Combining Edge Detection and Region Segmentation for Lip Contour Extraction

Usman Saeed and Jean-Luc Dugelay

Eurecom
2229 Routes des Cretes,
06560 Sophia Antipolis, France.
{Usman.Saeed, Jean-Luc.Dugelay}@Eurecom.fr

Abstract. The automatic detection of the lip contour is relatively a difficult problem in computer vision due to the variation amongst humans and environmental conditions. In this paper we improve upon the classical methods by introducing fusion. Two separate methods are first applied, one based on edge detection and the other on region segmentation to detect the outer lip contour, the results from them are then combined. Empirical evaluation of the detection process is also presented on an image subset of the Valid database, which contains lighting, pose and speech variation with promising results.

Keywords: Image Processing, Lip Detection.

1 Introduction

Lip detection is still an active topic of research; the significant interest in this topic originates from the numerous applications where lip detection either serves as a pre-processing step or directly provides visual information to improve performance. It has been applied successfully to Audio-Video Speech and Speaker recognition, where it has considerably improved recognition results, especially in the presence of noise. Another domain of application is gesture recognition for closely related fields of human computer interaction, affective computing. It has also been used in the analysis and synthesis of lips for talking head in video conferencing applications.

In this paper we propose a lip contour detection algorithm based on fusion of two independent methods, edge based and region based. The basic idea is that both techniques have different characteristics and thus exhibit distinct strengths and weaknesses. We also present empirical results on a dataset of considerable size with illumination and speech variation. The rest of the paper is divided as follows. In Section 2 we give the state of the art and in Section 3 we elaborate the proposed method, after that we report and comment our results in section 4 and finally we conclude this paper with remarks and future works in section 5.

2 State of the Art

Lip detection literature can be loosely classified in three categories. The first category of techniques directly uses image information, the second tries to build models and the third is a hybrid approach that combines the image and model based techniques to increase robustness.

2.1 Image Based Techniques

Image based techniques use the pixel information directly, the advantage is that they are computationally less expensive but are adversely affected by variation such as illumination.

Color Based Techniques. Several algorithms base the detection of lips directly on color difference between the lip and skin, but lack of contrast and illumination variation adversely affects these techniques. Some have also suggested color transforms that increase the contrast between skin and lip regions. [1] have reported that difference between red and green is greater for lips than skin and proposed a pseudo hue as a ratio of RGB values. [2] have also proposed a RGB value ratio based on the observation that blue color plays a subordinate role so suppressing it improves segmentation.

Color clustering has also been suggested by some, based on the assumption that there are only two classes i.e. skin and lips, this may not be completely true if facial hair or teeth are visible. Fuzzy clustering was applied for lip detection in [3] by combining color information and spatial distance between pixels in an elliptical shape function. [4] have used expectation maximization algorithm for unsupervised clustering of chromatic features for lip detection in normalized RGB color space. Markov random fields also been proposed to add spatial continuity to segmentation based on color, thus making segmentation more robust in [5].

Subspace Based Techniques. [6] have proposed a lip detector based on PCA, firstly outer lip contours are manually labelled on training data, PCA is then applied to extract the principal modes of contour shape variation, called eigencontour, finally linear regression was applied for detection. LDA has been employed in [7] to separate lip and skin pixels. [8] have proposed a method in which a Discrete Hartley Transform (DHT) is first applied to enhance contrast between lip and skin, then a multi scale wavelet edge detection is applied on the C3 component of DHT.

2.2 Model Based Techniques

Model based techniques are based on prior knowledge of the lip shape and can be quite robust. They are however computationally expensive as compared to image based techniques as they usually involve minimization of a cost function.

[9] have proposed a real time tracker that models the dynamic contours of lips using quadratic B-Splines learned from training data using maximum likelihood estimation algorithm. Tracking is then carried out using Kalman filtering for both frontal and profile view of the lips. [10] have proposed a model consisting of two parabolas for the upper lip and one for lower lip.

