A Comparative Study of 2D and 3D Lip Tracking Methods for AV ASR

Roland Göcke and Akshay Asthana

Department of Information Engineering, Research School of Information Sciences and Engineering,
Australian National University, Canberra, Australia

roland.goecke@anu.edu.au, aasthana@rsise.anu.edu.au

Abstract

Over the past two decades, many algorithms have been proposed to detect and track a human face and its facial features. Of particular interest to the Automatic Speech Recognition (ASR) community are algorithms that can track the shape of the lips, as such visual speech input can then be used in an auditory-visual (AV) ASR system to improve the recognition accuracy of traditional audio-only ASR systems, particularly in the presence of acoustic noise. Despite the large number of face and lip tracking algorithms that have been proposed over the years, there is a lack of a comparative study that evaluates such algorithms in the context of AV ASR performance. In this paper, the performance of various 2D and 3D lip tracking algorithms is compared from a point of view of AV ASR. In particular, the focus of this study is on algorithms that use explicit lip models. A number of variants of the recently popular Active Appearance Models (AAMs) are compared with a 3D lip tracking algorithm that uses stereo vision. All performance evaluations are made using the AVOZES data corpus.

Index Terms: Lip tracking, auditory-visual automatic speech recognition, active appearance model

1. Introduction

Human perception of the world is inherently multi-sensory because the information provided is multimodal. The perception of spoken language is no exception. Beside the auditory information, there is visual speech information as well, provided by the facial movements as a result of moving the articulators during speech production. Visual speech information contributes to speech perception in all kinds of audio conditions, but its effect is perhaps most readily noticed in noisy acoustic conditions. Various research groups around the world have found that auditory-visual automatic speech recognition (AV ASR) systems result in an improved recognition rate compared to audio-only systems, in particular in noisy audio conditions (for overviews, see [1, 2, 3], for example).

One of the central topics in visual and AV ASR is the process of detecting and tracking the lips and of extracting lip features and measures that capture both the static and dynamic characteristics of the lips moving during speech production. While it has been shown that also other parts of the lower face half contribute relevant visual information (cp. [1]), the lips carry the majority of this information and are the focus here. This paper provides an overview of various 2D and 3D methods for accurately detecting and tracking the lips and for automatically extracting lip contours and features which can then be used in visual and AV ASR. This study concentrates on the approaches that model the lips explicitly, rather than approaches that model them implicitly through image pixel values (e.g. [4, 5]). In particular, this comparative study focusses on various (2D) variants of the Active Appearance Model (AAM) approach, that has been very popular in recent years, as well as a highly accurate 3D lip tracking method that relies on a stereo vision approach [6]. While there are many more approaches in the literature, comparing all of them is beyond the scope of this study. It is hoped that the selected popular algorithms give the reader a useful comparison at hand from which he or she can make an informed decision about the best method to use for their specific situation. Experimental results of such uses are given on the example of AV ASR performance on a subset of the AVOZES data corpus.

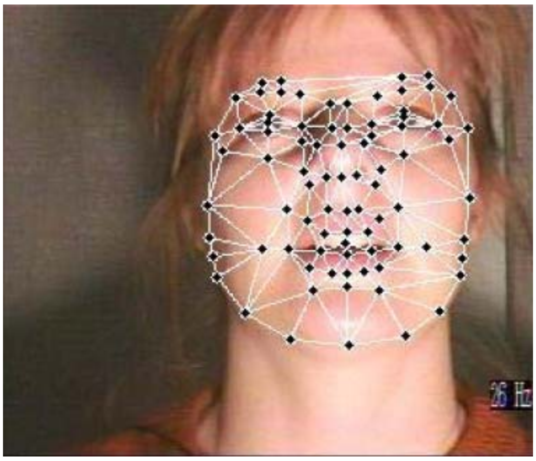


Figure 1: An example of an AAM fitted to a face.

