# Lip Tracking Towards an Automatic Lip Reading Approach

Luca Lombardi, Waqqas ur Rehman Butt, Marco Grecuccio
Department of Electrical Computer and Biomedical Engineering
University of Pavia
Pavia, Italy
luca.lombardi@unipv.it, {waqqasurrehman.butt01, grecuccio01}@ateneopv.it

***ABSTRACT****:Current era is to make the interaction between humans and their artificial partners (Computers) and make communication easier and more reliable. One of the actual tasks is the use of vocal interaction. Speech recognition may be improved by visual information of human face. In literature, the lip shape and its movement are referred to as lip reading. Lip reading computing plays a vital role in automatic speech recognition and is an important step towards accurate and robust speech recognition. In this paper we proposed method to detect the lips, shape and movement and use towards automatic lip tracking in detail by using Active Appearance Model (AAM). Purposed method shows the how accurate lip detection and tracking which is useful for speech recognition. Appearance-based methods consider the raw image for lip extraction. AAM uses for detection of speaker's lips features and feature points from faces automatically by appearance parameters of a speaker's lips. In this paper, we will describe the Lips shape, movement and appearance descriptions for lip detection and detailed information about the automatic lip tracking system. Individual characteristics and efficiency of both approaches towards the Lip reading is also discuss. The efficiency of purposed method will be evaluated individually by characteristics and their performances are then compared. The recognition process of common visual features (i.e. selection of lip, shape and extraction of movement) for an improved lip reading.*

**Keywords:** Speech Recognition, Active Appearance Model, Lip Reading, Lip Tracking

**Received:** 1 November 2013, Revised 16 December 2013, Accepted 21 December 2013

## 1. Introduction

In recent years speech recognition has attracted significant interest. Most studies have demonstrated that accuracy rate of automatic speech recognition (ASR) systems has been improved by visual information aided, especially in noise environment or multiple talkers. When a person intends to understand more about what he hears, he would subconsciously use visual information that came from the speaker's facial expression, the influence of lip movement and gesture is describes by the McGurk Effect. Sometime listener perceives something else than what is said by speaker acoustically due to the influence of conflicting visual cues so that Visual information is a key for lip reading, The McGurk effect shows that visual articulatory information is integrated into our perception of speech automatically and unconsciously. These observations provide a motivation for attempting to integrate vision (lip movement) with speech in a computer speech recognition system [4]. In reverse, the result would be influenced or disturbed for human speech perception. Visual information are extracted from visual features especially from the lip movement and recognized to get more accuracy and robust result in speech recognition. Visual Features such as eyes, mouth region and lip shapes extract from the image sequence and send to classifier where extracted features are compared with stored features in dataset to produce the final recognition result as shown in Figure 1.

The face is the most visible part in the human image and movements of lips are the visible aspect of speech production. Relationship between the speaker's lips and the resulting speech described in [2, 3] was used for speech recognition and speech enhancement. Performance of speech recognition from visual information in noisy environment would be a problem for satisfactory result. The main objective of the lip reading system is extracting motion information of the mouth effectively and correctly.

In this paper we split the description of our approach towards automatic lip reading in three parts. In the first part we will describe the selection and extraction of both face and lip movement from visual features of speaker's face by using appearance based approach with the reference of previous research and then we proposed our approach for lip tracking algorithm and perform experiments for lip shape, image enhancement and filtering to obtain precise result for tracking. In the third section we will discuss our final remarks.

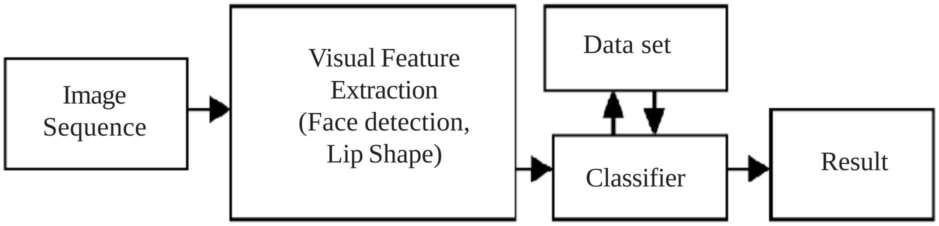


Figure 1. Visual Speech Recognition

## 2. Literature Review

Most of the work done on Visual Speech Reading (VSR) came through the development of Audio Video Speed Reading(AVSR) systems, as the visual signal completes the audio signal, and therefore enhances the performance of these systems [5]. Previous research proved that erroneous extractions because of the improper parameters, e.g., ulterior evolving curve with small local radius or proper evolving curve with large local radius. The prior knowledge about color information does not consider [8], which actually provides more information to improve the extraction performance, especially when the images are shadowed, shaded and highlighted [6, 7]. In this paper, we adopt a 16-point geometric deformable model proposed in [9] to model the lip shapes, which is more flexible and physically meaningful in comparison with the less points based lip models.

Lip reading is categorized in two steps, feature extraction and Visual speech feature recognition. In this paper we will consider the appearance based approach for feature extraction. Some considerations related to existing feature extraction approaches are described below.

