Lip Localization and Viseme Recognition from Video Sequences

Thejaswi. N. S and Somnath Sengupta
Computer Vision Lab, Department of Electronics and Electrical Communication Engineering,
Indian Institute of Technology, Kharagpur 721302, India
Email: tns@iitkgp.ac.in, ssg@ece.iitkgp.ernet.in

***Abstract*—Viseme (visual cue) recognition is one of the steps to be followed in building an automated lip-reading system. In order to recognize a viseme, one has to first detect the lips of the speaker from the video sequences and track them to extract the feature vectors for the final recognition. A novel method for lip-localization based on the color models has been proposed. Also, the basic possible lip-shapes depicting the visual-cues have been presented along with their mapping to the corresponding phonemes. In the next level, mapping of the feature vectors from the lip-localization algorithm to the visual cues has been performed.**

I. Introduction

Lip-Reading is a technique to understand speech by interpreting the visual information from the lip movements, face and tongue. When we try to "lip-read" we gather the information from all the above mentioned tools. This paper attempts to make this process automated. The main problem one faces in this process is the task of lip-localization, as the boundaries of the lips are not very well-defined. Several researchers have tried to solve this problem. One of the popular information used to localize the lips is the color information from the images, such as [2], [6], and [7] etc., whereas some use the gradient information, such as [1], [4], [8]. This work proposes a novel and robust color based lip detection algorithm to be performed in the YIQ domain of image representation. It also proposes the basic visemes which a human eye can distinguish during a normal conversation. The recognition system integrates these two components. First, the lip-localization and feature extraction is carried out, followed by the mapping of the feature vectors to the actual visemes.

The rest of this paper is arranged as follows. Section-2 introduces the lip-localization problem along with the algorithm for the same. The viseme-phoneme mapping and the process of mapping the feature vectors to the visemes are given in Section-3. Section-4 presents the results of the simulation of this model on the test cases. Section-5 concludes this paper.

II. Lip Localization

Lip detection and tracking are the most important and primary components in the process of viseme recognition. Described in this section is a novel method for lip-localization. It will be assumed throughout this paper that the input video sequences are in the RGB domain. The color model used internally is the 'YIQ' domain representation of the image. The equation for RGB-to-YIQ conversion is given below:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.311135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

The reason for switching to YIQ format is due to the observation which was performed on the face and lips of humans in different domains of color representation. It was observed that the lips of a human are generally brighter in the 'Q' space and the face of a human is generally brighter in the 'I' space. At the same time, the procedures which use the gradient information of the image for lip-localization suffer from the problem of inconspicuous lip information. This can be seen in Fig.1, which displays the gradient information of a human face in negative scale.

Thus initially, face-detection will be performed in the 'I' space and the final lip-localization will be performed in the 'Q' domain. This particular space-conversion is helpful as it de-couples the luminance information from the chrominance information, providing with an illumination-invariant approach (unless the illumination is so poor that the visibility is affected). An example image in 'I' and 'Q' domains can be seen in Fig.3.

Lip detection is done by the application of segmentation on the 'Q' domain information, before which the face of the speaker will be localized using the information in the 'I' domain. This is necessary as the localization of lips becomes easier if one has already localized the face. Once the lip-localization is successful, active-contours (snakes) have again been applied to track the motion of the lips. Using the contour for the previous frame, and the 'Q' domain information of the present frame as the energy information for the active contours, the tracking has been performed successfully. Since, 'Q' domain information provides brighter regions around the lip, using this as the image energy information will prevent the snakes from getting stuck in local minima. A detailed algorithm is given in Fig.4.

III. Viseme Recognition

Viseme is defined as a lip-shape which a speaker makes in-order to pronounce a phoneme. Viseme can be considered as the unit of visual information of the speech, as opposed to the phonemes for audio. Given in Table.1 are some of the phonemes (mono-phonemes to be precise) in the English Language. But viseme-phoneme mapping is a many-to-one mapping. That means, given a viseme, it can correspond to two or more phonemes. For example, the lip shape for pronouncing the phoneme 'p' (as in 'poor') is the same as that of for the phoneme 'b' (as in 'boor'). In this regard, 15 discernible visemes for English language have been distinguished and have been shown in the Table.1, along with their congruent phonemes and examples.