The remainder of this paper is organised as follows. Section 2 gives a brief overview of related work in the face detection, face tracking and lip tracking area. Next, Section 3 provides an overview of common and state-of-the-art variants of the AAM approach. In Section 4, the 3D stereo vision lip tracking algorithm is briefly described. Section 5 outlines the AVOZES data corpus that is used for the experiments in this study. Then, Section 6 details the experimental setup for the AV ASR performance comparison. The results are presented and discussed in Section 7. Finally, the conclusions are provided in Section 8.

2. Related Work

Fundamental to AV ASR are the abilities to, first of all, automatically find the face in an image and to track it over a sequence of video frames, and secondly, to extract useful parameters that describe the visible speech-related movements of the articulators. A common method for finding the face and then the lips in an image is based on skin colour detection [7, 8, 9, 10], where

Copyright © 2008 AVISA
Accepted after peer review of full paper

a model of the skin colour distribution in an appropriate colour space is built, often using colour histograms. The skin colour distribution of each individual is a multivariate normal distribution, with the parameters of the distribution accounting for differences among people and lighting conditions.

More recently, face detection by a cascade of weak classifiers have become very popular [11]. The detector is learnt from a large number of training examples and while this offline training step is computationally expensive and slow, the actual detection of objects, e.g. faces, at different scales is very fast (real-time capable) due to the concept of integral images. The Haar wavelet-like features themselves are limited but their performance can be drastically improved through a boosting process, e.g. AdaBoost. Further improvements and variants have been proposed since the original work, exploring different classifiers and boosting methods, but to list them here is beyond the scope of this paper.

Once the face has been detected and the overall head position is tracked, the issue of how to accurately find the lips and extract the lip contours and features arises. Early systems, e.g. [10, 12] used image-based methods such as integral projection and thresholding. However, such simple methods suffer from a lack of robustness to variations in pose and illumination. [13, 14] use artificially coloured lips to enhance the contrast and to facilitate the above mentioned image-based methods for feature extraction. Although that is a valid way of simplifying the feature extraction problem, wearing blue lipstick is a considerable step away from a natural, non-intrusive system for practical applications of visual and AV ASR systems. Artificial facial markers, for example infrared LEDs tracked by an infrared camera system (OPTOTRAK, Qualisys), offer another way of extracting feature points with high accuracy, but again it is an intrusive system which requires familiarisation for the speaker and is not a solution for real-world applications of AV ASR technology.

In the last 20 years, deformable models in 2D and 3D have gained significant popularity and can be seen as the current state of the art. Active shape models (ASM) [15] imply constraints on the shape of active contour models. ASMs were used by [16] to track the internal and external lip contour for AV ASR. An ASM can only deform in ways characteristic to the class of objects it represents. These characteristics are learned from a set of training images and stored in a point distribution model. In order to make the ASM even more robust, [17] developed the Active Appearance Model (AAM). AAMs model non-rigid shape and texture of a visual object using a low dimensional representation obtained from applying principle component analysis (PCA) to a set of labelled data. The power of the AAM stems from two fronts. Firstly, its compact representation as a linear combination of a small number of modes of shape and texture variation enables optimisation over a small number of parameters. Secondly, the use of a fixed linear parameter update model, in the original formulation, allows for an efficient calculation of parameter updates. [18] employed an AAM for AV ASR. 3D extensions of AAMs exists, known as 3D Morphable Models [19], but suffer from the large computational complexity.

Finally, fully model-based approaches have been used for lip tracking. [20] developed a system in which a 3D lip model is fitted to image data for lip tracking, speech recognition and visual speech animation purposes. The lip model consists of a 3D polynomial surface model controlled by three articulatory-oriented parameters learned on the speaker. A similar system based on the backprojection of a 3D model into 2D image space and adjusting the model parameters until the model fits the mouth shape in the image was developed by [21]. The model is built on physical and statistical information about permissible mouth shapes from training image sequences.

3. Active Appearance Models

In the past decade, model based approaches have become very popular because of the many benefits of having a model that can replicate the deformations of a non-rigid object, such as a face. In the original work of [17], the AAM is built by applying PCA on the set of labelled data to model the intrinsic variation in shape and texture of the object. A parameterised model is formed that is capable of representing large variation in shape and texture by small set of parameters.

