Automatic Extraction of Lip Feature Points

Roland Göcke[1], J Bruce Millar[1], Alexander Zelinsky[2], and Jordi Robert-Ribes[3]
[1]Computer Sciences Laboratory and [2]Robotic Systems Laboratory,
RSISE, Australian National University, Canberra ACT 0200, Australia
[3]Cable & Wireless Optus, 101 Miller St, North Sydney NSW 2060, Australia
E-Mail: Roland.Goecke@anu.edu.au

Abstract

We present a novel algorithm for the robust and reliable automatic extraction of lip feature points for speechreading. The algorithm uses a combination of colour information in the image data and knowledge about the structure of the mouth area to find certain feature points on the inner lip contour. A new confidence measure quantifying how well the feature extraction process worked is introduced. A parameter set describing the shape of the mouth is derived from the positions of the feature points. Using a stereo camera system, measurements are in 3D. Such a 3D parameter set is of great value for automatic speech-reading systems.

1 Introduction

Advances in technology in recent years have led to a widespread use of automatic speech recognition (ASR) systems in Human-Computer Interaction (HCI). Such systems usually employ statistical models of spoken language and enable continuous speech recognition in reasonably good acoustic conditions. However, these systems can fail unpredictably in noisy conditions. One way of overcoming some of the limitations of audio-only ASR systems is to use the additional visual information of the act of speaking, just like humans do. An automatic speech-reading, or lip-reading, system as part of an Audio-Visual Speech Processing (AVSP) system leads thus one step closer to more natural human-machine interactions. We are particularly interested in the correlation of the audio and the video speech signals.

Automatic speech-reading systems rely on the extraction of relevant features in the mouth area. Two main directions of feature extraction can be found. First, implicit feature extraction methods use the data of all pixels as input of a recognition engine (Hidden Markov Model, neural network) which learns the typical pixel value patterns associated with certain lip movements [Meier *et al.*, 1996]. A principal component analysis can be used to reduce the dimensionality of the input vector and to define the main directions of variation. Other implicit feature extraction methods are based on optical flow techniques [Mase and Pentland, 1991].

Secondly, explicit feature extraction methods use image processing techniques to find the position of certain feature points in the image, such as the lip corners, for example. Methods range from purely image-based approaches, such as thresholding [Petajan, 1984] or integral projection [Yang *et al.*, 1998], for example, to sophisticated model-based approaches, e.g. active contour models (or snakes) [Kass *et al.*, 1988], active shape models [Cootes *et al.*, 1995], or 3D lip models [Revéret and Benoît, 1998; Basu *et al.*, 1998].

We present an algorithm for the explicit extraction of the position of lip feature points. The algorithm uses a combination of colour information in the image data and knowledge about the structure of the mouth area to find certain feature points on the inner lip contour. Section 2 outlines the face tracking system used to find the mouth area. The lip feature points and the parameter set derived from their positions are explained in Section 3. The feature extraction algorithm is described in detail in Section 4. A new confidence measure quantifying the goodness of fit is introduced in Section 5. The results are discussed in Section 6. Finally, the conclusions and the outlook on future work are presented in Section 7.

2 Face Tracking System

Our lip feature point extraction algorithm builds on top of a stereo vision face tracking system [Newman *et al.*, 2000]. The system consists of two calibrated standard, colour analog NTSC video cameras whose outputs are multiplexed at half the vertical resolution into a single channel before being acquired by a Hitachi IP5005 video card on a Pentium II (300MHz CPU). The result is a 512×480 image in the YUV colour space, captured every 33ms, with the left camera image occupying the top half and the right camera image being in the bottom half

(Figure 1). The face tracking algorithm is based on template matching using normalised cross-correlation. The system is set up for a speaker at a distance of about 600mm from the cameras and is able to track the person's movements at a frame rate of 15-20Hz. The output of the face tracking system is an estimate of the head pose in 3D.

