UNSUPERVISED LIPS SEGMENTATION BASED ON ROI OPTIMISATION AND PARAMETRIC MODEL

Christian Bouvier[(1)], *Pierre-Yves Coulon*[(1)], *Xavier Maldague*[(2)]

GIPSA_lab[(1)], INPG, CNRS, UJF, U.Stendhal
46 av. F. Viallet, 38031 Grenoble, France
Christian.Bouvier, Pierre-Yves.Coulon@inpg.fr

LVSN[(2)], University Laval, Sainte-Foy,
Quebec, Canada, G1K7P4
maldagx@gel.ulaval.ca

ABSTRACT

Lips segmentation is a very important step in many applications such as automatic speech reading, MPEG-4 compression, special effects, facial analysis and emotion recognition. In this paper, we present a robust method for unsupervised lips segmentation. First the color of the lips area is estimated using expectation maximization and a membership map of the lips is computed from the skin color distribution. The region of interest (*ROI*) is then found by automatic thresholding on the membership map. Given a mask of the *ROI*, we initialize a snake that is fitted on the upper and lower contour of the mouth by multi level gradient flow maximization. Finally to find the mouth corners and the final contour of the mouth, we use a parametric model composed of cubic curves and Bezier curves.

Index Terms **—** Mouth, Lips Segmentation, snake, parametric model, color, features detection.

1. INTRODUCTION

Many previous works have been done on human face components detection such as mouth, lips, eyes, … .We are interested in lips detection but our purpose in this paper is to extract the lips contours as accurately as possible. This can be useful for example for speech recognition by adding visual information to the audio one [1] or for facial emotion detection [2]. The global problem in face features analysis is that the algorithms have to deal with unpredictable conditions such as the lighting conditions, different scales, different subjects, with different characteristics and different acquisition systems.

To achieve a good and robust segmentation, many techniques have been developed in the past. We can, mainly, classify those techniques in 2 families, the deterministic methods and statistical methods.

In the first class of method there is no prior knowledge and we can distinguish two different approaches, a region approach and a contour approach.

For example Liévin [3] is using Markov random field in a suitable space color and movement to segment the area of the mouth and then applies an active contour on the mask to extract the contour of the mouth. This method can give accurate results but the problem with the Markov random field is to initialize the color distributions for the relaxation to classify the pixels which are from the lips, from the face and from the background. Moreover, the final mask can lead to impossible results because the shape of the mouth is not constraint. Chung Liew and al [4] proposed a similar method using fuzzy clustering on color image and then extract the contour from the final mask. There are processing to constraint the global shape of the mouth, but the final contours are noisy.

Eveno [5] used a contour approach to extract the lips contour. He introduced a flexible polynomial model of the mouth composed of cubic curves and fits it using gradient information based on pseudo hue [6] and luminance. The model developed by Eveno is very interesting because it can be fitted on extreme shape of mouth, but the initialization of the jumping snake for the upper lip and the snake for the lower lip based on gradient mean flow maximization are not robust to change of lighting conditions and thus the model can be fitted on a false contour.

More recently, statistical methods have been developed to extract face features and particularly the mouth. Coots and al. [7] introduced active shape (ASM) and active appearance models (AAM). The shape and the appearance of the object of interest are learned from a training set of manually annotated images. To reduce the dimension of the model, a principal component analysis is run on the data collected. Using a cost function, the models are iteratively fitted to reduce the difference between the models and the real image. Gacon [8] used a similar approach to construct a multi-speakers model of the mouth area. The model is then initiated by color based segmentation. Then the goal is to find the parameters that will minimize the difference between the responses of Gaussian descriptors [9] of the real image and those from the statistical model. The results given by Gacon show good performances but one needs to manually annotate hundreds of pictures to construct the

statistical model. The advantage of the statistical model, which is to always give a "possible" result, is a problem in case of a subject or a shape too far from the mean model.

Figure 1 : Pictures of the lips and the corresponding picture after color transformation

Our Goal in this work is to achieve an unsupervised method to accomplish a robust and accurate segmentation of the mouth of color face pictures which can eventually be used after to initiate a more complex method, for example a method with ASM and AAM [7].

We make the hypothesis that the face have been detected by a preliminary processing and that the image is focused on the lower part of the face (Fig. 1).

The paper is organized as follows. In section 2 we will describe the extraction of the region of interest (*ROI*) by color segmentation using EM -expectation maximization. Section 3 describes the detection of the key points of the mouth and the contour extraction. Finally section 4 presents some experimental results and concludes this paper.