For constructing an AAM, the annotated training images are aligned into a common co-ordinate frame by Procrustes analysis. The modes of shape and texture variation are btained by applying PCA to the set of aligned images. Due to the correlation between the shape and texture variations, a PCA is applied to a concatenated vector of shape and texture to generate a combined, compact representation. The model can then be represented in terms of the appearance parameter.

The *AAM Fitting* process is equivalent to finding the model parameters that best fit the model to the image. It is performed by iteratively updating the model parameters via the update function. A number of AAM fitting algorithms have been proposed. The algorithms typically deal with the problem of fitting as a minimisation/maximisation of some measure between the model's texture and the warped image region.

In the following, we briefly characterise the AAM fitting algorithms investigated in this study. For an excellent in-depth review of these and other AAM methods, the interested reader is referred to [22].

3.1. Fixed Jacobian Method (FJ)

In the original AAM approach [17], also known as the fixed Jacobian method, the problem of fitting is treated as a minimisation of least squares error between the model's texture and the warped image region, where it is assumed that the Jacobian of the error is fixed for all settings of the model parameters. This enables a linear update model to be precomputed through a pseudo-inverse of the fixed Jacobian. Since the assumption of fixed Jacobian holds only loosely, the method requires the use of an adjustable step size, where at each iteration the predicted parameter updates are halved until a reduction in the appearance difference between the model and the cropped image is attained. This results in a rapid and reasonably efficient fitting to take place. However, if the object exhibits large variation in shape and texture, FJ struggles to perform because of the assumption of fixed linear update model which can be too restrictive.

3.2. Project-out Inverse Compositional Method (POIC)

This method is currently the fastest AAM fitting algorithm and belongs to the class of fitting methods using the inverse-compositional model, i.e. the role of image and model in the error function are reversed [23]. In POIC, the fitness function, that measures the difference between the model's appearance and the cropped image region, is grouped into two components: one which lies within the subspace of appearance deformations and another which is orthogonal to it. This procedure requires optimisation over the shape parameters only, assuming the optimal choice (in a maximum likelihood sense) of the appearance parameters is chosen at each iteration. Since the minimisation of the fitness function depends only on the subspace orthogonal to the texture variation, a fixed linear update model can be analytically computed over the shape parameters only. This better justifies the assumption of linear update model as compared the original formulation (FJ) and is also extremely fast. However, this approach only works well for person-dependent models.

3.3. Simultaneous Inverse Compositional Method (SIC)

This method is another adaptation of inverse compositional image alignment for AAM fitting that addresses the problem of the significant shape and texture variability by finding the optimal shape and texture parameters simultaneously [24]. Although the derivative of the warping function can be precomputed, the linear update model has to be recomputed at each iteration as it depends on the current appearance parameters, making this method comparatively slow. However, rather then recomputing the linear update model at very iteration using the current estimate of appearance parameters, it is evaluated using the mean appearance parameters, allowing the update model to be precomputed. Reusing the current appearance parameters is beneficial only if the current estimates are expected to be close to the actual parameters, for example, in continuous tracking, where the parameters from current frame may be closer to the parameters for the next frame.

3.4. Robust Inverse Compositional Method (RIC)

In [25], the idea of inverse compositional method for AAM fitting is extended further to use an M-estimator (robust penaliser) instead of the least squares fitting criterion and hence, results in an iteratively reweighted least squares fitting scheme. However, this method requires the normalisation of the mean subtracted error image with respect to the direction of appearance variability [24] which results in a high computational complexity. For this purpose, the error image is first projected onto the subspace of appearance variability. This projected error image is used for generating the model's appearance that is later subtracted from the error image to get the measure of the fitness function.

3.5. Iterative Error Bound Minimisation Methods (IEBM)

AAM fitting has the peculiarity that the warped image texture for a given parameter setting extracts only a subset of the information required to completely describe the optimal parameter setting. Consequently, it is difficult to build an update model which can accurately predict updates to parameters which depend on the missing information. In [26], a novel linear update scheme for AAM fitting was proposed that uses the optimality property of Support Vector Regression (SVR), i.e. each sample is adjusted to achieve its respective parameter setting where the error is minimised, giving priority to those samples that produce maximum error. IEBM focusses on building the update model by utilising the information from various combinations of parameter settings. Since this update model is learnt offline, the method is extremely efficient, achieves superior fitting results and has been shown to exhibit good generalisability.

