Nostril Detection for Robust Mouth Tracking

Luca Cappelletta[†] and **Naomi Harte**

Department of Electronic and Electrical Engineering
Trinity College Dublin,
IRELAND

E-mail: [†]cappelll@tcd.ie

Abstract — **Within an *Audio-Visual Speech Recognition* (AVSR) framework an important process is video feature extraction. Several methods are available, but all of them require mouth region extraction. To achieve this, a semi-automatic system based on nostril detection is presented. The system is designed to work on ordinary frontal videos and to be able to recover brief nostril occlusion. Using the nostril position a motion compensated *Accumulated Difference Image* (ADI) is generated. This ADI is less noisy than the non-compensated one, and this leads to better mouth region tracking. Results show that the ADI stage has good reliability, whereas the nostril detection stage may be further improved.**

Keywords — **Audio-Visual Speech Recognition, KLT, Nostrils Detection, ADI.**

I Introduction

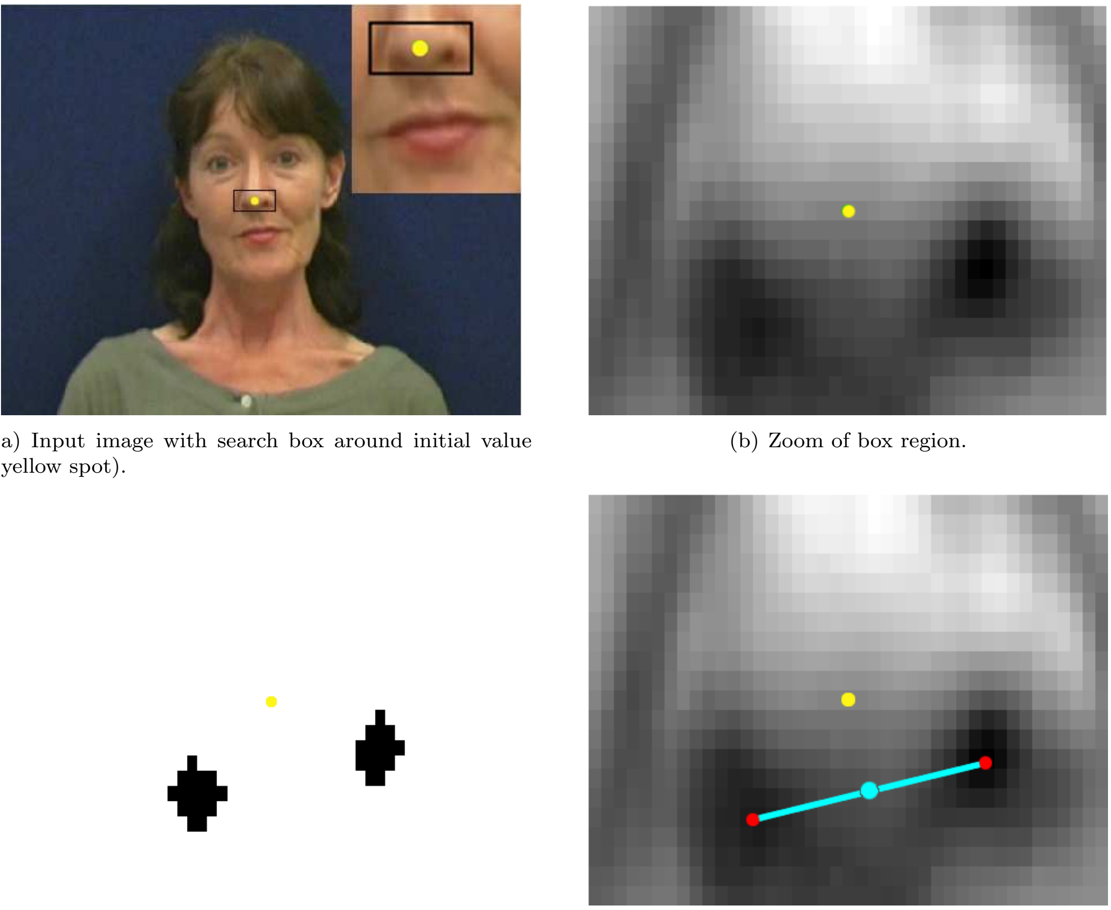
Recently a boost to *Automatic Speech Recognition* (ASR) technology has been provided by using both audio and visual cues present in speech, creating the new paradigm of *Audio-Visual Speech Recognition* (AVSR). It has been shown that AVSR improves the robustness of ASR systems [1].