- **Geometric Features-based:** In this approach mouth geometric information from the region (mouth shape, height, width, and area) are considered
- **Appearance-based:** The pixel values of the mouth region are considered and these methods are applied to both grey and colored images. Principal Component Analysis (PCA) is used for computational reduction of region of Interest (ROI)
- **Image-transformed-based:** They use features extraction by transforming the mouth image to a space of features, using some transform technique (discrete Fourier, discrete wavelet, and discrete cosine transforms DCT)
- **Hybrid approaches:** They exploit features from more than one approach

Lip movement is one of the best visual clues for recognizing when a person is speaking or is silent, since the lips move more than 80% of the time in human speech [14]. The existing relationship between lip movements and Voice activity detection is described in [15]. It showed that the lips shapes can be similar on voice activity and on silence.

## 3. Appearance Based Visual Feature Extraction

### 3.1 Face Detection

Face detection is the first step to extract visual features from an image and the main objective of face detection is to determine whether a face is present in the image. It depends upon many factors such as pose, presence or absence of structural components, expression, occlusion, image orientation and image condition. It also uses in many face-related applications such as face recognition and lips reading, so that reliability of lip reading is based on accurate detection of face. In 2002 Yang et. al., classified the face detection into four categories [9].

1. **Knowledge-based methods:** They require human knowledge about facial features.

2. **Feature invariant approaches:** Their goal is to find structural features that are not affected by the general problems as with the face detection process, e.g. pose and light conditions.

3. **Template matching methods:** They use patterns to describe a typical face and compare these patterns with image to find the best correlation between the pattern and targeted image's window. These templates can be predefined templates or deformable templates.

4. **Appearance-based methods:** They declare the template, that is learned from a set of images then the learned template is used for detection. A variety of methods fills in this gap.

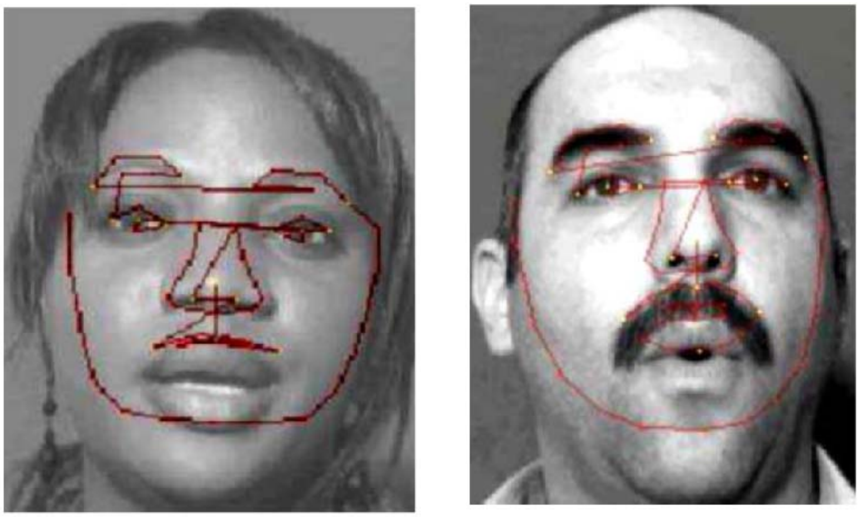
### 3.2 Lip Detection

Visual speech information is extracted from the most visual parts of the human face such as lips and mouth regions, therefore it is necessary to capture all the information from these visual parts for speech recognition. Many applications are concerned for processing/analysis in the lips and mouth area within human face and based on accurate and reliable detection of lips. The most deformable part of the face is lips, detecting them is a nontrivial problem, adding to the long list of factors that adversely affect the performance of image processing/analysis schemes, such as variations in lighting conditions, pose, head rotation, facial expressions and scaling [5]. There are two main techniques used for lip detection, model based and image based.

1) **Model-based:** "*Snakes*", Active Shape Models (ASM), Active Appearance Models (AAMs), and deformable templates.

2) **Image-based:** These include the use of spatial information, pixel color and intensity, lines, corners, edges, and motion.

Active Appearance Model (AAM) searches for the lips in any freshly input image. The best fit to the model, with respect to some prescribed criteria, is declared to be the location of the detected lips. In this paper we will reflect on AAM for lip detection, others methods such as Active Shape Model (ASM) and Snakes are not considered. Snake's require external forces to detect the lips and ASM is difficult to apply in real time especially having moustache and beard [5]. Fig. 1 shows these problems.



(a) Lip detection converges to local minima

(b) The effect of facial hair [5]

Figure 2. ASM Facial features detection

### 3.3 Active Appearance Model

Active Appearance Models (AAMs) are employed to extract the location of specific points on the face from image.

The AAM was introduced in [1]. This is an extension of the Active Shape Model and the combination of the shape information (i.e. ASM) and texture information (i.e. appearance based approach) is usedas searching scheme. Statistical models of variations of shape texture are created by AAM. The shape is determined considering a training set of samples shapes and these aligned by Generalized Procrustes Analysis. Each face sample is then warped so that the control points match the ones of the mean shape as shown in Figure 3 [11].