Snakes have been commonly used for lip segmentation [11] and achieve reasonable results but need to be properly initialized. Another problem faced by snakes is there inability to detect lip corners as they are located in low gradient regions. [12] have proposed a jumping snake that removes the limitations present in classical snake. It can be initialized far from the lip edge and the parameter adjustment is easy and intuitive.

[13] have proposed Active Shape Models (ASM) and Active Appearance Models (AAM), which learn the shape and appearance of lips from training data that has been manually annotated. Next PCA is applied to reduce the dimensionality and using cost functions, models are iteratively fitted to test images for lip detection. Deformable templates initially proposed by [14] has been extended and modified by several others. [15] have proposed a lip detection method based on Point Distribution Model (PDM) of the face.

2.3 Hybrid Techniques

These methods combine image based and model based techniques for lip detection. Image based techniques are considered computationally less expensive but not so robust to illumination and other types of variation. Model based techniques on the other hand are robust and accurate but are much more computationally complex. Thus majority of the hybrid techniques proposed in the literature use color based techniques for a quick and rough estimation of the candidate lip regions and then apply a model based approach to extract accurate lip contours.

[16] have proposed a hybrid technique that first applies a color transform to reduce the effect of lighting. Then horizontal and vertical projections of the lip are analyzed to detect the corner points and finally a geometric lip model is applied. [17] have combined a fuzzy clustering algorithm in CIELAB color space for rough estimation and then an ASM for accurate detection of lip contours. [18] have proposed a hybrid system that models the lip area by expectation maximization algorithm after a color transform in RGB space has been applied. Then a snake is initialized, which is fitted on the upper and lower contours of the mouth by a multi level gradient flow maximization. [19] have proposed a lip tracking by combining lip shape, color and motion information. The shape has been modeled using two parabolas, lip and skin color is modeled by Gaussian distribution and motion by modified Lucas-Kanade tracking.

3 Proposed Lip Detection

In this section we present a lip detection method to extract the outer lip contour that combines edge based and region based algorithms. The results from the two methods are then combined by AND/OR fusion. The novelty lies in the fusion of two methods, which have different characteristics and thus exhibit different type of strengths and weaknesses. The other significance of this study lies in the extensive testing and evaluation of the detection algorithm on a realistic database. Most previous studies either never carried out empirical comparisons to the ground truth at all or sufficed by using a limited dataset. Even if empirical testing was done by some studies [24], [8] they were limited to high resolution images with constant lighting conditions.

Figure 1 gives an overview of the lip detection algorithm. Given an image, it is assumed that a human face is present and already detected; the first step is to select the mouth Region of Interest (ROI) using the lower one third of the detected face. The next step involves the outer lip contour detection where the same mouth ROI is provided to the edge and region based methods. Finally the results from the two methods are fused to obtain the final outer lip contour.

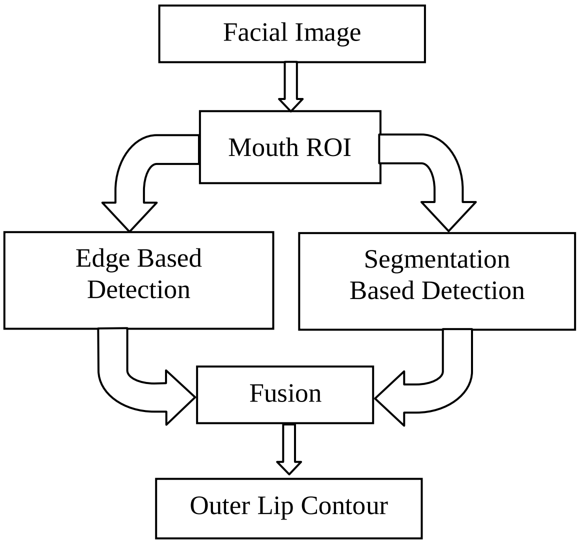


Figure 1. Overview of the proposed lip detection method.