2. ROI COLOR SEGMENTATION

2.1. Color transformation

The choice of a suitable color space in color-based approaches is essential to the robustness of the algorithm. Most of the time, conversion to spaces such as HIS or HSV gives noisy results because of the poor quality of the pictures. In this work we use the ratio $H=G/R$, with R and G, respectively, the red and green component of the *RGB* color space. This is a very simple and computationally fast processing and it is very efficient to enhance contrast (see Fig. 1) between skin color and lips color:

$$H = \begin{cases} 256 \times \dfrac{G}{R} & if \;\; R > G \\ 255 & otherwise \end{cases} \tag{1}$$

In [3] Liévin and Luthon give theoretical background to this transformation.

2.2. Lips Area Detection

In this work the hypothesis is that the color distribution of the picture can be approached by a gaussian mixture model. Let $HUE=[h_1 \ldots h_M]$ be a vector of the value of the pixel from H and $g_1 \ldots g_M$ a set of N gaussians. Using Expectation maximization, the parameters $\theta=\{\mu_1 \ldots \mu_N, \sigma_1 \ldots \sigma_N, w_1 \ldots w_N\}$ of the Gaussian mixture $P(h_i)$ are estimated:

$$P(h_i) = \sum_{j=1}^{N} w_j \times g(h_i, \mu_j, \sigma_j) \tag{2}$$

$$g(h_i, \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-(h_i - \mu_j)^2 / 2{\sigma_j}^2\right) \tag{3}$$

The EM algorithm is initiated by a K-means algorithm with N clusters, as we have no prior knowledge on the distributions parameters. We supposed that the pictures are focused on the lower part of the face and so most of the pixels are from the skin. From the Gaussian mixture we extract the Gaussian that has the highest weight w_j and it is associated with the skin. With the estimate distribution of the skin we compute a membership map of the lip pixels (Fig 2.):

$$MAP = 1 - \exp\left(-(h - \mu_{skin})^2 / 2{\sigma_{skin}}^2\right) \tag{4}$$

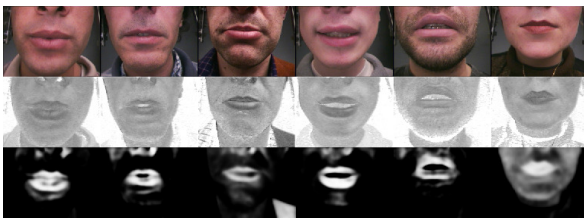

Figure 2 : Pictures of the lips, the corresponding hue picture H and membership map

The membership map is not computed using the direct estimation of the lips color distribution. Most of the time, there is note enough lips pixel (In case of very thin lips) and the K-means algorithm combined with the *EM* fails to estimate the lips color distribution. Instead we use the skin color distribution as it is more robust because most of the pixels in the image are from the skin.

A threshold t (5) must now be found to obtain a binary mask *MASK* of the lips.

$$\begin{aligned} &MASK = MAP > t \\ &t \in [0 \ldots 1] \end{aligned} \tag{5}$$

In order to get t, first we compute the gradients R_{top} and R_{bottom} as Eveno proposed in [5], assuming I, R, G are respectively the luminance, red and green components of the image with x, y coordinates:

$$R_{top}(x,y) = \nabla\left(\frac{R}{G}(x,y) - I(x,y)\right) \tag{6}$$

$$R_{bottom}(x,y) = \nabla\left(\frac{R}{G}(x,y)\right) \tag{7}$$

Figure 3: Pictures of the lips and the corresponding lips *ROI* detected

Secondly, the threshold t is computed by maximization of the coefficient α (10) which is the product between Φ (8), the mean flow of the gradients R_{top} and R_{bottom} through the contour (Cont) of the mask (5), and the ratio R (9) (Fig 3.). This ratio is high when pixels with strong membership are in the area covered by the mask.

$$\Phi = \left(\int_{Cont} R_{top}.dn + \int_{Cont} R_{bottom}.dn \right) \Big/ |Cont| \tag{8}$$

$$R = \iint_{MAP} MASK \times MAP.dxdy \Big/ \iint_{MAP} MAP.dxdy \tag{9}$$

$$\alpha = R \cdot \Phi \tag{10}$$

dn is the vector orthogonal to the contour.

2.3. Choice of *N*

The choice of the number of cluster N has an important influence on the lips area detection. Typically there are N=3 clusters that are looked for, the background, the skin and the lips. In our application, the amount of background pixels can be very small and N=2 clusters gives better results. So the preceding procedure is run for N from 2 to 3 and the result that maximizes α is chosen. The final mask is called *ROI*.

