

Continuous Pose-Invariant Lipreading

Patrick Lucey, Sridha Sridharan and David Dean

Speech, Audio, Image and Video Technology Laboratory
Queensland University of Technology, Brisbane, QLD, 4000, Australia

p.lucey@qut.edu.au, s.sridharan@qut.edu.au, ddean@ieee.org

Abstract

In audio-visual automatic speech recognition (AVASR), no research to date has been conducted into the problem of recognising visual speech whilst the speaker is moving their head. In this paper, we extend our current system to deal with this task, which we entitle *continuous pose-invariant lipreading*. By developing an AVASR system which can deal with such a scenario, we believe we are making the system effectively “real-world” as it requires little cooperation from the user and as such can be used in a host of realistic applications (e.g. mobile phones, in-vehicles etc.). In this proof of concept paper, we show via our experiments on the CUAVE database, that recognising visual speech whilst a speaker is moving their head during the utterance is feasible.

Index Terms: audio-visual automatic speech recognition (AVASR), lipreading, pose-invariance, pose-estimation.

1. Introduction

It is well known that visual speech information extracted from video of a speaker’s mouth region can improve performance of automatic speech recognisers, especially in the presence of acoustic noise. However, a major reason stymying progress in this area is the lack of work focussing on head position with nearly all of the previous work published in this area focussing on a speaker’s fully frontal face (see [1, 2] for overviews).

Having an audio-visual automatic speech recognition system (AVASR) able to recognise speech regardless of a speaker’s head position would be of great benefit in many situations as shown in Figure 1. In the first example (Figure 1(a)), having an in-vehicle AVASR system that could deal with random pose changes would be of most benefit due to the frequent movement of the driver’s head. This would be the same for AVASR in mobile phones (Figure 1(b)) as there would be no guarantee of where the speaker’s head/lips would be positioned. Another scenario of interest for AVASR would be in meetings and lectures inside smart rooms (Figure 1(c)). In this situation, pan-tilt-zoom (PTZ) cameras would be able to track the meeting speaker(s) providing high resolution views. However, like the previous examples, due to the camera being fixed, frontal speaker views cannot be guaranteed.

The smart-room scenario was the focus of our previous work in [3, 4, 5]. A major focus of this research was to develop a lipreading system which could recognise visual speech regardless of head pose [4, 5]. The experiments performed in this work were constrained just to the stationary scenario, where the speaker was fixed in one pose (i.e. frontal or profile) for the entire utterance and the pose of the speaker was assumed.

Even though this work provided a good start in the overall goal of achieving AVASR across multiple views, this stationary assumption hardly makes the lipreading system “real-world”. In



Figure 1: *Examples of practical scenarios where frontal AVASR is inadequate: (a) Driver data inside an automobile; (b) Mouth data captured from a mobile phone; (c) Data from a lecturer captured by a pan-tilt-zoom camera inside a smart room.*

an attempt to remedy this situation, in this paper we propose a *continuous pose-invariant lipreading system*, which can recognise visual speech whilst allowing the speaker to freely move their head around while speaking. In this paper, we concentrate entirely on lipreading (also known as visual speech recognition or automatic speechreading). We limited our work just to the visual modality to prevent the results being skewed from the audio signal.

To our best knowledge, this constitutes the first attempt in attacking the problem of continuous pose-invariant lipreading. The specific contributions emanating from this work are:

- Extending the novel pose-invariant lipreading paradigm to the continuous domain, which involves recognising visual speech regardless of head position (Section 2).
- Incorporating a pose estimator into the visual front-end¹, which was developed from a series of frontal and non-frontal face classifiers (Section 2 and Section 4).
- Developing a protocol for the CUAVE audio-visual database, to be used for the continuous pose-invariant lipreading paradigm (Section 3). As this is the only data freely available for such a task, we hope that this protocol can be used by future researchers so that comparisons can be made.

Following these contributions, we present the results from our continuous pose-invariant lipreading system (Section 5), which was developed via our current start-of-the-art system (Section 4). We believe that the work in this paper is a notable step in bringing the goal of achieving AVASR in real-world settings.

