Automatic Lip Tracking and Extraction of Lip Geometric Features for Lip Reading

Sunil S. Morade and B. Suprava Patnaik

Abstract—**Lip tracking is very crucial for visual lip reading recognition system. This paper presents a novel active contour guided geometrical feature extraction approach for lip reading. Three active contour methods are studied, namely snake, region scalable fitting energy method and loacalised active contour model. These methods are adopted for salient geometrical feature calculation. A joint feature model, obtained by combinatining inner area, height and width has been proposed. Results of experimentations on digit utterances are given to show the improvement achieved by visual speech recognition systems.**

Index Terms—**Localised active contour model, lip tracking, principal component analysis, region scalable fitting energy method.**

I. Introduction

It is known to the human from long time that there is useful information conveyed about speech in the facial movements of speaker. Hearing impaired listeners are able to use lip reading techniques very successfully. However, even for those with normal hearing, by seeing face of speaker intelligibility increases especially under noisy conditions. The importance of visual modality was well known back as in 1954 [1]. The intimate relation between the audio and visual sensory domains in human recognition can be demonstrated with audio video illusions such as the McGurg effect [2]. The primary advantage of visual information is that, it is not affected by the acoustic noise and cross talk among speakers. Thus Visual speech information is important.

In this work study and experimentations were carried out for lip feature extraction which are useful for visual speech recognition. Several methods are used for visual feature extraction. Out of these methods, shape based or appearance based feature extraction methods are important. In this paper we have implemented geometric features extraction method. Petajan developed one of the first audio visual systems. (Petajan et al., 1988) In his system mouth area, perimeters, height derived from lip were used as visual features.

G. Chiou and J. N. Hwang used Lip segmentation on color video [4]. Colour method of lip segmentation having drawback that it does not yield satisfactory results on image with weak colour contrast. This method is not robust to drastic variation in illumination. For visual speech information lip tracking is important. From lip shape or contour, different features are extracted. In this paper different active contour methods are used for lip segmentation or lip tracking.

The paper organized as follows Section II gives a description of lip tracking. Feature extraction of lip is described in Section III. Experimental results are set in Section IV. Finally paper is concluded in Section V.

Manuscript received December 29, 2012; revised April 26, 2013.

Sunil S. Morade is with the Department of E and TC Engineering, K K Wagh Institute of Engineering Education & Research, Nashik, India (e-mail: ssm.eltx@gmail.com).

B. Suparva Patnaik was with SVNIT, Surat, India. She is now with the Department of E and TC Engineering, Xavier Institute of Engineering, Mumbai, India (e-mail: suprava_patnaik@yahoo.com).

II. Lip Tracking

Once a region of interest of mouth is located, algorithms can be used for lip contour estimation. Here three methods of active contour for boundary extraction are compared. One of the most common methods of active contour models is snakes. Second method is based on Region scalable fitting energy method (RCFE). Third method is localized active contour model (LACM) which depends upon local(region) information. To our knowledge, in this work both RCFE and LACM methods are used for the first time for lip tracking.

A. Snake Method

A snake is an energy minimizing spline guided by external forces and influenced by image. Result of snake is line or edges. Snakes are active control models. The snake technique was first introduced by Kass, Witkin and Terzoupoulos [5]. We implemented this method for lip tracking in comparison with other methods. A serious problem with the snake is that usually the user must place the initial snake points close to the feature of interest. Also change from lip to teeth snake does not follow the proper path. Snake algorithm adapts the algorithm of Williams and Shah [6]. In this algorithm the overall energy term is reformulated as

$$E = \int (\propto (s)E_{cont} + \beta(s)E_{curve} + \gamma(s)E_{image})ds \quad (1)$$

where α, β and γ is used control relative importance of each term. Barnard and et. all implemented pattern matching snakes for lip tracking [7]. Drawback of this method is that initial contour selection. The lip shape being variations of parabola, rectangular, square or circular shape initialization is not an efficient.