Focusing on video only, a crucial process is the definition of features to describe lip motion. A variety of feature extraction algorithms are used, like the4 *Kanade Lucas Tomasi* (KLT) tracker [2], optical flow [3], image projection on Zernike functions [4][5] and *Active Appearance Models* (AAM) [6]. No matter which system is used, it is important to know where the mouth is located in the image. Hence it is useful explore this issue, in particular in relation to the KLT system. The goal of the KLT is to align a template image $T(x)$ to an input image $I(x)$. By this, the algorithm is able to track the position of some points, or features, determined in the first frame using the process proposed by Shi and Tomasi [7]. They assert that so-called *good features* can be located by examining the minimum eigenvalue of each 2 by 2 gradient matrix. This approach is a kind of corner detector and it works in textured areas, but it is not so good in flat regions, like lips. The KLT *good features* process is composed of a cascade of several filters and, because of edge effects from each filter stage, this reduces the image portion that can be used. The input image must then be wider than the mouth region. However, having too wide a region can cause all the feature points to be placed close to the nose, a more textured region. Of course feature points around the nose are not useful for lip tracking, so the correct balance is to have a wide enough input image but to know where the mouth is located too.

Therefore the purpose of this paper is to propose an algorithm for lip localization in video footage. This algorithm is based on knowledge of the nostrils position and it is composed of two stages. The first one is the nostrils detection and refinement process, and this is described in Section II. The second stage (Section III), is the use of nostril position to refine a motion compensated ADI image. The new algorithm is applied to data from the VidTIMIT database and results are discussed in section IV).

II Nostril Detection & Refinement

The nostrils position will be used as starting point for the mouth region detection. In the literature,



(a) Input image with search box around initial value (yellow spot).

(b) Zoom of box region.

(c) Box region after thresholding. Two clusters are present.

(d) The calculated nostril positions (red spots). The mean value of those (cyan spot) will be used as the initial point for the next frame.

Fig. 1: Nostrils Detection process. Starting from an input image and an initial position, figure (a), the nostril positions are calculated (d). Using those position are calculated two new parameter: the nostrils mean position (cyan spot) will be used as initial position in the next frame nostril detection, and the nostrils distance (cyan line) will be used in the nostrils tracking process of the current frame.

several papers on nostril detection are present, e.g. [8], but use an approach proposed by Petajan [9]. They assume the camera is not in at the level of the speaker's face (frontal position), but that it is at a lower level and angled (e.g. at the chest height). In this case nostrils are significantly more visible compared to a ordinary frontal framing. The goal of this paper is to perform nostril detection using an unconstrained camera position. Because of this, managing any nostril occlusion is crucial.

The nostril processing comprises two stages:

Detection: using an initial value of the nostrils position, detection determines the new position in the current frame. If the nostrils are not found, an error value is returned.

Refinement: takes the *detection* stage output, assigning the final value of the nostrils position. This extrapolates the current nostril position in cases where an error or unreasonable value was returned.

Both steps are performed for each frame. An initialization is required only for the first frame. These two stages are described fully in the next section.

a) Nostril Detection

Figure 1 shows how the nostrils detection process works. As input it requires an image and an initial point (see figure 1(a)). The initial point has to be reasonably close to the real nostrils position. Currently a manual initialization is used. The search is performed in a box around the initial point (see figure 1(b)). The box size has been heuristically defined according to the image dataset employed. The nostrils are the darkest spots in the middle of the face area [9], so it is possible to locate them using a 5% threshold on the search area histogram [8]. The result of this thresholding is shown in figure 1(c). This result includes a sequence of morphological filters implementing erosion and dilation. This is done to remove isolated pixels above the 5% threshold. Once the two dark regions are detected, it is possible to calculate their mean value yielding the current nostril position (see figure 1(d)).

This entire process is performed for each frame, but the manual initialization is required only for the first frame. For subsequent frames, the initial point is defined as the mean value of the nostrils position calculated in the previous frame, e.g. the cyan spot in figure 1(d) will be used as initial point in the next frame.

b) Nostril Refinement

The purpose of the refining stage is to handle errors or unreasonable values coming from the detection stage. The nostril detection process returns an error value if less than two clusters are found. That occurs when, for example, one nostril is much darker than the other because of the lighting conditions, or when a cluster is too small and is deleted by the erosion/dilation filter. In these cases at least one nostril is missing. Here the refining stage will obtain a new value from a linear extrapolation based on the previous two nostril positions, according to the equation:

$$N_n = N_{n-1} + \alpha \left(N_{n-1} - N_{n-2}\right) \tag{1}$$

where N_n is the position of a nostril (left or right) at the frame n. $\alpha = 1$ is the natural candidate for the parameter value in the above equation. However, from an experimental point of view it is more reasonable to use a smaller value, eg $\alpha = \frac{1}{2}$. This is because $\alpha = 1$ can lead the nostril very far from its real location in the case of several error values in a row.