¹In literature, there is some confusion to what the visual front-end refers to. In this paper, the visual front-end refers solely to the task of locating and tracking a speaker’s face and facial features

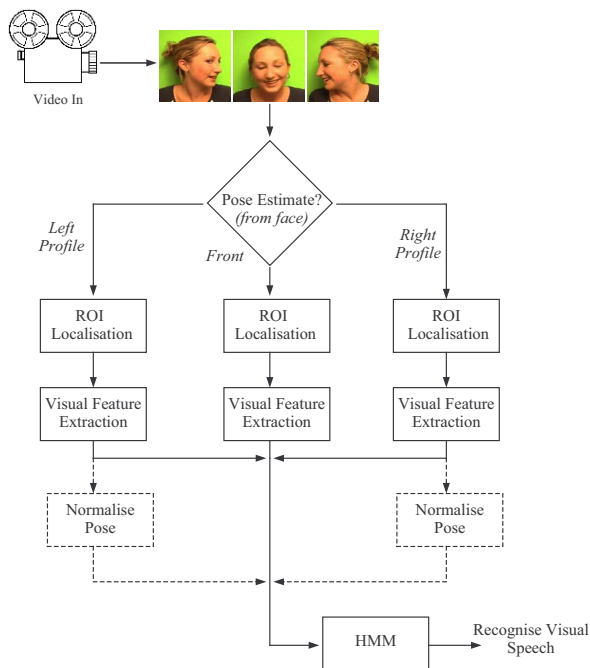


Figure 2: Block diagram of the continuous pose-invariant lipreading system.

2. Continuous pose-invariant lipreading

The deployment of a continuous pose-invariant lipreading system is very similar to a normal lipreading system, albeit with one modification. This modification is the inclusion of a *pose-estimator* at the front of the visual front-end, as depicted in Figure 2. It can be seen that once we gain an estimation of the pose of the speaker, we can use this estimation to direct the system to locate the region-of-interest (ROI) of that particular pose. Once the ROI has been extracted, we can perform visual feature extraction and the subsequent features can be combined into a single model or normalised into the frontal pose as described in our previous work [4, 5].

2.1. Pose estimation

For continuous pose-invariant lipreading, a multi-pose visual front-end paradigm has to be visited. This highlights a benefit of using a boosted cascade of simple classifiers as described by Viola and Jones [6], as it is able to accommodate for the multi-pose scenario by the inclusion of a pose-estimator, which still allows for extremely quick localisation of faces and facial features [7].

According to Jones and Viola [7], the multi-pose visual front-end depicted in Figure 2 is the preferred option compared to a holistic approach. A reason they gave was that a holistic approach, where a single classifier is trained to locate the face position of all poses, is unlearnable with existing classifiers. In their informal experiments they found that using the holistic approach yielded extremely inaccurate results, most probably due to over generalisation. Preliminary work in using this holistic approach also backs up this assertion. Another disadvantage of the holistic approach is no information about the speaker's pose is gained. This means that important information is lost, which could be used to improve the system (i.e. project the unwanted

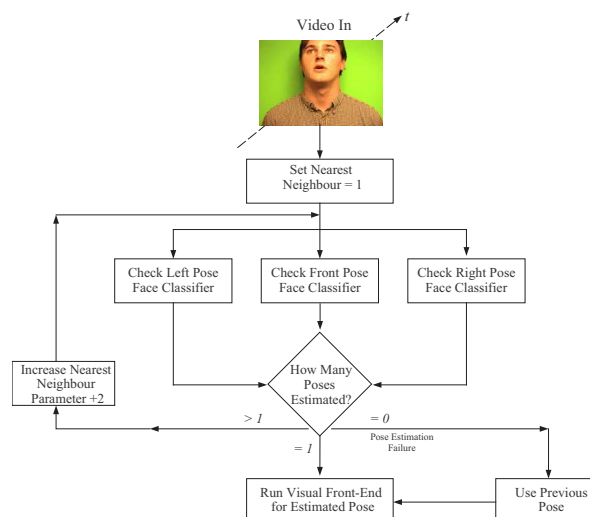


Figure 3: Block diagram of the pose estimator which incorporates the pose estimation with the face localisation.

pose into the desired pose – see [4, 5] for details).