The feature vector for the recognition is of length 16, whose first 15 elements are the variations of the angle of the tangents along the detected lip contour and the last element is a ratio which stands for the amount of opening of the mouth. It is defined as the ratio of number of pixels in the opened-mouth-region to the total number of pixels covered by the mouth region (including the lips and the opening). The process of mouth-open area is evaluated as follows. This is based on the observation performed on the images. A representation in the 'Q' domain shows that the mouth region (if open) will be predominantly dark. So, the process of evaluation of the area of open mouth region will be just to segment out this dark region formed inside the mouth and count the number of pixels inside this region. An example of an open-mouth is shown in Fig.2 below.

The tangent angle variations along the lip contour provide the lip shape, as shown in Fig.5, whereas the last element (the ratio) provides the information about the amount of mouth opening. In total, these 16 elements can totally represent the lip information for a given frame of the video. These features appeal to the present context, as the purpose here is to recognize the visemes based on the lip shapes. A detailed description of feature vector extraction has been shown in Fig.5. These feature vectors will be fed as inputs to the recognizer (Neural networks in this case) in order to recognize the viseme being generated in the frames.

Some of the strong points of this system are shown below. YIQ format decouples the luma from chroma information, thus, lip detection is illumination invariant, unless the illumination falls to such levels where the luma and chroma information can not be de-coupled. Chroma based detection procedure makes the detection algorithm position invariant as well as rotation and scale invariant [2]. Usage of 'Q' domain information as image energy removes the possibility of the presence of any local minima in case of active contour optimization. The viseme recognition is also based on the temporal information from the previous frames as well and hence, the viseme recognition algorithm models the real-world situation closely.

IV. Simulation and Results

The viseme recognition system has been simulated in MATLAB (Version 7.0). The test videos were captured at 15 fps, under moderate illumination conditions. Videos were recorded with two set of words, the first set of words cover all possible phonemes in the English language (shown in table.1) and the second set of words, known as the NATO phonetic alphabets, [9], covering all possible visemes in English language. The videos from the former set of words were used for training and the latter for testing the system's performance. Different network architectures were simulated and only the ones with best performances are mentioned here.

The algorithm for lip detection was tested with still images as well as videos. In this section some of the results are displayed. Shown in Fig.6 are the results of the lip-localization algorithm, which displays the accuracy of the detection and tracking abilities of the algorithm. Using these videos, the accuracy of lip-localization algorithm was found to be about 98%, as shown in Table.2. The errors were checked manually.

When any speaker utters a word, each viseme has its correspondence with its past and its future visemes, the feature vectors from the lip-localization algorithm have been concatenated for the purpose of recognition. The simulation has also been carried out with the concatenation of 0 (nothing) to 4 feature vectors (see Table.3). The networks offering the best performance for each of the buffer lengths are shown in the Table.3. All networks are of multi-layered-back-propagation type, with sigmoid non-linearity and were trained using 'Scaled Conjugate Gradient' algorithm with 2000 epochs. It can be inferred from the Table.3 that using a buffer length of '4' gives the highest possible accuracy.

It was also observed that the accuracy increases with the buffer length, which is logical as the increase in buffer length signifies the increase in the temporal information for the purpose of recognition. But, increased buffer length causes increased computational complexity. Also, increased buffer length implies an increased effect of time averaging on the samples. So, one has to make a trade-off between the accuracy and the computational burden. Fig.7a and Fig. 7b show the confusion matrix representation of the prediction accuracy for the buffer lengths from 0 till 4. The confusion matrix is a matrix in which the matrix element (i, j) represents the number of times the element 'i' being classified as 'j'. In this case, the confusion matrix has been row-normalized for the purpose of display. It can be seen that for the case of buffer length equal to '4', the accuracy is highest and the lateral diversion (misclassification) in the confusion matrix is minimal. Thus, a buffer length of '4' (i.e.

a feature vector of length 80) having an accuracy of about 88% suffices for all practical purposes.

V. Conclusion

A 'Viseme Recognition' system using video processing and neural networks has been proposed. A novel method for lip-localization and feature extraction has also been proposed by working on the YIQ domain information. The feature vectors generated for recognition purposes were based only on the lip contour information of the speaker. Also, reported in this paper are a set of basic visual-cues which a human eye can distinguish and the techniques for mapping the feature vectors to one of these visemes.

References

[1] Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon, "Key Points Based Segmentation of Lips", *IEEE International Conference on Multimedia and Expo, 2002, ICME '02. Proceedings. 2002*, Volume 2, 26-29 Aug. 2002 Page(s):125 - 128 vol.2.