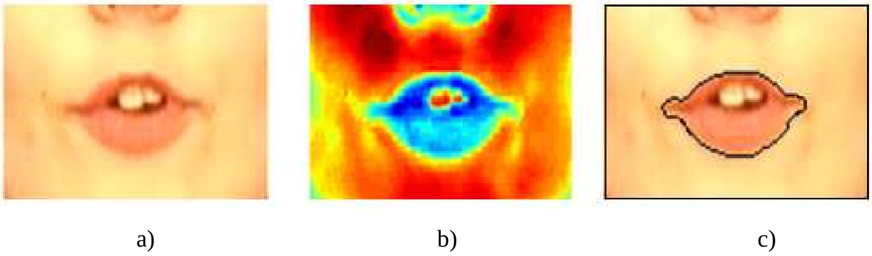
3.1 Edge Based Detection

The first algorithm is based on a well accepted edge detection method. It consists of two steps, the first one is a lip enhancing color transform and the second one is edge detection based on active contours. Several color transforms have already been proposed for either enhancing the lip region independently or with respect to the skin. Here, after evaluating several transforms we have selected the color transform proposed by [2]. It is based on the principle that blue component has reduced role in lip / skin color discrimination and is defined in eq. 1.

$$I = \frac{2G - R - 0.5B}{4}. \qquad (1)$$

Where R,G,B are the Red, Green and Blue components of the mouth ROI. The next step is the extraction of the outer lip contour, for this we have used active contours [20]. Active contours are an edge detection method based on the minimization of an energy associated to the contour. This energy is the sum of internal and external energies; the aim of the internal energy is to maintain the shape as regular and smooth as possible. The most straightforward approach grants high energy to elongated contours (elastic force) and to high curvature contours (rigid force). The external energy models the edge of the object and is supposed to be minimal when the active contours (snake) is at the object boundary. The simplest approach consists in using regularized gradient as the external energy. In our study the contour was initialized as an oval half the size of the ROI with node separation of four pixels.

Since we are have applied active contours which have the possibility of detecting multiple objects, on a ROI which may include other features such as the nose tip, an additional cleanup step needs to be carried out. This consists of selecting the largest detected object approximately in the middle of the image as the lip and discarding the rest of the detected objects.



a) b) c)

Figure 2: a) mouth *ROI,* b) Color Transform, c) Edge Detection.

3.2 Region Based Detection

In contrast to the edge based technique the second approach is region based after a color transform in the YIQ domain. As in the first approach we experimented with several color transform presented in the literature to find the one that is most appropriate for lip segmentation. [21] have presented that skin/lip discrimination can be achieved successfully in the YIQ domain, which firstly de-couples the luminance and chrominance information. They have also suggested that the I channel is most discriminant for skin detection and the Q channel for lip enhancement. Thus we transformed the mouth ROI form RGB to YIQ color space using the equation 2 and retained the Q channel for further processing.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.31135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \qquad (2)$$

In classical active contours the external energy is modeled as an edge detector using the gradient of the image, to stop the evolution of the curve on the boundary of the desired object while maintaining smoothness in the curve. This is a major limitation of the active contours as they can only detect objects with reasonably defined edges. Thus for the second method we selected a technique called "active contours without edges" [25], which models the intensities in different region of the image and uses it as the stopping term in active contours. More precisely this model [25] is based on Mumford–Shah functional and level sets. In the level set formulation, the problem becomes a mean-curvature flow evolving the active contour, which will stop on the desired boundary. However, the stopping term does not depend on the gradient of the image, as in the classical active contour models, but is instead based on Mumford–Shah functional for segmentation.