The pose-estimation of a speaker's face is essentially a *chicken or the egg* problem. Firstly, we have to locate the face to determine the pose, but we have to know the pose to find the face. A prudent strategy to achieve this would be to solve both of these problems simultaneously. To do this, we have to use a face classifier for each pose and then use this classifier to exhaustively search across each position and scale in the image. As this is extremely expensive in terms of computation, a rapid detection framework like the Viola-Jones framework has to be employed. In [7], Jones and Viola did such a thing by building different detectors for different poses of the face.

In this paper, we used a similar strategy to Jones and Viola to develop the pose estimator. A diagram of the devised pose estimator is depicted in Figure 3. From this figure it can be seen that given a frame of a speaker's face, we apply all the face classifiers to the image to determine the location of the face. Once a face has been located by a pose specific classifier, we then give this information to the continuous pose-invariant lipreading system which is described by Figure 2. In our experiments, we found that this procedure works well when only one of the poses is estimated. However, it gets complicated when there is more than one pose estimated as there is no way of knowing which pose is the correct one (e.g. both the frontal and right profile poses have been estimated on the same frame). To counteract this problem, we utilised the *nearest neighbour* variable. The nearest neighbour variable is a parameter in OpenCV's Viola-Jones based generic object detector [8], which essentially regulates how much an object has to look like the object of interest before it is recognised as that object. In our system, if we find only one face/pose, then we use that estimate. However, if more than one pose is estimated, we increase the nearest neighbour parameter by two to determine which is the more likely pose. This process is continued until only one face/pose is found. If no unique face/pose is found, the face and pose information from the previous frame is used. See Section 4.1 for full description of training and development of face and facial feature classifiers.



Figure 4: Examples of the CUAVE individual sequences.

3. CUAVE database

An audio-visual database which contains speakers talking in non-frontal poses, is the Clemson University Audio-Visual Experiments or CUAVE database [9]. The individual section of the CUAVE database was broken into 2 parts. The first was for isolated-digits and the second was the connected-digits. As no profile data was included in the connected-digits section, only the isolated-digits portion was used. Each isolated-digit sequence was broken into the following four tasks:

1. Normal, where each speaker spoke 50 digits whilst standing still naturally,
2. Moving, where each speaker was asked to move side-to-side, back-and-forth, or tilt the head while speaking 30 digits,
3. Right profile, where each speaker utters 10 digits in the right profile pose, and
4. Left profile, where each speaker utters 10 digits in the left profile pose.

Examples of these tasks are given in Figure 4. In addition to performing experiments on these four individual tasks, experiments on the combination of all the tasks were also undertaken. As such, continuous video data across all these tasks were required (i.e. speaker in shot at all times), which meant that we could only use 33 of the 36 speakers. As only one sequence was available per speaker, a *quasi* speaker-independent paradigm was used, which consisted of 10 different train/test sets, consisting of 25 speakers for training and 8 speakers for testing for each set. We termed this as *quasi* speaker-independent because it is not a fully speaker-independent task, as only one visual front-end was developed.

4. The lipreading system

4.1. Multi-pose mouth localisation and tracking

In these experiments, we used the Adaboost framework of Viola and Jones [6], later extended by Leinhardt and Maydt [10], to perform the multi-pose mouth region-of-interest (ROI) localisation and extraction. This framework allowed us to generate face and facial feature classifiers specific for each of the poses (i.e. frontal, left and right profile). These classifiers were generated using the OpenCV libraries [8].

As shown in Figure 2, the first step is to both estimate the face/pose of the speaker. If the frontal pose was estimated, the two eyes were located and then a coarse mouth region was located. From these estimates, we applied classifiers to located the lip corners which were then used to extract a normalised 32×32 pixel ROI. If the right profile face/pose was estimated, the left eye and nose were located. These located features were then used to estimate the position of the mouth center and left mouth corner. A normalised 32×32 pixel profile mouth ROI



Figure 5: Examples of face and facial feature localisation from the multi-pose visual front-end. The bottom row gives the associated examples of the extracted normalised 32×32 ROIs

was then extracted, based on the distance from the left mouth to the left eye. We used these two points as reference points, as they were the most reliable to located. The same procedure was used for the left profile, albeit with opposite features (i.e. right eye, right mouth corner etc.). As the Adaboost framework allows for extremely quick detection, we were able to perform this procedure on every frame and used median filtering to allow for smooth tracking.

