Visual Front-End Wars: Viola-Jones Face Detector vs Fourier Lucas-Kanade

Shahram Kalantari, Rajitha Navarathna, David Dean, Sridha Sridharan

Speech, Audio, Image and Video Technology Lab, Queensland University of Technology, Australia.

{s1.kalantari,r.navarathna,d.dean,s.sridharan}@qut.edu.au

Abstract

The performance of visual speech recognition (VSR) systems are significantly influenced by the accuracy of the visual front-end. The current state-of-the-art VSR systems use off-the-shelf face detectors such as Viola-Jones (VJ) which has limited reliability for changes in illumination and head poses. For a VSR system to perform well under these conditions, an accurate visual front end is required. This is an important problem to be solved in many practical implementations of audio visual speech recognition systems, for example in automotive environments for an efficient human-vehicle computer interface. In this paper, we re-examine the current state-of-the-art VSR by comparing off-the-shelf face detectors with the recently developed Fourier Lucas-Kanade (FLK) image alignment technique. A variety of image alignment and visual speech recognition experiments are performed on a clean dataset as well as with a challenging automotive audio-visual speech dataset. Our results indicate that the FLK image alignment technique can significantly outperform off-the shelf face detectors, but requires frequent fine-tuning.

Index Terms: Visual Front-ends, Viola-Jones, Fourier Lucas-Kanade, Visual Speech Recognition

1. Introduction

A VSR system has to be able to locate and track the visible articulators which cause human speech. It is well known that the majority of these visible articulators emanate from the region around a speaker's mouth [4]. Therefore in VSR, it is most important to have a robust visual front-end system which has the ability to track and detect the speaker's mouth region or *region of interest* (ROI). If the visual front-end performs poorly then the error in tracking/registration of the speaker's ROI will scatter throughout the system and causes erroneous results (as shown in Figure 1).

Having a very accurate visual front-end system is a very difficult task due to many challenges including the speaker's pose, facial expressions, change of illumination, and presence of structural components (i.e beards, moustaches and glasses of speakers) [19]. The choice of visual front-end is dependent on the type of application and the recording conditions of the captured video. A plethora of works has been conducted to develop visual front-end systems such as template matching, feature invariant and appearance based approaches [19].

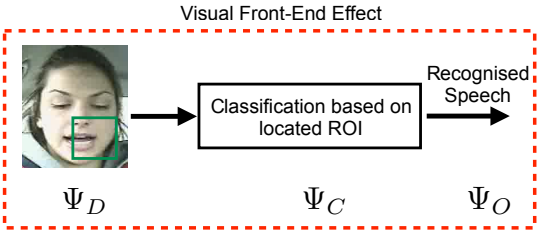

Figure 1: Depiction of the cascading front-end effect. If the region of interest (ROI) is located poorly, then this corrupt input will cascade throughout the system and will cause additional visual speech recognition errors.

Yuille et al. [21] first applied template matching methods for mouth and eye localisation using appearance and shape models. In this approach, deformable templates of the eyes and labial contour is fitted to an intensity model by calculating a cost function based on the gray-scale intensity edges around the template's boundaries. Unfortunately, this approach shows poor performance due to the heuristic nature of the shape and intensity models when applied across a large number of subjects. Under feature invariant approaches, methods based on edges [14], colour [16], as well as localised texture [7] have been used to generate geometric lip models. However, these approaches are problematic in conditions of poor illumination and speaker movement and require extremely precise localisation of lip features. In the VSR literature, appearance based approaches have been widely used in visual front-end systems [6, 18] due to them being well suited to many different objects (face, eyes, nose etc.) under varied conditions due to their probabilistic nature and have shown good performances compared with other approaches.

Even though there are complex deformable face alignment methods such as active appearance models (AMM) [2], constrained local models (CLM) [15] and so on, most VSR systems have been conducted using ROI-based or coarse alignment techniques due to their simplicity. The most common of these approaches is the haar-like feature matching approach of the off-the-