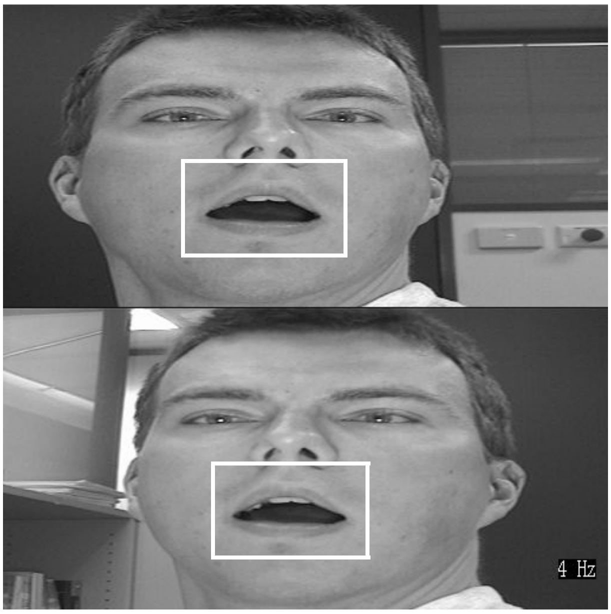

Figure 1: Stereo input frame with the mouth window.

A rectangular area containing the mouth area in both camera images is automatically determined during face tracking (Figure 1). The position of these mouth windows is based on the general head pose estimate. A generously sized area is chosen, so that the mouth is always completely inside.

3 Lip Features and Parameter Set

The lip feature points that we are interested in are the two lip corners and the mid-points of upper and lower lip (Figure 2). Since every speaker has different lips, we are interested in the inner lip contour so that the personal characteristic shape of the lips has minimal effect on the measurements. Furthermore, facial hair affects the visibility of the outer lip contour.

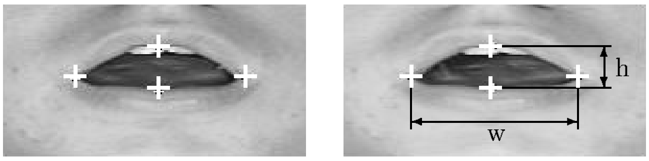

Figure 2: Lip feature points and mouth parameters.

If the mouth is fully closed, the inner lip contour line cannot be determined. In that case, the shadow line between the lips is taken as the inner lip contour line. This shadow line is always present, even for a directly illuminated face, as long as only one light source is used.

From these lip feature points, we derive a parameter set which describes the shape of the mouth during speech articulation. The parameter set consists of:

- mouth width w (= 3D distance from lip corner to lip corner),
- mouth height h (= 3D distance from mid-point upper lip to mid-point lower lip), and
- a protrusion factor (ratio of mouth width to mouth height).

Furthermore, we label each frame on the appearance of teeth from the upper jaw and/or lower jaw. More parameters can be derived but have not yet been looked at. In particular, we also want to examine how these parameters change over time, i.e. what are their velocity and acceleration patterns. This requires to look at the first and second derivative of the parameters.

We establish a discrete state model for the openness of the mouth with three states: *closed*, *partially open*, and *mouth wide open*. The classification of each frame into one of the states is based on the 3D mouth height. A measurement of less than 5mm indicates a closed mouth. If the mouth height is more than 15mm, the mouth is wide open. Otherwise, the mouth is partially open.

4 Feature Extraction Algorithm

The lip feature extraction algorithm combines colour information from the images with knowledge about the structure of the mouth area. Colour information is a very powerful cue in facial feature detection. However, the YUV signal from the NTSC cameras alone is of little use because the image signal is encoded into an intensity (Y) signal and two colour difference signals (U, V). The YUV signal can be transformed into a standard computer RGB signal. However, images in the RGB colour space are affected by changes in illumination, i.e. the RGB colour space is intensity-sensitive, which does not help image processing or object recognition. Using intensity-normalised RGB values can overcome this problem but such colour information does not assist lip feature point extraction. A better choice is the HSV colour space which separates hue (H) and saturation (S) from intensity (V) [Foley *et al.*, 1996].

The immediate idea of using the hue value to separate the lips and surrounding skin flesh from the oral cavity does not work because the hue of the oral cavity is still very close to the hue ("red") of other face parts although a human observer hardly perceives it that way. However, there is a clear difference in the saturation values for skin/lips, teeth, and oral cavity. The dark oral cavity exhibits the largest saturation values and the teeth the smallest, while the skin values lie between these two extremes. A combination of intensity (Y) and saturation (S) values is therefore used throughout the algorithm.