For the training of the pose-estimator and pose specific visual front-ends, only the frontal, left-profile and right-profile poses were considered. The face and facial feature classifiers for each pose were trained up on 500 manually annotated positive examples and 2000 negative examples. The set of 500 positive examples for each pose were taken from all the 33 subjects. We did this because there were not enough speakers to create classifiers to achieve accurate localisation for the ten different train/test sets devised. As such, only one variant of the pose-estimator and visual front-ends were developed for these experiments. The set of positive examples for each pose were augmented by including rotations of $\pm 5^\circ$, $\pm 10^\circ$, providing a set of 2500 positive examples. A separate validation set of 39 annotated images for each specific pose were used to test the pose-estimator and pose specific visual front-ends. Examples of the face/pose and facial feature localisation and the extracted ROIs for the various poses are shown in Figure 5. From experiments conducted on the manually annotated validation set, our pose estimator correctly estimated approximately 90% of the poses.

4.2. Visual feature extraction

For both frontal and profile poses, the same visual feature extraction process was applied. Our implementation is similar to that of Potamianos et al. [2], however in the continuous pose-invariant lipreading paradigm, no feature normalisation is used due to the constant changing of poses throughout the sequence which we found introduced more error. Following ROI extraction, a two-dimensional, separable, discrete cosine transform (DCT) was then applied on the resulting mean-removed ROI, with the $M = 100$ top DCT coefficients retained, according to a zig-zag pattern. An intra-frame linear discriminant analysis (LDA) step was then used to project the features down to $N = 30$ dimensions, resulting in a “static” visual feature vector. Subsequently, in order to incorporate dynamic speech information, five of these neighboring static feature vectors over $\pm J$ adjacent frames were concatenated, and were projected via an inter-frame LDA step to yield a “dynamic” visual feature vector, extracted at the video frame rate of 30 Hz, resulting in a

<i>Task</i>	<i>WER (%)</i>
Normal	46.88
Moving	67.26
Right Profile	71.95
Left Profile	71.54
<i>Average Individual</i>	<i>57.97</i>
Continuous	61.20

Table 1: *The upper part of the table shows the average lipreading performance for each individual task, whilst the bottom part compares the average of the individual tasks against the continuous pose-invariant results.*

final feature vector of dimension 40. We found this to be optimal configuration via heuristic and empirical evidence. For the LDA matrix calculation, we used the HMM states as classes which we based on the forced alignment of the audio-only channel of the database.

4.3. Speech recognition system

In this paper, we employed a hidden Markov model (HMM) based ASR system. In particular, for the connected-digit recognition task considered here, eleven nine-state, left-to-right, whole-word models are used, one for each digit (both “oh” and “zero” are included), with seven Gaussian mixtures per state. A silence and short-pause model are also employed. All models are bootstrapped from a segmentation of the audio channel of the database, obtained by an audio-only HMM with identical topology, and trained by the expectation-maximization algorithm. For testing, Viterbi decoding is used with no grammar or language model present (i.e., no constraints are imposed on the digit string length). The HTK toolkit is utilized for both system training and testing [11].

5. Lipreading results

The experiments were broken up into two sections. The first section investigated the lipreading performance of the four individual tasks; normal, moving, right profile and left profile. For each of these individual tasks, individual HMM models were trained and tested solely on the data which referred to their respective task. In the second section, we implemented the system depicted in Figure 2, with one HMM model which was trained up on all the different tasks was used for testing. This was termed the “continuous” result. It should be noted that all the features were extracted from the same multi-pose mouth locator and tracker, which uses the pose estimator.