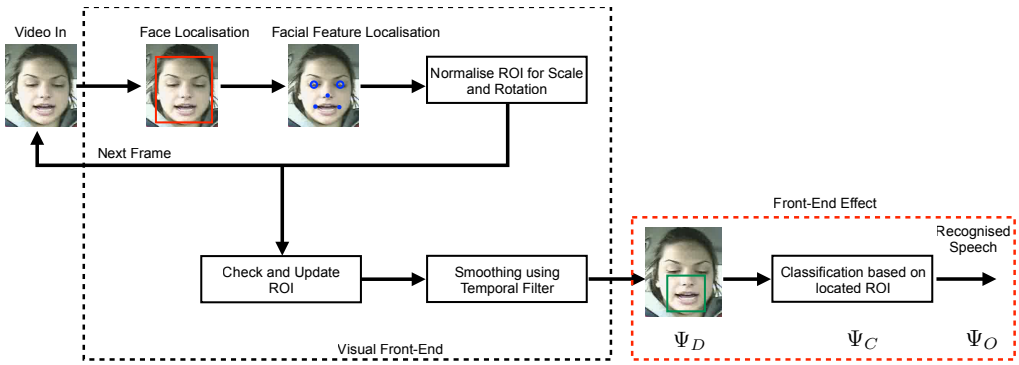

Figure 2: Block diagram of visual front-end system for VSR and the cascading of the front-end effect.

shelf Viola Jones (VJ) object detector (available in the OpenCV image processing libraries [17]. Since the original development of VJ object detectors, there have been many advances in computer vision and image alignment and it is therefore worthwhile to re-examine most recent image alignment techniques in comparison with VJ approach. In this work we will be comparing a modern image alignment technique based on the Fourier Lucas-Kanade (FLK) method proposed by Lucey et al. [9] to the coarse alignment approach of using typical VJ object detectors. We will perform a variety of image fitting and VSR experiments on both a clean dataset and a challenging automotive speech dataset to fully evaluate both techniques.

2. Visual speech feature extraction

A typical VSR system consists of a visual-front, followed by feature extraction and classification stages.

2.1. Visual front-end

In VSR, the most important stage is to reliably track and detect the speaker's ROI. The majority of these visible articulators emanate from the region around a speaker's mouth. The success of the entire system depends on designing a robust visual front-end which will be able to locate and track the speaker's face and facial features across a variety of conditions (i.e. illumination and head pose). If the visual front-end is not accurate, it will have a detrimental effect on feature extraction and classification stages. This error from the visual-front end will cascade throughout the system and will cause error in visual speech recognition. This effect is known as the *front-end effect*, and can be formally written as,

$$\Psi_O = \Psi_D \times \Psi_C \tag{1}$$

where Ψ_D is the probability that the ROI has been correctly located, Ψ_C is the decision probability given the located ROI and Ψ_O is the overall probability that the system recognizes the correct speech. An overview of the visual-front-end process with the *front-end effect* is depicted in Figure 2.

2.2. Visual speech feature extraction

Visual speech is best discriminated by the movements of the visual articulators (i.e mouth, lips and jaw) [13]. Visual feature extraction seeks to find representations of the given observations that provide discrimination between the various speech units whilst providing invariance to irrelevant transforms within the same speech units.

Cascading appearance-based features, devised by Potamianos et al. [13] have been established as the state of the art for visual feature extraction as they contain information about the visible articulators such as tongue, teeth, and the muscles around the jaw and can be computed very quickly, lending themselves to real-time implementation. Essentially, this process is broken into two sections: static and dynamic feature extraction.

2.2.1. Static visual features

Typically, following ROI tracking from the visual front-end, the ROI images are converted to gray-scale and image-mean normalization is performed to help attenuate any irrelevant information, such as illumination or long-term variations in speaker appearance. Then a two-dimensional, separable, DCT is applied to the mean-removed image. The top $\mathbf{m}$ higher energy components according to a zig-zag pattern from the top-left containing the most variability in the tracked ROI, are then used as static features to represent the visual speech information within each frame of the ROI.

2.2.2. Dynamic visual features

The best features for representing visual speech are generally focused on the movement of the features, rather than the features within each frame [13]. One technique which has shown good performance is the use of linear discriminant analysis (LDA) to extract the relevant dynamic speech features from the ROI [12]. In order to incorporate dynamic speech information the static features in multiple successive frames are concentrated before

speech-class based LDA is performed based on the force alignment of acoustic models with the known transcription. The transformation matrix is found from the concatenation of $\pm J$ frames surrounding the current frame. Each input frame to the LDA step can be represented as Equation 2 and the resulted feature vector is size of $(2J+1)\mathbf{M}$.

$$\mathbf{O}^C t = [(\mathbf{O}^s_{t-J})', \ldots, (\mathbf{O}^s_t)', \ldots, (\mathbf{O}^s_{t+J})']' \quad (2)$$

The obtained static features can then be projected via an inter-frame LDA stage, where the LDA transformation is trained on acoustically-aligned subword units, to yield a $\mathbf{q}$ dimensional dynamic visual feature vector.

