Automatic Hybrid Approach for Lip POI Localization: Application for Lip-reading System

Salah Werda, Walid Mahdi and Abdelmajid Ben Hamadou
MIRACL: Multimedia Information systems and Advanced Computing Laboratory
Higher Institute of Computer Science and Multimedia, Sfax, Tunisia
Salah_werda@yahoo.fr; {walid.mahd benhamadou.abdelmajid}@isims.rnu.tn

Abstract

Automatic Lip-reading system is one of the different assistive technologies for hearing impaired or elderly people. We can imagine, for example, a dependent person ordering a machine with an easy lip movement or by a simple visemes (visual phoneme) pronunciation. The need for an automatic lip-reading system is ever increasing. The lip-reading system is decomposed in three subsystems, first we have the lip localization system, which localizes the lips in the digital input, then there is the feature extracting system, and finally we have the classification system, which maps feature vectors to visemes. The major difficulty of the lip-reading system is the extraction of the visual speech descriptors. In fact, to ensure this task it is necessary to carry out an automatic localization and tracking of the labial gestures. We present in this paper a new automatic approach for lip POI localization on a speaker's face based on the color information of mouth and a geometrical model of lips. This hybrid solution makes our method more tolerant to noise and artefacts in the image. The algorithm works in natural condition and experiments revealed that our lip POI localization approach for lip-reading purpose is possible using the proposed approach.

1. Introduction

The disadvantages and the social exclusion faced by dependant people are considerably increasing because of their physical and cognitive situations, the lack of supports, and of accessible environments. Communication technologies have not always been attentive to the needs of people who are deaf or severely hard of hearing. A case in point is the Human-Computer interaction based on automatic speech recognition (ASR), which for almost a century inadvertently excluded most of this population from its use. On the other hand we can imagine for this population a communication system based on lip movements. This system proposes a hand free terminal with integration of vocal technologies (using visual speech recognition approach) for the classification and the services of personalized repertory. It can also allow a protected and confidential network access and reinforce the robustness of the vocal communication of famous mobile telephony very degraded by the noise of coding and the environmental noise. Today, many works in the literature, from the oldest [1] and [4] until the most recent ones [2], [3] and [11] have proved that movements of the mouth can be used as one of the speech recognition channels. Recognizing the content of speech based on observing the speaker's lip movements is called 'lip-reading'. It requires converting the mouth movements to a reliable mathematical index for possible visual recognition. It is around this thematic that our ALiFE (Automatic Lip Feature Extraction) prototype appears. ALiFE allows visemes recognition from a video locution sequence, and then these visemes can correspond to any machine commands. More precisely, ALiFE prototype implements our approach which is composed of three steps: At first, it proceeds by localizing lips and some Point Of Interest (POI). The second step consists on tracking these POI throughout the speech sequence and extracting of precise and pertinent visual features from the speaker's lip region. At the end, the extracted features are used for visemes (visual phoneme) classification and recognition. In this paper we detail the first step of our ALiFE system.

2. Labial Segmentation Methods: an overview

Several research works stressed their objectives in the research on automatic and semi-automatic methods for the extraction of visual indices, necessary to recognize visual speech (lip-reading) [1], [6] and [10]. Two types of approaches have been used for lip-reading depending on the descriptors they use for the recognition of the viseme:

- The low-level approach (Image-based approaches) [5] and [6], use directly the mouth region. This approach supposes that the lip pixels have a different colour feature compared to the ones of skin pixels. Theoretically, the segmentation can therefore be done while identifying and separating the lips and skin classes. In practice, methods of this type allow rapid locations of the interest zones and

make some very simple measures of it (width and height of the lips, for example). However, they do not permit to carry out a precise detection of the lip edges.
- The high level approach (Model-based approaches) [7], [8], [9] and [2], which is directed by physical distance extraction, uses a model. For example, we can mention the active contour, which were widely used in lip segmentation. These approaches also exploit the pixel information of the image, but they integrate regularity constraints. The big deformability of these techniques allows them to be easily adapted to a variety of forms. This property is very interesting when it is a matter of segmenting objects whose form cannot be predicted in advance (sanguine vessels, clouds...), but it appears more as a handicap when the object structure is already known (mouth, face, hand...). In the following sections, we will present a new hybrid approach of lip extraction. Our approach exploits both the color and the geometric information of the lips to automatically localize a set of POI on the speaker's lips. These POI will be tracked throughout the speech sequence. This tracking will carry out visual information describing the lip movements among the locution video sequence. Finally, this visual information will be used to recognize the uttered viseme.

