other attempt to solve this particular problem has been made. As such, we believe this paper to be the first real attempt in determining how much visual speech information can be automatically extracted from profile views, and to compare this with visual speech information obtained from frontal images.

The task of recognizing visual speech from profile views is in principle very similar to that of frontal views, requiring to first detect and track the mouth region and subsequently to extract visual features. However, this problem is far more complicated than the frontal case, because the facial features we require to detect and track the mouth lie in a much more limited spatial plane, as can be viewed in Fig. 1. Clearly, much less data is available compared to that of a fully frontal face, as many of the facial features that we are interested in (eyes, nostrils, mouth, chin area etc.) are fully or partially occluded. In addition, the search region for all visible features is approximately halved, as the remaining features are compactly confined within the profile facial region. These facts remove redundancy in the facial feature search problem, and

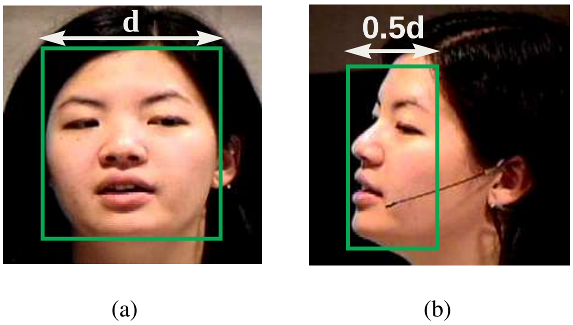


Fig. 1. Synchronous (a) frontal and (b) profile views of a subject recorded in the IBM smart room (see Section IV). In the latter, visible facial features are "compacted" within approximately half the area compared to the frontal face case, thus increasing tracking difficulty.

* Work conducted during Patrick Lucey's internship with the IBM T.J. Watson Research Center.

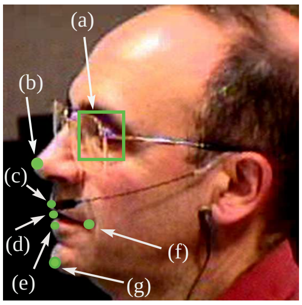


Fig. 2. Example of the points labeled on the face: (a) left eye, (b) nose, (c) top mouth, (d) mouth center, (e) bottom mouth, (f) left mouth corner, and (g) chin. The center of depicted bounding box around the eye defines the actual feature location.

therefore make robust mouth tracking a much more difficult endeavor.

Nevertheless, one can still achieve mouth region tracking by employing techniques similar to frontal facial feature detection. In particular, in the AVASR literature, most systems use appearance-based methods for face and facial feature detection. Some are based on "strong" classifiers, such as neural networks [7], support vector machines (SVMs) [8], eigenfaces [1], hidden Markov models (HMMs) [9], or Gaussian mixture models (GMMs) [10], and others on cascades of "weak" classifiers, such as the Adaboost framework of Viola and Jones [11], later extended by Leinhart and Maydt [12]. In this paper, we use the latter approach to first detect the face and subsequently the facial features in profile views, as described in more detail in Section II.

Following that, Section III focuses on the AVASR system description, namely visual feature extraction based on the tracked mouth region and audio-visual fusion. In addition, details of a number of systems used in our experiments are given, including a baseline frontal-view AVASR system that has been refined in our previous work [1]. Section IV presents our experimental results, and, finally, Section V concludes the paper with a summary and a few remarks.

II. Mouth Detection and Region-of-Interest Extraction from Profile Views

For the task of mouth detection and *region-of-interest* (ROI) extraction, we devised a similar strategy to that of Cristinacce et al. [13], where we used a boosted cascade of classifiers based on simple Haar-like features to detect the face and subsequent facial features. These classifiers were generated using OpenCV libraries [14].