4. 3D Lip Tracking via Stereo Vision

The lip tracking algorithm described in [6] is a three-step process. The first and second steps are applied separately to both the left and right camera images. Once the 2D image positions of the lip corners in both views are known, their 3D positions can be calculated. This is called solving the point correspondence problem and incorrectly identified correspondences lead to incorrect 3D coordinates. As the mouth shape is changing rapidly during speech, static methods such as template matching do not work well. Therefore, a combination of colour information and structural knowledge is used.

The first step determines the general degree of mouth openness. The lip tracking algorithm must be able to handle mouth shapes during speech ranging from a completely closed mouth to a wide open mouth. No single image processing technique would give good results for all possible mouth shapes. By pre-classifying mouth shapes into one of three categories based on mouth openness (closed, partially open, wide open), specific techniques individually targeted at each category can be applied to give better results.

In the second step, the lip corners are found. Here, the *a priori* knowledge about the structure of the mouth area becomes useful. For example, if the mouth is closed, teeth will not be visible, so the shadow line between upper and lower lip is the outstanding feature. Various definitions of what constitutes the inner lip contour of a closed mouth are possible. In [6], the shadow line between the lips was considered to be part of the inner lip contour. Therefore, the algorithm looks for this line. When the mouth is open, it is very likely that either or both the upper and lower teeth are visible, so the algorithm looks for them as well as for the oral cavity. By tailoring the algorithm in this way to fit a particular situation, more accurate results can be obtained than from a general-purpose, 'one-size-fits-all' algorithm. The result is then used in the third and final step, in which the positions of the lip midpoints are determined.

5. AVOZES

The AVOZES (*A*udio-*V*ideo *OZ*stralian *E*nglish *S*peech) data corpus is used in the experiments [27]. The data corpus is novel in that a calibrated stereo camera system was used for the video recordings, which allows for the testing of 3D lip tracking approaches. The video output from the two cameras is compressed into one video frame by halving the vertical resolution (see Figure 2). The double frame has a resolution of 512×480 pixels ($= 512 \times 240$ per camera) and the frame rate is 30Hz.

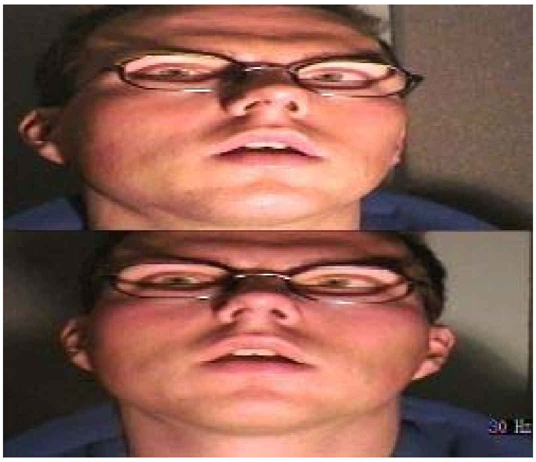
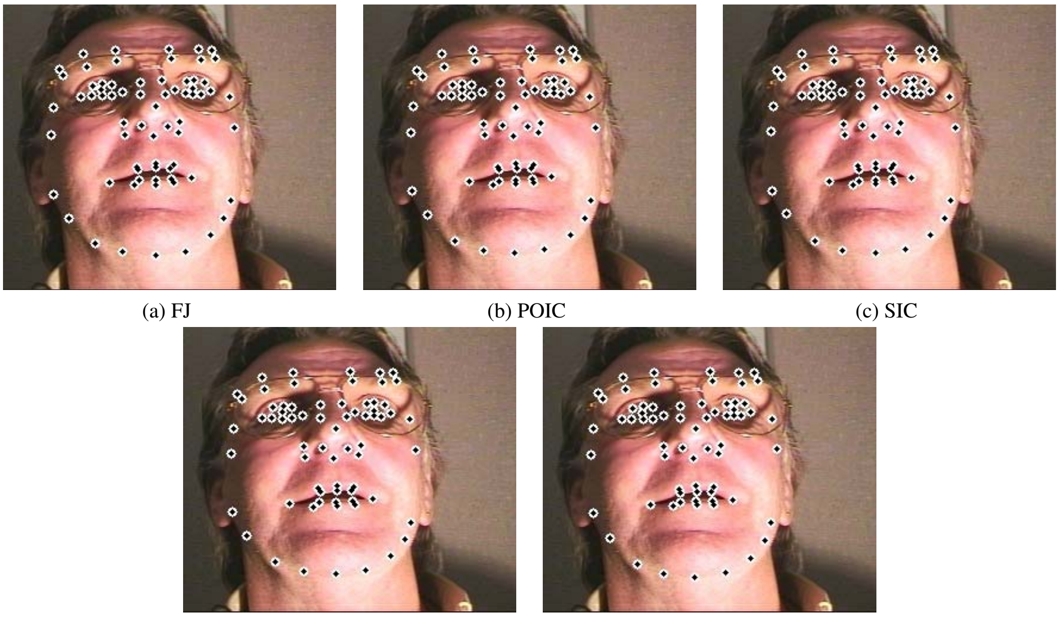