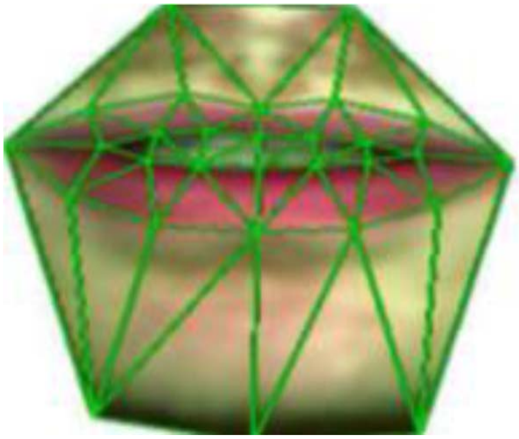


Figure 3. The mean shape and the mean texture used for AAM [11]

The grey level texture model is built by warping each training image so that the landmark points lie on the mean shape positions. AAM extracted the face information from face image which is useful for the specific parameters, that describes the appearance of facial features. Figure 4 showed 16 points after selection and Parameterization of a lip within the image. This will be done in following steps:

**Key points Selection:** Select the points for capturing the lip movement.

**Parameterization:** Convert the selected points into representative forms.

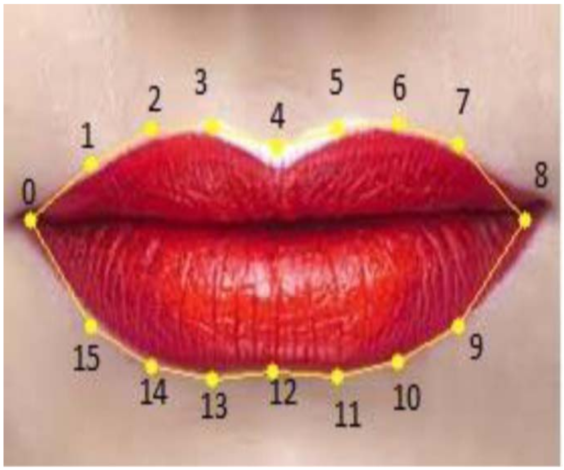


Figure 4. Selection of 16 landmark points

Parameterized key points gives the most important of aspects for lip movement recognition as described in table 1.

| Selection point no. | Characteristics |
|---|---|
| 0 | Leftmost point of the lip |
| 1-7, 9-15 | Used to calculate the lip area |
| 1, 2, 3, 4, 5, 6, 7 | Upper edges of upper Lip |
| 8 | Rightmost point 9,10,11,12,13,14,15 Lower edges of lower lip |

Table 1. Key point description

We can obtain the exact shape of lips, position and movement by using these key points. Figure 4 shows no lip movement and mouth status closed so that area, height and width can be calculated.

Mouth height will be obtained by calculate the distance between pair of points (2, 14), (3, 13), (4, 12), (5, 11), and (6, 10). Lip width can be obtained by the calculating the distance between the points 0 and 8. The mouth area is assumed to be made of 16 triangles formed by these 16 points and every triangle made from a center point C. The area of the mouth can be found by summing up the areas of 16 triangles. The sample triangle assumed to be enclosed in a rectangle is shown in figure 5a.

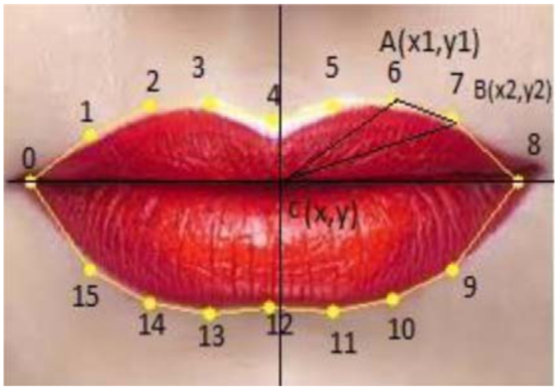


Figure 5. a) Mouth Status

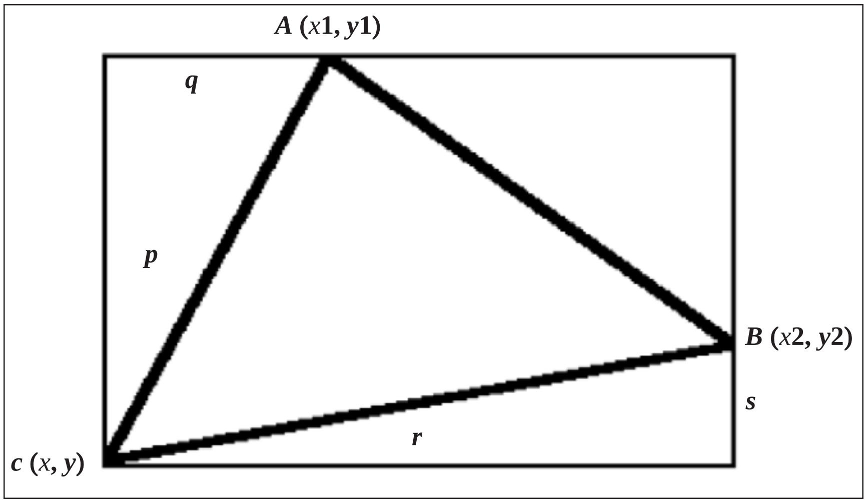


Figure 5. b) Sample triangle

Sample triangle area can be found by the equation 1.

$$area\ i = \frac{p * r}{2} - \frac{q * s}{2} \tag{1}$$

Where, $p$ is the vertical distance between point $C$ and point $A$ (i.e. $p = y - y1$), $q$ is the horizontal distance between point $A$ and point $C$ (i.e. $q = x1 - x$), similarly r is the horizontal and s is the vertical distance between point $B$ and point $C$. The total area of lips is the sum of area of 12 such rectangles. Thus, the lip area can be calculated using the following equation 2 [12].

$$lips\ area = \sum_{i=1}^{12} area\ i \tag{2}$$

### 4. Proposed Method for Lip Tracking and Experiment Result

This algorithm works with RGB images in .png format. We did not use compress images etc. JPEG because compression introduces many noisy artifacts. First of all load the RGB image (original file) as shown in the figure 6a. After load original image a filter media to obtain the correct illumination effect. Illumination equalization has played an important role in image analysis and processing. Liew et al. [21] has introduced an effective way to reduce the effects of uneven illumination provided that the illumination falls along the vertical direction.