B. Minimization of Region-Scalable Fitting Energy Method

In level set method, a contour C belong to domain Ω is represented by the zero level set of a Lipschitz function Φ: $\Omega \rightarrow$ R , which is called a level set function. A. Sayeed *et al.* used lip contour detection using the level set segmentation method [9]. Lip contour is given by zero level set when the iterative loop of level set evoluation is complete. Minimization of Region-Scalable Fitting Energy method was introduced by C. li *et al.*, used for biomedical image

segmentation. In this paper we used this method for lip tracking. The below equation (2) is the level set evolution equation. The term is derived from the data fitting energy; this term plays a key role in the model, since it is responsible for driving the active contour toward object boundaries fitting term [8].

$$\frac{d\phi}{dt} = -\delta_\epsilon(\phi)(\lambda_1 e_1 - \lambda_2 e_2) \quad (2)$$

$$e_i(x) = \int K_\sigma (y-x)|I(x) - f_i(x)|^2 dy \;\; i = 1,2 \quad (3)$$

$$f_i(x) = \frac{K_\sigma (x) * [M_i \phi(x)) I(x)]}{K_\sigma (x) * [M_i \phi(x))]} \;\; , \;\; i = 1,2 \quad (4)$$

Variable $K\sigma$ is guassian kernal. $I(x)$ is a function of image intensity. λ is a constant. This equations are implemented for lip tracking. The value of σ =3, $\lambda_1 = \lambda_2$=1, $M1=H(\Phi)$, $M2=1-H(\Phi)$where $H(\Phi)$ is Heaviside function. δ(Φ) smooth dirac delta function (derivative of $H(\Phi)$).

C. Localised Active Control Method (LACM)

Shawn Plankton and Allen Tannenbaum proposed a natural framework that allows any region-based segmentation energy to be re-formulated in a local way. They consider local rather than global image statistics and evolve a contour based on local information. Localized contours are capable of segmenting objects with heterogeneous feature profiles that would be difficult to capture correctly using a standard global method. This method is used for lip tracking and equations used are as follows [9].

$$\frac{d\emptyset}{dt}(x) = \delta\emptyset(x) \int_\sigma D(x,y) F(I(y), \emptyset(y) dy + K\big(\emptyset(x)\big) \quad (5)$$

$$K(\emptyset(x)) = \propto (\emptyset(x) div \left(\frac{(\nabla(\emptyset(x)}{(|\nabla(\emptyset(x)|}\right) \quad (6)$$

$$\nabla_{\emptyset(y)} F = (\delta\emptyset(y)((I(y) - u_x)^2 - (I(y) - v_x)^2) \quad (7)$$

In equation (5) x and y are independent variable each representing single point in domain Ω of image. I represent single image from frame. F is generic internal energy measure to represent local adherence. 'I' is image in two dimensional. In the energy models the foreground and background as constant intensities represented by their means of u and v. $\delta(\Phi)$ is a smooth version dirac delta . α is weight of smoothing term. F is the energy function of uniform modeling energy. Energy function is given by equation (7). $D(x, y)$ is a mask local region whose value is 1 inside radius r and is zero outsider.

Results of contour depend on number of iteration, Localization Radius (in pixels) and Alpha Weight of smoothing term. For radius equal to 3 and 4 result are better. Smaller the radius more local energy, bigger the radius more global energy is considered. Higher the value of alpha smoother is contour. From the contour different geometrical features are extracted.

III. Feature Extraction

A. Geometric Parameter Extraction

In the geometric feature approach we first appropriately normalize and rotate the outer lip contours, in order to compensate for relative location variations between the subject and camera. Geometric features are extracted from the contour C. Features, namely lip height (H), width (W), area (A) and angle (ɵ) are the most informative for automatic speech-reading.

Let $f(x, y) = 1$ if (x, y) ϵ C_{inside}.

$= 0$ otherwise.