Although the face tracking can cope with both rotational and translational head movements, the speakers were asked to maintain an approximately frontal and upright head pose towards the cameras to avoid any additional head movement which could effect the accuracy of the lip feature extraction process. The subject's face was illuminated by a single light source positioned directly below the cameras. Using a well-illuminated face simplifies the task by avoiding shadows on the face and thus supports a high accuracy in the positioning of the feature points. This is a valid simplification because we are interested in accurate measurements for investigating the correlation between audio and video speech signals. Visual degradations of the environment through changes in illumination could be added at a later stage.

The algorithm consists of three main steps, each of which will be explained in detail in the following.

Step 1. Determine the general degree of mouth openness (4.2).

Step 2. Find the lip corners (4.3).

Step 3. Refine position of mid-lip feature points based on lip corner positions (4.4).

Steps 1 and 2 are applied separately to both the left and right images. The results are then combined to calculate the 3D position of the lip corners which are used for the refinement of the initial estimate of the mid-lip feature points in Step 3.

4.1 Algorithm Principles

Two modules, that are used at several points in the algorithm, test for the *shadow line* between the lips and for the visibility of *teeth* in the image data. The shadow line is detected by the typical high saturation and low intensity values. Teeth can be distinguished from other parts of the face by their characteristic low saturation and high intensity values. However, since some skin parts can show similar values, the teeth check must also meet an edge detection test which looks for the horizontal edge between the lip and the teeth using the Sobel operator:

$$K = \begin{bmatrix} +1 & +1 & +1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}. \qquad (1)$$

As a principle throughout the algorithm, whenever a threshold operation takes place, the *threshold is dynamically determined* at that time, instead of using hard-coded threshold values, to improve robustness. That is, the starting value of the threshold is chosen to be overly strict, so that no pixel value in the area of interest will pass it. The threshold is then iteratively changed until the value has been found, at which pixel values start passing the threshold. The algorithm then continues to use this threshold value for the rest of the processing of that frame. Secondly, whenever a particular pixel position is tested, not only the pixel value of that position is checked but also of at least two other pixels in the neighbourhood. A *voting* takes place and only when two out of three pixel positions indicate that a threshold has been passed, that position is accepted as being correct.

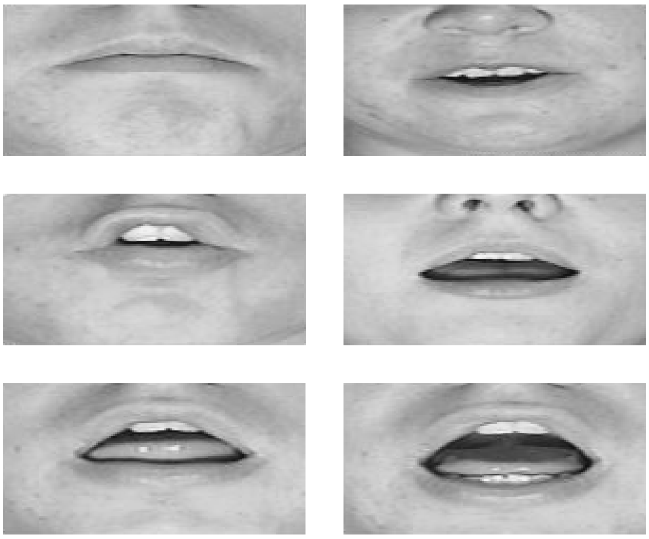

Figure 3: Different degrees of mouth openness and of teeth and tongue visibility.

4.2 Step 1: Determine Mouth Openness

To determine the openness of the mouth (Figure 3), the vertical positions of the mid-lip feature points must be determined. For the following calculations, the horizontal position of the mid-points is (temporarily) considered to be at the middle between the left and right boundaries of the mouth window. This estimate is close enough to the true position to start the algorithm for finding the lip corners (Step 2) but the position will be recalculated (Step 3) after the lip corner positions were found.

Horizontal integral projection on the intensity values of the image data is used to give a starting estimate for the vertical positions. The sometimes rough estimates may need to be refined. If the mouth is closed, either the shadow line between the lips (correct) or the external lip contour (incorrect) has been found. If the mouth is open, either or both upper and lower teeth are visible and the horizontal integral projection has picked up either the edge between lip flesh and teeth (correct), or between teeth and oral cavity (incorrect).