Figure 2: An example of a frame of stereo video from AVOZES.

The design of AVOZES follows a modular framework in accordance with the design methodology proposed in [28]. According to the framework, any AV speech data corpus must con-



(a) FJ (b) POIC (c) SIC

(d) RIC (e) IEBM

Figure 3: Example (frame 53, /Babe/ sequence, subject m2) fitting results for the five AAM methods investigated in this study.

tain at least three mandatory modules, covering the recording setup without and with speakers, as well as the actual speech material sequences which should contain the phonemes and visemes of a language. Additional optional modules cover specific issues, e.g. different levels of illumination, different head poses, or different levels of acoustic noise. Of the six AVOZES modules available, this study makes use of :

- the scene with speaker, head turning;
- 'calibration sequences' exhibiting horizontal and vertical lip movements during speech production; and
- CVC- and VCV-words in a carrier phrase covering the phonemes and visemes of AuE.

Each utterance was recorded once using a clip-on microphone. Recordings were made in clean audio conditions. The subjects were free to move their head in a natural way while speaking, as long as they kept it within the viewable area of the cameras. No markers needed to be used on the faces.

AVOZES contains recordings of 20 native speakers of AuE (10 female + 10 male speakers). 6 speakers wear glasses, 3 wear lip make-up, 2 have beards. At the time of the recordings, the age of the speakers varied between 23 and 56 years.

6. Experimental Setup

Ideally, the accuracy of any lip tracking algorithm should be directly quantified by comparing the tracking results with a set of ground truth data and thus computing error statistics, such as the average error in landmark location. However, it turns out that such ground truth is virtually non-existent for any of the commonly used face or AV speech data corpora, which is not surprising given how tedious and time consuming the process of manually annotating faces is. Approaches to automatically generate ground truth (e.g. [29]) are still in their infancy and have not yet matured to a level where one could reasonably trust the results without checking the annotations.

As a consequence, the performance of lip tracking algorithms therefore needs to be measured indirectly through some other criterion. In this study, the performance of an AV ASR system is used as this criterion. To this end, the word error rates (WER) of the AV ASR system, where the input of the visual speech data stems from the various lip tracking algorithms, are compared (see Section 7 for results), while all other parameters are left unchanged.

The experiments were performed on a subset of AVOZES, consisting of the 10 male speakers and the CVC-word utterances, i.e. the words containing the 18 vocalic phonemes of AuE. Using HTK, a 3-state left-right HMM with no skips is built for each monophone. On the audio side, 13 MFCC parameters and their Δ and $\Delta\Delta$ parameters were used as input. White noise was added to the audio parameters to achieve a SNR of 0db, which represents a significant amount of acoustic noise. As is known from the literature, the advantage of AV ASR systems over audio-only systems is most emphasised in noisy acoustic conditions, where the additional visual speech information helps the system to perform better. Both training and testing were performed at the 0db SNR setting.