Color information, which actually provides more information to improve the extraction performance, especially when the images are shadowed, shaded and highlighted [25, 26]. In our purposed method we use the color space HSV and work on channel *V* which showed the more accurate result as compare to using channel U which is used by [20]. Convert the image in the HSV space color. The filter consist in a 3 × 3 mask that moves on the image to flatten the pixel values. Often the result cannot be seen with the naked eye because the image is already in good condition as shown in the figure 6b.

The presence of teeth, in image can cause disturbance. The next step creates mask to delete the teeth from the image. Color can provide additional significant information that is not available in gray-level cases. Lip Image Analysis by Colour space HSV CIELAB and CIELUV received more attention. This mask work with image in CIELab and CIELuv color space [20]. Mask applied going to control the range of the pixels in the channels "*a*" and "*u*". Figure 6c and 6d is showing Cielab and Cieluv images and figure 6e and 6f are showing the image after applying the teeth mask. If there are no teeth presence mask will not apply and if teeth appeared then the mask will remove the teeth by changing the teeth pixels value to 0 and go to next step for remove the shadow. In this experiment mouth status is closed so there is no teeth appearance so teeth mask will be skip.

Another disorder is represented by the presence of shadows in the image. Shadow mostly appears under the lower lip and due to some lightening effect, which gives the incorrect extraction of lip boundaries and provides incorrect result. In purposed method we remove this shadow by convert the image in gray scale and find the vertical axis where the shadow started. Divided the image in two sub images (left and right sub image) as shown in the figures6g, 6h. For each sub image apply the stretch contrast of the gray pixel value and merge the both images and we obtained a convolved image figure 6i. The vertical axis represent the index where the shadow started. Mostly shadowed appears down the lower lip or some time by lighting effect.

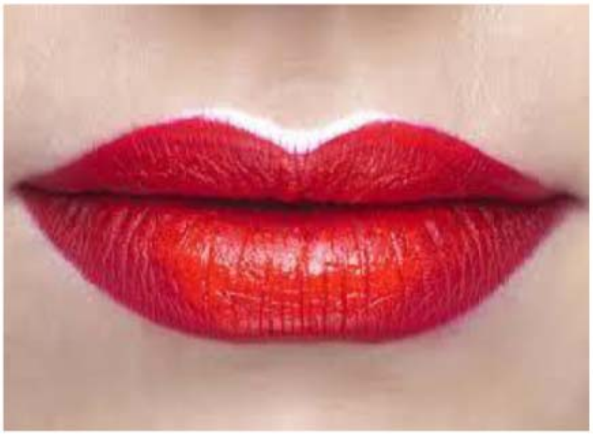


Figure 6 (a). Original RGB image

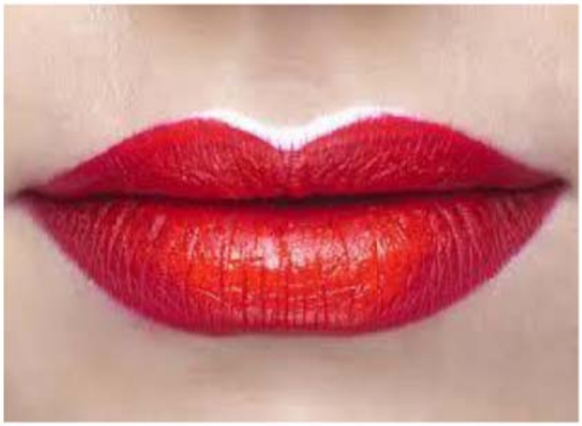


Figure 6 (b). Equalization of illumination

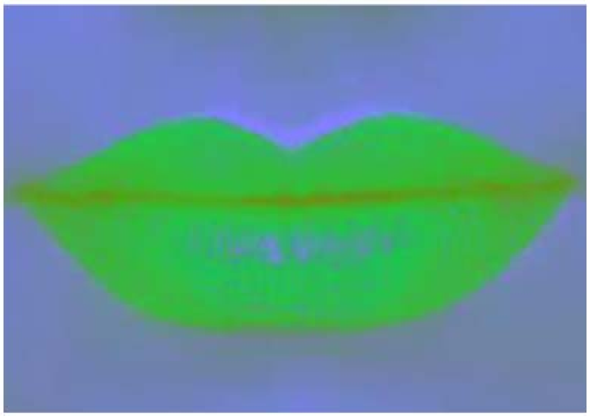


Figure 6 (c). Cielab original image (Teeth Filter)