The positive examples used for training these classifiers were obtained from a set of 847 training images, with 17 manually labeled points for each face. Due to the compactness of the facial features within the dataset, we initially utilized only 7 of the 17 manually labeled points, namely the left eye, nose, top of the mouth, center of mouth, bottom of the mouth, left mouth corner and chin, as depicted in Fig. 2. This provided 847 positive examples for all 7 facial features. Approximately 5000 negative examples were used for each facial feature. These negative examples consisted of images of the other facial features that surrounded its location, as these areas would be the most likely to cause false alarms. The face training set was further augmented by including rotations in the image plane by ±5 and ±10 degrees, providing 4235 positive examples. A similar amount of negative examples of the background were also employed in the training scheme. As the facial features were located so close to each other (a matter of pixels in some cases), we opted not to include rotated examples of the facial features.

TABLE I
Facial feature point detection accuracy results

Facial Feature	Accuracy (%)
Left Eye	**86.49**
Nose	81.08
Top Mouth	78.37
Center Mouth	81.08
Bottom Mouth	72.97
Left Mouth Corner	**86.49**
Chin	62.16

A dilemma we experienced was on selecting appropriate facial feature points to use for image normalization. In the frontal face scenario, eyes are predominately used for this task, but in the profile-view case, we don't have the luxury of choosing two geometrically aligned features. We instead chose to use the nose and the chin, with a normalized constant (K) distance of 64 pixels between them. By using the distance between these two points, the problem of head pose variation was minimized compared to the other possibilities (such as eye to nose distance etc.). The top mouth, center mouth, bottom mouth and left mouth corner were trained on templates of size 10×10 pixels, based on normalized training faces. Both nose and chin classifiers were trained on templates of size 15×15 pixels, and the eye templates were larger, 20×20 pixels. The normalized positive face examples were templates of size 16×16.

Due to the lack of manually labeled faces available, all classifiers were tested on a small validation set of 37 images, which gave us an indication on what particular features would give us the best chance of reliably tracking the detected features (see Table I). A feature was not considered detected if the location error is larger than 10% of the annotated distance between the nose and the chin. As the left eye and left mouth corner yielded the best results, we decided to use these two points for scale normalization. The only difference between using the left eye and left mouth corner, compared to the nose and chin is changing the scaling factor K from 64 to 45. The face detection accuracy on this test set was 100%. As no manual labels for the face bounding box was available, the accuracy was determined upon inspection.

The profile mouth detection and tracking of our system is outlined in Fig. 3. Given the video of a spoken utterance,

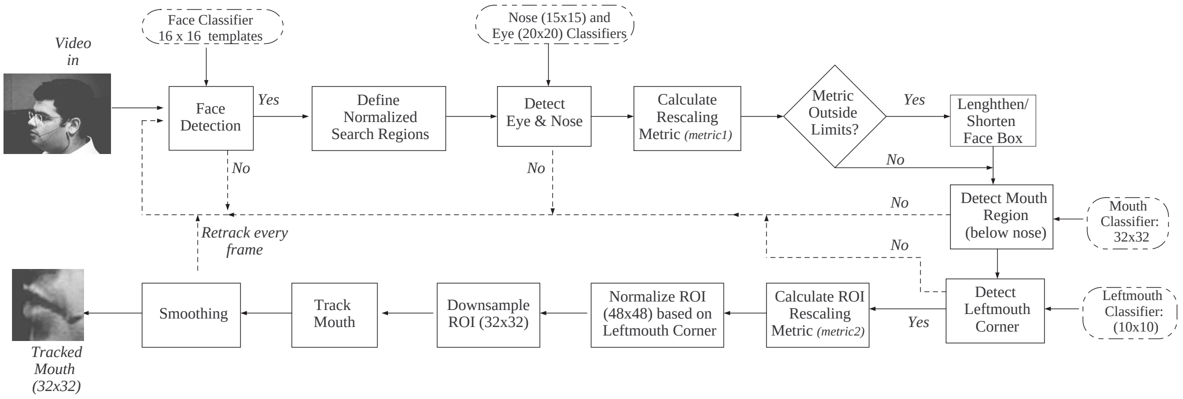

Video in
Face Classifier 16 x 16 templates
Face Detection
Yes
No
Define Normalized Search Regions
Nose (15x15) and Eye (20x20) Classifiers
Detect Eye & Nose
No
Calculate Rescaling Metric (metric1)
Metric Outside Limits?
Yes
No
Lenghthen/ Shorten Face Box
Detect Mouth Region (below nose)
Mouth Classifier: 32x32
No
Retrack every frame
No
Detect Leftmouth Corner
Leftmouth Classifier: (10x10)
Yes
Calculate ROI Rescaling Metric (metric2)
Normalize ROI (48x48) based on Leftmouth Corner
Downsample ROI (32x32)
Track Mouth
Smoothing
Tracked Mouth (32x32)