On the video side, two different sets of video parameters were extracted. For the AAMs, the landmark locations of the lower face half were used, which mark the position of the lips (inner and outer contour) and the chin line. The locations were normalised with respects to the 2D position of nose tip. For the 3D lip tracking method, the mouth width and height (inner lip contour), protrusion of upper lip, protrusion of lower lip, and relative teeth count were used as video parameters [6]. In both case, the Δ and $\Delta\Delta$ variables were also included.

From the monophone HMMs, context-dependent triphone HMMs were built by simply cloning the monophone HMMs and re-estimating them using triphone transcriptions. An early (feature) fusion approach was taken here, i.e. the values of the video feature vectors were added to the audio feature vectors.

For learning the AAMs, 50 images from the head turning and calibration sequences were manually annotated for each subject before building the models. Each of the AAM algorithms investigated here was then trained with these annotated images to produce a person-dependent model. At each of the steps, 95% of the variation was kept. For the 3D lip tracking method, a manual face model building step needed to be completed first, in which a number of facial landmarks were selected from three video frames of the head turning sequence. For the tracking phase, the AAMs were manually initialised to be roughly within ± 10 pixels from the correct solution.

While person-dependent models were used for the lip tracking, the audio input data for the HMM building phase was taken from all speakers, due to the insufficient amount of training data for individual speakers. As this study is concerned with evaluating the performance of lip tracking algorithms, this approach poses no problem. In all experiments, only the visual speech input was changed, while all other variables were left unchanged.

7. Results and Discussion

Table 1 presents the results for the experiments described in the previous section. Shown are the WERs for each of the 10 male speakers and for each of the lip tracking algorithms. In the first results column, the WERs for the audio-only ASR system are also shown for comparison.

As can be seen from the table, the audio-only ASR system does not perform well in the 0db SNR scenario. Up to half of the words are not recognised correctly. Given the strong acoustic noise, the results are not surprising.

All of the AV ASR results show the significant improvement that can be gained by incorporating visual speech information. Among the AAM algorithms, the Fixed Jacobian method performed worst. This is due to the assumption of a fixed linear update model which is too restrictive for the fitting process to work correctly when the initialisation of the model is not close enough to the solution. Figure 3 shows an example of fitting results for the AAM methods. Note the subtle problems the FJ method has in finding the inner lip contour correctly.

All three inverse-compositional methods performed similar in terms of WERs. One of the reasons for this is the experimental setup with person-specific AAMs, such that the advantages of the Simultaneous and Robust Inverse-Compositional methods in terms of generalisability for person-independent models do not come into play. On a side note, the Project-Out Inverse-Compositional method is still the fastest AAM fitting method to date (real-time capable).

The Iterative Error Bound Minimisation method performs best among the AAM methods and is also overall the best performing method in this study. This method has been shown to have better convergence properties than other AAM methods [26], while retaining high computational efficiency due to the pre-computed update model. In particular, the IEBM method can handle a larger amount of initialisation error better than the other methods.

The 3D lip tracking algorithm based on stereo vision performs better than the Fixed Jacobian method, but generally not as well as the other AAM methods. One of the reasons is that in the case of the AAM methods, information from the entire lower face was used, whereas only information about the lip movements was used in the AV ASR system for the 3D lip tracking method. It is known from the literature that other parts of the face, such as the jaw, also carry useful visual speech information. However, for sequences that contained a large amount of head movement by the speaker, the 3D method outperformed all of the 2D (AAM) methods as its strength in handling head rotations out of the image plane came into play.