3. Visual front-ends

3.1. Viola Jones

Initially, an efficient visual front end system which is able to track and locate the speaker's face and mouth ROI was developed using the VJ algorithm [17]. Given the video of a speaker, initially the system detects the face using the face classifier. Once the face was located, we then locate the eyes and based on these locations, the face was normalised with respect to scale, rotation and translation based on an inter-ocular distance of 40 pixels. We then applied a mouth classifier and extracted a ROI to be used in visual speech recognition. The extracted mouth region mostly contains jaw and cheeks and it was down-sampled to 40×40 pixels to keep the dimensionality low. The location of 40×40 pixel mouth region is smoothed using a mean filter. Following the ROI localisation, this process was performed over consecutive frames. The previous ROI location is used if the detected ROI is too far away from previous frame.

3.2. Fourier Lucas-Kanade template update

Recently, Lucey et al. [9] proposed an extension to the Lucas-Kande (LK) [8] algorithm for fitting a template across multiple filter responses in the Fourier domain which they referred as Fourier Lucas-Kanade (FLK). The goal of the LK algorithm is to find the parametric warp $\mathbf{p}$ that minimizes the sum of squared difference (SSD) between a template image T and a warped source image I. This can be written as,

$$\arg\min_{\mathbf{p}} \parallel I(\mathbf{p}) - T(\mathbf{0}) \parallel^2, \quad (3)$$

where $I(\mathbf{p})$ represents the warped input image using the warp specified by the parameters $\mathbf{p}$, while $T(\mathbf{0})$ represents the un-warped template image. The FLK method reformulates Equation 3 by solving:

$$\arg\min_{\mathbf{p}} \parallel \hat{I}(\mathbf{p}) - \hat{T}(\mathbf{0}) \parallel^2_{\mathbf{S}}, \quad (4)$$

where,

$$\mathbf{S} = \sum_{i=1}^{M} (\mathrm{diag}(\hat{\mathbf{g}}_i))^T \mathrm{diag}(\hat{\mathbf{g}}_i), \quad (5)$$

and $\hat{I}(\mathbf{p}), \hat{T}(\mathbf{0}), \hat{\mathbf{g}}_i$ are the $2D$ Fourier transforms of vectorized images $I(\mathbf{p}), T(\mathbf{0})$ and choice of filters $\mathbf{g}_i$ respectively. The Equation 4 can then be represented by a matrix $\mathbf{F}$ which contains the Fourier basis vectors. This can be seen in the following FLK objective function,

$$\arg\min_{\mathbf{p}} \parallel I(\mathbf{p}) - T(\mathbf{0}) \parallel^2_{\mathbf{F}^T\mathbf{S}\mathbf{F}} . \quad (6)$$

The minimization of the Equation 6 is a non-linear optimization task with respect to $\mathbf{p}$. This can be linearised by performing Taylor series,

$$\arg\min_{\Delta\mathbf{p}} \parallel I(\mathbf{p}) + \mathbf{J}\Delta\mathbf{p} - T(\mathbf{0}) \parallel^2_{\mathbf{F}^T\mathbf{S}\mathbf{F}}, \quad (7)$$

where the Jacobain matrix of $I(\mathbf{p})$ can be written as $\mathbf{J} = \left(\frac{\partial I(\mathbf{p})}{\partial \mathbf{p}}^T\right)$. The explicit solution for $\Delta\mathbf{p}$ can be written as

$$\Delta\mathbf{p} = \mathbf{H}^{-1}\mathbf{J}^T\mathbf{F}^T\mathbf{S}\mathbf{F}\left[T(\mathbf{0}) - I(\mathbf{p})\right], \quad (8)$$

where the pseudo Hessian matrix is defined by

$$\mathbf{H} = \mathbf{J}^T\mathbf{J} \ . \quad (9)$$

However, the above formulation is problematic due to the re-estimation of the Hessian matrix at each iteration.