3. Lip POI Localization and tracking

In this phase, we start with the localization of the external contours of the lips on the first image of the video sequence. Then, we identify on these contours a set of POI that will be followed throughout the video locution sequence. Thus, there are two problems: (1) the lip and POI localization, and (2) POI tracking in video sequence. The details of our approach are presented in the following sections.

3.1. Lip and POI Localization

Our approach for lip POI localization is to proceed first by detecting a lip contour and secondly by using this contour to identify a set of POI. In the following sub-sections we detail in the first step our mouth corners localization technique, in order to assure a good initialization of the snake. Secondly we specify the lip contour extraction method based on a geometrical model. Finally POI will be localized on the extracted contour.

3.1.1. Initialization stage: Mouth corners localization

Various works have been made to extract facial regions and facial organs using colour information as clues especially for the localization of mouth knowing that colour of lips is different to skin colour. Among the colour systems used to localize the mouth position we quote the HSV colour system and the rg chromaticity diagram [12]. These colour systems are relatively widely used to separate the skin and the mouth map colour. Yasuyuki Nakata and Moritoshi Ando in [13] represent the colour distribution for each facial organ based on the relationship between RGB values normalized for brightness values in order to address changes in lighting. We have exploited this idea in our mouth localization approach and we apply a morphological operation to detect the position of the mouth gravity centre. The details of our approach are presented in the following sections. As mentioned above, our approaches begin by representing the image in ($R_nG_nB_n$) color system, defined by the following equation (1):

$$R_n = 255*\frac{R}{Y}, G_n = 255*\frac{G}{Y}, B_n = 255*\frac{B}{Y}. \quad (1)$$

With Y the intensity value.
After reducing the lighting effect by this color system conversion (Figure 2a) we apply a binary threshold based on the R_n value, knowing that the R_n is the most dominant component in lip region. The results of binarization are showing in figure (3a). After the binarization step, we apply on the image an oilify filter (Figure 2c). This filter makes the image look like an oil painting and it works by replacing the pixel at (x,y) with the value that occurs most often in its region. This region is named structuring element (SE). Precisely, we proceed in this step by eliminating the false positive skin pixels which have a dominant R_n value. Thus, we use in this phase a diamond-shaped (SE) (Figure 1). The aims goal is to maintain on the final result, only lip pixels. The width (*w*) and the height (*h*) of the SE are set according to the focal camera distance. In our experiments, we have fixed these measures (*w*) and (*h*) respectively to 30 and 10 pixels.

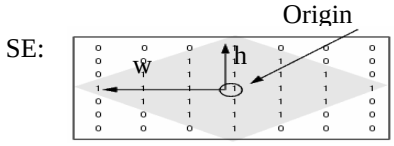

Figure 1 : diamond-shaped structuring element (SE).

Finally, we calculate the gravity center of the lip pixels; it represents the mouth center (Figure 2d). We remark on this first mouth localisation step that the final result is very sensitive to the noise which can be caused by the red component dominance in some skin pixels other than lip pixels. Thus, the centre of the mouth which has been detected is not rather precise, so, it will be considered in this second step of our mouth corner localization process as the effective centre of the Mouth Region (MR) and not as the centre of the mouth (Figure 3). Knowing that the corners and the interior of mouth constitute the darkest zone in MR, we use in this step the saturation component from the original image in order to localize the mouth corners. Precisely, we proceed by the projection of the

pixel saturation values from the MR on the vertical axis. This projection allows the detection of the darkest axis (D_KA_x) in the mouth region (Figure 3).