The other case when the refining process acts is when unreasonable nostril positions are returned. This happens when they don't satisfy two criteria. The fundamental parameter for these criteria is the distance between the nostrils (cyan line, figure 1(d), defined as:

$$D_n = dist_{eucl}(N_n^{left}, N_n^{right}) \tag{2}$$

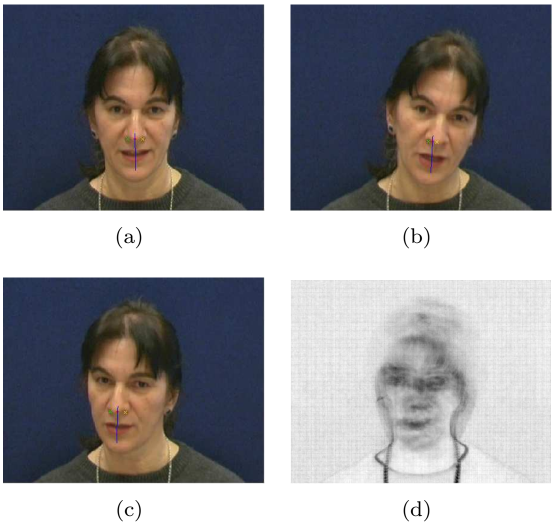
Indexing the current frame as n, the first criterion constrains the current nostrils distance (D_n) does not vary too much from the previous one (D_{n-1}):

$$0.8 \cdot D_{n-1} \leq D_n \leq 1.2 \cdot D_{n-1} \tag{3}$$

The second criterion limits the displacement of a single nostril between two consecutive frames.

$$dist_{eucl}(N_n^{left}, N_{n-1}^{left}) \leq \frac{D_{n-1}}{2} \tag{4}$$

The same constraint is applied to N_n^{right}. If the nostril position fails one of the two criteria, a new position will be computed according to equation 1.


(a) (b)

(c) (d)

Fig. 2: Example of moving head speaker. Figure (d) shows the ADI **without** motion compensation, performed using equation 5.

III Accumulated Difference Image

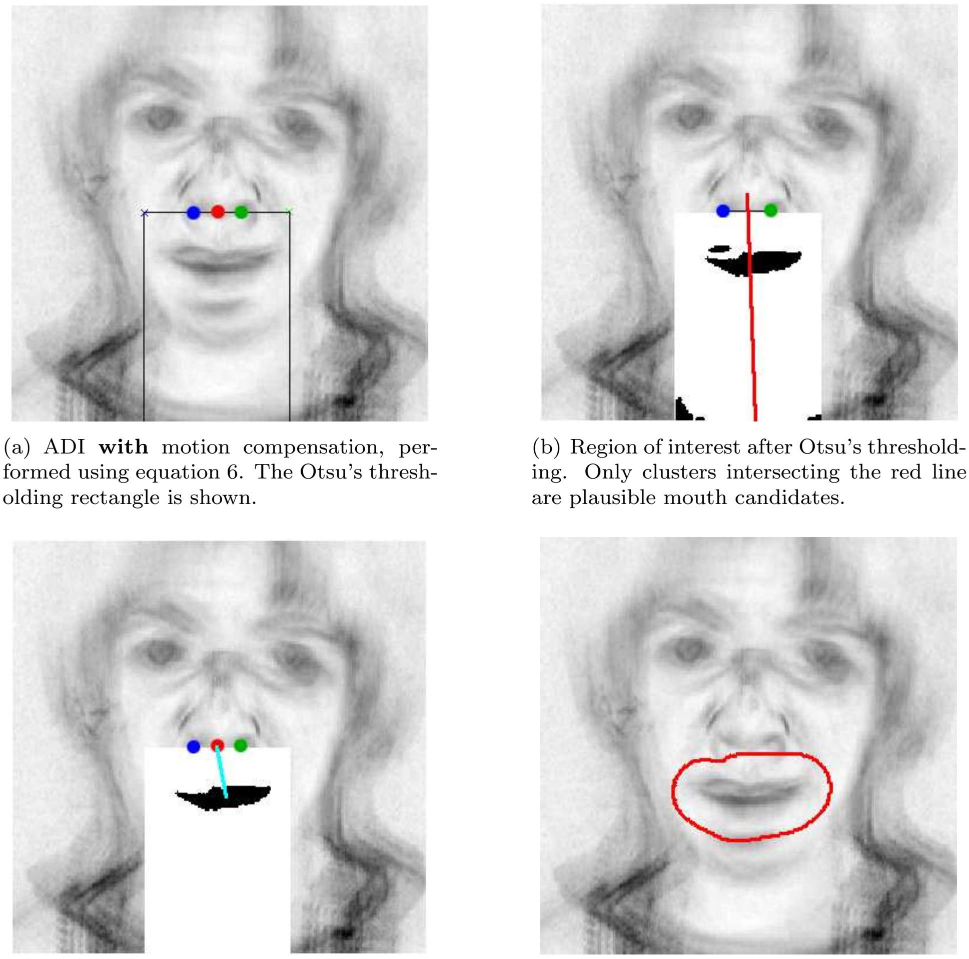
To detect the mouth area the *Accumulated Difference Image* (ADI) [10][11] is used. The ADI is built by summing the absolute value of consecutive red component (R) images difference over a series of frames, as in the following equation:

$$ADI(i,j) = \sum_{k=2}^{K} |R_k(i,j) - R_{k-1}(i,j)| \tag{5}$$

with $(i,j) \in [1,N]\times[1,M]$, where N and M are the image dimensions, and K is the number of frames. The problem in this approach is that even if the speaker is silent he may move his head, causing an increasing of ADI value in the non-lips area. This problem is demonstrated in figure 2. To avoid this problem it is preferable to perform the ADI not on the whole image, but only in a window moving with the speaker's head. To do this, the knowledge about the nostrils position can be useful, because for each frame the windows will be centered on the nostrils mid point. This can be defined as a *motion compensated* ADI.

$$ADI(i,j) = \sum_{k=2}^{K} |R_k(i,j|\Omega_k) - R_{k-1}(i,j|\Omega_{k-1})| \tag{6}$$

where Ω_k is the moving window at the k-th frame. The size of Ω is the same for every frame, only its center changes in each frame. Window size has been heuristically chosen in order to include the whole face. The difference in results of the two ADI approaches are shown in figure and figure 2(d) 3(a). Besides this improvement in the ADI definition, several expedients have been used.



(a) ADI **with** motion compensation, performed using equation 6. The Otsu's thresholding rectangle is shown.

(b) Region of interest after Otsu's thresholding. Only clusters intersecting the red line are plausible mouth candidates.

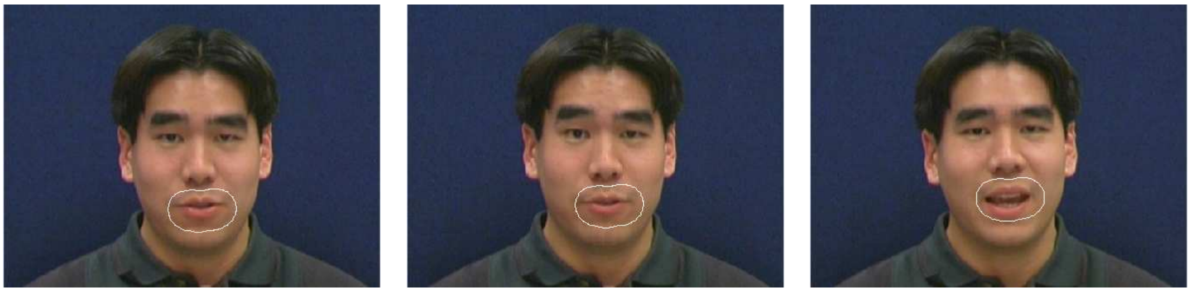
(c) The mouth region is determined. The cyan line represents the parameter to use for the region dilation.

(d) The edge of the mouth region after the dilation is shown over the original ADI image.

Fig. 3: Several stages of the ADI process. Blue and green spots are the mean position for left and right nostrils, the red spot is their mean value.