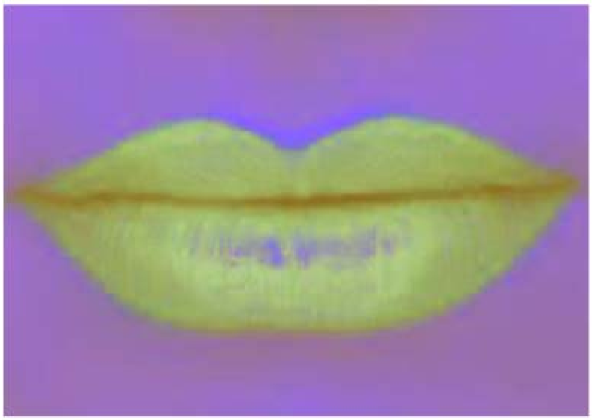


Figure 6 (d). Cieluv original image (Teeth Filter)

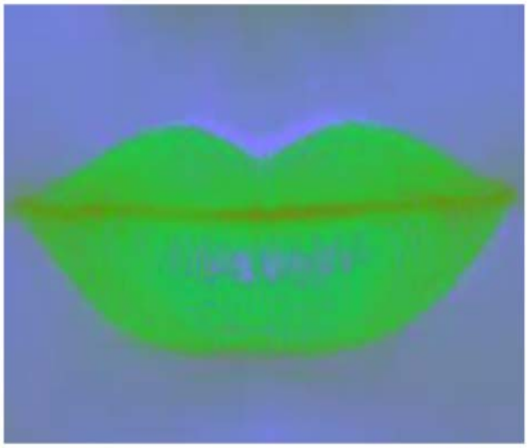


Figure 6 (e). New cielab after applying Teeth Filter

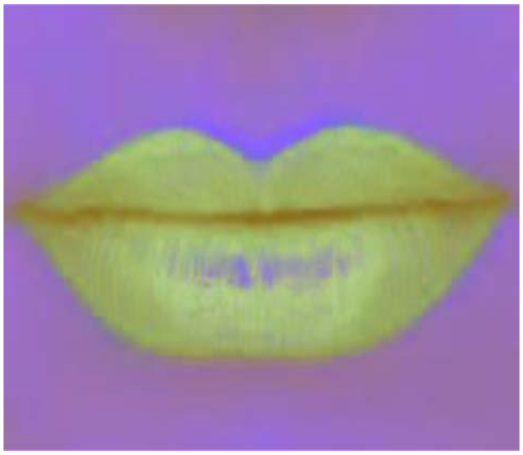


Figure 6 (f). New cielab after applying Teeth Filter

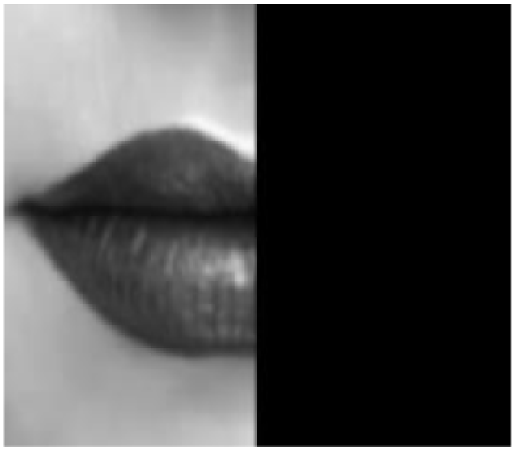


Figure 6 (g). Left sub image

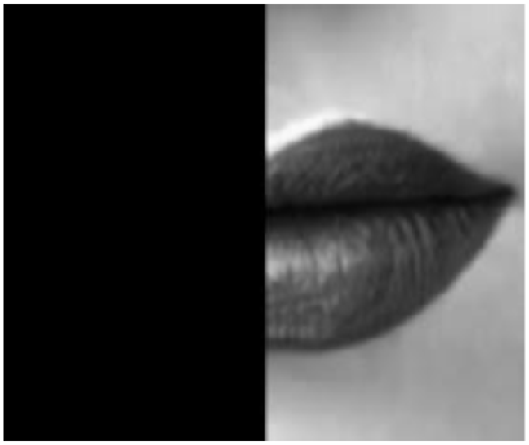


Figure 6 (h). Right sub Image

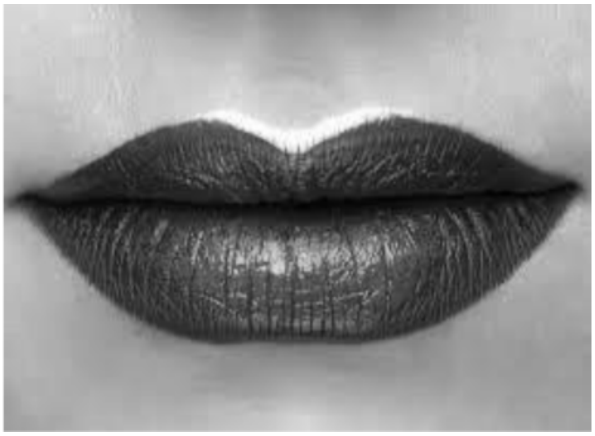


Figure 6 (i). Convolved image

In the next, we select 4 Crucial Points (La, Lb, Va, Vb) described in [20], In our method we found the 4 Crucial Points (La, Lb, Va, Vb) with horizontal and vertical lines. We found the center point C. In the intersection of these two line as shown in the Figure 5a. We retrieve the lips area, shapes height and width as well as movement by this Center point C. In the stretched gray scale images search for the median horizontal axis of the lip (minimum row value) and search the crucial points. For La, search the max value variations in the left sub image and same for the Lb, search in the right sub image. For the Va, search in upper image (0 to median horizontal axis). For Vb, search in lower image (median horizontal axis to height) as described in figures 6j.