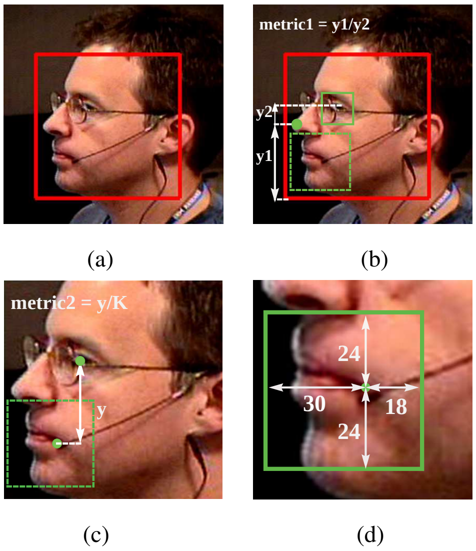
Fig. 3. Block diagram of the face and mouth detection and tracking system for profile views.

face detection is first applied to estimate the location of the speaker's face at different scales as the face size is unknown. Once the face was detected, the left eye and nose were searched over specific regions of the face (based on training data statistics). During developing this system, we found that the bottom of the face bounding box was often far below the bottom of the subject's actual face, or well above it. As the face box defines the search region for the various facial features, this caused the system to miss detecting the lower regions of the face. To overcome this, we used the ratio (*metric1*) of the vertical eye to nose distance, over the vertical nose to bottom of the face bounding box distance. If *metric1* was below a fuzzy threshold (again determined by training statistics), the box was lengthened, or if it was above the threshold then it was shortened. We found this greatly improved the detection of the generalized mouth area (trained on normalized 32×32 mouth images), which was located next. This step is illustrated in Fig. 4(b).

Once the generalized mouth region was found, the left mouth corner was detected. The next step was to define a scaling metric, so that all ROI images would be normalized to the same size. As mentioned previously, the ratio (*metric2*) of the vertical left eye to left mouth corner distance over some constant K (45) was used to achieve this (see Fig. 4). A $(48 \times 48) \cdot metric2$ normalized ROI based on the left mouth corner was extracted (see Fig. 4). The ROI was then downsampled to 32×32, for use in our AVASR system (see III).

Following ROI detection, the ROI is tracked over consecutive frames. If the detected ROI is too far away from the previous frame, then it is regarded as a detection failure and the previous ROI location is used. A mean filter is then used to smooth the tracking. Due to the speed of the boosted cascade of classifiers, this detection and tracking scheme is used for every frame.

Overall, the accuracy of the ROI detection and tracking system was very good, with only a very few number of poorly or mistracked ROI's in the dataset. A major factor affecting performance was due to random head movement and some head pose variability, where subjects exhibit a somewhat more frontal pose than the profile view of the majority of the subjects – see also Fig. 5, where examples of accurately and poorly tracked ROIs are depicted. The latter is also the reason why we were not able to employ any rotation normalization. Many different configurations were experimented with, however, they seemed to cause more problems than they solved. We tried to rotate the ROI according to the left eye to left mouth corner angle, however, the many different head poses made this very problematic. Another attempt was to rotate the ROI using the angle between the mouth center and the left mouth center. This also failed, as the distance between these two points was too small (20 pixels), and any slight mistake



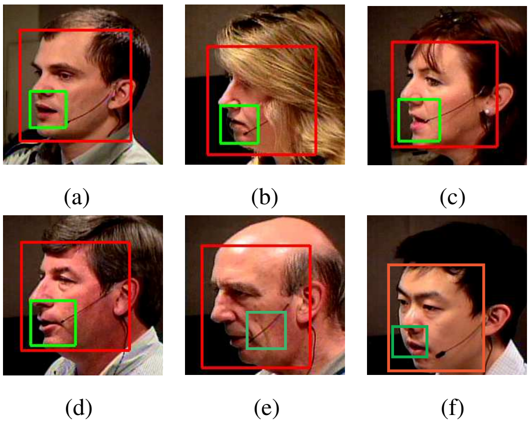
(a)

