STEREO VISION LIP-TRACKING FOR AUDIO-VIDEO SPEECH PROCESSING

Roland Goecke[1]*, J Bruce Millar*[1]*, Alexander Zelinsky*[2] *and Jordi Robert-Ribes*[3]

[1]Computer Sciences Laboratory and [2]Robotic Systems Laboratory,
Research School of Information Sciences and Engineering,
Australian National University, Canberra ACT 0200, Australia
[3]Cable & Wireless Optus, 101 Miller St, North Sydney NSW 2060, Australia
Contact: Roland.Goecke@anu.edu.au URL: http://cslab.anu.edu.au/∼rgoecke

ABSTRACT

We present the first results from applying a recently proposed novel algorithm for the robust and reliable automatic extraction of lip feature points to an audio-video speech data corpus. This corpus comprises 10 native speakers uttering sequences that cover the range of phonemes and visemes in Australian English. The lip-tracking algorithm is based on stereo vision which has the advantage of measurements being in real-world (3D) coordinates, instead of image (2D) coordinates. Certain lip feature points on the inner lip contour such as the lip corners and the mid-points of upper and lower lip are automatically tracked. Parameters describing the shape of the mouth are derived from these points. The results obtained so far show that there is a correlation between width and height of the mouth opening as well as between the protrusion parameters of upper and lower lips.

1. INTRODUCTION

Lip-tracking has a wide range of applications in the field of human-computer interaction (HCI), for example, in animation, expression recognition [1], and audio-video (AV) speech processing [2, 3, 4]. Although automatic speech recognition (ASR) systems have become common tools in HCI, they still have some limitations with respect to the environment in which they can be used. Current commercially available ASR systems employ statistical models of spoken language and enable continuous speech recognition in reasonably good acoustic conditions. However, they 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 which requires a way to track the lips in a video sequence.

Various lip-tracking techniques have been explored, ranging from purely image-based approaches [4, 5], to sophisticated model-based approaches [6, 7, 8, 9]. All these approaches use video data from a single camera, thus limiting direct measurements to the 2D image space. However, the mouth is a 3D structure which deforms in all three dimensions. It should therefore also be tracked in 3D.

We recently proposed a novel algorithm for the explicit extraction of lip feature points based on a stereo vision head tracking system [10]. To our knowledge, it is the first time that a stereo vision algorithm has been applied to lip-tracking for AV speech processing. The head tracking system is explained in Section 2, followed by details of the lip-tracking algorithm in Section 3. Section 4 describes the AV speech data corpus for our experiments. In Section 5, the first results from applying the lip-tracking algorithm to the data corpus are presented and discussed. Finally, we look at the conclusions and future work in Section 6.

2. HEAD TRACKING SYSTEM

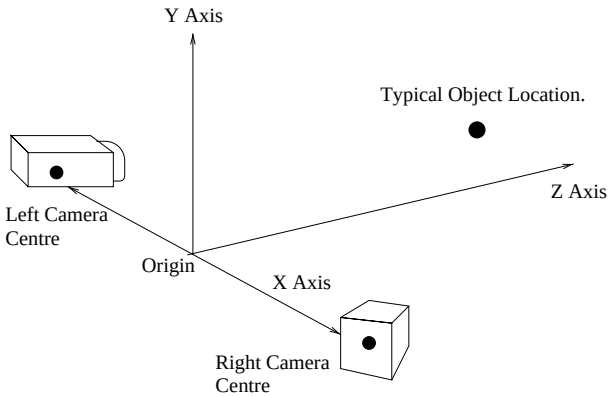

Fig. 1. Stereo camera arrangement.

Our lip-tracking algorithm builds on top of a stereo vision head tracking system [11] which is completely non-intrusive and does not require facial markers. The layout of the system is shown in Figure 1. The system consists of two calibrated standard colour analog NTSC video cameras which are positioned equidistant from the origin and