White vertical line, when shadow starts and black horizontal line, the median axis of the lip.

After having these points draw the Ellipse in RGB Image around the detected lips. Figure 6h showing two mid ellipse for the upper lip and for the lower lip. Finally draw the 16 points for lip model, for this first convert the image in HSV color space as shown the in figure 6j because this is the correct color space to represent the human skin. We create the graph, from RGB image, that represent the structure of the model as showing figure 6i.

To obtain the 16 point lip model we performed three steps, firstly initialize the model search and referring to the graph. In the first step each point is given by intersection with the ellipse. In the second step, some points of the initial model are not in the correct position. We correct the position of the points looking variations of red color, in the channel H of the HSV image. We retrieve more precise result after second step threshold. Final step, validate the position of each point. Search if some points is in incorrect position. If yes then we recalculate of the position as in step 2. These three steps for 16 point lip model is showing in figure 6k, 6l, and 6m.

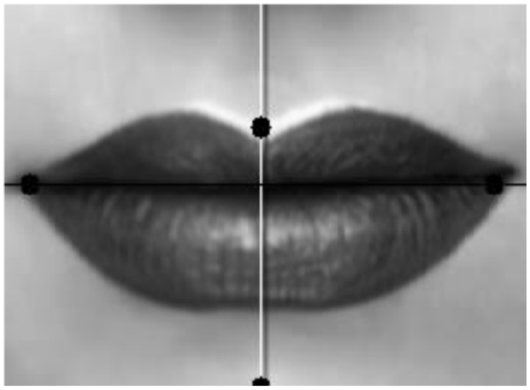


Figure 6 (j). Finding 4 crucial points

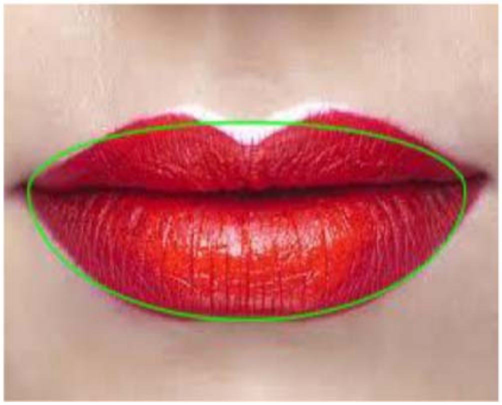


Figure 6 (h). Draw Ellipse

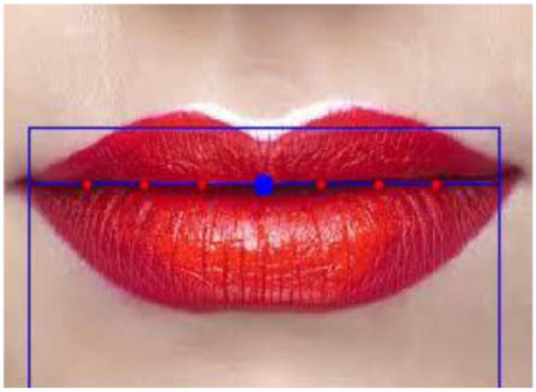


Figure 6 (i). 16 points given by horizontal axis

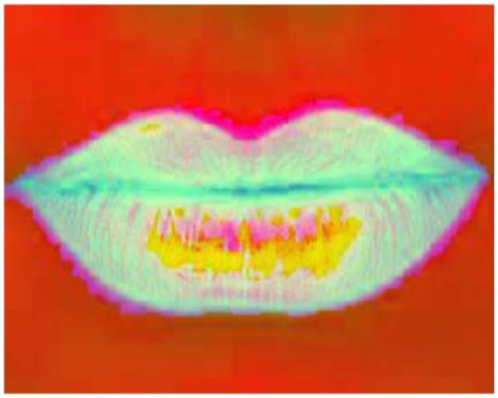


Figure 6 (j). HSV image

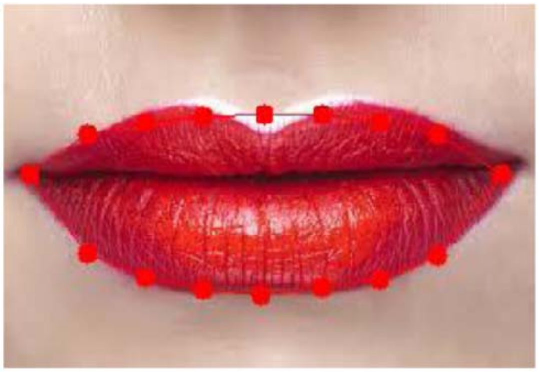


Figure 6 (k). step 1 initialize lip model

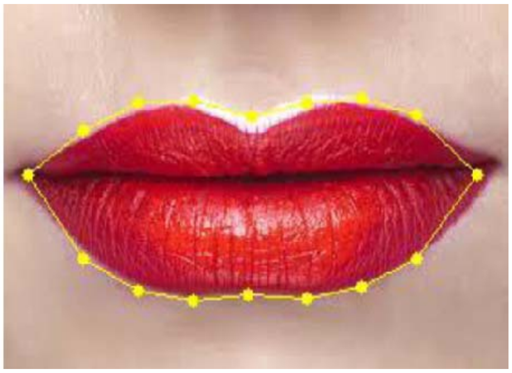


Figure 6 (l). step 2 Threshold lip model

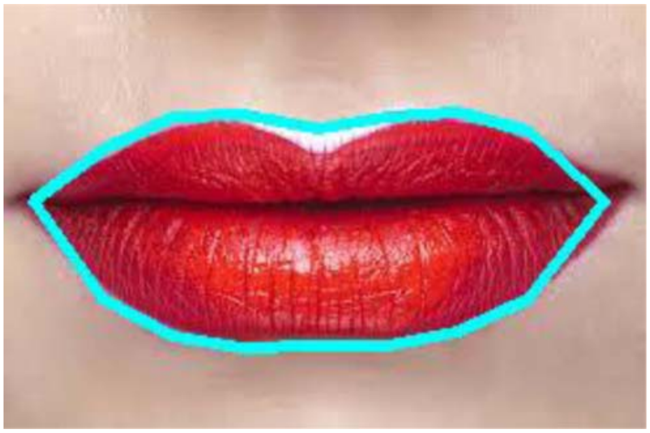


Figure 6 (m). Step 3 validation of the model

It is found that lip extraction result by using conventional region-based ACM extraction, LACM based extraction result with the improper parameters is not correct, by ignoring the complex appearances of an image that may arise outside the local region and also gives not accurate result having moustache [20]. We select a 16-point model which is proposed in [22]. Purposed method gives model the lip shapes more precise and more flexible and physically meaningful in comparison with the less points based lip models as compare to 4 points model and 6 points model, which shows the extracted lip contours have lack of geometric constraint [23, 24]. C. Yiu-ming et al. also used 16 points deformable model with evolving curve and regional energies, but in our method we use some filters such as teeth removal, shadowing enhancement of image and equalization illumination before extraction the lip contour and get 16 points model with more precise result.

## 5. Recognition Phase

Hidden Markov Models (HMMs) are statistical models which particularly are used for describing speech events and actual recognition of lip movement. HMM relies on the Baum-Welch forward-backward re-estimation algorithm (which is a form of expectation maximization) to increase the probability of training data [12]. For the training dataset, different shapes of lip movement to be stored in dataset in term of distributed points on the lips. Lips pattern are divided into several