3.2.1. Fourier inverse compositional algorithm

The main point in this algorithm is linearized $T(\Delta\mathbf{p})$ rather than $I(\mathbf{p}+\Delta\mathbf{p})$ results in the following linearized objective function [1] ,

$$\arg\min_{\Delta\mathbf{p}} \parallel I(\mathbf{p}) - T(\mathbf{0}) - \mathbf{J}_{flk(ic)}\Delta\mathbf{p} \parallel^2_{\mathbf{F}^T\mathbf{S}\mathbf{F}} \ . \quad (10)$$

Since $\mathbf{J}_{flk(ic)} = \frac{\partial T(\mathbf{0})}{\partial \mathbf{p}}$ depends only on the template image, $\mathbf{J}_{flk(ic)}$ remains constant across all iterations. In addition to $\mathbf{J}_{ic}$, the pseudo-Hessian $\mathbf{H}_{flk(ic)} = \mathbf{J}^T_{flk(ic)}\mathbf{F}^T\mathbf{S}\mathbf{F}\mathbf{J}_{flk(ic)}$ also remains static for all iterations. The solution to the Equation 10 can be written as

$$\Delta\mathbf{p} = \mathbf{H}^{-1}_{flk(ic)}\mathbf{J}^T_{flk(ic)}\mathbf{F}^T\mathbf{S}\mathbf{F}\left[I(\mathbf{p}) - T(\mathbf{0})\right]. \quad (11)$$

The current warp parameters are iteratively updated using[1] $\mathbf{p} \leftarrow \mathbf{p} \circ \Delta\mathbf{p}^{-1}$.

We implemented the FLK inverse compositional template tracking method to find the best match to the template mouth region in every subsequent frame in the given video. The template is updated in every frame and by manually inspecting the extracted ROI, if it is too far from

[1] The operation $\circ$ represents the composition of two warps

(a) (b)

Figure 3: Visual examples from (a) CUAVE database and (b) AVICAR database

the actual mouth region we manually re-initialise the current template. In our experiments, we realized that on CUAVE database, this manual update needed to be performed in every 400 frames, while for AVICAR database, it was every 100 frames. We defined the weighting matrix $\mathbf{S}$ in Equation 5 using a bank of Gabor filters with 9 scales times 8 orientations [3].

4. Experimental setup

4.1. The databases

4.1.1. The CUAVE database

The CUAVE database [11] is a publicly available audio-visual database which contains speakers talking in frontal and non-frontal poses. This database consists of two sections, with the first being the individual and the second being the group section. The individual section of the CUAVE database consists of 36 speakers (19 male and 17 female speakers). All the recorded speech is in English with 29.97 frames per second and resolution of 720×480 pixels. The database is captured in a clean environment.

We selected frontal pose of 31 subjects for the experiments where, each speaker spoke 50 digits whilst standing still naturally. Some examples from the individual section for the frontal pose is given in Figure 3 (a).

4.1.2. The AVICAR database

The AVICAR database is a publicly available in-car speech corpus containing multi-channel audio and video recordings [5]. It consists of audio and video recording of 100 speakers (50 male and 50 female). Most of the speakers are American English speakers, with the remainder of speakers from Latin America, Europe and East or South Asia. The database is recorded across five distinct recording conditions in English, as shown in Table 1. The audio-visual speech was captured using an array of 8 microphones on the passenger's sun-visor and a 4-camera array positioned on the dash, with each camera aimed towards the passenger to capture the different views of the face. The video streams are combined using a multiplexer in order to be stored in a single file for each utterance with 29.97 frames per second and each camera having an individual resolution of 360×240 pixels.

Noise	Description
35U	Car travelling at 35mph and windows closed
35D	Car travelling at 35mph and windows open
55U	Car travelling at 55mph and windows closed
55D	Car travelling at 55mph and windows open
IDL	Car stopped and engine idling

Table 1: Noise Conditions in the AVICAR database

We selected 31 subjects from the phone numbers portion of the AVICAR database according to the protocol developed by Navrathna et al [10]. This portion consisted of two sessions of a 10 digit utterances for each speaker and noise condition. The phone number digit sequences were identical across all subjects with all digits used for each 10 digit phone number. Subjects were instructed to pronounce the digit 0 as 'zero' in session 1 and 'oh' in session 2. We used data which was captured from left-top camera. Visual examples are shown in Figure 3 (b).

4.2. Visual speech recognition

Following the tracking of the mouth ROI, the visual features were extracted using the dynamic visual speech feature extraction process described earlier. Image mean normalization was performed to remove any irrelevant information, such as illumination or speaker variances. Then a 2D-DCT was applied to the mean-removed image and the top $\mathbf{M} = 100$ higher-energy components were selected in zig-zag pattern to capture the static visual speech information.

