AUTOMATIC LIP LOCALIZATION AND FEATURE EXTRACTION FOR LIP-READING

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

Keywords: Visual information, Lip-reading system, Human-Machine interaction, Spatial-temporal tracking.

Abstract: In recent year, lip-reading systems have received a great attention, since it plays an important role in human communication with computer especially for hearing impaired or elderly people. The need for an automatic lip-reading system is ever increasing. Today, extraction and reliable analysis of facial movements make up an important part in many multimedia systems such as videoconference, low communication systems, lip-reading systems. We can imagine, for example, a dependent person ordering a machine with an easy lip movement or by a simple syllable pronunciation. We present in this paper a new approach for lip localization and feature extraction in a speaker's face. The extracted visual information is then classified in order to recognize the uttered viseme (visual phoneme). To check our system performance we have developed our Automatic Lip Feature Extraction prototype (ALiFE). Experiments revealed that our system recognizes 70.95 % of French digits uttered under natural conditions.

1 INTRODUCTION

Today, extraction and reliable analysis of facial movements make up an important part in many multimedia systems such as videoconference, low communication systems and even for biometric system especially in noisy environments. In this context, many works in the literature, from the oldest (Petajan et al., 1988) and (McGurck and Mcdonald, 1976) until the most recent ones (Daubias, 2002) and (Goecke, 2004) 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 visual speech recognition from a video locution sequence. More precisely, it implements our approach which is composed of four 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. Extraction of precise and pertinent visual features from the speaker's lip region will be then achieved in the third step. At the end, the extracted features are used for visemes (visual phoneme) classification and recognition. Our ALiFE approach presented in this paper covers the totality of the visual speech recognition steps shown in figure1.

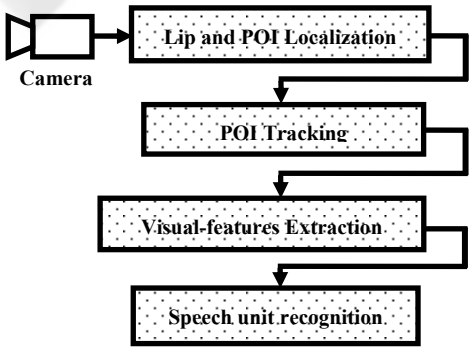


Figure 1: Overview of the complete ALiFE System for visual speech recognition.

In section (2) we present an overview on labial segmentation methods proposed in the literature. Section (3) details out our lip localization and lip tracking methods. In section (4), we present the different features which will be used for the recognition. In section (5), we evaluate our ALiFE prototype for the visual recognition of French digits.

Werda S., Mahdi W. and Ben Hamadou A. (2007).
AUTOMATIC LIP LOCALIZATION AND FEATURE EXTRACTION FOR LIP-READING.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 268-275

Rates of visual digit recognition as well as a matrix of confusion between these digits will be shown.

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 (Petajan et al., 1988), (Meier et al., 1999) and (Potamianos et al., 1998). Two types of approaches have been used for lip-reading:

- The low-level approach (Matthews et al., 1996) and (Meier et al., 1999), use directly the image of 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 (Prasad et al., 1993), (Rao and Mersereau, 1995), (Delmas, 2000) and (Daubias, 2002), 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...). According to speech specialists (Daubias, 2002), the pertinent features of verbal communication expression are: the heights, widths and inter-labial surface. From this interpretation we notice that it will be judicious to opt for an extraction method of these features based on the detection and the tracking of some "Points Of Interest" (POI) sufficient to characterize labial movements. Therefore, the problem of labial segmentation is to detect some POI on the lips and to track them throughout the speech sequence.

In the following sections, we will present a new hybrid approach of lip feature extraction. Our approach applies in the first stage the active contour method to automatically localize the lip feature points in the speaker's face. In the second stage, we propose a spatial-temporal tracking method of these points. This POI tracking will carry out visual information describing the lip movements among the locution video sequence. Finally, this visual information will be used to classify and recognize the uttered viseme.

3 ALIFE: 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.

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. One of the most efficient solutions to detect lip contour in the lip region, is the active contour techniques, commonly named "Snakes" (Eveno, 2003), and (Eveno et al., 2004). This method meets a lot of successes thanks to its capacity to mix the two classic stages of detection of contours (extraction and chaining). On the other hand, snake method imposes a prior knowledge of the mouth position. This constraint guarantees a good convergence of the final result of the snake. In fact, we proceed in the first step of our lip POI localization by detecting the mouth corners. These corners will indicate the position of the initialization of our snake.

3.1.1 Initialization Stage: Mouth Corners Localization

Mouth is the part of the lips visible on the human face. Various works have been made to extract facial regions and facial organs using color information as clues especially for the localization of mouth knowing that color of lips is different to skin color. Among the color systems used to localize the mouth position we quote the HSV color system and the rg chromaticity diagram (Miyawaki et al., 1989). These color systems are widely used to separate the skin and the mouth map color. Yasuyuki Nakata and Moritoshi Ando in (Nakata et al., 2004) represent the color 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 gravity centre of the mouth. 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 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 (2a).

After that, we apply on the image an oilify filter. This filter works by replacing the pixel at (x,y) with the value that occurs most often in its N x M region. The aim objective consists in eliminating the false positive skin pixels which have a dominant R_n value. Thus, we use in this phase a diamond-shaped structuring element (SE) (Figure 2) the aims goal is to maintain on the final result, only lip pixels. The width and the height of the SE are set according to the distance of the locator to the camera. In our experiments, we have fixed these measures respectively to 30 and 10 pixels.

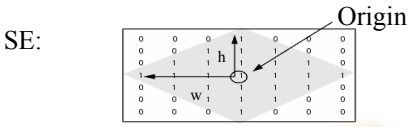


Figure 2 : diamond-shaped structuring element (SE).

Finally, we calculate the gravity center of the lip pixels, it represents the mouth center (Figure 3). 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 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 process 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 4).

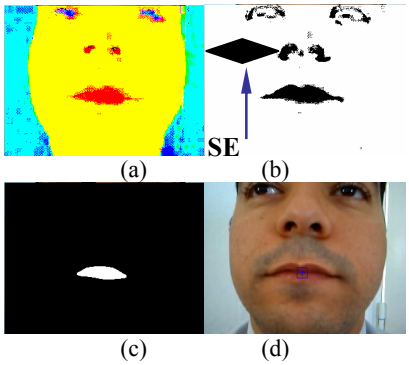


Figure 3: First step mouth localization : (a) original image after the conversion in $R_nG_nB_n$ system (b) after the binarization step (c) Image after the oilify filter.

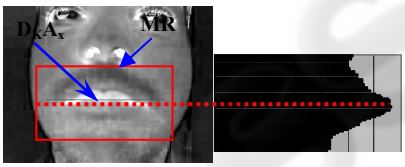


Figure 4: Second step mouth corners localization: projection of the saturation values in the mouth region (MR) and the localization of the Darkest Axis D_KA_x.

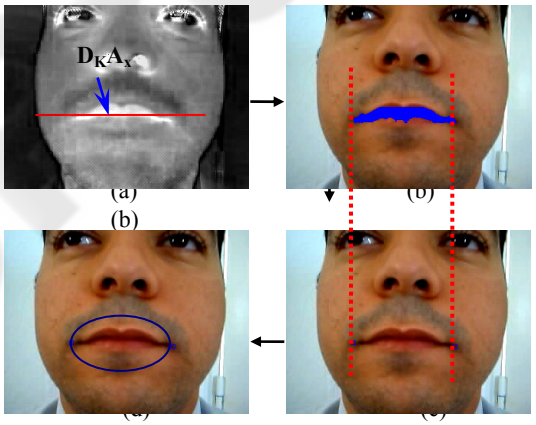


Figure 5: (a) Scanning different pixels behind the axis (b) results of scanning step (c) projection of local maxima on the horizontal axis and diction of the right and the left corner (d) initialization of the snake.

In figure 5a 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. Figure 5 shows results of our corners localization method. Finally, the detected corners will be the basis of the snake initialization (Figure 5c).

3.1.2 POI Localization Based on the Active Contour Method

The active contours (or snakes) are deformable curves evolving in order to minimize functional energy, which are associated to them (Delmas, 2000). They move within the image of an initial position toward a final configuration that depends on the influence of the various terms of energy.

The initialization of our snake is based on the mouth corners detected in section (3.1.1). Figure 5d shows the initialization of the active contour by an ellipse. We consider that our snake is composed of (*n) Vi* points with (i≤n), and that "*s"* is the parameter of spatial evolution in the contours image, for example the curvilinear abscissa.

- The internal Energy: E_{int} is going to depend only on the shape of the snake. It is a regularity constraint curve. We calculate it according to Equation 2.

$$E_{int} = (a(s)*|V'(s)| + (b(s)*|V''(s)|) \quad (2)$$

Where a and b are respectively the weights of the first and second derivative V " and V'. We will adjust a and b to find a flexible contour.

- A potential energy imposed by the image: E_{ext} is characterized by a strong gradient depicted by Equation 3.

$$E_{ext} = -|\nabla I(x,y)|^2 \quad (3)$$

- The constraint energy: E_{cont} is often defined by the user, according to the specificities of the problem. One of the cores of our contribution is the definition of E_{cont}. For us, E_{cont} aims at pushing the evolution of the snake toward the gravity centre G (x_g,y_g) of the active contour. It represents the Euclidian distance between G and V_i computed as follows:

$$E_{cont} = \sqrt{((x_s - x_g)^2 + (y_s - y_g)^2)} \quad (4)$$

With (x_s,y_s) and (x_g,y_g) the respective Cartesian coordinates of snake's points (s) and gravity center of the snake (G).

The principal goal of this energy is to ensure the evolution of the snake in the picture zones having weak gradient values. Therefore the total energy of the snake E_{tot} can be computed as follows:

In our implementation our snake is composed of two ellipse halves. So we define for the snake progression two E_{tot}, one for the upper and one for the lower part of the snake. The main idea of this constraint is to allow a more reasonable weights affectation to different terms of energy.

For example, knowing that the upper lip has a strong gradient value, and it isn't the case for the lower lip, the weight (λ_1) of the E_{cont} with the lower part of the snake will be higher compared to that (λ_2)

$$E_{tot}(V_{i-1}, V_i, V_{i+1}) = \sum_{i=1\to n} (Ei_{tot}) = \sum_{i=1\to n} \alpha * E_{int}(V_{i-1}, V_i, V_{i+1}) + \beta * E_{ext}(V_i) + \lambda * E_{cont}(V_i) \quad (5)$$

of the upper part of the snake. This constraint guarantees a good convergence of the final snake, especially in regions having weak gradient values. After the definition of active contour energies, the snake is going to evolve progressively in order to minimize its total energy E_{tot}. In order to maintain the initial form of our snake we interpolate the snake points to two ellipse halves, one for the upper lip and one for the lower lip. The snake progression will be stopped, when E_{tot} reaches its minimal value (Figure 6).

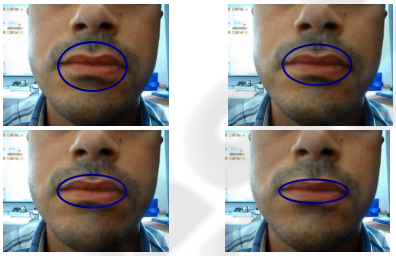


Figure 6: Snake evolution according to the energy minimisation principle (The snake progression will be stopped when the E_{tot} reaches the minimum value).

Once the external contours of the lips are extracted we employ a horizontal and vertical projection of the snake points, to detect different POI. Figure 7 illustrates this localization process.

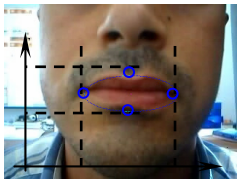


Figure 7: Points of interest detection by the projection of the final contour on horizontal and vertical axis (H and V).

3.2 Lip Tracking

The problem of POI tracking (in our context each POI is defined by a block of size w*w 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. Our approach of POI tracking is an alternative of the Template Matching technique exploiting the spatial-temporal indices of the video. The originality of our technique of labial-movement tracking lies in the case of being limited to a set of POI. The details of our spatial-temporal voting approach of POI tracking are presented in (Werda et al., 2005).

4 ALIFE: LIP FEATURE EXTRACTION AND CLASSIFICATION

In this section we present the different visual descriptors which we use for the characterization of the labial movements.

4.1 Lip Features Extraction

In this section we describe our hybrid features. We can classify these descriptors in two categories: the low-level feature using directly the image of the mouth region and high level feature which is directed by physical distance extraction. The extraction of these descriptors will be based on the tracking of POI already presented in section 3.

4.1.2 High Level Features

In this section we detail the intelligibility of the different high level features.

- Vertical Distance (Upper lip / lower lip: DV'): The variation of the vertical distance from the upper and lower lip gives a clear idea on the opening degree of the mouth during the syllabic sequence. This measure is very significant for the recognition of the syllables containing the vowels which open the mouth for example /ba/.

- Vertical Distance (Corner axis / lower lip: DV''): The variation of the second vertical distance between the lower lip and the horizontal axis (formed by the left and right corners of the lip), is a very important parameter especially for the recognition of labial-dental visemes. Precisely, one of the labial-dental visemes characteristic is that the lower lip is in contact with the incisive tooth (like /*fa*/) so this feature (DV'') perfectly describe the position of the lower lip.

- Horizontal Distance (DH): The variation of the Horizontal distance between the right and the left lip corners describes the stretching intensity of the lips during the locution sequence. This measure is very significant for the recognition of the visemes containing vowels which stretch the mouth for example /bi/.

- Opening Degree (OD): In addition to the variation of the vertical and horizontal distances that give a clear idea on the opening level of the mouth during the syllabic sequence we calculate the angle (γ). This measure (γ) characterizes the Opening Degree (OD) of the mouth according to the position of the lower and the upper lip and the right or the left corners. This variation is very high with vowels which open the mouth (Figure 8).

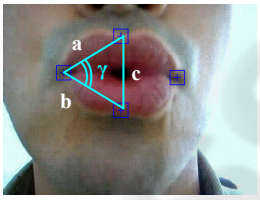


Figure 8: Mouth Opening Degree parameters (OD).

The angle (γ) will be calculated according to the following equation:

$$\gamma = \arccos \frac{a^2 + b^2 - c^2}{2ab} \quad (6)$$

With *a*, *b* and *c* are the different distances between two pixels $P_1(x_1,y_1)$ and $P_2(x_2,y_2)$. For example, the distance *a* is computed as follows:

$$a = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \quad (7)$$

4.1.3 Low Level Features:

The extraction of the low level parameters makes it possible to take into account many characteristics such as the appearance of the tooth or a particular mouth area.

- Dark Area (DA):
Speech is the result of vocal combinations which have a symbolic value that constitutes a language. The speech happens after a stop in the expiry. This is a compression of the intra-thoracic air with closing of the glottis, then opening of the glottis and emission of the air, thus we hear the voice. The air expelled by the lungs crosses the larynx where it is put in vibration by successive opening/closing of the vocal cords or, more probably, by the undulation of the mucous membrane which covers them. Precisely, according to the configuration that the mouth organs can take, the air ejected by the lungs until outside, will pass through these organs which

will produce the speech. So we had the idea to develop this measure, named the dark area (DA). The dark area thus will define the inter-labial surface from which the air will be ejected towards the outside of the mouth. We will see in our experiments that this descriptor is very relevant and constitutes discriminating criteria for the configuration of several visemes like /bou/.

To extract the dark pixels which are inside the mouth, we will try to find these pixels in the region of interest (ROI) described by a polygonal form. This region is formed by the four POI. The main problem is to separate between the dark and non-dark pixels. Here is a question of finding a method which can operate at various conditions of elocution sequence acquisition, different configurations and colours of the vermilion (which is not regular for all speakers). With this intention we propose an extraction of dark areas method with an adaptive threshold (S_{dark}).

$$S_{dark} = \alpha \times \frac{\sum_{i=1}^{n} I(x_i, y_i)}{n} \quad (8)$$

With n the number of pixel within the ROI.

α is a threshold fixed at 0.3 according to experimental results' that we carried out on our audio visual corpus. The application of this dark area detection reveals that we have some false DA detection (Figure 10b). Generally, knowing the physiognomy of the head we think that it is very natural that this problem occurs. Both the physiognomy and the light condition will generate an important shadow effect in some regions within the ROI. This shadow will affect the efficiency of our adaptive threshold (S_{dark}), since the (S_{dark}) is calculated on the whole ROI. To resolve this problem, we apply a spatial adaptive threshold (SAT_{dark}). Precisely, the idea consists to dividing the ROI to a three sub-regions (Figure 9).

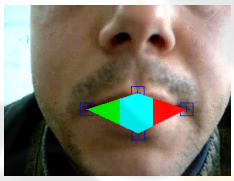
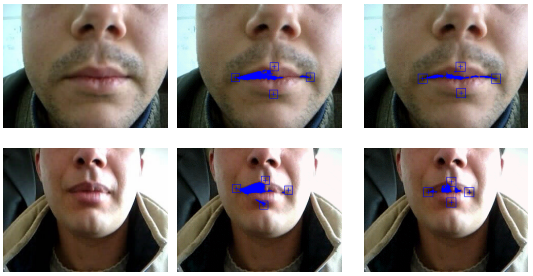


Figure 9: Division of the ROI in three sub-regions.

The detection of dark pixels is made in such a way that we calculate for every sub-region an (SAT_{dark}). This improvement largely reduces the shadow effect in the image. The result of our dark area detection approach is shown in figure 10.

In fact, the number of dark pixels is not so discriminating between the various configurations of each visemes.



(a) (b) (c)

Figure 10: Result of the dark area detection with (SAT_{dark}) improvement. (a) Original image, (b) Dark area detection with (S_{dark}), (c) Dark area detection with (SAT_{dark}).

The spatial position of these pixels inside the ROI is more interesting and more relevant. To develop this criterion (spatial position) we proceed by a weighing of the dark pixels on their position compared to the ROI midpoint. The values of the dark area feature (V [DA]) will be calculated in the following equation:

$$V\,[DA] = \sqrt{(X - X_c)^2 + (Y - Y_c)^2} \quad (9)$$

With Xc and Yc Cartesian coordinates of the ROI gravity centre.

With such approach, we exploit the spatial distribution of the dark pixels inside the ROI.

- Teeth Area (TA):

The descriptor "Teeth Area" characterizes the visibility of tooth during the locution sequence. For example, the inter-dental phonemes /T/ and /D/ can be satisfactorily produced by either protruding the tongue through the teeth, or placing the tongue behind the teeth of the upper jaw.

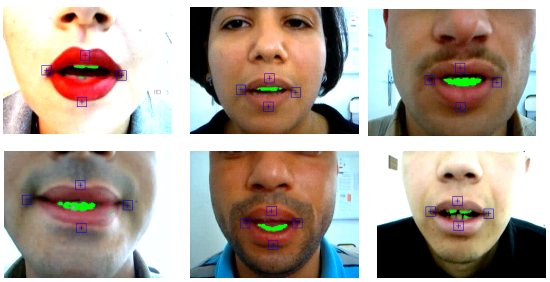


Figure 11: Result of the teeth area detection with various speakers under different lighting conditions.

However, teeth can be distinguished from other parts of the face by their characteristic low saturation. Saturation measures the white quantity in a colour. Then, the more the saturation is low (near

to 0) the more the colour is white or pastel. On the other hand higher saturation values (near to 1) indicate that the colour is pure. Therefore, Zhang in (Zhang et al., 2002) detects teeth by forming a bounding box around the inner mouth area and testing pixels for white tooth color: $S<S_0$, where S_0 is fixed to 0.35, and S is between 0 and 1. However, the visibility of the teeth was detected by a color search within the ROI. In our teeth area extraction process, we exploit this idea to find pixels which have low saturation values (Figure 11).

4.2 Lip Features Classification and Recognition: Application in French Digits

ALiFE prototype is evaluated with multiple speakers under natural conditions. We have created a specific audio-visual (AV) corpus constituted by the different French digits. The capture is done with one CCD camera; the resolution is 0.35 Mega of pixels and with 25 frames/s (fps). This cadence is wide enough to capture the major important lip movement. Our recognition system ALiFE, like the majority of recognition system, is composed by tow sub-systems, one for training and one for recognition. The training stage consists of building the recognition models for each viseme in our corpus. In the second stage, we classify the viseme descriptors by the comparison with the models built during the first stage. On the other hand, we note that the syllable sequence duration is not necessarily unvaried for all speakers. Moreover, we do not have the same mouth size for all people. So, to construct a robust recognition system we apply a spatial-temporal normalization on the distance variation curves. The detail of our own lip feature classification method is presented in (Werda et al., 2006) and (Werda et al., 2006).

5 EXPERIMENTAL RESULTS

In this section we present the experimentation results for the evaluation of our ALiFE system for the visual speech recognition. We perform our visual speech recognition system using our own audiovisual database. The database includes ten test subjects (three females, seven males) speaking isolated words repeated 10 times. In our experiment, we use the data set for ten French digits.

We conducted tests for only speaker dependent using the six visual parameters described in section 4. The rates of recognition of each viseme as well as a matrix of confusion between these viseme 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.

Table 1: Recognition rate of French digits.

Input	Recognition Rate
1	83.33%
2	83.33%
3	83.33%
4	66.67%
5	66.67%
6	66.67%
7	50.00%
8	83.33%
9	83.33%
0	42.86%
Recognition Rate	**70.95%**

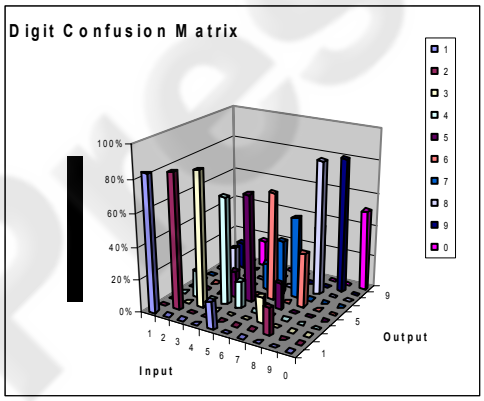


Figure 12: Experimental results of digits confusion matrix.

It can also be seen from figure 12 that the latter three words were mainly confused with another. The poor recognition rate for digit '*Zero*' (42.86%) is due to the specific syllabic composition of the digit '*Zero*'. This digit is composed of two visemes /ɛ/ and /*o*/. So, it is important to take in to consideration this constraint in future work.

6 CONCLUSION AND FUTURE WORK

Many works in the literature, from the oldest (Petajan et al., 1988) until the most recent ones (Goecke, 2004), 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 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 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 our ALiFE system of visual speech recognition. ALiFE is a system for the extraction of visual speech features and their modeling for visual speech recognition. The system includes three principle parts: lip localization and tracking, lip feature extraction, and the classification and recognition of the viseme. This system has been tested with success on our audio-visual corpus, for the tracking of characteristic points on lip contours and for the recognition of the viseme.

However, more work should be carried out to improve the efficacy of our lip-reading system. As a perspective of this work, we propose to add other consistent features. We also propose to enhance the recognition stage by the adequate definition of the feature coefficients for each viseme (use of the principal component analysis ACP). Finally, we plan to enlarge the content of our audio-visual corpus to cover other French language visemes and why not to discover other languages.

REFERENCES

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.

Daubias P., 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.

Goecke R., 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.

McGurck et Mcdonald J., Hearing lips and seeing voice. Nature, 264 : 746-748, Decb 1976.

Matthews I., 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.

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

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.

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

Delmas P., 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.

Potamianos G., 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.

Matthews I., J. Andrew Bangham, and Stephen J. Cox. A comparaison of active shape models and scale decomposition based features for visual speech recognition. *LNCS, 1407 514-528,* 1998.

Eveno N., "Segmentation des lèvres par un modèle déformable analytique", *Thèse de doctorat de l'INPG, Grenoble,* Novembre 2003.

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

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

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

Zhang X., Mersereau R., Clements M. and Broun C., Visual Speech feature extractionfor improved speech recognition. In Proc. ICASSP, Volume II, pages 1993-1996, Orlondo,FL, USA, May 13-17 2002.

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

Werda S., Mahdi W., Tmar M. and Benhamadou A., "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.

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