Then the four features of interest are defined as follows

$$W = \max_y \sum_x f(x,y), \quad H = \max_x \sum_y f(x,y) \quad (8)$$

$$\mu_{10} = \sum_x \sum_y x f(x,y), \quad \mu_{01} = \sum_x \sum_y y f(x,y), \quad (9)$$

$\max_y$ indicates maximum height in y direction and $\max_x$ indicates maximum height in x direction.

'A' denotes Area of lip contour

$$A = \sum_x \sum_y f(x,y) \quad (10)$$

Angle for contour is found by using following formula

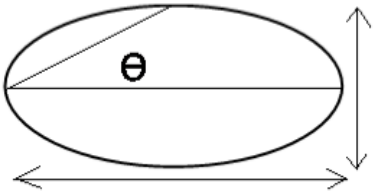

Fig. 1. Angle calculation.

$$\text{Angle } (ɵ) = \tan\text{-}1(H/W) \quad (11)$$

B. Feature Extractions Using PCA

Kaynak et al. compared different geometrical features of lip. Lip features that give more accurate result are used for digit recognition. They observed that combination of width, height, and variation in angle is important and give more accurate result as compared to single parameter [11]. In his paper I. Matthews et. all discuss different way to Extract visual features for lipreading [12]. The combination of geometric parameter is converted into a lip feature vector by using principal component analysis.

For selecting the visual feature combinations, PCA is used to obtain optimal combinations from statical point of view. In PCA technique first variance of input matrix is calculated (for combination of geometric parameter). Then Eigen vectors are calculated. Eigen vectors associated with higher eigen values are selected. Using PCA technique, size of Final feature matrix reduces. The visual features combinations were obtained by increasing the dimension of the feature vector. Important combination of geometrical parameters is tested by using angle distance method. using feature vectors. Equation (12) is used for Angle distance method calculation.

$$d(x,y) = -\cos(x,y) = \frac{\sum_1^n x_i y_i}{\sqrt{\sum_1^n x_i^2 \sum_1^n y_i^2}} \quad (12)$$

IV. Experimentation

A. Database for Lip Reading

G. Potamianos et al. created their Speaker independent audio-visual database for bimodal ASR. In this experiment we have created our own database and used [3].

As the available video database is either in American English accent or other foreign languages. Video Data base is created in Indian English accent. In data base video recording male and female both are used .Recording distance is kept constant. No head movement is allowed. Background used is blue. Video was recorded for digit 0-9 number. For this professional camera used having specification 25frames/s and audio sampling frequency 48000 is used. For each person six videos are recorded in sequence of 0-9 and random manner also. Videos are recorded in normal light. Volunteers are selected from the age group of 22-25.

Professional camera (SonyHVR-Z7 HDV) is used. Each image frame has resolution of 720x526. Aspect Ratio used is 4:3. While recording the data care is taken such that there is no head movement and subject has normal expression. Video data is recorded for numbers (0-9) pronunciation with a pause in between the digits. The video data is stored in MPEG format.

Visual features are usually extracted from lips of video frames. At initial stage face localization and mouth localization is done manually. For localised active contour method to work properly exact localization mouth is necessary. Initial contour may be square or ellipse of covered with mouth area and result tested by using both the method. For computation of the next frame contour previous frame contour is used as an initial contour. Using previous contour as a starting contour, computational efficiency increases. Results of contour depend on number of iteration, Localization Radius (in pixels) and Alpha, weight of smoothing term.

B. Results of Lip Tracking

Three methods are tested with own database. While snake method requires exact initial features and is shown Fig. 2 .Resultant contour of snake method is not exactly touching with lip boundary. Fig. 3 shows resultant contour with RCF method. It is also not accurate as contour crossover lip boundary and corners are not exact. Fig. 4 shows lip contour obtained by LACM method which is matching with lip boundary. Fig. 5 shows lip contour comparison of three methods.

For implementing proposed lip tracking Matlab Ver. 7.8 is used on intel core 2 duo CPU 2.0GHz and 2 GB RAM. In snake and region-scalable fitting energy method are giving less accuracy as compare to LACM. Also for these methods we are not able to use previous contour for the next frame. LACM gives better accuracy. For LACM method first frame execution time for track the lip is more (151s) because more number of iterations are required but next frame to track the lip required less time (61s). Region scalable energy method requires more time for each frame (73s). Total computation time for entire digit lip tracking using LACM method is less. So LACM is computationally more efficient**.** Geometrical parameters of lip are calculated based on the result of LACM.

C. Results of Geometric Features Extraction

On an average 16 frames are required for a digit to represent visually. From 16 frames 16 contour are extracted. This information is stored for further computation.

Fig. 2. Frame with lip contour by RCF energy method.

Fig. 3. Frame with lip contour by RCF energy method.

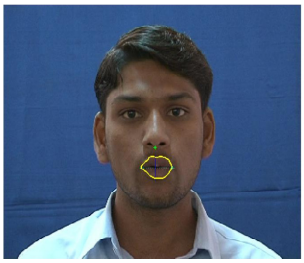

Fig. 4. Frame with lip contour by LACM energy method.

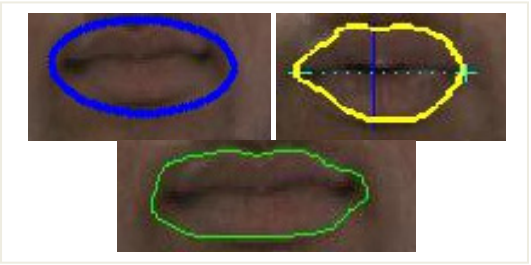

Fig. 5. Comparison of three lip contours blue-Snake, Yellow-LACM, Green-RCF energy method.

From contours of lip height (H), width (W), angle, area (A) are calculated for each frame. These parameters are converted into normalized form. Table I shows only height information in normalized form for digit 0 to 4. Similar way W, A and θ is calculated. More information is obtained from height as compared to width because for digit utterance lip vertical direction movement is more as compare to horizontal, hence is more informative. From the values table I average values of H is calculated. Similar way all other (W, A, θ) average numerical values are calculated. From this value W/H and standard deviation of area (std) A are calculated. Table II shows average value of height, width height ratio (W/H), std area, and angle for digit 0 to 4. The fitness rate of

the lip movement feature is calculated by combining the feature data in Table II. Higher the fitness rate, selected frames can be used for computation. Fitness rate is useful for real time application.

TABLE I: NORMALIZED HEIGHT FOR LIP FRAME 4 TO 9 FOR DIGIT 0 TO 5

Digit	*Average H,W/H, θ ,S.A for 16 frames*			
	Height (H)	***width height(W/H)ratio***	***Angle(θ)***	***Std_Area***
Zero	0.896	0.871	0.504	42.53
One	0.888	0.865	0.521	124.21
Two	0.889	0.944	0.4882	74.076
Three	0.925	0.939	0.484	47.74
Four	0.925	0.939	0.484	47.74

TABLE II: NORMALIZED HEIGHT FOR LIP FRAME 4 TO 9 FOR DIGIT 0 TO 5

Digit	*H value for frames 4,5,6,7,8,9*					
	4	*5*	*6*	*7*	*8*	*9*
zero	0.766	0.733	0.766	0.766	0.833	0.866
One	0.833	0.866	0.933	1.000	1.000	0.933
Three	0.866	0.900	0.933	1.000	1.000	1.000
Four	0.800	0.833	0.833	0.833	0.900	0.866
Five	0.933	0.900	0.900	0.833	0.833	0.900

TABLE III: DIGIT RECOGNITION RESULT USING COMBINATION OF PARAMETER HEIGHT, AREA AND ANGLE FOR DIGIT 2,3 AND 4.

Test Input PCA	***Train Input PCA***					
	Zero	***One***	***Two***	***Three***	***Four***	***five***
Two	0.811	0.823	**0.435**	0.640	0.737	0.483
Three	0.861	0.905	0.857	**0.754**	0.838	0.802
Four	0.433	0.248	0.279	0.306	**0.121**	0.218

D. Results of Lip Features Extraction

Principal component analysis is applied for combination of lip parameters. Eigen vectors are found . Eigen vectors of testing number are compared with Eigen vectors training number by distance method. Digit recognition results are shown in the table III. The first column in table III is the testing input and other column indicates comparison result. Minimum value along the rows of table III indicates that digit is recognized or distance between Eigen vectors is minimum. We are getting better result as shown in Table III, by a combination of Height, angle and area parameters. So these are the high impact geometrical features for lip reading.

V. CONCLUSIONS

In this paper we have implemented lip tracking by LACM, RCFE and Snake method. LACM and RCFE methods are used first time for lip tracking. These methods are generally used for biomedical application.

Feature vectors are found by principal component analysis technique. The experimental results are performed to determine important statistical parameters of geometrical method. Area, height and angle of lip are found to be useful parameters for feature vectors during lip reading process.

REFERENCES

[1] W. H. Summy and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America,* vol. 26, pp. 212-215,1954.

[2] H. Mcgurk and J. Macdonald "hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[3] G. Potamianos, E. Cosatto, H. P. Graf and D. B. Roe "Speaker independent audio-visual database for bimodal ASR," in *Proc. of the European Tutorial and Research Workshop on Audio-Visual Speech Processing,* pp.65-68,1997.

[4] G. Chiou and J. N. Hwang. "Lipreading from color video," IEE*E Trans. Image Process.*, vol. 6, no. 8, pp. 1192–1195, 1997.

[5] Kass, Witkin and Terzoupoulos,"Snakes :Active contour models," *International Journal of computer vision*, pp. 321-331,1988.

[6] D. J. Williams and M. Shah,"A fast algorithm for activecontours and curvature estimation," *CVGIP Image Understanding*, vol. 55, no. 1, no.14–26,1991.

[7] M. Barnard , E. Holden, and R. Owens"Lip tracking using pattern matching snakes," *ACCV2002,* pp. 1-5, 2002.

[8] C. Li, C. Kao, J. C. Gore, and Z. Ding, "Minimization of Region-Scalable Fitting Energy for Image Segmentation," *IEEE transactions on image processing*, vol. 17, pp. 1940-1949, 2008.

[9] A. Sayeed, Md. Sohail, and P. Bhattacharya, "Automated lip contour detection using the level set segmentation method," *International Conference on Image Analysis and Processing* (ICIAP 2007), pp. 425-430, 2007.

[10] Shawn lankton ,allen Tannenbaum " Localizing region based active contours," *IEEE transactions on image processing*, vol. 17, pp. 2029-2039, 2008.

[11] M. N. Kaynak , Q. Zhi , A. D. Cheok , K. Sengupta, Z. Jian, and K. Chi Chung, "Lip geometric features for human–computer interaction using bimodal speech recognition: comparison and analysis," *Speech Communication*, vol. 43, pp. 1–16, 2004.

[12] I. Matthews, T. Cootes, and J. Bangham, "Extraction of visual features for lipreading," *IEEE Trans. on Pattern Analysis and Machine Vision*, pp. 198-213, 2002.

Sunil S. Morade received his M.Tech. in Electronics from VNIT, Nagpur in 1993. He is working as an Associate Professor at KKWIEER. Nashik and he is doing his PhD from SVNIT, Surat, India. His research interests are in image and video processing, Embedded System and Signal Processing. He is a member of IETE and ISTE (India).

Suprava Patnaik obtained B.Tech and M.Tech. degree from NIT Rourkela. She received Ph.D from IIT Kharagpur in 2004. She was a Professor in Electronics Department of SVNIT, Surat. Currently she is working as a Professor at Xavier Institute of Engineering, Mahim, Mumbai. Her research interests include machine learning, computer vision and image processing.