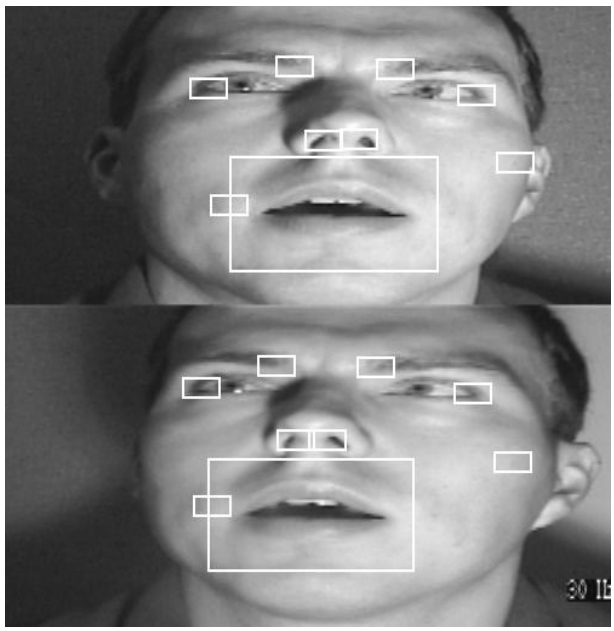

Fig. 2. Stereo image with templates and mouth windows.

are verged towards the origin in the horizontal plane at about 5°. This layout is designed to give the best measurements for a speaker at a distance of about 600mm from the cameras. The camera outputs are multiplexed at half the vertical resolution into a single 512x480 image (Figure 2) before being acquired by a Hitachi IP5005 video card on a Pentium II (300MHz CPU) every 33ms. The head tracking system is based on template matching using normalised cross-correlation. It is able to track head movements at a frame rate of 15-20Hz. (See http://www.seeingmachines.com for an improved version of the head tracking system.)

3. LIP FEATURE EXTRACTION

A generously sized rectangular area containing the mouth area in both camera images is automatically determined during head tracking (Figure 2). The position of these mouth windows is based on the general head pose estimate from the head tracker. The lip feature point extraction algorithm is then applied to these areas.

The feature points that we are interested in are the two lip corners and the mid-points of upper and lower lip (Figure 3). Since every person has differently shaped lips, we use the inner lip contour so that the personal characteristic shape of the lips has minimal effect on the measurements. Furthermore, facial hair can affect the visibility of the outer lip contour. 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.

The lip feature extraction algorithm combines colour information from the images with knowledge about the structure of the mouth area. The YUV signal from the cameras is transformed into HSI colour space which separates hue (H) and saturation (S) from intensity (I). By using all three values, our algorithm is more robust to changes in illumination. The algorithm consists of three main steps:

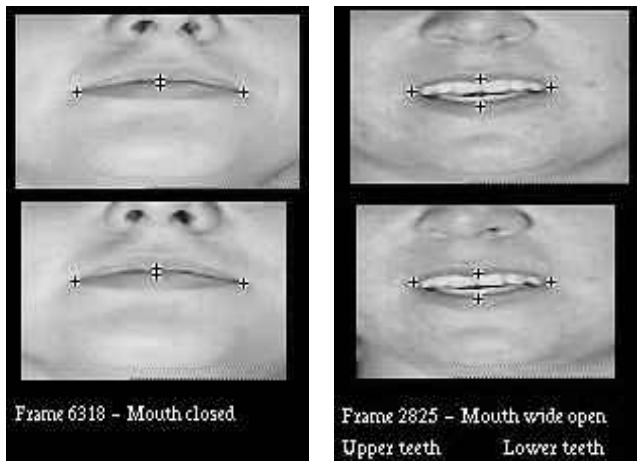

Fig. 3. Lip-tracking results.

Step 1. Determine degree of mouth openness.

Step 2. Find lip corners.

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

Step 1 uses horizontal integral projection of I to find the vertical positions of the mid-points of upper and lower lip from which the general degree of mouth openness is determined according to our mouth model which has three discrete states: *closed*, *partially open*, and *wide open*. In Step 2, the lip corners are found by vertical integral projection of I and a procedure that follows the shadow line between upper and lower lips. Steps 1 and 2 are applied separately to both the left and right images. The results are then combined to calculate the 3D positions of the lip corners. These are used to determine the horizontal positions of the mid-lip feature points in Step 3. Finally, small automatic adjustments to the vertical positions of these mid-lip feature points can be necessary at their new horizontal positions.