(b)

(c)

(d)

Fig. 4. (a) An example of face detection. (b) Based on the face detection result, a search area to detect the left eye and nose is obtained. The face box is lengthened or shortened according to *metric1*. (c) The left mouth corner is detected within the generalized mouth region. The ratio (*metric2*) is then used for normalizing the ROI. (d) An example of the scaled normalized detected ROI of size $(48 \times 48) \cdot metric2$ pixels.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Examples of accurate (a-d) and inaccurate (e,f) results of the detection and tracking system. In (f), it can be seen that the subject exhibits a somewhat more frontal pose compared to the profile view of the other subjects.

in the detection phase gave large errors.

III. The AVASR System

We now proceed to briefly describe the remaining components of the AVASR system, following detection of the mouth ROI. There exist two main such components, overviewed in the next two subsections: (a) feature extraction, which includes the visual features that complete the visual front end subsystem, and of course the audio feature extraction step; and (b): the audio-visual fusion (integration) step. In this work, neither component exhibits significant differences between the introduced profile-view AVASR system and our baseline frontal-view AVASR system refined in previous work [1]. These systems will be compared in Section IV.B. Furthermore, performance of a combined AVASR system that uses *both* profile and frontal views will also be discussed there. Specifics of all three AVASR systems are briefly overviewed in Section III.C.

A. Feature Extraction

Following ROI extraction, a two-dimensional, separable, *discrete cosine transform* (DCT) is applied to it, with the 100 top-energy DCT coefficients retained. The resulting 100-dimensional vectors are available at the video rate (30 Hz). In order to simplify integration with audio and to improve system robustness, the vectors are interpolated to the audio feature frame rate of 100 Hz, and are mean-normalized, independently over each utterance. Furthermore, for dimensionality reduction, an *intra*-frame cascade of *linear discriminant analysis* (LDA) followed by a *maximum-likelihood linear transform* (MLLT) is applied, resulting to 30-dimensional "static" visual features. Subsequently, to incorporate dynamic speech information, 15 neighboring such features over ±7 adjacent frames are concatenated, and are projected via an *inter*-frame LDA/MLLT cascade to 41-dimensional "dynamic" visual feature vectors. More details can be found in [1].

In parallel to visual feature extraction, 24-dimensional *mel-frequency cepstral coefficients* (MFCCs) are extracted at a 100 Hz frame rate, based on the audio signal. After mean normalization, the features are processed by an inter-frame LDA/MLLT cascade over ±5 frames to produce 60-dimensional acoustic features.

B. Audio-Visual Integration

Following feature extraction, time-synchronous audio and visual features are available at 100 Hz with dimensions 60 and 41, respectively. In our prior work, we have investigated a range of audio-visual integration techniques [1]. In this preliminary version of the paper, we will be reporting results using a *feature fusion* approach: In this technique, the bimodal feature vectors are concatenated, resulting to 101-dimensional features that are subsequently projected onto 60 dimensions using an LDA/MLLT cascade (note that this equals the audio feature vector dimensionality). The feature fusion approach has been selected due to its simplicity in producing experimental results quickly. In the final version of the paper, results using *decision fusion* based on multi-stream HMMs will be reported. The latter method is well known to yield significantly better results than the adopted feature fusion technique, but requires optimizing the modality integration weights (typically on held-out data). Notice that the feature fusion mechanism will also be used in our experiments to combine the profile- and frontal-view visual-only ASR (lipreading) systems into a "multi-view" lipreading system, as discussed next.