points, therefore lip movement are recognized by a process in which these points are observed in an appropriate order and accurate alignment e.g. top to bottom. Once the HMM is trained, the output result of recognition units determines the class to which it's belong in dataset. Hence since there is no one-to-one mapping from the phoneme set to the viseme set researchers define the visemes by clustering together the phonemes which have a similar visual outcome. The decision about the similarity of the outcome is based on the degree of confusion and distinction human subjects are attaining [10], [11]. Using only one shape parameter together with its delta coefficient and delta scale gave the best recognition rate and also indicate that the training set was not large enough to model more than the first main shape mode reliably [13].

## 6. Conclusion

In this paper our goal was to detect, track and analyze the lip movement to reach the goal of an automatic and robust lip reading. Experiment result showed that our purposed 16 points lip model we can have more precise lip shape that can be used in lip reading as well as recognition. The proposed lip contour extraction method has achieved a more satisfactory result with the smallest value of contour matching performance. This shows that the proposed approach gives the more precise lip boundaries in comparison with the previous methods. By using teeth mask filter, shadowing, equalization Illumination we can retrieve robust results. The Active Appearance Model appears to be one of the most promising tools and was more consistent with the detection of the non-speech sections containing complex lip movements and it should be noted that the reliance on prior information restricts the AAM method to be person specific (a generic method is currently being investigated) [12]. Every model has its own benefits and disadvantages [10]. Finally, we retrieve the more precise lip detection and tracking with 16 key points of lip model. Purposed method also provides the useful in the presence of mustache, lighting condition and shadows. As future work, we will intent to explore more observable features (key points) and make the data sets for 16 points lip model to produce the recognition result. Recognition phase we will apply HMM for recognition of lip movements.

## References

[1] Cootes, T. F., Edwards, G. J., Taylor, C. J. (2001). Active Appearance Models, *IEEE Trans on Pattern Analysis and Machine Intelligence*, 23 (6) 681–685.

[2] Potamianos, G., Chalapathy, N., Luettin, J., Matthews, I. (2005). Chapter 10: Audio-Visual Automatic Speech Recognition: An Overview, *In*: Audio-Visual Speech Processing, MIT Press, September.

[3] Girin, L., Schwartz, J. L., Feng, G. (2001). Audio-Visual Enhancement of Speech in Noise, *J. Acoust. Soc. Am*., 109 (6) 3007–3020.

[4] Yuhas, B. P., Goldstein, M. H., Sejnowski, T. J. (1989). Integration of Acoustic and Visual Speech Signals using Neural Networks, *IEEE Communications Magazine*, p. 75-81.

[5] Ahmad, B. A., Hassanat. (2011). Visual Speech Recognition, Speech and Language Technologies, Prof. Ivo Ipsic (Ed.), InTech, Available from: http://www.intechopen.com/books/speech-andlanguage-technologies/visual-speech-recognition.