From these four lip feature points, we derive a parameter set which describes the shape of the mouth during speech articulation in real-world distances:

- *Mouth width* (3D distance from lip corner to lip corner),
- *Mouth height* (3D distance from mid-point upper lip to mid-point lower lip),

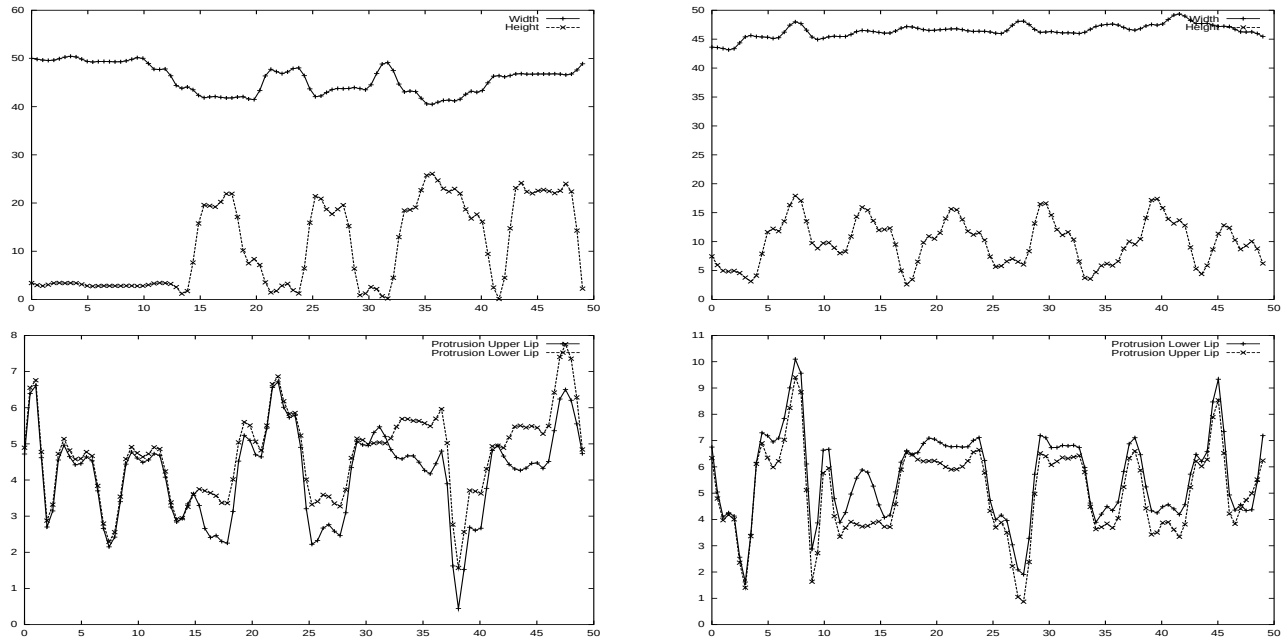

Fig. 4. Parameter curves for two speakers: Width and height (top) and protrusion of upper and lower lip (bottom).

- *Protrusion of upper lip* (3D distance from mid-point upper lip to mid-point between lip corners), and
- *Protrusion of lower lip* (3D distance from mid-point lower lip to mid-point between lip corners.

Furthermore, the algorithm labels each frame with the appearance of teeth from the upper jaw and/or lower jaw. Figure 3 shows two examples of lip-tracking results. Video clips of the lip-tracking process can be found at our homepage (see URL in authors' contact details). A detailed description of all steps of the algorithm was presented in [10].

The accuracy of the algorithm has been validated by comparing its results with the results from a manual selection of the lip feature points [12]. Although the manual process is itself error-prone, it is the best available 'ground truth' for determining the accuracy of the lip-tracking algorithm. The comparison showed that the manual and automatic procedures yielded similar results. Internal mouth width and height results differed by about 1-2mm only which is very accurate given that we use a completely non-intrusive system. The protrusion parameters were less accurate with errors of 3-8mm. Work continues to improve the accuracy of the protrusion parameters because they are important for describing the 3D shape of the mouth.

4. AV SPEECH DATA CORPUS

We have recorded an AV speech data corpus for Australian English using the head tracking system described in Section 2. The data corpus comprises 10 native speakers (5 female and 5 male speakers) [12]. It was designed to cover all phonemes and visemes in Australian English because we investigate the correlation between parameters describing the audio and video representations of spoken language.