C. The Speech Recognition Systems

In our experiments below, we will be comparing three AVASR systems: The introduced profile-view AVASR system, a baseline AVASR system based on frontal views [1], and a combination of the two, namely a "multi-view" AVASR system. Furthermore, audio-only and visual-only systems will also be compared. All such systems are designed in this work to recognize connected-digit sequences (10-word vocabulary with no grammar), and they are based on single-stream HMMs operating on sequences of 60-dimensional features – with the exception of the visual-only single-view systems that use 41-dimensional features, as discussed in Section III.A. All HMMs are trained by employing the expectation-maximization algorithm over an available training set (see Section IV.A), and have an identical topology, containing 104 context-dependent states and approximately 1.7k Gaussian mixture components.

Before moving on to our experiments, we should emphasize a few differences between the compared systems: The "multi-view" visual-only system operates on 60-dimensional visual features that result from an LDA/MLLT cascade applied on the concatenated single-view (frontal + profile) visual-only feature vectors having a combined dimension of 82 (=41+41). This is in contrast to the single-view (profile, or frontal) visual-only systems, that use 41-dimensional features. In addition, one should note that the visual front end sub-systems of the frontal- and profile-view AVASR systems differ in two aspects: One concerns the face and mouth region tracking algorithm, where the frontal view system tracking is based on a set of "strong" classifiers, as described in detail in [1], [10]. The second difference lies on the size of the extracted ROIs, before DCT feature extraction is applied. It is 32×32 pixels for the profile-view system, but 64×64 pixels for frontal-view AVASR.

IV. Experimental Results

We now proceed to report a number of experimental results on the performance of the developed profile-view AVASR system. The experiments are conducted on a multi-sensory audio-visual database, recorded in the IBM smart room, that is briefly described next.

A. The Audio-Visual Database

A total of 38 subjects uttering connected digit strings have been recorded inside the IBM smart room, using two microphones and three pan-tilt-zoom (PTZ) cameras. Of the two microphones, one is head-mounted (close-talking channel – see also Fig. 1) and the other is omni-directional, located on a wall close to the recorded subject (far-field channel). The three PTZ cameras record frontal and two side views of the subject, and feed a single video channel into a laptop via a quad-splitter and an S-video–to–DV converter. As a result, two synchronous audio streams at 22kHz and three visual streams at 30 Hz and 368×240-pixel frames are available. Among these available streams, in the reported experiments we utilize the far-field audio channel and two video views: the frontal and one of the two side views, namely the one that consistently provides views closest to the profile pose (see also Fig. 1). A total of 1598 utterances are used in our experiments, partitioned using a multi-speaker paradigm into 1198 sequences for training, 242 for testing, and 158 sequences that are allocated to a held-out set (for future multi-stream HMM speech recognition experiments).

B. Recognition Results

In the first experiment, we report the *visual-only* system performance on this dataset. The *word error rate* (WER) of the baseline frontal-view system [1] on the test set is 23.4%, however the developed profile-view system achieves a significantly worse performance of 50.4% WER. This is more than double the error of the frontal-view system, but much better than pure chance. Hence, the profile system is capable of recognizing visual speech, but of course less so than the frontal system, in line with human lipreading experiments reported in [6]. Interestingly, by combining the two systems, the resulting "multi-view" visual-only performance becomes 21.7%, which demonstrates that there exists some information in the profile view, not captured by the frontal-view system (possibly that of lip protrusion).

When combining the three systems with the far-field audio channel, the *audio-only* system performance of 1.86% WER improves somewhat to an *audio-visual* WER of 1.76%, 1.62%, and 1.53%, when incorporating the frontal-, profile-, and "multi-view" visual information, respectively. These differences are however not significant due to the small database size. Of course, they become more pronounced, if we corrupt the audio channel by additive noise; in our experiments, "speech babble" is used for this purpose. The results are depicted in Fig. 6 and further verify the experimental observations of the previous paragraph. As expected, in high noise environments, the visual modality benefit to ASR is dramatic, with the "multi-view" system demonstrating the biggest gains, mostly due to the contribution of the frontal view video.