The results for the continuous pose-invariant lipreading system are given in Table 1. For the individual tasks, the normal task achieved the best performance with an average lipreading WER of 46.88%. This was to be expected as this was the easiest task to perform, due to the speaker being relatively stationary. Even though the moving task had the speaker in the frontal pose, having the speaker move their head back and forth whilst speaking degraded the lipreading performance markedly to 67.26%. As this task had the speaker moving their head quite fast, it can be assumed that a major reason for this poor performance is due to poor tracking of the ROI. The left and right profile tasks achieved even worse WERs of 71.54% and 71.95% respectively. The main reason for this is due to the relatively small size of the CUAVE database, which effectively meant that there was not enough speech data to adequately train the models for each task. This would be the case especially for the profile models, as only 250 words were available to train the models.

The average individual WER over the 4 tasks was 57.97%.

Our continuous pose-invariant system achieved a WER of 61.20% which is close to the average individual value indicating feasibility of our approach. Even though the WER of the proposed system is slightly higher (this result would expected since each individual task has its own pose specific model, compared to our continuous system having to generalise across all poses) the results warrant further research into improving this performance by better design of the pose estimator and the modelling process.

6. Conclusion

In this paper we have extended the current research of AVASR to include the problem of recognising visual speech whilst the speaker is changing their head pose. The key module in our novel system is the pose estimator, which we developed in conjunction with the face localiser. The results indicate that the goal of continuous pose-invariant lipreading is indeed attainable as an achieved WER of 61.20% was obtained with this first attempt which was quite close to the average result obtained when individually trained HMMs are used to for each pose. The development of a continuous pose-invariant lipreading system is the next step which will facilitate the deploying a fully functional “real-world” AVASR system and the key research challenge of this task is developing a robust visual front-end which has an improved pose-estimator.

7. Acknowledgements

This research was supported by the Australian Research Council Grant No:LP0562101. Thanks also to Clemson University for freely supplying their database.

8. References

- [1] Chibelushi, C. C., Deravi, F., & Mason J. S. D., “A review of speech-based bimodal recognition”, *IEEE Transactions on Multimedia*, 4, 23–37, 2002.
- [2] Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W., “Recent advances in the automatic recognition of audio-visual speech”, In *Proceedings of the IEEE*, 91(9), 1306–1326, 2003.
- [3] Lucey, P., & Potamianos, G., “Lipreading using profile versus frontal views”, In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing – MMSP*, (Victoria, Canada), 24–28, 2006.
- [4] Lucey, P., Potamianos, G., & Sridharan, S., “A unified approach to multi-pose audio-visual ASR”, In *Proceedings of the Conference of the International Speech Communication Association – Interspeech*, (Antwerp, Belgium), 650–653, 2007.
- [5] Lucey, P., Potamianos, G., & Sridharan, S., “An extended pose-invariant lipreading system”, In *Proceedings of the International Workshop on Auditory-Visual Speech Processing – AVSP*, (Hilvarenbeek, The Netherlands), 2007.
- [6] Viola, P., & Jones, M., “Rapid object detection using a boosted cascade of simple features”, In *Proceedings of the International Conference on Computer Vision and Pattern Recognition – CVPR*, (Kauai, HI, USA), 511–518, 2001.
- [7] Jones, M., & Viola, P., “Fast multi-view face detection”, Tech. Rep. TR2003-96, MERL, June 2003.
- [8] OpenCV: Open Source Computer Vision Library. [online] <http://sourceforge.net/projects/opencvlibrary>
- [9] Patterson, E., Gurbuz, S., Tufekci, Z., & Gowdy, J., “CUAVE: A new audio-visual database for multimodal human-computer interface research”, In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing – ICASSP*, (Orlando, FL, USA), 2002.
- [10] Leinhardt, R., & Maydt, J., “An extended set of Haar-like features”, In *Proceedings of the International Conference on Image Processing – ICIP*, (Rochester, NY, USA), 900–903, 2002.
- [11] Young, S., Everman, G., Hain, T., Kershaw, D. Moore, G., Odell, J., et al. *The HTK Book (for HTK Version 3.2.1)*. Entropic Ltd, 2002.