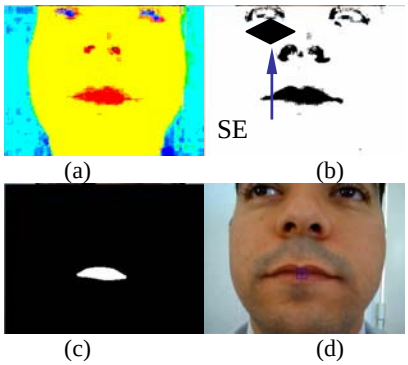

Figure 2 : First step mouth localization : (a) original image after the conversion in $R_nG_nB_n$ system (b) after the binarization step (c) Image after the oilify filter

In figure 4a we remark that the mouth corners are not on the detected D_KA_x, it is very normal according to the physiognomy of lips. So, we proceed by scanning different pixels along the D_KA_x to localize local maxima saturation values. Extremas of these detected local maxima pixels will be defined as the left and the right corners of the mouth ($X_{RCorner}$, $Y_{RCorner}$) and ($X_{LCorner}$, $Y_{LCorner}$). Figure 4 shows results of our corners localization method. Finally, the detected corners will be the basis of the lip and POI localization detailed in the next section.

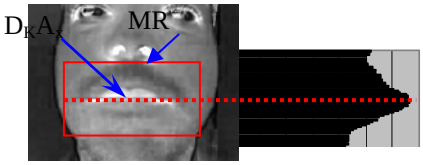

Figure 3 : Second step mouth corners localization: projection of the saturation values in the mouth region (MR) and the localization of the Darkest Axis D_KA_x.

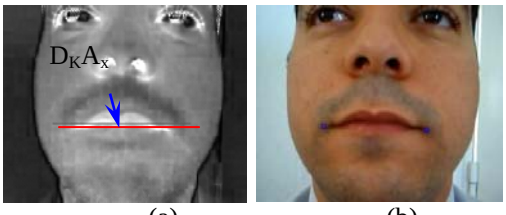

Figure 4 : (a) Scanning different pixels behind the axis (b) projection of local maxima and detection of the right and the left corner.

3.1.2. POI Localization based on geometrical model

In order to locolaize the different POI, we process by the extraction of the edge of the mouth image; the objective is the localisation of the lip contour. The use of edge-based approach for lip extraction is problematic since the edge maps obtained are usually very noisy with many false edges. Moreover, edges are often absent on the lip boundary or they are very low in magnitude and can often be overwhelmed by strong false edges not associated with lip boundary. In view of these difficulties, our lip extraction approach does not rely only on detecting edges on the lip boundary but we use also a geometric lip model. The model enables a priori knowledge about the expected lip shape to be incorporated. Since the allowable shape is pre-constrained in a geometric model, such a model is more robust compared to the unconstrained shape model such as snake.

In what follows, we will elaborate our lip model derived from the resolution of the quadratic equations given by [14].

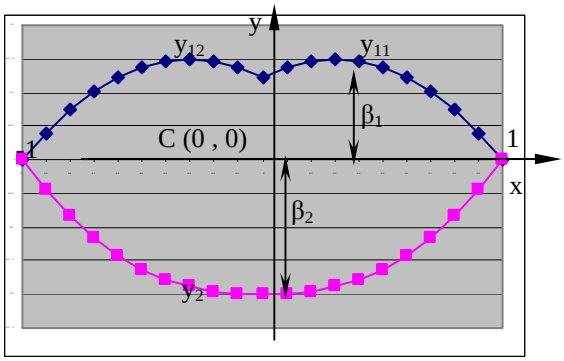

Figure 5 : The geometric model

The geometric model as shown in Figure 5 allows the description of the lip shape by a small set of parameters with clear physical interpretation of each parameter. These parameters as well as the geometric lip model are described by the following equations.

$$y_{11} = -((-\beta_1 + 2\alpha_1\beta_1 + x^2\beta_1 + 2x\alpha_1\beta_1) / (-2\alpha_1 + \alpha_1^2 + 2x\beta_1\varepsilon_1 + 2\beta_1\alpha_1\varepsilon_1 + 1)) \quad (2)$$

$$y_{12} = ((\beta_1 + 2\alpha_1\beta_1 + x^2\beta_1 + 2x\alpha_1\beta_1) / (-2\alpha_1 + \alpha_1^2 + 2x\beta_1\varepsilon_1 - 2\beta_1\alpha_1\varepsilon_1 + 1)) \quad (3)$$

$$y_2 = ((\beta_2 - \left(x^2\right)^{\alpha_2} * \beta_2)/(2 * \alpha_2 * \beta_2 * \varepsilon_2 - 1)) \quad (4)$$

The equation (2) and (3) describe respectively the right and left higher sub-model (y_{11} and y_{12}) and the equation (4) describe the lower sub-model (y_2). For $x \in [-1; 1]$ with the origin at C (0; 0). The physical interpretations of

β_1 and β_2 are shown in Figure 5. The parameters α_1 and α_2 describes the skewness of the lip shape and it allows fitting lip with different degree of curvature. Although a speaker's lip may not be symmetric in general, the parameters ε_1 and ε_2 describes the deviation of the curve and it allows capturing some of the asymmetry.

With such parameters we cover the totality of lip shape that can be occurring. Figure 6 shows different lip model occurrences on the speaker's image. Experimentally, we have fixed an interval of variation for each parameters $\beta 1$, $\beta 2$, α_1, α_2, ε_1 and ε_2. At all we have used 1800 models for the lower lip and 756 for the upper one (Figure 6). We implement this geometrical model on the speaker's image by replacing x and y respectively with (C_x + x**Unit*) and (C_y + y**Unit*). Where, (C_x, C_y) is the Cartesian coordinate of the middle point of the left and the right corners of the mouth ($X_{RCorner}$, $Y_{RCorner}$) and ($X_{LCorner}$, $Y_{LCorner}$) and *Unit* is the half distance between these two corners. In our implementation the problem of inclined lip is resolved by the use of fixed camera on the speaker face. It guarantees a fixed and reasonable disinclined plan of the lip's region.

Figure 6 : Different lip model occurrences on the speaker's image

The next process aims to find the optimum lip model that maximise the extern energy of the model given the lip image of the speaker. This energy is based on the gradient of the image.

(a) (b)

Figure 7 : (a) Original image (b) gradient image

So, we calculate separately for every upper and lower lip model (defined respectively by parameters β_1, α_1, ε_1 and β_2, α_2, ε_2) its E_{ext} (L_{Mod}) according to the following equation:

$$E_{ext}(L_{Mod}) = \sum_{s=1}^{n} \left| \nabla I(s) \right| \qquad (5)$$

With *n* number of points in the lip model L_{Mod}.

Finally, we vote for models (upper and lower lip models) which have the higher E_{ext} (L_{Mod}). Figure 8a illustrate the selected models for the lower and higher lip.

Once the external contours of the lips are extracted, we proceed to the detection and the initialization of the different POI. Here we intend to employ a projection technique (horizontal and vertical) of the various points of the snake, to detect different POI. More precisely, the maximum projection on the horizontal axis indicates the position of two corners of the lips and the maximum's projection on the vertical axis indicates the position of the lower lip and the Cupidon bow. Figure 8b show this localization process.

(a) (b)

Figure 8 : (a) Selected model (b) Final result and POI initialization

3.2. Lip Tracking

The problem of POI tracking (in our context each POI is defined by a block formed by its w*w neighbour pixels) is to detect these POI on the successive images of the video sequence. This problem is to look for the block (j) on the image (i) which has the maximum of similarity with the block (j) detected on the image (i-1) knowing that i is the number of image in the video sequence and j is the number of block which defines the different POI. Several algorithms and measurements of similarity were presented in the literature to deal with the problem of pattern tracking. However, we notice that there are some difficulties to adapt these algorithms to our problems for the reason that the movements of the lips are very complex [15]. Our approach of POI tracking is an alternative of the Template Matching technique exploiting the spatial-temporal indices of the video. The principle of this approach consists in seeking in a gray level image $I_i(x, y)$ the most similar block to the block pattern forming a point of interest (POI) defined in section 3.2. Our algorithm of tracking is based on two principle steps: in the first step POI tracking is done in the different directions of the Freeman coding to localize the candidate points describing the potential POI movements. In the second steps a vote technique will be used to identify among all the candidate points, the one that most much the origin POI. The details of our spatial-temporal voting approach of POI tracking are presented in [15].

4. Experimental Results

In this section we present the experimentation results for the evaluation of our ALiFE system for the visual speech recognition. The proposed lip contour extraction algorithm was tested with success on a large number of speakers. Figure 9 shows some lip localization results

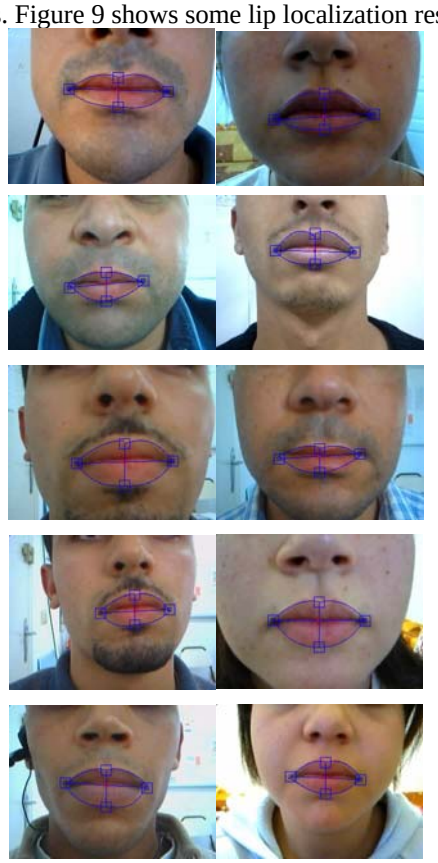

Figure 9. Experimental results of the lip localization process.

We perform our visual speech recognition system using our own audiovisual database. The database includes ten test subjects (three females, seven males) speaking isolated visemes repeated ten times. In our experiment, we use the data set for eight French visemes. We conducted tests for only '*speaker dependent*' using the six visual features and the classification system described in [16] and [17]. The recognition rate of each viseme as well as a matrix of confusion between these visemes will be shown. The test was set up by using a leave-one-out procedure, i.e., for each person, five repetitions were used for training and five for testing. This was repeated ten times for each speaker in our database. The recognition rate was averaged over the ten tests and again over all ten speakers. The experimental results are presented in Table 1. In these results we notice that we reach a good performance with K=8 and the weighted features method. But, we also remark that the recognition rate varies considerably for different words (like viseme /ba/, 71.43%).

Input	Recognition Rate
/i/	100.00%
/o/	100.00%
/ba/	71.43%
/fi/	100.00%
/cha/	100.00%
/la/	100.00%
/ta/	100.00%
/so/	85.71%
Recognition Rate	94.64%

Table 1. Recognition rate of French visemes

It can also be seen from figure 10 that the poor recognition rate for viseme /ba/ (71.43%) is due to the big confusion with viseme /ta/. So, we should think in future work to resolve this viseme confusion.

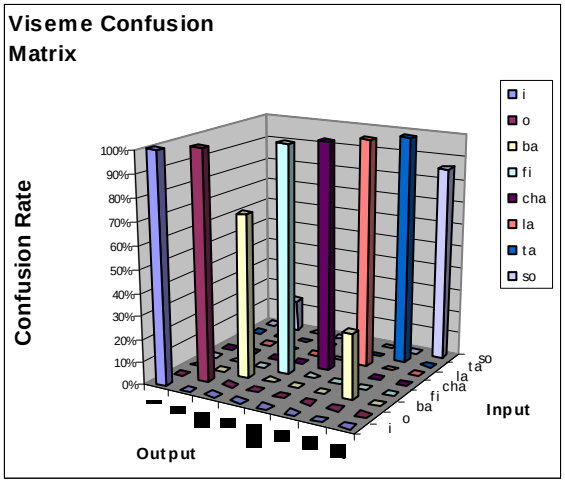

Figure 10. Experimental results of viseme confusion matrix

5. Conclusion and future work

Many works in the literature, from the oldest [1] until the most recent ones [3], proved the efficiency of the visual speech-recognition system, particularly in noisy audio conditions. Our research tasks relate to the use of visual information for the automatic speech recognition. The final objective is to develop a simple human computer interaction system based on lip-reading. This system

allows depending people to order his machine with easy lip movement or by simple viseme pronunciation. The major difficulty of the lip-reading system is the extraction of the visual speech descriptors. In fact, to ensure this task it is necessary to carry out an automatic tracking of the labial gestures. The lip localization and tracking constitutes in itself an important difficulty. This complexity consists in the capacity to treat the immense variability of the lip movement for the same speaker and the various lip configurations between different speakers.
In this paper, we have presented a new approach for automatic lip POI localization. Our approach uses both the color information of mouth and a geometrical model for lips. This hybrid solution makes our method more tolerant to noise and artefacts in the image. The algorithm has been tested with success in our ALiFE system of visual speech recognition. As a perspective to this work, we propose to develop a lip-board system that can be used in Human-Computer interaction application based on automatic visual speech recognition (AVSR), but we must in this case resolve the confusion problem between some visemes.

6. References

[1] Petajan, E. D., Bischoff, B., Bodoff, D., and Brooke, N. M., "*An improved automatic lipreading system to enhance* speech *recognition*," *CHI 88*, pp. 19-25, 1988.

[2] Philippe Daubias, Modèles a posteriori de la forme et de l'apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle. Thèse à l'Université de Maine France 05-12-2002.

[3] Roland Goecke, A Stereo Vision Lip Tracking Algorithm and Subsequent Statistical Analyses of the Audio-Video Correlation in Australian English. Thesis Research School of Information Sciences and Engineering. *The Australian National University Canberra, Australia*, January 2004.

[4] McGurck et John Mcdonald. Hearing lips and seeing voice. Nature, 264 : 746-748, Decb 1976.

[5] Iain Matthews, J. Andrew Bangham, and Stephen J. Cox. Audiovisual speech recognition using multiscale nonlinear image decomposition. *Proc . 4th ICSLP, volume1, page 38-41*, Philadelphia, PA, USA, Octob 1996.

[6] Uwe Meier, Rainer Stiefelhagen, Jie Yang et Alex Waibe. Towards unrestricted lip reading. *Proc 2nd International conference on multimodal Interfaces (ICMI)*, Hong-kong, Jan 1999.

[7] Prasad, K., Stork, D., and Wolff, G., "Preprocessing video images for neural learning of lipreading," *Technical Report* CRC-*TR-9326, Ricoh California Research Center*, September 1993.

[8] Rao, R., and Mersereau, R., "On merging hidden Markov models with deformable templates," *ICIP 95, Washington D.C.*, 1995.

[9] Patrice Delmas, Extraction des contours des lèvres d'un visage parlant par contours actif (Application à la communication multimodale). *Thèse à l'Institut National de polytechnique de Grenoble*, 12-04-2000.

[10] Gerasimos Potamianos, Hans Peter Graft et eric Gosatto. An Image transform approach For HM based automatic lipreading. *Proc, ICIP, Volume III, pages 173-177, Chicago, IL, USA* Octb 1998.

[11] N. Eveno, A. Caplier, and P-Y Coulon, "Accurate and Quasi-Automatic Lip Tracking" , *IEEE Transaction on circuits and video technology*, Mai 2004.

[12] Miyawaki T, Ishihashi I, Kishino F. Region separation in color images using color information. Tech Rep IEICE 1989;IE89-50.

[13] Nakata Y, Ando M. "Lipreading Method Using Color Extraction Method and Eigenspace Technique", Systems and Computers in Japan, Vol. 35, No. 3, 2004

[14] Alan Wee-Chung Liewa, Shu Hung Leungb, Wing Hong Laua, "Lip contour extraction from color images using a deformable model", The Journal Of Pattern Recognition Society, Nov 2002.

[15] S. Werda, W. Mahdi and A. Benhamadou, "A Spatial-Temporal technique of Viseme Extraction: Application in Speech Recognition ", *SITIS 05, IEEE,*

[16] S. Werda, W. Mahdi, M. Tmar and A. Benhamadou, "ALiFE: Automatic Lip Feature Extraction: A New Approach for Speech Recognition Application ", *the 2nd IEEE* International *Conference on Information & Communication Technologies: from Theory to Applications - ICTTA'06 - Damascus, Syria*. 2006.

[17] S. Werda, W. Mahdi, and A. Benhamadou, "LipLocalization and Viseme Classification for Visual Speech Recognition", International Journal of Computing & Information Sciences. Vol.4, No.1, October 2006.