In order to reduce noise level, an Otsu's thresholding [12] is performed to $|R_k - R_{k-1}| \quad \forall k$ in equation 6. Now the ADI has to be segmented in order to obtain the high motion regions in the image. To obtain this an Otsu's thresholding is applied twice, but not on the whole ADI, just on a limited region. This region is the rectangle having as upper limit the nostrils position, the lower limit is the image edge, and the width is three times the distance between the mean position of the left and right nostril (see figure 3(a)). From this logical image, possible holes inside the clusters are filled using a filling process [13] and small objects are removed using an erosion filter (see figure 3(b)). Unfortunately several clusters are still present after the thresholding. To select the mouth cluster an assumption is made: the mouth region has to lie right below the nostrils. Hence the mouth cluster must intersect the red line shown in figure 3(b). This line starts from the nostrils mid point (red spot) and it is perpendicular to the black line going from the left nostril (blue spot) to the right nostril (green spot). Finally, within the set of selected clusters (in figure 3(b)) the one closest to the red spot is selected as mouth region. Unfortunately the upper lip moves less than the lower one, so it is harder to distinguish it in the ADI.Thus it may not be included in the selected mouth region. Because of this, the selected mouth region (see figure 3(c)) has to be dilated using an asymmetrical mask. To create the mask, the distance between the nostrils mid position (figure 3(c), red spot) and the selected region centroid (figure 3(c), yellow spot) is used as parameter. In this way the mask dilation does not have to be customized for each speaker.

The output of the ADI process is a mask having the same size as Ω (see equation 6) and having `true` value only in the calculated mouth region. Now this mask is applied to each frame, changing its position according to the nostrils position in each frame (see figure 4).



(a) Frame 41.

(b) Frame 51.

(c) Frame 86.

Fig. 4: ADI region edge superimposed on some video frames.

IV Experimental Results

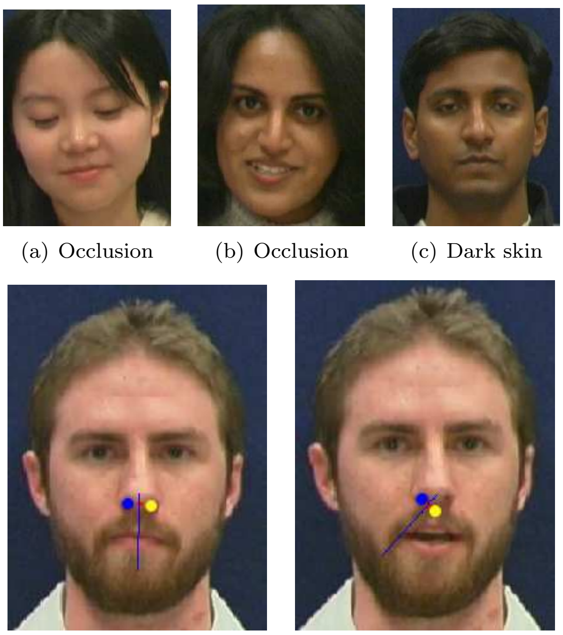
a) VidTIMIT Dataset

The VidTIMIT dataset [14] is comprised of the video and corresponding audio recordings of 43 people, reciting short sentences. Most of the speakers (33 on 43) are caucasian, 6 are asiatic and 4 black. The speakers are well balanced in gender (24 male and 19 female) and 5 of the male speakers have facial hair. The sentences were chosen from the test section of the TIMIT [15] corpus. The selection of sentences in VidTIMIT is designed to cover all the 11 visemes [16] present in the English language. The recording was done in an office environment using a broadcast quality digital video camera. The video of each person is stored as a numbered sequence of JPEG images with a resolution of 512 x 384 pixels. 90% quality setting was used during the creation of the JPEG images.

b) Nostril Detection Stage

In more than 60 % of the footage analyzed the nostril detection stage achieves good results. A *good result* means that the algorithm is able to follow the nostril position during all the video and recover it in case of brief loss of track.

The algorithm has difficulties with some categories of speakers/video. For example, male speakers with facial hair, in particular moustaches, are very difficult to analyze because in this case one of the assumptions fails: the nostril is not the darkest spot in the middle of the face. In this case it is very likely the nostril position slides down on the moustache (see figure 5(d) and 5(e)). Another problem is skin colour: speakers with dark skin don't have enough contrast between the nostrils and nose/lips, so the process may fail (see figure 5(c)). A problem that is not possible to overcome is long term nostril occlusion. Because of the nose shape or the head position it is possible the nostrils are not visible (see figure 5(b) and 5(a)). If this condition persist for the whole video or for a big part of it, the algorithm is not able to find to nostril position.