3. KEY POINTS DETECTION AND CONTOUR EXTRACTION

3.1. Detection of the upper and lower contour

To detect the upper and lower contour, because the contours of the mouth are often not well defined, we choose a multi scale approach similar to [9] where Lindeberg propose a methodology to detect features with automatic scale selection. Using the *ROI* detected previously, we will compute an edges map to initiate snakes [10] for the upper and lower contour. Then those contours will be fitted to the mouth using multi-scale gradient information.

The idea to use multi-scale is to be able to deal with different resolution and subject. On Fig 4 we have computed $|R_{bottom}|$ for s=1…3. Gradients fields have been normalized for display. We can see that the lower contour of the mouth is well defined for higher scales (s=3).

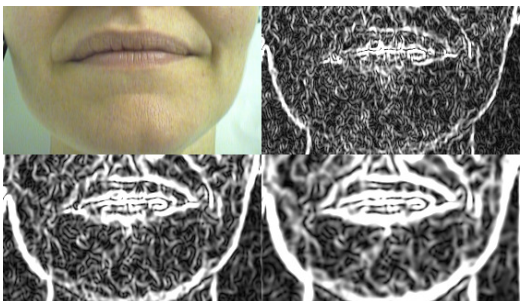

Figure 4: Result of $|R_{bottom}(s)|$ for s=1…3.

First we compute the normalized [9] R_{top} and R_{bottom} for different scales s :

$$R_{top}(x,y,s) = \nabla\left(\left(\frac{R}{G}(x,y) - I(x,y) \right) * g(x,y,s) \right) \cdot s \tag{11}$$

$$R_{bottom}(x,y,s) = \nabla\left(\frac{R}{G}(x,y) * g(x,y,s) \right) \cdot s \tag{12}$$

With * the convolution operator and $g(x,y,s)$ a gaussian window with the standard deviation s. In this work s=1…3.

After that we compute edges maps for the upper lip and for the lower lip for s=1…3:

$$\begin{aligned} &EDGE_k(s) = \left| R_k(x,y,s) \right| > \beta_k(s) \\ &\beta_k(s) = mean\left(ROI \times \left| R_k(x,y,s) \right| \right) \\ &k = (top, bottom) \end{aligned} \tag{13}$$

A labeling is done on the edges maps for s=1…3 and k=(top,$bottom$) and we keep the edges that are found overall the scales (Fig 4(c)). Finally we obtain two binary images composed of the selected edges $EDGE_{top}$ and $EDGE_{bottom}$ and we add them:

$$EDGE = (EDGE_{top} + EDGE_{bottom}) \times ROI \tag{14}$$

An active contour is then run on that binary image composed of the selected edges (Fig 5(c)) and is sampled with a fixed step. This gives us a set of the upper points M_{top} that are used as an initial state of the upper contour and a set of the lower points M_{bottom}, used as an initial state of the lower contour of the lips (Fig 5(c)).

The fitting procedure for the upper contour is then: for each point $M_{top}(m)$, the horizontal position is fixed, and the best vertical position is found by maximizing the $R_{top}(s)$ mean flow through a Bezier curve for s=1…3. The curve is interpolated using [$M_{top}(m-1)$, $M_{top}(m)$, $M_{top}(m+1)$] as control points for a limited number of vertical neighbors of $M_{top}(m)$ (Fig 4.). For the extreme points, all the points M_{top} are used to compute the Bezier curve.

The procedure for the lower contour of the lips is the same using $R_{bottom}(s)$ for s=1…3 and the set of points M_{bottom}.

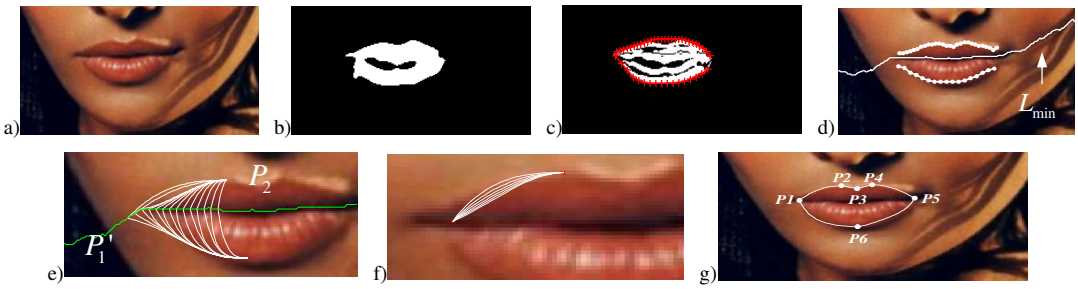

Figure 5: Overview of the complete processing, a) Picture of a mouth, b) ROI mask of the lips, c) Edge map and the initialization of M_{top} and M_{bottom}, d) The contours after optimization, e) Left contours for different left mouth corners tested on L_{min}, f) Cubic curve of the upper-left contour for different slopes in the mouth corner tested, g) The final contour of the mouth.