[2] Wen-Nung Lie and Hung-Chih Hsieh, "Lips Detection by Morphological Image Processing", *4th International Conference on Signal Processing Proceedings, 1998. ICSP '98. 1998*, Volume 2, 12-16 Oct. 1998 Page(s):1084 - 1087 vol.2.

[3] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active contour models", *Int. J. Computer Vision*, 1(4):321–331, 1987.

[4] Richard Harvey, Iain Matthews, J. Andrew Bangham, Stephen Cox, "Lip reading from scale-space measurements", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997*, 17-19 June 1997, Page(s):582 – 587.

[5] Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon, "Jumping snakes and parametric model for lip segmentation", *International Conference on Image Processing, ICIP 2003*, Volume 2, 14-17 Sept. 2003, Page(s):II – 867-870.

[6] Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon, "Accurate and quasi-automatic lip tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 14, Issue 5, May 2004, Page(s):706 – 715.

[7] Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon, *"New color transformation for lips segmentation", IEEE Fourth Workshop on Multimedia Signal Processing, 2001*, 3-5 Oct. 2001, Page(s):3 – 8.

[8] M. Lievin, F. Luthon, "Unsupervised lip segmentation under natural conditions", *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP '99*, Volume 6, 15-19 March 1999 Page(s):3065 – 3068.

[9] Authors acknowledge the site http://www.iaco.int/ for their 'Nato-Phonetic Alphabets' which have been used in preparing the test videos.

[10] Patrick Lucey, Terrence Martin and Sridha Sridharan, "Confusability of Phonemes Grouped According to their Viseme Classes in Noisy Environments", *Proceedings of the 10th Australian International Conference on Speech Science & Technology Macquarie University, Sydney, December 8 to 10, 2004*, Page(s):265-270.

[11] Yang, J.; Xiao, J.; Ritter, M., "Automatic selection of visemes for image-based visual speech synthesis", *IEEE International Conference on Multimedia and Expo, 2000. ICME 2000. 2000* Volume 2, 30 July-2 Aug. 2000 Page(s):1081 – 1084.

[12] Martin Fodslette Moller, **"**A scaled conjugate gradient algorithm for fast supervised learning", *Source Neural Networks*, Year of Publication: 1993, Volume-6, Issue 4, Pages 525-53

Fig.1. Gradient of the image (in negative scale).

Fig.2. Image of open-mouth in 'Q' domain.

Fig.3. Image in RGB, I and Q domains respectively.

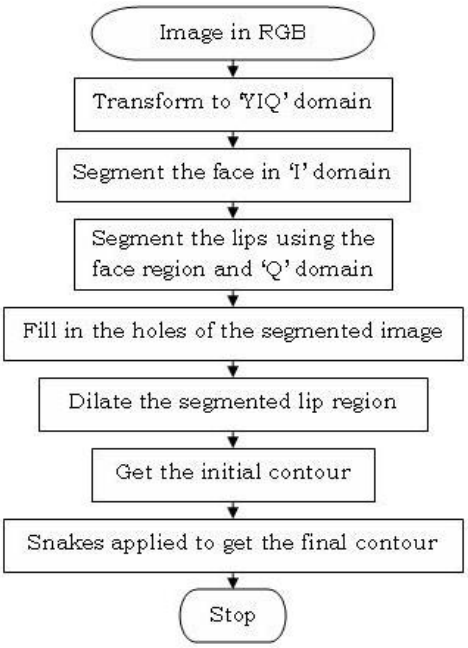
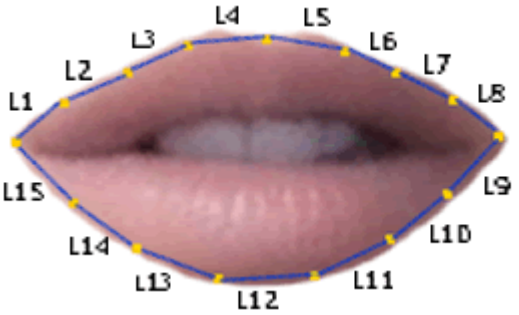


Fig.4. Block Diagram for the Lip-Localization algorithm.



α_i = *angle between* L_i *and* L_{i+1}

(addition operator being circular in nature)

The feature vector will then be:

$$V = [\alpha_1 \quad \alpha_2 \quad . \quad . \quad \alpha_{15} \quad r]$$

$$r = \frac{open-mouth-area}{total-area-of-mouth} \quad (measured\ in\ pixels)$$

Fig.5. Feature Vector Extraction

Fig.6. Results of applying the lip localization algorithm

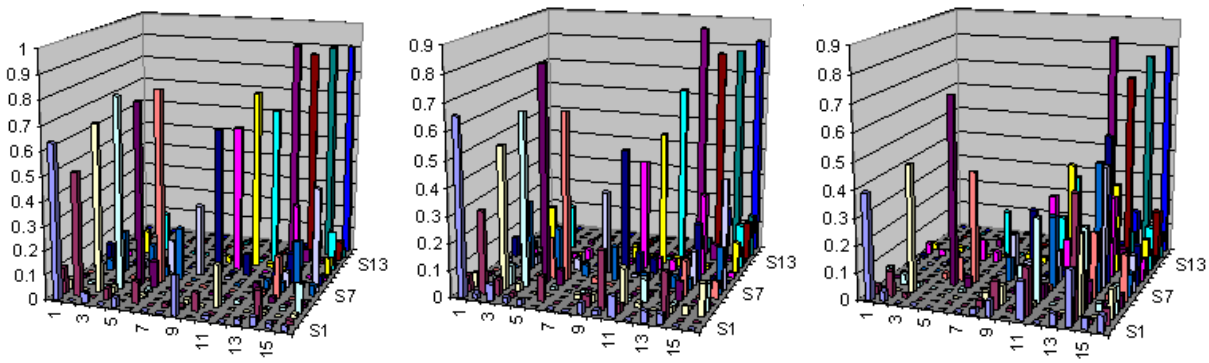


Fig.7a. Confusion matrix for the recognition for buffer lengths from '0', '1' and '2' (left-to-right)
(Series follow the same order as that of the visemes given in Table.1)

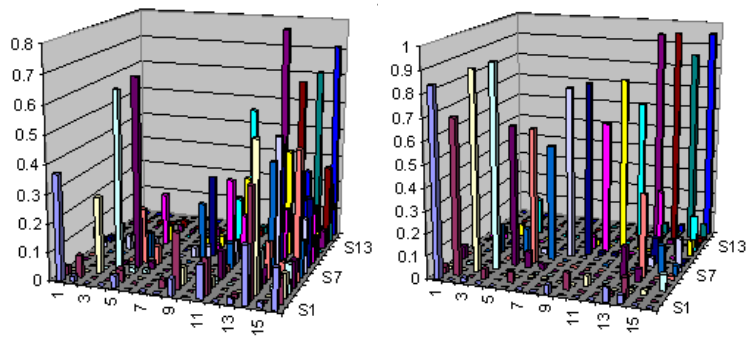


Fig.7b. Confusion matrix for the recognition for buffer lengths from '3' and '4' (left-to-right)
(Series follow the same order as that of the visemes given in Table.1)

Type	Phoneme	Example	Viseme	Type	Phoneme	Example	Viseme
Long	AH	Back	AH	**Explodents**	K	Cane, leek	t
	A	Tale	A		G	Gain, league	t
	E	Each	E	**Continuants**	F	Fat, safe	f
	AW	All	AW		V	Vat, save	f
	O	Oak	O		TH	Thigh, wreath	t
	OO	Ooze	OO		Th	Thy, wreathe	t
Short	a	At	a		S	Seal, base	t
	e	Etch	e		Z	Zeal, baize	t
	i	It	i		SH	She, dash	t
	o	Odd	o		ZH	Treasure, vision	t
	u	Tub	u	**Nasals**	M	Met, seem	p
	oo	Book	oo		N	Net, seen	t
Explodents	P	Post, rope	p		NG	Kingly, long	t
	B	Boast, robe	p	**Liquids**	L	Light, tile	t
	T	Tip, fate	t		R	Tire, right	t
	D	Dip, fade	t	**Coalescents**	W	Wet, away	f
	CH	Chest, etch	t		Y	Yet, ayah	t
	J	Jest, edge	t	**Aspirate**	H	High, adhere	t

Table.1. Phoneme-Viseme chart (Phoneme-Courtesy: Pitman Shorthand by Sir. Issac Pitman)

Total no. of frames tested	No. of erroneously tracked frames	Success Rate (in %)
2329	25	98.93

Table.2. Success of Lip-Localization Algorithm

Buffer length	Feature Vector Length	Network Structure	Success Rate (in %)
0	16	16-15-10-4	51.70
1	32	32-16-8-4	59.04
2	48	48-24-12-4	69.51
3	64	64-32-16-4	80.03
4	80	80-40-20-4	88.24

Table.3. Success rate of Viseme recognition as a function of temporal buffer length.