(a) Occlusion

(b) Occlusion

(c) Dark skin

(d) Moustache, frame 8

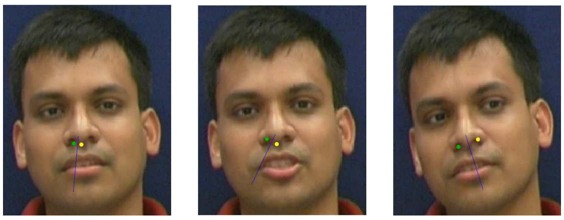
(e) Moustache, frame 97

Fig. 5: Example of nostril detection failure.

However the process has the ability to recover the nostril position in case of temporary failure or non-correct initialization. Figure 6 demonstrates how the process can recover the position of both nostrils.

c) Combined ADI

For all the videos, if the nostril detection stage tracks the nostrils during the video, the ADI stage will detect and track the mouth region as well. The tracked mouth is also significantly less noisy than only using the ADI. Sample results are shown in figure 4. Clearly it is difficult to fully appreciate the continuous tasking of the mouth for a full sentence from a small number of frames. Sample videos are available to view at `www.mee.tcd.ie/~sigmedia/Research/LucaPage`. The main weakness of the algorithm is its inability to recover from poor nostril tracking. Poor ADI generation was consistently caused by poor nostril



(a) Frame 10. (b) Frame 37. (c) Frame 59.

Fig. 6: Nostril position recovery. The process recovers the position of both nostrils

detection. The definition of the ADI region depends directly on the nostril position and whilst significant leeway is possible the bisector must intersect the mouth region. Another issue is the requirement for manual initialization in the first frame. This requires the definition of a point relatively close to the nostrils. The manual initialization could potentially be replaced by a face detection stage and face template matching. For this work the question of interest was how to track the mouth and this issue was not further investigated.

V Conclusions

The nostril tracking is very accurate in the majority of videos and can recover from short losses in nostrils position. This has removed the need for a constrained camera position. The combination of a motion compensated ADI and nostril tracking gives a smooth track of the speakers mouth. The sensitivity to accurate nostril tracking is a definite weakness. However, it may be possible to introduce this idea as part of a larger system where this may not cause overall system failure. Other improvements include extending the motion compensation model by including rotation; only translation is currently considered. Currently a single 5% threshold is used for skin thresholding. This could automatically be adjusted according to the skin colour.

References

[1] G. Potamianos *et al.*, "Recent advances in the autormatic recognition of audio-visual speech," *Proceeding of the IEEE*, vol. 91, no. 9, 2003.

[2] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of Imaging Understanding Workshop*, 1981.

[3] S. Tamura *et al.*, "Multi-modal speech recognition using optical-flow analysis for lip images," *The Journal of VLSI Signal Processing*, vol. 36, no. 2, pp. 117–124, 2004.

[4] Y. Wai Chee, "Visual speech recognition method using translation, scale and rotation invariant features," K. Dinesh Kant and A. Sridhar Poosapadi, Eds., vol. 0, 2006, pp. 63–63.

[5] W. C. Yau *et al.*, "Visual speech recognition using motion features and hidden markov models," in *CAIP 2007*, S.-V. B. Heidelberg, Ed., 2007.

[6] T. F. Cootes *et al.*, "Active appearance model," *Proc European Conference on Computer Vision*, vol. 2, 1998.

[7] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 1994, pp. 593–600.

[8] F. Bourel *et al.*, "Robust facial feature tracking," in *British Machine Vision Conference*, Bristol, 2000, pp. 232 – 241.

[9] E. Petajan and H. P. Graf, "Robust face feature analysis for automatic speechreading and character animation." IEEE Computer Society, 1996, p. 357.

[10] X. Zhang *et al.*, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1228–1247, 2002.

[11] C. Berry and N. Harte, "Region of interest extraction using colour based methods on the cuave database," in *Irish Signals and Systems Conference, 2009. IET*, 2009.

[12] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 1, pp. 62–66, 1979.

[13] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

[14] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM-Verlag, 2008.

[15] J. S. Garofolo *et al.*, "Darpa timit acoustic phonetic continuous speech corpus cdrom," NIST, Tech. Rep., 1993.

[16] J. Janet and B. Margaret, *Speechreading (Lipreading)*. Charles C Thomas Pub Ltd, 1971.