The core part consists of 40 words in consonant-vowel-consonant (CVC) and vowel-consonant-vowel (VCV) contexts to study the phonemes and visemes free of phonological or lexical restrictions. The vowel context is the wide open "ar" and the consonant context is the bi-labial /b/. Having a bi-labial opening and closing simplifies the visual analysis. On a negative side, a bi-labial context causes strong coarticulation effects in the formants. However, these effects are quite predictable for /b/ and we believe the advantages outway this disadvantage. To overcome the typical articulation patterns associated with reading words from a list, each word was put in the carrier phrase "You grab *word* beer." which emphasizes the bi-labial context. In addition to these sequences, every speaker also uttered sequences with the digits from 0 to 9 in the same carrier phrase as well as three sentences covering all phonemes and visemes in a continuous speech stream.

5. RESULTS AND DISCUSSION

It has been shown that incorporating information about visible speech articulation in an ASR system can improve the recognition rate in adverse conditions (cp. [13]). This suggests that there is redundant information in the two modali-

ties that can be used if one of the two channels is affected by noise (acoustic noise, visual noise). Ultimately, we want to develop an adaptive AV ASR system which flexibly relies more strongly on one or the other modality depending on their respective noise levels. This requires a-priori knowledge of the varying correlation between audio and video speech parameters for a variety of speakers. Little work has been done on which audio and video parameters correlate best. We are in the process of establishing this knowledge using our AV speech data corpus.

We present here the first results of applying our lip-tracking algorithm to the recorded sequences. Figure 4 shows the resulting parameter curves for two speakers on a calibration sequence of "ba ba ba ...". The left panels show a speaker with strong visible speech articulation, the right panels one with less visual expressiveness. The vertical movement of the lips depicted in the parameter curve 'height' clearly dominates which is not surprising for this sequence. The mouth width changes less dramatically but at least for the group of visual expressive speakers, there is a correlation ($\bar{r} = -0.68$) between mouth width and height due to the fact that the lip flesh can only stretch in a limited way. Hence, a larger mouth height leads to a smaller mouth width and vice versa. Thus, one parameter value could be inferred from the other if only one can be measured accurately by the AV ASR system. However, for speakers with less visible speech articulation, there is no correlation.

The protrusion parameters of upper and lower lip exhibit a high degree of similarity (Figure 4). The average correlation coefficient across all 10 speakers is $\bar{r} = 0.93$. This is not surprising because when the lips are protruded in normal speech articulation, typically both lips are moved in a similar way. Hence, measuring only one of the two protrusion parameters would suffice in an AV ASR system.

6. CONCLUSIONS AND FURTHER WORK

The first results from applying a novel stereo-vision lip-tracking algorithm to part of our AV speech data corpus show that there is redundant information in the parameter set. We found a strong correlation ($\bar{r} = 0.93$) between the protrusion parameters of upper and lower lips across all speakers. The group of visually more expressive speakers also showed a correlation, though weaker, between mouth width and height. We will continue analysing the data for all phonemes and visemes and will report further on the correlation of audio and video speech parameters.

7. REFERENCES

[1] M. Pantic and L.J.M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

[2] A. Adjoudani and C. Benoît, "On the Integration of Auditory and Visual Parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke, Eds. 1996, vol. 150 of *NATO ASI Series*, pp. 461–471, Springer-Verlag.

[3] C. Bregler and Y. Konig, ""Eigenlips" for Robust Speech Recognition," in *Proceedings of ICASSP'94*, Adelaide, Australia, 1994, vol. II, pp. 669–672.

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

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

[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and applications," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[8] L. Revéret and C. Benoît, "A new 3D Lip Model for Analysis and Synthesis of Lip Motion," in *Proceedings of AVSP'98*, Terrigal, Australia, 1998, pp. 207–212.

[9] S. Basu, N. Oliver, and A. Pentland, "3d lip shapes from video: A combined physical-statistical model," *Speech Communication*, vol. 26, no. 1–2, pp. 131–148, 1998.

[10] R. Goecke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes, "Automatic Extraction of Lip Feature Points," in *Proceedings of ACRA2000*, Melbourne, Australia, 2000, pp. 31–36.

[11] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky, "Real-time stereo tracking for head pose and gaze estimation," in *Proceedings of Automatic Face and Gesture Recognition FG2000*, Grenoble, France, 2000.

[12] R. Goecke, Q.N. Tran, J.B. Millar, A. Zelinsky, and J. Robert-Ribes, "Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing," in *Proceedings of SST2000*, Canberra, Australia, 2000, in print.

[13] D.G. Stork and M.E. Hennecke, Eds., *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series*, Springer-Verlag, 1996.