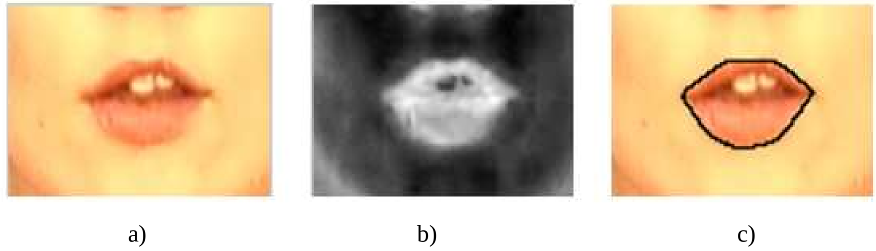


Figure 3: a) Mouth *ROI,* b) Color Transform, c) Region Detection

3.3 Fusion

Lip detection being an intricate problem is prone to errors, especially the lower lip as reported by [22]. We faced two types of errors and propose appropriate error detection and correction techniques. The first type of error, which was commonly observed in the edge based method, was caused when the lip was missed altogether and some other feature was selected. This error can easily be detected by applying feature value and locality constraints such as the lip cannot be connected to the ROI's boundary and cannot have an area value less than one-third of the average area value in the entire video sequence. If this error was observed, the detection results were discarded.

The second type occurs when the lip is not detected in its entirety, e.g. missing the lower lip, such errors are difficult to detect thus we proposed to use fusion as a corrective measure, under the assumption that both the detection techniques will not fail simultaneously.

The detected outer lip contours from the above described methods are then used to create binary masks which describe the interior and the exterior of the outer lip contour. These masks are then fused using AND and OR logical operators.

4 Experiments and Results

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on a subset of the Valid Database [23], which consists of 106 subjects. The database contains five sessions for each subject where one session has been recorded in studio conditions while the others are in uncontrolled environments such as the office or corridors. In each session the subjects repeat the same sentence, “Joe took father's green shoe bench out”. One image was extracted from each of the five videos to create a database of 530 facial images. The reason for selecting one image per video was that the database did not contain any ground truth for lip detection, so ground truth had to be created manually, which is a time consuming task. The images contained both illumination and shape variation; illumination from the fact that they were extracted from all five videos, and shape as they were extracted from random frames of speaker videos.

As already described above the database did not contain any ground truth with respect to the outer lip contour. Thus the ground truth was established manually by a single operator using Adobe Photoshop. The outer lip contour was marked using the magnetic lasso tool which separated the interior and exterior of the outer lip contour by setting the exterior to zero and the interior to one.

To evaluate the lip detection algorithm we used the following two measures proposed by [8], the first measure determines the percentage of overlap (OL) between the segmented lip region *A* and the ground truth A_G. It is defined by eq. 3.

$$OL = \frac{2(A \cap A_G)}{A + A_G} * 100. \tag{3}$$

Using this measure, total agreement will have an overlap of 100%. The second measure is the segmentation error (SE) defined by eq. 4.

$$SE = \frac{OLE + ILE}{2 * TL} * 100. \tag{4}$$

OLE (outer lip error) is the number of non-lip pixels being classified as lip pixels and *ILE* (inner lip error) is the number of lip-pixels classified as non-lip ones. *TL* denotes the number of lip-pixels in the ground truth. Total agreement will have an SE of 0%.