In order to incorporate dynamic speech information, seven of these neighbouring static feature vectors over ± 3 adjacent frames were concatenated, and were projected via an inter-frame LDA step to yield a $\mathbf{Q} = 50$ dimensional 'dynamic' visual feature vector. The classes used for LDA matrix calculation were HMM states, based on forced alignment of a separately-trained audio-only HMM. For each word in the database, A separate word-based visual HMM was trained with different number of states and mixtures and the best HMM was selected for test. All speech recognition results quoted in this paper are HTK-style [20] word accuracies (%).

5. Experiments and results

5.1. Tracking performance

In order to measure the performance of the VJ and FLK tracking approaches, portions of the CUAVE and AVICAR databases were manually annotated with ground-truth face location. Errors in tracking were calculated as the root mean square (RMS) of the difference between the location of the tracking and ground truth. Initially, images of the first subject in CUAVE dataset are

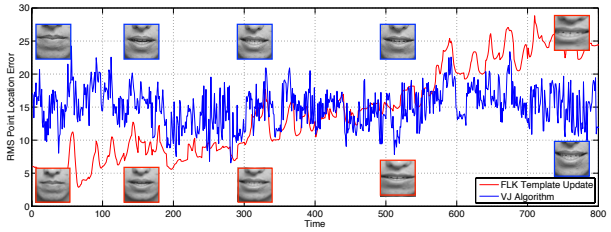

Figure 4: FLK mouth detection degradation based on RMS of point location errors over time

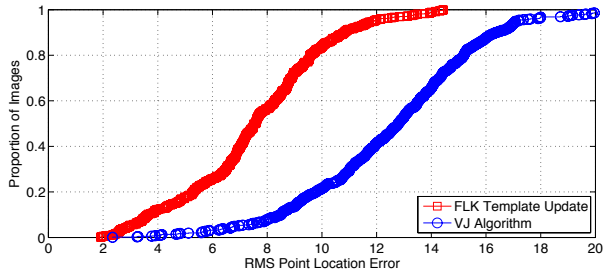

Figure 5: Proportion of images that have RMS of point location errors of mouth region less than specified values

passed to both VJ and FLK mouth detection procedure to see how well they fit compared to ground-truth annotations, with the result shown in Figure 4.

As Figure 4 shows, the FLK approach starts tracking the mouth region with higher accuracy than VJ. However, as it continues, it reaches the mean RMS error of VJ at around the 400th frame and continues to degrade. This suggests that the template image needs to be updated regularly to avoid failing. Therefore, FLK approach needs manual inspection of the extracted images so that if there was a corruption in extracted mouth region, the template image becomes manually updated to avoid subsequent failures. This semi-automatic FLK approach is used in the following experiments.

In order to evaluate the semi-automatic FLK, 600 frames of CUAVE database are used, with the first 100 images of videos of the first 6 subjects annotated manually to provide the coordinates of 4 corners of their mouth region. Figure 5 shows the proportion of the images that reside below the specified RMS errors with the two approaches.

Figure 5 shows that FLK approach has clearly provides better performance in fitting the ground truth. Particularly, we can see that all of the extracted mouth images using FLK approach are fitted with the actual mouth regions at the cost of 14.3 RMS of point location errors respectively. Whereas in case of VJ approach, this is achieved at the value of 20 for RMS of point location errors. In addition, half of the extracted mouth images using FLK approach could be fitted with ground truth at RMS value of less than 8. However, only less than 10% of the VJ extracted face images could be fitted at this RMS value. All in all, this figure shows that ROI extraction of FLK approach is more precise than VJ in fitting application.

	VSR accuracy(%)	
	VJ	FLK
AVICAR	46.73	49.22
CUAVE	59.30	59.07

Table 2: VSR accuracy of FLK based vs VJ based front-end on AVICAR and CUAVE databases

	VSR accuracy(%)				
	35D	35U	55D	55U	IDL
VJ	47.98	49.56	42.37	49.25	49.61
FLK	50.57	50.81	45.60	50.00	53.65

Table 3: Word recognition accuracy of FLK-based vs VJ based front-end on AVICAR database organized into different noise conditions

5.2. VSR performance

In this experiment, we compare FLK-based mouth detection with VJ approach in terms of VSR accuracy. In order to do that, different visual HMMs are trained with different number of states and mixtures for both approaches. For CUAVE database, videos of 25 subjects were used for training and videos of the remaining 6 subjects are used for testing. For AVICAR, 70% of the video data were used for training and the remaining 30% were used for testing.