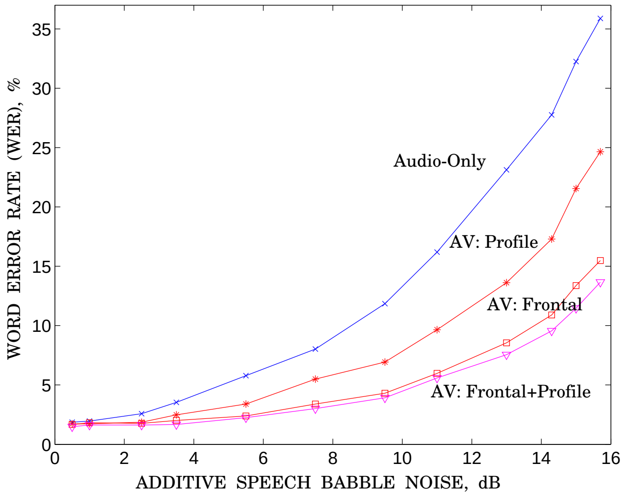


Fig. 6. Test set WER, %, of audio-only and audio-visual ASR. Three AVASR systems are depicted, based on profile view, frontal view, and both views ("multi-view" system). Additive speech babble noise at various dBs has been applied on the far-field audio channel.

We plan to extend our experimental results by also considering the close-talking audio channel of the data, as a contrast to the far-field condition. Furthermore, and as already mentioned in Section III.B, we plan to use multi-stream HMMs, when reporting fusion results. We expect this approach to further enhance system performance, when combining modalities or views.

V. Conclusions and Future Work

In this paper, we presented an AVASR system capable of extracting visual speech information from profile views. To our knowledge, this is the first serious attempt to lipreading from side views that allows quantifying the performance degradation as compared to lipreading from the traditional frontal view of the speaker's mouth. In our experiments, we demonstrated that profile views contain significant visual speech information, sufficient to improve ASR robustness to noise. Such benefit is of course less pronounced than when using the frontal view, however is not totally redundant to the frontal video, as the "multi-view" experiments demonstrated.

In further work, which we hope to include in the final version of this paper, we plan to improve the ROI extraction of the proposed profile-view system. This abstract represents our first attempt in developing such, and a number of refinements could very well boost its performance, as compared to our much more "stable" frontal-view AVASR system. In particular, we hope to successfully address the ROI rotation normalization problem. Another difficulty lies in the background: In the case of frontal faces, a slight detection/tracking error does not cause significant appearance change, due to the somewhat uniform background around the lips (i.e., the skin of the speaker). In contrast, in the profile view case, part of the ROI may capture the background behind the speaker's profile. It's non-uniformity may be a cause of significant performance degradation. We hope that improved tracking will reduce such effect.

References

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. of the IEEE*, vol. 91, no. 9, 2003.

[2] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, 2005.

[3] A. Pentland, "Smart rooms, smart clothes," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, vol. 2, 1998.

[4] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio visual speech recognition using lip movement extracted from side-face images," in *Proc. Auditory Visual Speech Processing (AVSP)*, 2003, pp. 117–120.

[5] ——, "Audio visual speech recognition using new lip features extracted from side-face images," in *Proc. ROBUST2004*, 2004.

[6] T. R. Jordan and S. M. Thomas, "Effects of horizontal viewing angle on visual and audiovisual speech recognition," in *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 6, 2001, pp. 1386–1403.

[7] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, 1998.

[8] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. Nefian, "Speaker independent audio-visual continuous speech recognition," in *Proc. Int. Conf. on Multimedia and Expo*, vol. 2, August 2002, pp. 25–28.

[9] A. Nefian and M. Hayes, "Face detection and recognition using hidden Markov models," in *Proc. Int. Conf. on Image Processing*, 1998, pp. 141–145.

[10] J. Jiang, G. Potamianos, H. Nock, G. Iyengar, and C. Neti, "Improved face and feature finding for audio-visual speech recognition in visually challenging environments," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 5, 2004, pp. 873–876.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[12] R. Leinhart and J. Maydt, "An extended set of Haar-like features," in *Proc. Int. Conf. on Image Processing*, 2002, pp. 900–903.

[13] D. Cristinacce, T. Cootes, and I. Scott, "A multi-stage approach to facial feature detection," in 15^{th} *British Machine Vision Conference, London, England*, 2004, pp. 277–286.

[14] Open Source Computer Vision Library,
`http://www.intel.com/research/mrl/research/opencv`