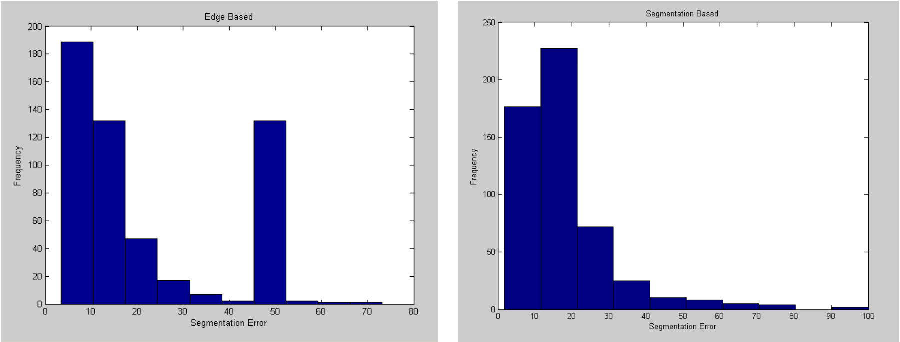


Figure 4: Histograms for Segmentation Errors

Initially we calculated the overlap and segmentation errors for edge and region based methods individually, and it was visually observed that edges based method was more accurate but not robust and on several occasions missed almost half of the lip. This can also be observed in the histogram of segmentation errors; although the majority of lips are detected with 10% or less error but a large number of lip images exhibit approximately 50% of segmentation error. On the other hand region based method was less accurate as majority of lips detected are with 20% error but was quite robust and always succeeded in detecting the lip.

Table 1: Lip detection Results

Lip Detection Method	Mean Segmentation Error (SE) %	Mean Overlap (OL) %
Segmentation Based	17.8225	83.6419
Edge Based	22.3665	65.6430
OR Fusion	**15.6524**	83.9321
AND Fusion	18.4067	84.2452
OR Fusion on 1st Video	13.9964	87.1492

Table 1 describes the results obtained, the best results were observed for OR fusion with 15.65% mean segmentation error. "OR Fusion on 1st Video" are the results that were obtained when OR fusion was applied to only the images from the first video, which are recorded in studio conditions.

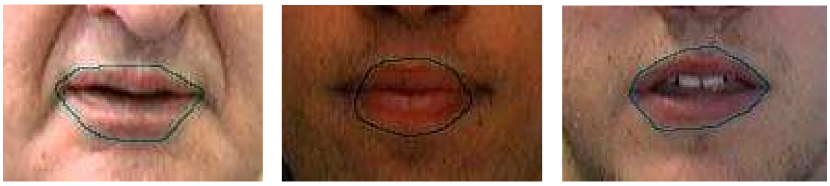


Figure 5: Example of Images with approximately 15 % Segmentation Error

The minimum segmentation error obtained was 15.65%, which might seem quite large, but on visual inspection of Figure 5, it is evident that missing the lip corners or including a bit of the skin region can lead to this level of error. Another aspect of the experiment that must be kept in mind is the ground truth, although every effort was made to establish an ideal ground truth but due to limited time and resources some compromises had to be made.

5 Conclusions

In this paper we have presented a novel lip detection method based on the fusion of edge based and region based methods, along with empirical results on a dataset of considerable size with illumination and speech variation. We observed that the edge based technique is comparatively more accurate, but is not so robust and fails if lighting conditions are not favorable, thus it ends up selecting some other facial feature. On the other hand the region based method is robust to lighting but is not as accurate as the edge based method. Thus by fusing the results from the two techniques we achieve comparatively better results than using only one method. The proposed methods were tested on a real world database with adequate results.

Although the results achieved are quite promising, there is still some room for improvements. Currently we compensate for errors by fusion, we would like to automatically evaluate the results from the independent methods and detect failure, then propose an appropriate fusion approach. We have only tested two fusion approaches; it would be interesting to study others such as a post-classifier.

6 References

1. Hulbert, A., Poggio, T.: Synthesizing a Color Algorithm from Examples. Science. vol. 239, pp. 482-485 (1998)
2. Canzlerm, U., Dziurzyk, T.: Extraction of Non Manual Features for Video based Sign Language Recognition. In: Proceedings of IAPR Workshop, pp. 318-321 (2002)
3. Leung, S.-H., Wang, S-L., Lau, W.-H.: Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. In: IEEE Transactions on Image Processing, vol.13, no.1, pp.51-62 (2004)
4. Lucey, S., Sridharan, S., Chandran, V.: Adaptive mouth segmentation using chromatic features. In: Pattern Recogn. Lett, vol. 23, pp. 1293-1302 (2002)
5. Zhang, X., Mersereau, R. M.: Lip feature extraction toward an automatic speechreading system. In: Proc. IEEE Int. Conf. Image Processing, vol. 3, pp. 226–229 (2000)
6. Lucey, S., Sridharan, S., Chandran, V.: Initialised eigenlip estimator for fast lip tracking using linear regression. In: Proceedings. 15th International Conference on Pattern Recognition, vol.3, pp.178-181 (2000)
7. Nefian, A., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A couple HMM for audio-visual speech recognition. In: Proc. ICASSP, pp. 2013–2016 (2002)
8. Guan, Y.-P.: Automatic extraction of lips based on multi-scale wavelet edge detection. In: IET Computer Vision, vol.2, no.1, pp.23-33 (2008)