8. Conclusions

In this paper, a comparative study of the performance of various 2D and 3D lip tracking algorithms was performed. In particular, a number of variants of the recently popular AAM approach (2D) were compared with a 3D lip tracking algorithm based on stereo vision. As a way of comparing the performance, visual speech features were extracted and formed the input into an AV ASR system, the performance of which was assessed by WER. Experiments were performed on the male subject CVC-word subset of AVOZES and the results showed that the more sophisticated AAM methods (e.g. IEBM) outperformed all other methods (including the simpler AAM methods, e.g. FJ) for the sequences where the speaker's face was frontal or mostly frontal with respect to the camera. However, the 3D lip tracking method performed better for sequences, which contain a lot of head rotation away from the camera plane which is not surprising given the nature of this algorithm.

Given the fact that the AAM methods can work well for single camera input and thus require far less equipment than the 3D lip tracking algorithm used here, future work will focus on further improving the recent AAM methods in terms of accuracy while retaining their efficiency during tracking. Furthermore, the question of generalisability of AAMs is still an open research issue.

9. Acknowledgements

The authors would like to thank Jason Saragih (ANU / CMU) for the use of his DeMoLib software package for the training and fitting of AAMs.

10. References

[1] D. Stork and M. Hennecke, Eds., *Speechreading by Humans and Machines*, ser. NATO ASI Series. Berlin, Germany: Springer-Verlag, 1996, vol. 150.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[3] R. Goecke, "Current Trends in Joint Audio-Video Signal Processing: A Review," in *Proceedings of the IEEE 8th International Symposium on Signal Processing and Its Applications (ISSPA 2005)*, vol. 1. Sydney, Australia: IEEE, Aug. 2005, pp. 70–73.

[4] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A Cascade Image Transform for Speaker Independent Automatic Speechreading," in *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo*, vol. 2. New York (NY), USA: IEEE, Aug. 2000, pp. 1097–1100.

[5] P. Scanlon and R. Reilly, "Lessons From Speechreading," in *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo (ICME'01)*. Tokyo, Japan: IEEE Computer Society, Aug. 2001, pp. 736–739.

[6] R. Goecke, "3D Lip Tracking and Co-inertia Analysis for Improved Robustness of Audio-Video Automatic Speech Recognition," in *Proceedings of the Auditory-Visual Speech Processing Workshop (AVSP 2005)*, E. Vatikiotis-Bateson, D. Burnham, and S. Fels, Eds. Vancouver Island (BC), Canada: ISCA, Jul. 2005, pp. 109–114.

[7] H. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multi-Modal System for Locating Heads and Faces," in *Proceedings of the Second International Conference on Automatic Face*

Speaker	Audio-only	FJ	POIC	SIC	RIC	IEBM	3D
m1	44.44	38.89	16.67	16.67	16.67	16.67	27.78
m2	50.00	38.89	16.67	16.67	22.22	16.67	33.33
m3	44.44	33.33	16.67	22.22	16.67	16.67	27.78
m4	38.89	27.78	16.67	16.67	16.67	11.11	27.78
m5	44.44	33.33	16.67	22.22	22.22	22.22	27.78
m6	50.00	38.89	27.78	22.22	27.78	22.22	33.33
m7	50.00	38.89	27.78	22.22	22.22	22.22	27.78
m8	50.00	38.89	22.22	27.78	27.78	22.22	33.33
m9	38.89	27.78	16.67	16.67	16.67	16.67	27.78
m10	44.44	33.33	22.22	16.67	16.67	16.67	33.33
Average	45.55	35.00	20.02	20.00	20.56	18.33	30.00

Table 1: Word Error Rate (%) results for the AV ASR system trained for the lip tracking algorithms compared in this study. Shown are the average WERs across all 18 CVC-words.

and Gesture Recognition FG'96. Killington (VT), USA: IEEE, Oct. 1996, pp. 88–93.

[8] R. Kjeldsen and J. Kender, "Finding Skin in Color Images," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96*. Killington (VT), USA: IEEE, Oct. 1996, pp. 312–317.

[9] M. Vogt, "Fast Matching of a Dynamic Lip Model to Color Video Sequences Under Regular Illumination Conditions," in *Speechreading by Humans and Machines*, ser. NATO ASI Series, D. Stork and M. Hennecke, Eds., vol. 150. Berlin, Germany: Springer-Verlag, 1996, pp. 399–407.