Table 2 presents the best VSR accuracy achieved by the two approaches on the AVICAR and CUAVE databases. As can be seen, the FLK approach outperformed VJ on the AVICAR database with 5.3% relative improvement in VSR accuracy. However, these two approaches had very similar performance on the CUAVE database.

Table 3 shows the accuracy of the best visual HMM using the two approaches achieved on AVICAR test data organized into the different noise conditions. In all of the 5 noise conditions, FLK-based VSR has a better performance than VJ approach. FLK approach is designed for tracking purposes, as it tries to find the ROI based on the previous frame, while VJ detects the ROI in each frame independently. This gives FLK approach the advantage of being more accurate in situations where image pose changes during the recording. This could be the reason why Table2 shows that in controlled recording conditions such as the one for CUVAE frontal videos, the accuracy of VJ and FLK approaches does not differ too much. However, in situations where there are pose changes and illumination changes, like for AVICAR database, FLK approach has a better performance.

6. Conclusion

According to the experiments, we can conclude that the FLK approach for tracking mouth region has the potential to improve upon the VJ approach for VSR applications. However, one of the disadvantages of FLK approach is that the first frame of each video needs to be annotated manually to create the template image for tracking. In addition, after a number of frames, the template may need to be updated, as the detected ROI may be erroneously extracted and may cause errors in finding the ROI in subsequent frames. However, as it is designed for tracking purposes, it has better tracking performance in situations where there are pose changes and illumination changes. Future work will focus on looking at allowing the FLK tracking to be automatically initialised and improving the ability of FLK to recover from bad tracking without causing subsequent frames to continue to worsen.

7. Acknowledgements

The authors would like to thank Australian Cooperative Research Center for Smart Services for supporting this research.

8. References

[1] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[2] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *In Proceedings European Conference on Computer Vision*, 1998.

[3] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers (London)*, 93(III):429–457, 1946.

[4] F. Lavagetto. Converting speech into lip movements: a multimedia telephone for hard of hearing people. *IEEE Transactions on Rehabilitation Engineering*, pages 90–102, 1995.

[5] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang. AVICAR: An audio visual speech corpus in a car environment. *In Proc. Interspeech 2004*, pages 2489–2492, Jeju Island, Korea.

[6] S. Li, J. Sherrah, and H. Liddell. Multi-view face detection using support vector machines and eigenspace modelling. *International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, 2000.

[7] M. Lievin and F. Luthon. Unsupervised lip segmentation under natural conditions. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1999.

[8] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *In Proceeding of the International Joint Conference on Artifical Intelligence*, 1981.

[9] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan. Fourier lucas-kanade algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 2013.

[10] R. Navarathna, D. Dean, P. Lucey, C. Fookes, and S. Sridharan. Recognizing audio-visual speech in vehicles using the AVICAR database. *In Australasian International Conference on Speech Science and Technology (SST)*, pages 110–113, 2010.

[11] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. 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)*, pages 217–2020, 2002.

[12] G. Potamianos and C. Neti. Audio-visual speech recognition in challenging environments. *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1293–1296, 2003.

[13] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu. A cascade image transform for speaker independent automatic speechreading. *International Conference on Multimedia and Expo (ICME)*, 2:1097–1100, 2000.

[14] R. Rao and R. Mersereau. Lip modelling for visual speech recognition. *In Proceedings of the Asilomar Conference on Signals, Systems and Computers*, pages 587–590, 1994.

[15] J. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting with a mixture of local experts. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2248 – 2255, 2009.

[16] Y. Tian, T. Kanade, and J. Cohn. Robust lip tracking by combining shape color and motion. *Proceedings of the Asian Conference on Computer Vision*, page 10401045, 2000.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *In CVPR*, pages 511–518, 2001.

[18] M. Yang, N. Abuja, and D. Kriegman. Mixtures of linear subspaces for face detection. *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 70–76, 2000.

[19] M. Yang, D. Kriegman, and Ahuja.N. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 34–58, 2002.

[20] S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, V. V. Ollason, D. D. Povey, and P. Woodland. *The HTK Book (for HTK Version 3.2.1)*, 2002.

[21] A. Yuille, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, pages 99–111, 1992.