9. Kaucic, R., Dalton, B., Blake, A.: Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications. In: Proceedings of the 4th European Conference on Computer Vision, vol. II (1996)
10. Coianiz, T., Torresani, L., Caprile, B.: 2D deformable models for visual speech analysis. NATO Advanced Study Institute: Speech reading by Man and Machine, pp. 391–398 (1995)
11. Aleksic, P. S., Williams, J.J., Wu, Z., Katsaggelos, A.K.: Audiovisual speech recognition using MPEG-4 compliant visual features. EURASIP J. Appl. Signal Processing. pp. 1213–1227 (2002)
12. Eveno, N., Caplier, A., Coulon, P.: Accurate and quasi-automatic lip tracking. In: IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pp. 706 – 715 (2004)
13. Cootes, T. F.: Statistical Models of Appearance for Computer Vision. Technical report, University of Manchester (2004)
14. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. Int. J. Comput. Vision, vol. 8, pp. 99–111 (1992)
15. Huang, C.L., Huang, Y.M.: Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. Journal of Visual Communication and Image Representation, vol. 8, pp. 278-290 (1997)
16. Werda, S., Mahdi, W., Ben-Hamadou, A.: Colour and Geometric based Model for Lip Localisation: Application for Lip-reading System. In: 14th International Conference on Image Analysis and Processing, pp.9-14 (2007)
17. Mok, L.L., Lau, W.H., Leung, S.H., Wang, S.L., Yan, H.: Person authentication using ASM based lip shape and intensity information. In: International Conference on Image Processing, vol.1, pp. 561-564 (2004)
18. Bouvier, C., Coulon, P.-Y., Maldague, X.: Unsupervised Lips Segmentation Based on ROI Optimisation and Parametric Model. In: IEEE International Conference on Image Processing, vol.4, pp. 301-304 (2007)
19. Tian, Y., Kanade, T., Cohn, J.: Robust lip tracking by combining shape, color and motion. In: Proc. ACCV, pp. 1040–1045 (2000)
20. Michael, K., Andrew, W., Demetri, T.: Snakes: active Contour models. International Journal of Computer Vision, vol. 1, pp. 259-268 (1987)
21. Thejaswi N. S. Sengupta, S.: Lip Localization and Viseme Recognition from Video Sequences. In: Fourteenth National Conference on Communications, (2008)
22. Bourel, F., Chibelushi, C.C., Low, A.A.: Robust Facial Feature Tracking. In: Proceedings of the 11th British Machine Vision Conference, vol. 1, pp. 232–241. UK (2000)
23. Fox, N.A., O'Mullane, B., Reilly, R.B.: The realistic multi-modal VALID database and visual speaker identification comparison experiments. In: 5th International Conference on Audio- and Video-Based Biometric Person Authentication (2005)
24. Liew, A.W-C., Shu Hung, L., Wing Hong, L.: Segmentation of color lip images by spatial fuzzy clustering.In: IEEE Transactions on Fuzzy Systems, vol.11, no.4, pp. 542-549 (2003)
25. Chan, T.F., Vese, L.A.: Active contours without edges. In: IEEE Transactions on Image Processing, vol.10, no.2, pp.266-277 (2001)