[10] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel, "Real-time Face and Facial Feature Tracking and Applications," in *Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98*, D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, Eds., Terrigal, Australia, Dec. 1998, pp. 79–84.

[11] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR2001*, vol. 1. Kauai, USA: IEEE, Dec. 2001, pp. 511–518.

[12] E. Petajan, "Automatic Lipreading to Enhance Speech Recognition," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1984.

[13] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, "Which Components of the Face Do Humans and Machines Best Speechread?" in *Speechreading by Humans and Machines*, ser. NATO ASI Series, D. Stork and M. Hennecke, Eds., vol. 150. Berlin, Germany: Springer-Verlag, 1996, pp. 315–328.

[14] H.-H. Bothe, "Relations of Audio and Visual Speech Signals in a Physical Feature Space: Implications for the Hearing Impaired," in *Speechreading by Humans and Machines*, ser. NATO ASI Series, D. Stork and M. Hennecke, Eds., vol. 150. Berlin, Germany: Springer-Verlag, 1996, pp. 445–460.

[15] 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, Jan. 1995.

[16] J. Luettin, N. Thacker, and S. Beet, "Active Shape Models for Visual Speech Feature Extraction," in *Speechreading by Humans and Machines*, ser. NATO ASI Series, D. Stork and M. Hennecke, Eds., vol. 150. Berlin, Germany: Springer-Verlag, 1996, pp. 383–390.

[17] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," in *Proceedings of the European Conference on Computer Vision ECCV'98*, ser. Lecture Notes in Computer Science 1406, H. Burkhardt and B. Neumann, Eds., vol. 2. Freiburg, Germany: Springer, Jun. 1998, pp. 484–498.

[18] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. Bangham, "Lipreading using shape, shading and scale," in *Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98*, D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, Eds., Terrigal, Australia, Dec. 1998, pp. 73–78.

[19] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[20] L. Revéret and C. Benoît, "A new 3D Lip Model for Analysis and Synthesis of Lip Motion," in *Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98*, D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, Eds., Terrigal, Australia, Dec. 1998, pp. 207–212.

[21] 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, Oct. 1998.

[22] J. Saragih, "The Generative Learning and Discriminative Fitting of Linear Deformable Models," Ph.D. dissertation, Australian National University, Canberra, Australia, Aug. 2008.

[23] S. Baker and I. Matthews, "Equivalence and Efficiency of Image Alignment Algorithms," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR2001*, vol. 1. Kauai (HI), USA: IEEE Computer Society, Dec. 2001, pp. 1090–1097.

[24] S. Baker, R. Gross, and I. Matthews, "Lucas-kanade 20 years on: A unifying framework: Part 3," Robotics Institute, Carnegie Mellon University, Pittsburgh (PA), USA, Tech. Rep. CMU-RITR-03-35, Nov. 2003.

[25] R. Gross, I. Matthews, and S. Baker, "Constructing and Fitting Active Appearance Models With Occlusion," in *Proc. IEEE Workshop on Face Processing in Video*. Washington (DC), USA: IEEE, Jun. 2004, DOI 10.1109/CVPR.2004.43.

[26] J. Saragih and R. Goecke, "Iterative Error Bound Minimisation for AAM Alignment," in *Proceedings of the 18th International Conference on Pattern Recognition ICPR 2006*, vol. 2. Hong Kong: IEEE, Aug. 2006, pp. 1192–1195.

[27] R. Goecke and J. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES," in *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004*, vol. III, Jeju, Korea, Oct. 2004, pp. 2525–2528.

[28] J. Millar, M. Wagner, and R. Goecke, "Aspects of Speaking-Face Data Corpus Design Methodology," in *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004*, vol. II, Jeju, Korea, Oct. 2004, pp. 1157–1160.

[29] J. Saragih and R. Goecke, "Monocular and Stereo Methods for AAM Learning from Video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2007*. Minneapolis (MN), USA: IEEE Computer Society, Jun. 2007, DOI: 10.1109/ICCV.2007.4409106.