Let us first look at correcting the mid-point of the lower lip. To test if a correction is necessary, the vertical line from the lower boundary of the mouth window to the mid-point of the upper lip at the estimated horizontal position is followed upwards. The lower lip feature point cannot lie above the upper lip feature point. While walking along the line, check for either the shadow line between the lips or for the appearance of teeth. If found and the position is different from the one obtained from the previous horizontal integral projection, the position is updated. As described above, the saturation and intensity thresholds used are determined dynamically and three horizontal pixel positions (x–10, x, x+10) are checked at each vertical step (Figure 4 left). Finally, the vertical position of the lower lip mid-point is accurately placed on the lip pixel bordering the oral cavity.

Secondly, the vertical position of the mid-point of the upper lip is corrected if necessary. The algorithm tests for the appearance of teeth above the position estimated from the horizontal integral projection. If found, the position is moved upwards until no further teeth pixels are above the current position. Again, the vertical position of the upper lip mid-point is finally accurately placed on the lip pixel forming the edge to the oral cavity.

4.3 Step 2: Find Lip Corners

So far, the vertical positions of the mid-lip feature points have been established at their estimated horizontal positions. From the 2D positions of each mid-lip feature in the left and right stereo images, their 3D positions can be computed using the parameters known from the camera calibration. The 3D (Euclidean) distance between the two feature points is then calculated. If the distance is less than 15mm, the mouth is either fully closed or only partially open. Otherwise, the mouth is wide open. These different degrees of mouth openness require different techniques to extract the desired lip feature points. The threshold of 15mm is a heuristic determined through a number of experiments.

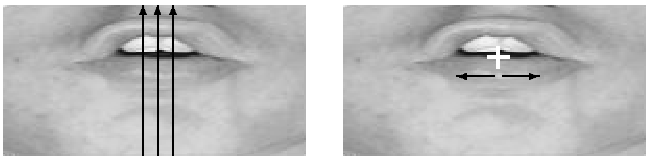

Figure 4: Left: Finding vertical positions of mid-lip feature points (Step 1); Right: Moving along shadow line to find lip corners (Step 2)

Mouth Closed or Partially Open

If the mouth is fully closed or only partially open, a vertical integral projection would not yield enough information to find the lip corners reliably. Thus, starting from the current position of the lower lip mid-point, a search along the shadow line to either side (Figure 4 right) is performed through a cycle of two tests. Let's consider the speaker's right lip corner. Starting from the mid-point (x, y) of the lower lip, the algorithm moves left in image space. Testing the five pixel positions $(x-1, y), \ldots, (x-5, y)$ allows to jump over pixels where the shadow line is discontinued due to image noise. If one of the pixels indicates the continuation of the shadow line, the current position is moved to $(x-1, y)$. If not, the test is repeated for $y+1$ first and then for $y-1$. In these two cases and positive tests, the current position is moved to $(x–1, y+1)$ and $(x–1, y–1)$, respectively. The reason for testing different vertical positions y is that the inner lip contour, which corresponds to the shadow line, of the lower lip is not necessarily a straight line but can be curved up or down depending on the generic mouth shape of the speaker.

If there are no more shadow line pixels ahead, the algorithm looks at the five pixels $(x, y-1), \ldots, (x, y-5)$ above the current position (x, y). If a shadow line pixel is found, move the current position to $(x, y–1)$ and repeat the test cycle. Otherwise, the lip corner has been found. The speaker's left lip corner is found similarly, except for moving in the opposite direction.

In rare cases, the shadow line will be discontinued for more than five pixels. Therefore, the found lip corner positions are checked to be at least 10mm away from the mid-point of the lower lip. Otherwise, the search along the shadow line is restarted from a point 10mm away from the mid-lip feature point.

Mouth Wide Open

If the mouth is wide open, vertical integral projection on the intensity values of the image will give reliable estimates of the horizontal positions of the lip corners. The vertical positions of the mid-points of upper and lower lip, determined in Step 1, define the vertical range for the integral projection. Once the horizontal positions of the lip corners have roughly been found, a search along the (vertical) pixel columns through these horizontal positions looks for the pixel with the lowest intensity value and the highest saturation value which corresponds to the internal lip contour in the lip corner. The two results are averaged to yield the estimate of the vertical position which makes the algorithm more robust against misleading pixel values. Finally, the found positions are refined by using the search technique described above for the closed or partially open mouth but with the current estimated positions as starting points.

