Eng. & Tech. Journal ,Vol.29, No.14 , 2011

Lipes Tracking Using Active Contour Method

Dr. Abdul Monem S. Rahma & Abdulhamza A. Abdul karim**

Received on: 6/2/2011
Accepted on: 8/9/2011

Abstract

Speech recognition based on visual information such as the lip shape and its movement is referred to as lip reading. The visual features are derived according to the frame rate of the video sequence. The proposed work adopted in this paper based upon the lower part of human face to extract the speaker sound relevant features accurately and robustly from the inner edge of lips, using biometric to verify a person's identity by drawing their relevant physiological or behavioral characteristics curves. Lips contain a large volume of unique features. The results are promising and offer a good reaction (even with reducing the number of tested frames). The recognition rate with only audio: 86% - 100%, with only visual: 73% - 100%, and with both (audio - visual) recognition rate is: 92% - 100%.

Keywords: Feature Extraction, visual features, Human Lip Tracking.

تعقب حركة الشفتين باستخدام طرق الكونتورات الفعاله

الخلاصة

تمييز الاصوات اعتمادا على المعلومات المرئيه مثال شكل الشفه وحركتها اثناء الكلام يعزى قراءة الشفه وتستمد الخواص المرئيه طبقا لمعدل اطر المقطع الفديوي. العمل المتبنى في هذا البحث ركز على الجزء الاسفل من الوجه البشري لانتزاع ميزات المتكلم الصحيحه ذات العلاقه بدقه وبشده من الحافه الداخليه للشفتين استخدم التقييس الاحيائي للتحقق من هوية الاشخاص وذلك برسم منحنيات الخواص الفيزيائيه او السلوكيه لذي العلاقه .

تحتوي الشفاه على حجم كبير من الخواص الفريده , النتائج واعده وعرضت رد فعل جيد (حتى بتقليل عدد الاطر المفحوصه). معدل التمييز باعتماد الصوت لوحده تراوحة بين 86% الى 100%, وباعتماد الصور لوحدها تراوحة بين 73% الى 100% اما باعتماد كلا العاملين (الصوت والصوره) فكانت النتائج تتراوح بين 92% الى 100%.

Introduction

Several researchers' works on the nature of visual information in human speech perception [1, 9] they have shown that the incorporation of visual information into acoustic speech recognizers improves recognition performance, especially in acoustically noisy environments. The visual signal is unaffected by the presence of background noise or cross-talk among speakers. Thus the promise of audio-visual speech recognition lies in its ability to extend computer speech recognition to adverse environments Such as offices, airports, train station, etc [7].

Such systems must be capable of tracking the lips (both inner and outer contour) [2], and reasoning about the presence/absence and position of the teeth and tongue on unconstrained speakers because the principal in tracking the inner mouth contour is the erratic appearance and disappearance

*Computer Science Department, University of Technology/ Baghdad
** Control and system Engineering Department, University of Technology/ Baghdad

https://doi.org/10.30684/etj.29.14.10
2412-0758/University of Technology-Iraq, Baghdad, Iraq

of the teeth. When the teeth are obscured by the lips, there is both an edge and intensity valley along the inner lip contour [3, 11], but when the teeth are visible; there are numerous edges inside the mouth which serve to distract the tracker as in figure (2).

In previous work using cosmetically assisted lips [4], it was demonstrated that visual information extracted from the outer lip contour could be used to provide robust recognition of speech in the presence of acoustic noise. In (Hidden markov model) based recognition, the inclusion of the visual signal tends to stabilize the Vertebra state alignment [5]. This is demonstrated that acoustic information alone is inadequate to accurately identify noisy speech. Biometric in general involves measuring unique biological characteristics for the purpose of comparing unknown samples against known samples, usually with the goal of confirming some one's identity. This technology has attracted a great deal of attention in many regions of the world because it has potential for the security industry as well as other areas of human effort.

Human Lip Tracking and Sound

Visual analysis system is used to track the position of the mouth through the sequence, and extract a meaningful parameter set for shape of the mouth. The parameter extraction may employ either a classification strategy where the input image is classified to one of several possible types, or measuring dimensions such as the width and height of the mouth [6]. Human speech is bimodal both in production and perception. Human speech is produced by the vibration in the vocal tract that are composed of articulator organs including the pharynx, the nasal cavity, the tongue, teeth, velum, and lips, together with the muscles that generate facial expressions; a speaker produces speech [7].

1. The two dimension outlines of the lips are parameterized by quadratic Spline which permits 5 parse representations of image data. Motion of the lips is represented by the x and y coordinates of B-Spline control point $(x_{(t)},y_{(t)})$,varying over time, a contour is then grown around the area identified as the inner mouth.. It is well known that human speech perception is enhanced by seeing the speakers face and lips even in normal hearing adults [2, 9] .To handle the variable frame length of the word sequence, by representing each visual feature using a B-Spline curve, thus transforming the discrete time measurement to the continuous domain [5].

2. Most of face recognition research is often based on static face image by assuming a neutral facial expression. However, the appearance of a face can change considerably during speech due to facial expressions [8]. It is also well – known that visual modality of speaker's mouth region provides additional speech information which can lead to improve speaker recognition and verification system performance. In general, visual features for automatic lip-reading can be grouped into three categories which are lip contour (shape) based features, pixel (appearance) based features and a combination of both.

For the lip contour based features, inner and outer lip contour are extracted for geometrics such as mouth, height and width are used in pixel based category, the entire image containing the speaker's mouth (Region of interest – ROI) is

considered as informative lip-reading [3].

In most automatic lip-reading system, the ROI is a square containing the image pixels of the speaker's mouth region. The ROI can also include larger parts of lower face, such as the jaw or even the entire face.

The proposed work based on lip and voice tracking

Start with building software system covers the needing functions like image algebra the arithmetic and logic operations, the spatial filters both mean filters and median filters, and the enhancement filters. Then the edge detection operators and all histogram modifications (stretch, shrink, and slide) are used. We have employed more than ten people with different ages, skin color, and different face shape.

Portrayed the face of each subject as he start saying the same sentence which is in the name of God the merciful, so we will have more than ten movies, to find out the active contour of each subject. Ulead-Video Studio is used to isolate the sounds from objects films so as to use it as external factor. The same montage programs applications are used in object segmentations and tracking in image sequences which are an important problem [6], it involves the isolation of a single object from the rest of the images that may include other objects and background. Meanly interested in the edges of the lip images, Edge detection is one of the fundamental operations in image processing; the edge of items in an image holds much of the information in the image. Figure (1) and Figure (2) shows the segmented image and its edge detected using Ulead-Video Studio, Figure (6) to Figure (11) shows the (off-line work, loading lip image, converting to gray level, edge detection, active contour, and feature detections) windows using proposed system software.

The number of masks used for edge detection is almost limited less. Many type of masks like Kirsch, Prewitt, and Soble are used.

The Sobol operator is selected to be the edge detection method used for finding edge points because of its clear wide edge for this reason, all the information of the edges will not be lost.

Derivatives the first-order derivative of choice in image processing is the gradient. The gradient of a 2 D function, f (x,y), is defined as the vector

$$\nabla F = \begin{vmatrix} g^x \\ g^y \end{vmatrix} = \begin{vmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{vmatrix} \qquad (1)$$

The magnitude of this vector is:

$$= \nabla f = mag(\nabla f) = [G_x^2 + G_y^2]^{1/2} = [(\frac{\partial f}{\partial t})^2 + (\frac{\partial f}{\partial y})^2]^{1/2} \qquad (2)$$

To simplify computation, this quantity is approximated sometimes by omitting the square – root.

$$\nabla f \approx G_x^2 + G_y^2 \qquad (3)$$

A fundamental property of the gradient vector is that it points in the direction of the maximum rate of change of f at coordinates (x,y). The angle at which this maximum of change occurs is

$$\alpha(x,y) = \tan^{-1}\left(\frac{Gy}{Gx}\right) \qquad (4)$$

One of the key issues is how to estimate the derivatives GX and G Y digitally.

Second – order derivatives in image processing are generally computed using the Laplacian of a 2-D function f (x,Y) which is formed from second – order derivatives as follows:

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \quad (5)$$

The digital image processing is used depends on the image needed and on what we need from the processing. The inner lip edge is detected using digital image processing Figure (2) shows detecting edges.

Active contour (a set of the coordinates of control points on the contour) is defined parametrically as:

V(s) = x(s), Y(s)

X(s) and y(s) are x, y coordinates past the

Contour(s) which is the normalized index of the control point [10].

After detecting the edge of inner mouth for each donor we have to extract their lip feature that means now we have to extract the active contour of each subject. Therefore B-Spline is used to perform their features.

Calculate the time, the number of frame, and number of tested frames for each subject. The subject take from 1.41 sec to 2.08 sec to say the same sentence since there is 25 frames for each sec so there are about 34 to 50 frames for each subject, after segmentation 9 to 13 frames are to be tested. The time taken by each subject to say that sentence is divided into 10 equal parts. Then by measure the dimension of the mouth in each part of time and for each frame i.e. measure the height of the mouth during moving time for each frame of that subject.

The characteristic curve is plotted for the tested frames of each subject. Table (1) shows the measuring distance (the height y and the width x) at different time of mouth movement for subject (5) lips. Figure (3) shows the curves of subject-5 frames.

Table (2) shows the same measuring distance of inner lips edge of subject-6 frames, Figure (4) shows the characteristic curve of subject-6 frames, and Table (3) shows the measuring distance of inner edge of subject-10 frames, and Figure (5) represented characteristic curve of subject-10 frames.

Conclusions

This paper presents technique to extract Information from digital sequence images of lips, and describes a new approach for utilizing active contours of lip - tracking based on Spline. Suitable geometric features are extracted from speaker's lip shapes, the work focuses on the lip shapes and movement. All subject said the same sentence, some of them like (subject 5&6) had the same time to say this sentence (1.66 sec) so both have the same numbers of tested frames but each one has its own characteristic as seen in Figure (3) and Figure (4) which gives different shape for theirs tested frame, that is mean even if the subject takes the same time to said the same sentence their characteristic is different, others had different time to say the same sentence, so has different numbers of tested frame and different characteristic curve too, Subject (10) had different time to complete the same sentence so had different number of tested frames and different characteristic curve as seen in Figure (5).

The result proves that each subject has his unique characteristic.

References

[1] M. Yazdi, M.Seyfi, A. Rafati, M. Asadi "Real-time Lip Contour Tracking for Audio-Visual Speech Recognition Applications" World

Academy of Science. Engineering and Technology 40 2008

[2] H.Mehrotra, G. Agrawal and M.C. Srivastava "Automatic Lip Contour Tracking and Visual Character Recognition for Computerized Lip reading" International Journal Electrical and Computer Engineering 4:1 2009.

[3] Z. Wu, J. Wu, and H. M. Meng "The use of Dynamic Deformable Templates for Lip Tracking in an Audio-Visual Corpus with Large Variations in head pose, face illumination and lip shapes" 978-1-4244-2942-4/081 $25.00 © 2008 IEEE.

[4] R. Kaucic, B. Dalton, and A.Blake "Real time Lip tracking for audio-visual speech recognition applications" In Proc. 4th European Conf. Computer Vision, PP, 376-387, Cambridge, England. Apr. 1996.

[5] R. Kaucic. "Lip Tracking for Audio-Visual Speech Recognition." PhD thesis, University of Oxford, 1997.

[6] S. Stillittano and A. Caplier "Inner Lip Segmentation by Combining Active Contours and Parametric Models" visa pp 2008-International Conference on Computer Vision Theory and Applications.

[7] T. Chen and R. Rao "Audio – Visual Interaction in Multimedia "Ken cooper / the image bank. 8755-3996/96/ $4.00© 1995IEEE.

[8] M. Hoch, P. C. Litwinowicz "A practical Solution for Tracking Edges in Image Sequences with Snakes" ©The Visual Computer, vol. no 12, no 2, 1996, PP 75-83.

[9] H. Shirgahi, S. Shamshirband, H. Motameni and P. Valipour "A new Approach for Detection by Movement of Lips Base on Image Processing and Fuzzy Decision" World Applied Sciences Journal 3(2):323-329, 2008 ISSN 1818-4952 ©IDOSI Publications, 2008.

[10] *A. Blake and M. I*sard "Active Contours" Springer 1997.

[11] C. Bouvier, P.Coulon, X. Maldague "Unsupervised Lips Segmentation Based on ROI optimization and Parametric Model" hal-00372142 version 1 – 31 Mar 2009.

Table (1) Measuring inner edge of subject 5 frames

Time Sec	Frame 1		Frame 2		Frame 3		Frame 4		Frame 5	
	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm
0.000	0.000	0.000	0.000	0.000	0.00	0.000	0.000	0.000	0.000	0.000
0.166	3.25	0.00	3.50	0.25	3.50	2.0	3.50	3.10	3.50	2.50
0.332	6.50	0.00	7.0	0.50	7.00	2.50	7.00	5.50	7.00	5.00
0.498	9.75	0.00	10.50	1.0	10.50	6.00	10.50	9.00	10.50	6.00
0.664	13.00	0.00	14.00	1.0	14.00	8.50	14.00	11.00	14.00	7.50
0.830	16.25	0.00	17.50	1.0	17.50	9.00	17.50	11.00	17.50	7.50
0.996	19.50	0.00	21.00	1.0	21.00	9.00	21.00	11.00	21.00	7.00
1.162	22.75	0.00	24.50	1.0	24.50	8.00	24.50	10.00	24.50	6.00
1.328	26.00	0.00	28.00	1.0	28.00	7.00	28.00	8.00	28.00	4.00
1.494	29.25	0.00	31.50	1.0	31.50	2.50	31.50	4.00	31.50	1.50
1.660	32.50	0.00	35.00	1.5	35.00	0.50	35.00	1.00	35.00	0.50
Sec	**Frame 6**		**Frame 7**		**Frame 8**		**Frame 9**		**Frame 10**	
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.166	3.50	0.000	3.250	2.50	3.20	3.50	3.00	0.000	3.70	0.000
0.332	7.00	0.000	7.00	4.00	6.40	5.00	6.00	4.00	7.40	0.00
0.498	10.50	0.000	10.50	7.00	9.60	6.00	9.00	6.00	11.10	0.00
0.664	14.00	0.000	14.00	9.00	12.80	8.00	12.00	7.00	14.80	0.00
0.830	17.50	0.000	17.50	10.00	16.00	8.00	15.00	6.0	18.50	0.00
0.996	21.00	0.000	21.00	10.50	19.20	7.50	18.00	9.0	22.20	0.00
1.162	24.00	0.000	24.00	9.50	22.40	7.50	21.00	10.0	25.90	0.00
1.328	28.00	0.000	28.00	6.50	25.60	7.00	24.00	6.50	29.60	0.00
1.494	31.50	0.000	31.50	2.00	28.80	5.50	27.00	4.00	33.30	0.00
1.660	35.00	0.000	35.00	0.50	32.00	0.30	30.00	0.50	37.00	0.00

Table (2) Measuring inner edge of subject6 frames

Time Sec	Frame 1		Frame 2		Frame 3		Frame 4		Frame 5	
	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.166	3.00	0.00	3.40	3.00	3.30	0.00	3.20	5.50	3.30	3.50
0.332	6.00	0.00	6.80	3.00	6.60	0.00	6.40	7.00	6.60	5.50
0.498	9.00	0.00	10.20	4.00	9.90	0.00	9.60	11.00	9.90	9.00
0.664	12.0	0.00	13.60	5.50	13.20	0.00	12.80	11.00	13.20	9.50
0.830	15.0	0.00	17.00	5.50	16.50	0.00	16.00	12.00	16.50	10.0
0.996	18.0	0.00	20.40	4.50	19.80	0.00	16.20	12.00	19.80	9.0
1.162	21.0	0.00	23.80	4.00	23.10	0.00	22.40	11.00	23.10	8.50
1.328	24.0	0.00	2720	3.50	26.40	0.00	25.60	10.00	26.40	8.50
1.494	27.0	0.00	31.60	2.00	29.70	0.00	28.80	9.00	29.70	6.00
1.660	30.0	0.00	34.0	0.20	33.0	0.00	32.00	1.5	33.00	1.50
Sec	Frame 6		Frame 7		Frame 8		Frame 9		Frame 10	
0.000	0.000	0.000	0.000	0.50	0.000	0.000	0.000	0.20	0.00	0.00
0.166	3.20	4.00	3.00	2.50	3.30	3.00	3.0	2.00	3.40	5.00
0.332	6.40	8.00	6.00	5.00	6.60	6.00	6.0	4.50	6.80	5.50
0.498	9.60	9.00	9.00	6.00	9.90	8.00	9.0	7.00	10.20	7.00
0.664	12.80	10.00	12.00	6.50	13.20	7.50	12.0	8.00	13.60	7.50
0.830	16.00	10.00	15.00	7.50	16.50	7.50	15.0	8.00	17.00	9.00
0.996	19.20	10.00	18.00	6.00	19.80	7.00	18.0	7.00	20.40	8.00
1.162	22.40	9.00	21.00	6.00	23.10	7.00	21.0	5.50	23.80	7.50
1.328	25.60	9.00	24.00	5.00	26.40	6.00	24.0	5.00	27.20	7.00
1.494	28.80	7.00	27.00	3.50	29.70	4.00	27.0	4.00	30.60	4.00
	32.00	1.5	30.00	0.50	33.0	0.50	30.0	1.00	34.0	1.00

Table (3) Measuring inner edge of subject 10 frames

Time Sec	Frame 1		Frame 2		Frame 3		Frame 4		Frame 5	
	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm	Xmm	Ymm
0.000	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.50	0.00	0.50
0.158	3.50	0.00	3.50	0.00	3.30	2.00	3.30	3.00	3.20	4.00
0.316	7.00	0.00	7.00	1.00	6.60	5.00	6.60	5.00	6.40	7.00
0.474	10.50	0.00	10.50	2.00	9.90	8.50	9.90	7.00	9.60	9.50
0.632	14.00	0.00	14.00	2.00	13.20	9.00	13.20	8.50	12.80	9.50
0.790	17.50	0.00	17.50	2.00	16.50	9.00	16.50	8.50	16.00	9.00
0.948	21.00	0.00	21.00	2.00	19.80	8.50	19.80	8.50	19.20	9.00
1.106	24.50	0.00	24.50	2.00	23.10	7.00	23.10	7.00	22.40	7.00
1.264	28.00	0.00	28.00	1.00	26.40	5.00	26.40	6.00	25.60	6.00
1422	31.50	0.00	31.50	0.00	29.70	2.00	29.70	3.00	28.80	3.00
1.580	35.00	0.00	35.00	0.00	33.0	1.00	33.0	0.50	32.0	0.50
Sec	Frame 6		Frame 7		Frame 8		Frame 9		Frame 10	
0.000	0.00	0.00	0.00	0.20	0.000	0.50	0.000	0.000	0.000	0.000
0.158	3.50	0.00	3.60	2.00	3.40	2.50	3.50	0.000	3.50	0.000
0.316	7.00	0.00	7.20	5.50	6.80	6.00	7.00	0.000	7.00	0.000
0.474	10.50	2.50	10.80	9.00	10.20	8.00	10.50	0.000	10.50	0.000
0.632	14.00	2.50	14.40	9.00	13.60	8.50	14.00	0.000	14.00	0.000
0.790	17.50	2.50	18.00	8.50	17.00	8.50	17.50	0.000	17.50	0.000
0.948	21.00	2.50	21.60	8.50	20.40	9.00	21.00	0.000	21.00	0.000
1.06	24.50	1.00	25.20	7.50	23.80	8.00	24.50	0.000	24.50	0.000
1.264	28.00	0.00	28.80	5.00	27.20	6.50	28.00	0.000	28.00	0.000

1422	31.50	0.00	32.40	1.50	30.60	3.00	31.50	0.000	31.50	0.000
1.580	35.0	0.00	36.0	0.20	34.0	0.5	35.0	0.000	35.0	0.000

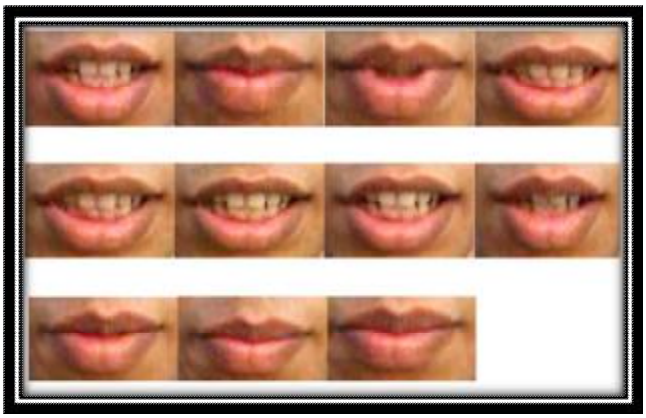

Figure (1) Segmented subject image (frames)

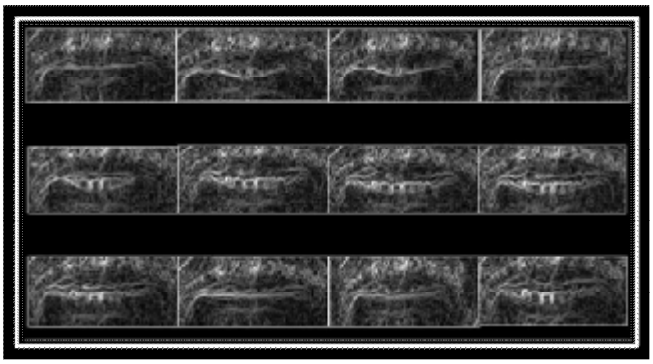

Figure (2) Inner lip edge detection

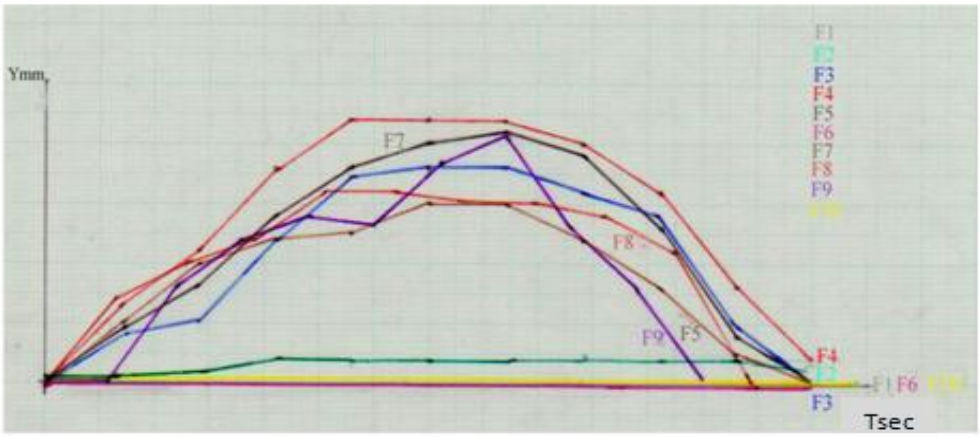

Figure (3) subject –5 frames curve

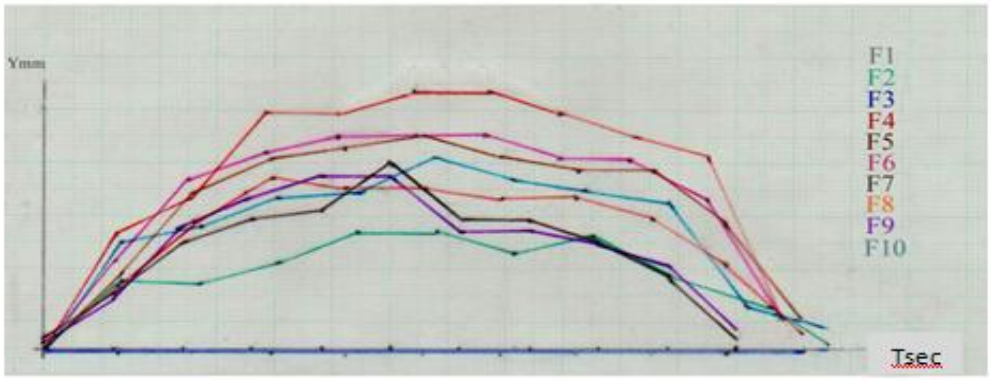

Figure (4) subject –6 frames curve

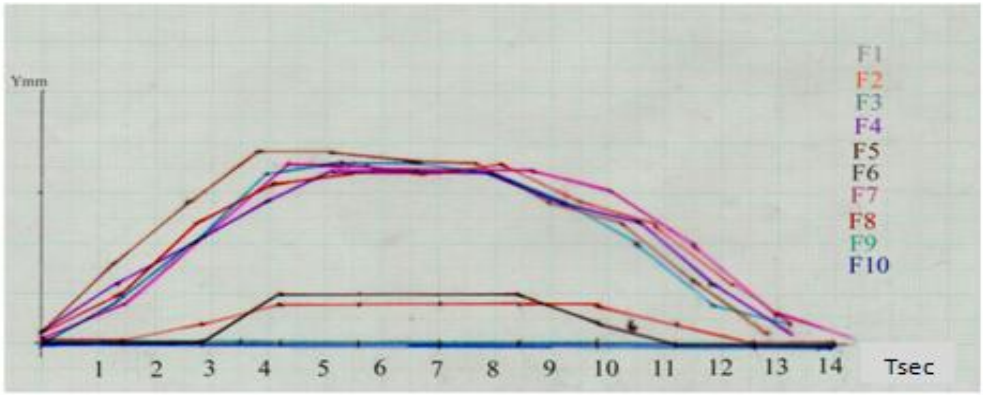

Figure (5) subject –10 frames curve

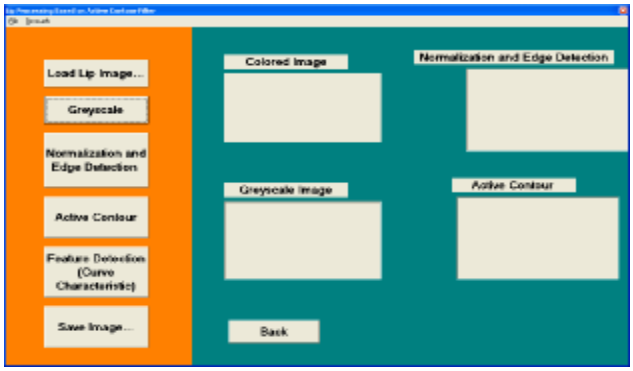

Figure (6) the Off-line Work Window

Figure (7) the Open Image Window

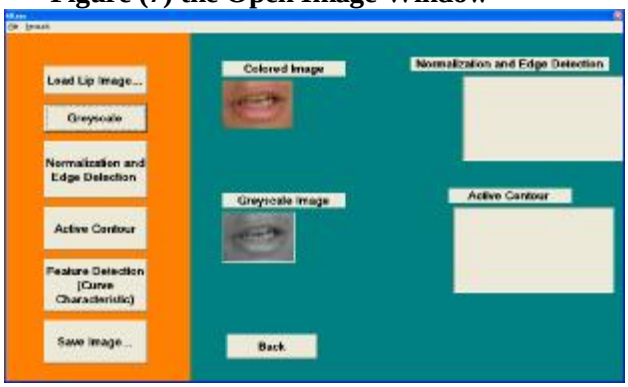

Figure (8) Convert the Color Image to Gray Level Window.

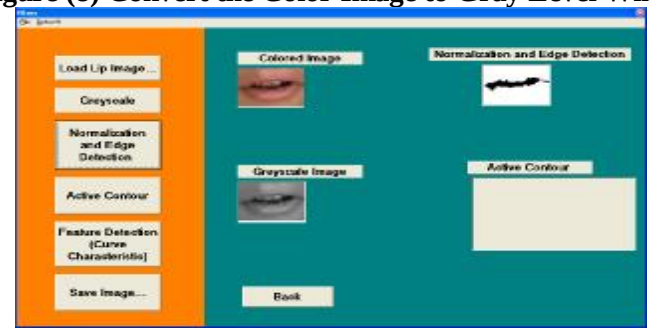

Figure (9) Normalization and Edge Detection Window.

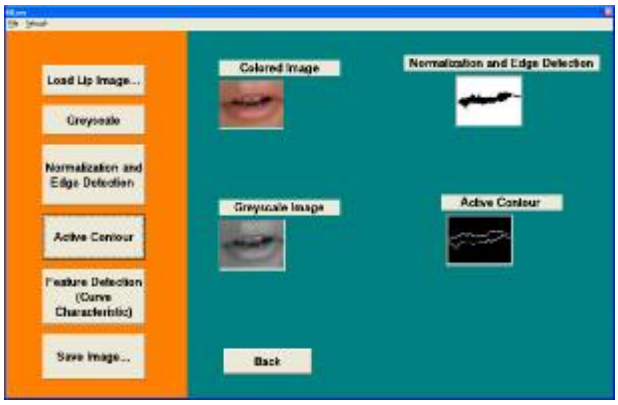

Figure (10) Active Contour Window.

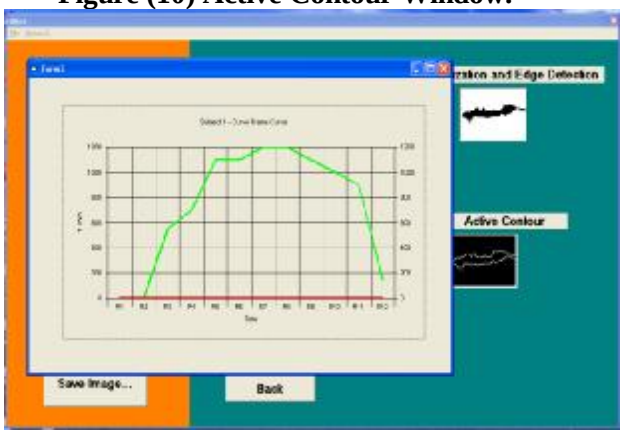

Figure (11) Feature Detection Window