3.2. Contour extraction

The final contour extraction is done by fitting a parametric model of the mouth using the contour detected before. Eveno and al [5] introduce a flexible model composed of cubic curves. In this work we use a similar model excepted for the lower contour for which we use Bezier curves (Fig 5(g)). Six key points are used to compute the final contour $P_{i=1...6}$. The points $P_{2,3,4}$ are extracted from M_{top}. P_6 is found by searching for the lowest point on M_{bottom} between P_2 and P_4.

The curve fitting and the detection of the mouth corners P_1 and P_5 are done at the same time. We make the hypothesis that the mouth corners are on the minimum luminance line [5] L_{min} (Fig 5(d)).

To detect the left mouth corner, we search for the point P_1 that will that maximized the $R_{top}(s)$ mean flow throughout the upper and lower left curves (Fig 5(e)). For each tested point a finite number of curves with different slopes are test for the upper contour (Fig 5(f)). For the right side of the mouth we use the same procedure.

The final contour of the Cupid's bow is given by plotting linear curve between P_2 and P_3 and between P_3 and P_4 (Fig 5(g)).

4. RESULTS AND CONCLUSION

We Have tested our algorithm on a database of 450 pictures from 12 different subject manually annotated without changing any parameter. The results are the error between the detected key points $P_{i=1...6}$ and the annotated key points normalized by the width of the mouth. For comparison we give the results obtained on the same database with the method from [5], a deterministic method, and from [8] in which the algorithm is trained on the database. The errors are given in percentage of the width of the mouth with the standard deviation (Table 1). We can see that the results of the proposed algorithm for the outer are good, close to the results from [8].

We have presented an algorithm for the extraction of the outer contour of the lips and we also presented quantitative results of the presented work compared to two different kinds of method, a deterministic method and a statistical method. The next improvement will be to add an inner contour extraction module.

Algorithm	Our Algorithm	Eveno Algorithm	Gacon Algorithm
Error (%)	4.1 ± 5.1	7.5 ± 13	2.7 ± 1.3

Table 1: Quantitative results of our algorithm

5. REFERENCES

[1] B. Le Goff, T. Guiard-Marigny and C. Benoît. "Read my Lips… and my Jaws! How Intelligible are the Components of a Speaker's Face", *In Proc. Of the European Conf. On Speech Communication and Technology*, pp. 291-294, Madrid, Espagne, 1995.

[2] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut, "Facial Expressions Recognition Based on The Belief Theory: Comparison with Different Classifiers," *Proc. 13th International Conference on Image Analysis and Processing*, Cagliari, Italy, 2005.

[3] M. Liévin, F. Luthon, "Nonlinear Color Space and Spatiotemporal MRF for Hierarchical Segmentation of Face Features in Video", *IEEE Transactions on Image Processing*, Volume 13, No. 1, pp. 63-71, 2004.

[4] A.W.C. Liew, S.H. Leung and W.H. Lau, "Segmentation of Color Lip Images by Spatial Fuzzy Clustering", *IEEE Transactions on Fuzzy Systems,* Volume 11, No. 1, pp. 542-549, 2003.

[5] N.Eveno, A. Caplier, and P-Y Coulon, "Accurate and Quasi-Automatic Lip Tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 14, No.5, pp. 706-715, May 2004.

[6] T. Poggio, and A. Hulbert, "Synthesizing a Color Algorithm From Examples", Science, Volume 239, pp. 482-485, 1998.

[7] T. F. Cootes. "Statistical Models of Appearance for Computer Vision", Technical report, free to download on http://www.isbe.man.ac.uk/bim/refs.html, 2004.

[8] P. Gacon, P.-Y. Coulon, G. Bailly. "Non-Linear Active Model for Mouth Inner and Outer Contours Detection", *2005 European Signal Processing Conference (EUSIPCO'05)*, Antalya, Turquie, 2005.

[9] T. Lindeberg, "Feature Detection with Automatic Scale Detection", *IJVC*, Volume 30, No.2, pp. 77-116, 1998.

[10] M. Kass, A. Witkin et D. Terzopoulos, "Snakes: Active Contour Models", *International Journal of Computer Vision*, pp. 321-331, 1987.