[6] Yang, L., Meer, P., Foran, D. J. (2005). Unsupervised segmentation based on robust estimation and color active contour models, *IEEE Transactions on Information Technology in Biomedicine*, 9 (3) 475–486.

[7] Cremers, D., Rousson, M., Deriche, R. (2007). A review of statistical approaches to level set segmentation: integrating color, motion and shape, *International Journal of Computer Vision*, 72 (2) 195–215.

[8] Lankton, S., Tannenbaum, A. (2008). Localizing region-based active contours, *IEEE Transactions on Image Processing*, 17 (11) 2029–2039

[9] Wang, S., Lau, W., Leung, S. (2004). Automatic lip contour extraction from color images, *Pattern Recognition*, 37 (12) 2375–2387.

[10] Yang, M. H., Kriegman, D., Ahuja, N. (2002). Detecting Faces in Images: A Survey, IEEE Trans. Pattern Anal. Mach. Intell., 24, p. 34 – 58.

[11] Butt, W. U. R., Lombardi, L. (2013). Comparisons of Visual Features Extraction Towards Automatic Lip Reading, Paper Accepted in EDULEARN13 (5th International Conference on Education and New Learning Technologies), Barcelona, Spain.

[12] Permit, P. (2007). An Automated Visual Speech Reading System, PhD Thesis, Department of Computing, Communications Technology and Mathematics London Metropolitan University.

[13] Rabiner, L. R., Jung, B.-H. (1993). Fundamentals of Speech Recognition. Prentice Hall.

[14] Williams, J. J., Rutledge, J. C., Garsteckiy, D. C., Katsaggelos, A. K. (1997). Frame rate and viseme analysis for multimedia applications, *In*: Proc. IEEE Works. Multimedia Signal Process, p. 13–18, Princeton.

[15] Williams, J. J., Rutledge, J. C., Aggelos Katsaggelos, K., Garstecki, D. C. (2004). Frame Rate and Viseme Analysis for Multimedia Applications to Assist Speechreading, *The Journal of VLSI Signal Processing*, 20 (1–2) 7-23, November 29.

[16] Aubrey, A., Rivet, B., Hicks, Y., Girin, L., Chambers, J., Jutten, C. (2007). Two novel visual voice activity detectors based on appearance models and retinal filltering. *In*: Proceedings of the 15th European Signal Processing Conference, EUSIPCO.

[17] Juergen, L., Neil A. Thacker, Steve W. Beet. (1996). Visual Speech Recognition Using Active Shape Models and Hidden Markov Models, *In*: proc. IEEE Int. Conf. on acoustics, speech and signal processing.

[18] Wang, L., Wang, X., Xu, J. (2010). Lip detection and tracking using variance based haar-like features and kalman filter, in International Conference on Frontier of Computer Science and Technology, p. 608 –612.

[19] Sodoyer, D., Rivet, B., Girin, L., Savariaux, C., Schwartz, J. -L. (2009). A study of lip movements during spontaneous dialog and its application to voice activity detection, *Journal of the Acoustical Society of America*, 125 (2)1184–1196, February.

[20] Yiu-ming, C., Xin, L., Xinge, Y. (2012). A local region based approach to lip tracking, *Pattern Recognition*, 45 (2012) 3336–3347, 45 (9), September.

[21] Liew, A. W. C., Leung, S. H., Lau, W. H. (2002). Lip contour extraction from color images using a deformable model, *Pattern Recognition*, 35 (12) 2949–2962.

[22] Wang, S., Lau, W., Leung, S. (2004). Automatic lip contour extraction from color images, *Pattern Recognition*, 37 (12) 2375–2387.

[23] Tian, Y., Kanade, T., Cohn, J. (2000). Robust lip tracking by combining shape, color and motion, *In*: Proceedings of the Asian Conference on Computer Vision, p. 1040–1045.

[24] Eveno, N., Caplier, A., Coulon, P. Y. (2004). Accurate and quasi-automatic lip tracking, *IEEE Transactions on Circuits and Systems for Video Technology*, 14 (5) 706–715.

[25] Narayanan, P., Nayar, S., Shum, H. -Y., Jian, Y. -D., Chang, W. -Y., Chen, C. -S. (2006). Attractor-guided particle filtering for lip contour tracking, *In*: Proceedings of the Asian Conference on Computer Vision, 3851, p. 653–663.

[26] Ong, E., Bowden, R. (2008). Robust lip-tracking using rigid flocks of selected linear predictors, *In*: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition.

**Authors Bibliographies**

**Luca Lombardi** received the Laurea (cum laude) in Electronics Engineering from the University of Pavia, in 1986 and the Ph. D. in 1990. From 1991 he held a position as Ricercatore (Assistant Professor) at the Faculty of Engineering of the University of Pavia. From 2000 he held a position as Associated Professor at the Faculty of Engineering of the University of Pavia. His fields of interest concern image processing, machine learning and architectures for image processing.

**Waqqas ur Rehman Butt** received the M.S. degree from University of Agriculture Faisalabad, Pakistan, in 2003. He held a position of Lecturer from 2004 at Centre of Excellence in WRE, University of Engineering and Technology, Lahore, Pakistan. He is currently the PhD Student at the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy. His research interests include image processing, computer vision, and pattern recognition. Marco Grecuccio: He is currently the MS Student at the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy.