UNIQUE LIPS FEATURES EXTRACTION

Prof. Dr. Abdul Monem S. Rahma* and Dr. Abdul Hamza A. Abdul Karim**

Received on: 25/5/2011
Accepted on: 25/8/2011

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 is based upon the lower part of the human face to extract the speaker sound relevant features accurately and robustly from the inner edge of the lips and trace it acoustically to prove its unique features and the possibility of merging it with sound features by measuring their physiological or behavioral characteristics curves. The results were promising and offered a good reaction: 94% - 100%.

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

انتزاع ميزات الشفاه الفريدة

الخلاصة:

تمييز الاصوات اعتمادا على المعلومات المرئية مثل شكل الشفاه وحركتها يشار لها بقرأة الشفاه. الخواص المرئية اشتقت طبقا لنسبة الاطر الفديوية المتسلسلة . العمل المتبنى استند على الجزء الاسفل من الوجه البشري لاستخلاص ميزات المتكلم الفريدة ذات العلاقة بدقة وشدة من الاثر السمعي والحدود الاخلية للشفتين ,لاثبات خواصها الفريدة وامكانية الدمج مع خواص الصوت باستخدام التقييس الاحيائي فيزيائيا و سلوكيا والتعرف على فوارق منحنيات الخواص الناتجة . كانت النتائج جيدة وواعدة عرضت نسبة تمييز جيدة : 94%-- 100% .

*Computer Science Dept., University of Technology
** Electromechanical Eng. Dept., University of Technology

1. Introduction

Several researchers work on the nature of visual information in human speech perception [1, 2, 3, and 4]. 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 [2].

Such systems must be capable of tracking the lips (both inner and outer contour) [5], 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 of the teeth. Lip image required image enhancement, threshold, and image segmentation figure (1) [6]. When the teeth are obscured by the lips, there is both an edge and intensity valley along the inner lip contour [5, 7], but when the teeth are visible; there are numerous edges inside the mouth which serve to distract the tracker as in figure (2).

In a previous work using cosmetically assisted lips [5], 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 HMM-based recognition, the inclusion of the visual signal tends to stabilize the Vertebra state alignment [8]. This demonstrates that acoustic information alone is inadequate to accurately identify noisy speech. Biometrics 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 [9].

2. Human Lip Tracking and Sound

Visual analysis system is used to track the position of the mouth through the image sequence, and extract a meaningful parameter set for the 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 [8]. Human speech is bimodal both in production and perception. Human speech is produced by the vibration in the vocal tract that is 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 [10].

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, 11] .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 [10,4].

2) Most of face recognition research is often based on static face image by assuming a neutral facial expression. However, the appearance of the face can change considerably during speech due to facial expressions [3,12]. It is also well-known that visual modality of speaker's mouth region provides additional speech information which can lead to improving 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 [6].

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 [6].

3. The proposed work based on lip and voice tracking

We start with building software system covering the needed functions like image algebra, the arithmetic and logic operations. The spatial filters mean both 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.

We portrayed the face of each subject as he/she starts saying the same sentence which is in the name of God the merciful (to make sure that there is a unique lip feature even with saying the same sentence or the same number of tested frames when more than one subject needed the same time to complete the sentence, 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 program applications are used in object segmentations and tracking in image sequences which are an important problem [8], involving the isolation of a single object from the rest of the images that may include other objects and background. Mainly 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) show the segmented image and its edge detected using Ulead-Video Studio. Our work followed the sequence below:

Present Biometric → Capture → Process → Store.

During the enrollment and verification of the subject, we used a biometric device, in our case the biometric device is a digital camera to provide a biometric sample. Preprocessing stage included all the operation required for making the features easy to extract. The biometric system will extract feature information from the biometric sample. In enrollment process, the biometric feature information is formatted into a template. This template is then stored in a centralized biometric database. While in verification process, the biometric feature information is formatted into a temporary template to be used to perform a search or verification. The template and temporary template are now submitted to the biometric engine (the identification or authentication process) to make a comparison against the suspected identity and generate a pass/fail output.

The number of masks used for edge detection is almost limitless. Many types of masks like Laplacin, Prewitt, and Soble are used. The Laplacin is seldom used by itself for edge detection, as a second – order derivative, it is unacceptably sensitive to noise, its magnitude produces double edges. The Sobel edge detection masks look for edges in both the horizontal and vertical directions and then combine this information into a single metric. The Prewitt is similar to the Sobel, but with different mask coefficients.

The digital image processing is used depending 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 [12].

After detecting the edge of inner mouth for each subject 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.

A threshold is used to segment images in two objects (background and lines) that identify each person. There is a way to extract selection of a threshold (T) that separates the modes. Then any point (x, y) for which f(x, y) > T is called an object point, otherwise, the point is called background point [13].

A threshold image g(x, y) is defined as:

$$g(x, y) = \left(\begin{array}{ll} 1 & \text{if } f(x, y) > T \\ 0 & \text{if } f(x, y) < T \\ & \end{array} \right)$$

where f(x, y) is a gray level value at location (x, y).

Calculate the time, the number of frame, and number of tested frames for each subject. The subjects are taken from 1.41 sec to 2.08 sec to say the same sentence since there are 25 frames for each sec so there are about 34 to 50 frames to be segmented, after segmentation, 9 to 13 frames are to be tested,

Table (1) shows the time taken by each subject, the number of frames, and the frames to be tested. The time taken by each subject to say that sentence is divided into 10 equal parts. Then by measuring the distance between the upper and lower lips 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 (***m*** open), then measuring H1,H2 (the upper and lower lips height) as in figure (3).

Tables (2), (3), and (4) show the measuring result of ***m*** open, H1, H2, the ten divided times, and the width of ***m*** open. Then we draw the characteristic curves of times against distances for each subject (mentioned above). Now by finding out the acoustic trace for each subject and fixing it under that curve as in Figure (4) and (5). We can note that each subject has his unique characteristic curve and his acoustic trace (the shape and its absolute value).

So the relation between visual lip movement (represented by ***m*** open, H1, H2, and width) and the sound produced by each subject (represented by acoustic trace itself or its absolute value), give the unique characteristic performed from merging sound and visual image. The merged relation is between different values of ***m*** open, H1, H2, and width of talking mouth (each of which is ten values for each tested frame) corresponding to a different absolute value of acoustic trace (10*tested frames).

4. Conclusions

This paper presents a technique to extract Information from digital sequence images of lips, and describes a new approach for utilizing active contours of lip-tracking based on B-Spline. Suitable geometric features are extracted from speaker's lip shapes, the work focuses on the lip shapes and movement, and features can be recovered from speakers lip shape. Both subject (8&11) take the same time (1.41 sec) to say the same sentence, so both have the same numbers of tested frames and time divisions, but each one has its own characteristic as seen in Figure (4) and Figure (5) which give different shapes for theirs tested frame and different absolute value. That means even if the subjects take the same time to say the same sentence their characteristic is different.

The result percentage changes depending upon the image resolution and free of acoustic distortion.

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 2008

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

[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 © 2008 IEEE.

[4] K. Nalini V. Ratha "Advance in Biometrics"©Springer-Verlag London limited 2008.

[5] 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.

[6] S. Stillittano and A. Caplier "Inner Lip Segmentation by Combining Active Contours and Parametric Models" VISAPP 2008-Int. Conf. on Computer Vision Theory and Applications

[7] 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.

[8] C. Bouvier, P.Coulon, X. Maldague "Unsupervised Lips Segmentation Based on ROI optimization and Parametric Model" hal-0037

[9] *Z*. Wu, J. Wu, and H. 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" 978-1-4244-2942-4/08/©2008 IEEE

[10] 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.

[11] S. Stillittano and A. Caplier "Inner Lip Segmentation by Combining Active Contours and Parametric Models" visa pp 2008-International.

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

[13] G. Rafael Gonszalez, E. Richard Woods, and I. Steven Eddine “Digital Image Processing Using Matlab“ Prentice Hall 2005.

Table (1) Times, Frames, and Tested frames for each subject

subjects number	Time/ sec	Frames	Tested frames
1	2.08	50	13
2	2.08	50	13
3	1.83	44	11
4	1.75	42	11
5	1.66	40	10
6	1.66	40	10
7	1.50	36	9
8	1.41	34	9
9	1.58	38	10
10	1.58	38	10
11	1.41	34	9
12	1.91	46	12

Table (2) Time, width and height for subject -8 tested frame 9

Frame No	Time Sec	Width mm	Height mouth open	H1 Upper lip open	H2 Lower lip open
1	0.157	35.0	1.00	5.50	10.0
2	0.314	32.2	4.00	4.50	10.0
3	0.471	30.0	4.00	4.00	12.0
4	0.8	32.0	5.0	7.00	12.0
5	0.785	32.0	1.50	7.00	10.0
6	0.942	33.0	6.0	6.00	13.0
7	1.099	30.0	4.00	7.00	7.50
8	1.256	30.0	5.00	6.50	12.50
9	1.41	33.0	0.00	4.00	11.00
10	-	-			

Table (3) Time, width and height for subject -10 tested frame 10

Frame No	Time Sec	Width mm	Height mouth open	H1 Upper lip open	H2 Lower lip open
1	0.158	36.0	0.00	6.00	12.50
2	0.316	36.0	3.00	6.0	12.50
3	0.474	3.30	8.00	8.0	14.0
4	0.627	35.0	8.50	7.50	14.0
5	0.790	33.0	9.0	8.0	14.0
6	0.948	36.0	3.50	6.0	14.5
7	1.106	35.0	8.0	7.5	13.5
8	1.264	33.0	9.0	7.0	9.0
9	1.422	33.0	0.00	5.0	7.0
10	1.580	36.0	0.00	6.0	10.0

Table (4) Time, width and height for subject -11 tested frames 9

Frame No	Time Sec	Width mm	Height mouth open	H1 Upper lip open	H2 Lower lip open
1	0.157	33.0	0.00	3.00	6.00
2	0.314	30.0	6.00	3.00	8.00
3	0.471	29.0	7.00	4.00	9.00
4	0.628	29.0	5.50	5.00	9.00
5	0.785	29.0	4.50	4.00	7.00
6	0.942	29.0	6.50	5.50	11.00
7	1.099	30.0	8.00	4.50	8.50
8	1.256	33.0	1.50	3.00	5.50
9	1.41	30.0	0.00	2.50	5.00

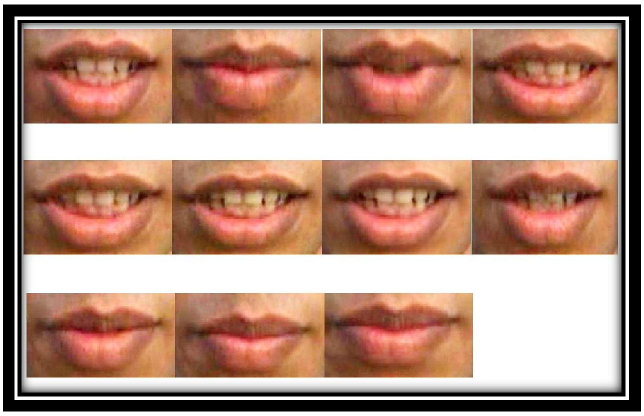

Figure (1) Segmented subject image (frames)

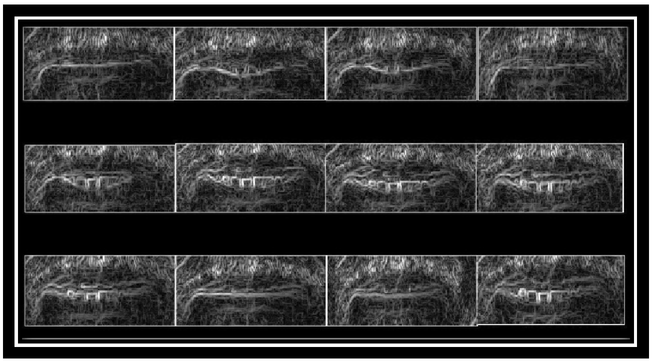

Figure (2) Inner lip edge detection

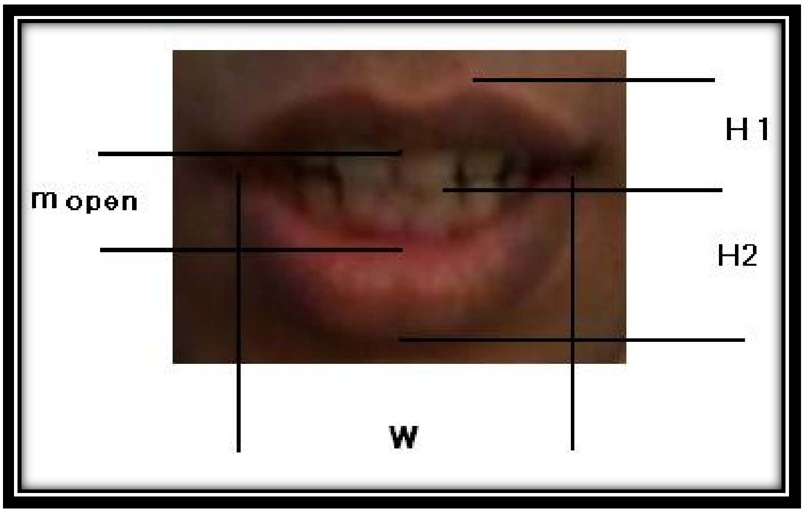

Figure (3) **m** open, upper, and lower lips highs

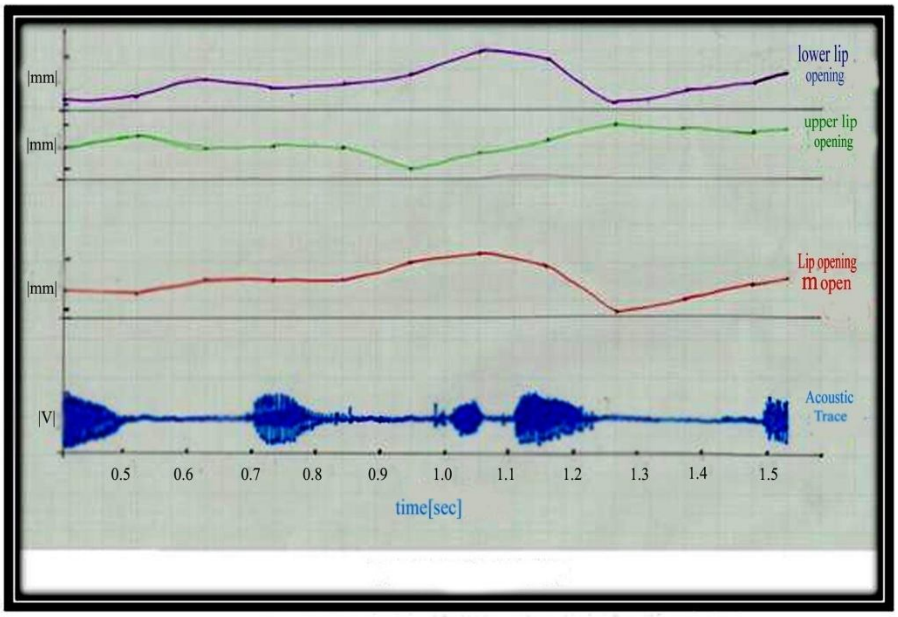

Figure (4)lips apening and acoustic trace for subject(8)

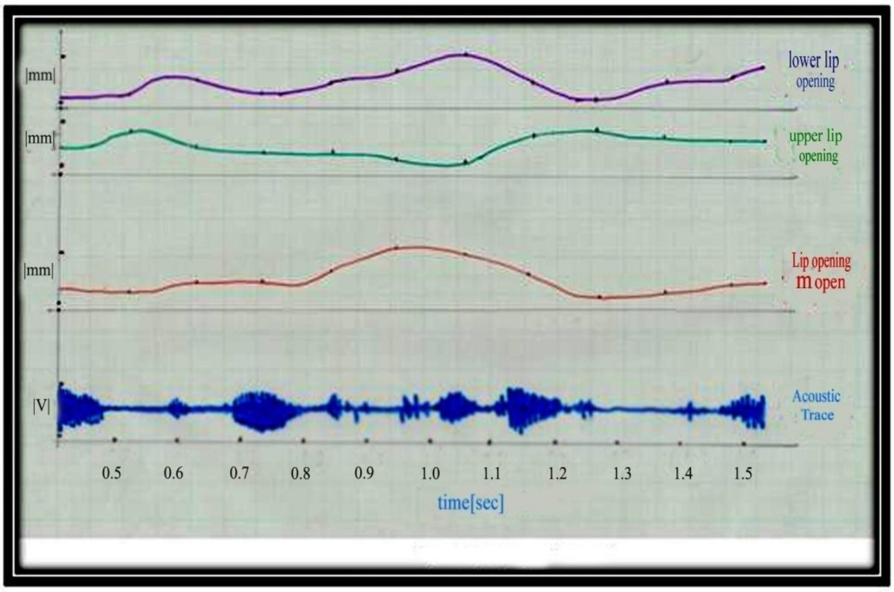

Figure (5) lips opening and acustic trace for subject(11)