4.4 Step 3: Refine Mid-lip Feature Points

Now that the lip corner positions are established, the horizontal position of the mid-lip feature points, which has so far simply been the centre between left and right boundaries of the mouth window, needs to be recalculated. Knowing the 2D positions of the lip corners in the left and right stereo images, their 3D positions $\vec{l}$, $\vec{r}$ are calculated using the known camera parameters. Based on the 3D positions of the lip corners, the centre point $\vec{c}$ between these points is computed:

$$\vec{c} = \frac{\vec{l} + \vec{r}}{2}. \tag{2}$$

Since the lip corner positions could be wrong, we use a linear combination of the previous mid-point estimates and the newly computed centre point $\vec{c}$ to determine the likely centre point $\vec{c}'$. The linear factor is the width confidence measure described in Section 5. We use information about the head pose from the general face tracker to define a normal vector perpendicular to an imaginary face plane and pointing away from the face. We then move the centre point $\vec{c}'$ 5mm along this vector. The reason is that the mid-lip feature points protrude about 5mm more than the lip corners. The centre point $\vec{c}'$ is then back projected into image space and the x coordinate of that point taken as the horizontal position of the mid-lip feature points in each of the two images.

After recalculating the horizontal position, small adjustments to the previously found vertical positions are highly likely and can be made with the same techniques as described in Step 1. Finally, the 2D positions of the mid-points of each lip in each image are combined to give their respective 3D positions.

5 A New Confidence Measure

Once the four lip feature points have been found, a confidence measure is computed. This measure is based on the difference between the corresponding 2D mouth width and mouth height distances in the left and right images, respectively. If the 2D measures in the two images do not agree, then that is an indication that something has gone wrong. However, we do not know which of the 2D measures represents the correct value and, hence, we can only mark the results as unreliable and not correct them. The chances that the algorithm fails in both images in the same way are very small.

Two separate confidence measures are calculated: One for the 3D mouth width and one for the 3D mouth height. Both are defined as:

$$C = 1 - \frac{|L - R|}{Max} \tag{3}$$

where C denotes the confidence measure, L either the 2D mouth width or 2D mouth height in the left image, R either the 2D mouth width or 2D mouth height in the right image, and Max either the width or the height of the mouth window. The resulting values lie in the range from 0 to 1. $C > 0.9$ indicates a reliable feature extraction process.

6 Results and Discussion

So far, the algorithm has been tested on three speakers. Each subject was asked to speak three sequences:

1. ba ba ba ...
2. e o e o e ...
3. Joe took father's green shoe bench out.

The first sequence maximises vertical mouth movement, while the second sequence emphasises lip rounding and stretching. The third sequence is an example of continuous speech taken from the design of the XM2VTSDB database [Messer *et al.*, 1999]. It was designed to cover all viseme and phoneme categories in the English language. Each sequence is about 4s long.

At the moment, the automatic lip feature point extraction is performed offline for analytical reasons, i.e. the face tracking systems determines the mouth windows in each stereo frame and then stores all the pixel data of the mouth windows in a data file, which is then accessed for the feature extraction. However, the feature extraction algorithm can run online in which case the frame rate decreases from 15-20Hz for the face tracking system alone to 5-10Hz for the system incorporating the lip feature point extraction.

Visual inspection of the extracted feature positions shows a high degree of accuracy. The correct positions are not found only in a few frames. Figure 5 shows correct and incorrect results. Incorrect positions are well detected by the confidence measure. In order to quantify the error, a ground-truth would be required but cannot be obtained for practical reasons. However, a software tool was developed which allows the comparison of the automatically extracted feature positions with the results from a manual extraction in which the user clicks on the feature positions [Tran, 2000]. While this process is tedious for long video sequences, it indicates the accuracy of the automatic extraction algorithm.

The comparison shows that the manual and automatic feature extractions yield very similar results. They only differ at about 1-2mm for the mouth width which is less than 4% for an average mouth width of 50mm. The difference for the mouth height is about 1mm on average. The classification into discrete mouth states has an error rate of about 3%. Misclassifications typically appear when measurements are close to the experimentally chosen thresholds of 5mm and 15mm, respectively, and are only in very few cases related to a bad lip feature extraction. The algorithm has an error rate of about 6% for determining the visibility of teeth. The errors can be attributed to the cases when only a narrow line from the teeth is visible and the rest is occluded by the lips (Figure 5 top right image). While the human observer detects the teeth easily, these cases proved to be difficult for the algorithm.

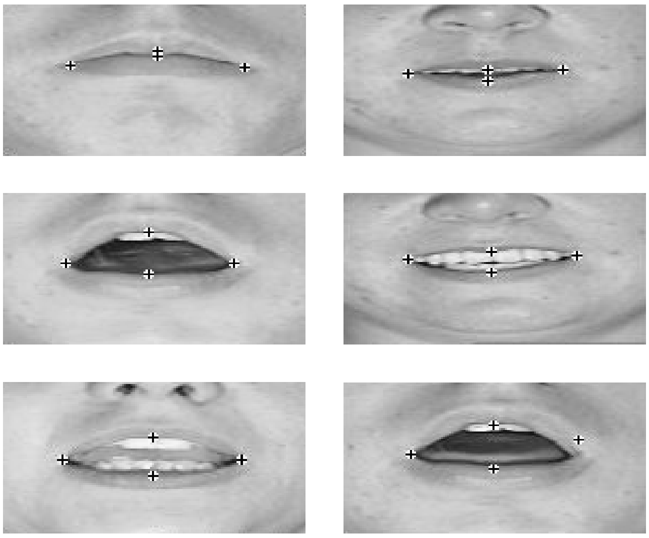

Figure 5: Correct and incorrect feature positions. The speaker's left lip corner in the lower right image was not found correctly.

7 Conclusions and Future Work

A novel algorithm to automatically extract the positions of certain lip feature points using a combination of colour information and knowledge about the structure of the mouth area has been presented. It is based on a stereo vision face tracking system and allows measurements in 3D. The automatic extraction algorithm shows a high degree of accuracy when compared to the results of a manual feature extraction procedure with differences in the range of a 1-2mm.

While the results have been promising so far, the algorithm needs to be tested on a larger number of speakers. In particular, we need to examine the effect of different personal characteristics of the mouth area, such as facial hair, lip thickness, missing teeth etc. We currently build an audio-video speech database with 10 speakers covering the phonemes and visemes in Australian English.

At the current stage, the algorithm only uses information from one frame at a time. Work is in progress to predict the 3D feature positions with a Kalman filter and to combine the measured and predicted feature positions using the confidence measure which is expected to further improve the accuracy.

References

[Basu *et al.*, 1998] S. Basu, N. Oliver, and A. Pentland. 3d lip shapes from video: A combined physical-statistical model. *Speech Communication*, 26(1–2):131–148, October 1998.

[Cootes *et al.*, 1995] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and applications. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[Foley *et al.*, 1996] J.D. Foley, A. van Dam, S.K. Feiner, and J.F. Hughes. *Computer Graphics - Principles and Practice.* Addison-Wesley, Reading MA, 1996.

[Kass *et al.*, 1988] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal on Computer Vision*, 1(4):321–331, 1988.

[Mase and Pentland, 1991] K. Mase and A. Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computer in Japan*, 22(6):67–76, 1991.

[Meier *et al.*, 1996] U. Meier, W. Huerst, and P. Duchnowski. Adaptive Bimodal Sensor Fusion For Automatic Speechreading. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP'96*, 1996.

[Messer *et al.*, 1999] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication AVBPA'99*, Washington D.C., 1999.

[Newman *et al.*, 2000] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. Real-time stereo tracking for head pose and gaze estimation. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition FG2000*, Grenoble, France, 2000.

[Petajan, 1984] E.D. Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois at Urbana-Champaign, 1984.

[Revéret and Benoît, 1998] L. Revéret and C. Benoît. A new 3d lip model for analysis and synthesis of lip motion. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, *Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98*, pages 207–212, 1998.

[Tran, 2000] Q.N. Tran. Show me your lips. Technical report, Computer Sciences Laboratory, RSISE, ANU, 2000.

[Yang *et al.*, 1998] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel. Real-time face and facial feature tracking and applications. In *Proceedings of AVSP'98*, pages 79–84, Terrigal, Australia, 1998.