A LIP LOCALIZATION BASED VISUAL FEATURE EXTRACTION METHOD

Namrata Dave[1]
[1]Department of Computer Engineering, Gujarat Technological University, India

ABSTRACT

This paper presents a lip localization based visual feature extraction method to segment lip region from image or video in real time. Lip localization and tracking is useful in many applications such as lip reading, lip synchronization, visual speech recognition, facial animation etc. To synchronize lip movements with input audio we need to first segment lip region from input image or video frame. This paper presents a novel color based approach for localization of lips, which is an early stage for tracking lips in real time. A phoneme is a basic unit of speech and a viseme is a visual representation of phoneme or shape of mouth while utterance of particular phoneme. The main goal of our work is to implement a system for synchronizing lips with the input speech. To extract visual features i.e. visemes from input video frame or image we have used HSV and YCbCr color model along with various morphological operations. We have developed algorithm to work with normal lighting conditions and natural facial images of female and male.

KEYWORDS

Lip localization, viseme extraction, visual feature extraction, HSV color model, YCbCr color model.

1. INTRODUCTION

Visual features of speech i.e. visemes are useful in various fields, such as visual speech recognition [1], [2], audio-video synchronization, or speaker identification/localization in audio-video scenes [3]. It is observed that mostly visual information related to speech is contained in the movements of lip. However, a most difficult part in visual speech recognition is to find an efficient method for extracting visual speech features. It is very difficult to extract visual features from input image or video which is taken in different situations under varying lighting conditions, different pose and for different skin tones. Many algorithms have been proposed for lip localization and viseme extraction. Several groups are working with template models based on active shape models [4], deformable models [5] and dynamic contours [6].

A lot of work is already done on this topic since years. Many novel approaches have been proposed in this field. Most approaches are divided into two main groups. First group is using shape based sophisticated methods like active shape model, deformable models, snake etc. to extract features from input image or video. Another group is extracting features using different colour models using different spatial methods which work on pixel intensity, edge and colour information of lip and/or facial skin.

In this paper, we presented image based approach to extract mouth region and then to discriminate visual features from it. Here we have tried to extract visemes for digits one to ten.

2. EXISTING METHODS

The Lip segmentation has become an important issue in facial animation, automatic VSR processing and automatic face recognition. In such systems, the region of interest (ROI) which is

lips in our case must be detected in each frame. This procedure is normally done by face detection and extraction of the lip region.

Earlier systems performed the lip segmentation in conjunction with the application of artificial markers (lipstick) on the lips. The application of lipstick enables the system to detect precisely the lips in the image data, but this procedure is inappropriate since it is uncomfortable for users and such systems can be operated only in specific constraint based applications. Thus, the main research efforts have been concentrated in the development of visual features based algorithms for lip segmentation. Many studies have shown that colour information of particular region can be applied to identify the skin or face in digital images. The main idea behind this approach is to transform the RGB signal into other representation where the mouth region can efficiently separate from other details in given image, so that segmentation of mouth can be done efficiently. Due to this fact, a large number of colour representations have been proposed. Coinaiz et al [7] used the hue component of the HSV representation to highlight the red colour which is assumed to be associated with the lips in the image. Later, the HSV colour space is used [8] for lip detection. They used the hue signal to locate the position of the lips in mouth region. The interior and exterior boundaries of lips are extracted using colour and edge information using a Markov Random Field (MRF) framework based on the lip area found. Some of the approaches used YCrCb colour space for the lips detection [9, 10].

Eveno et al [11] proposed a new colour mixture and chromatic transformation for lip segmentation in 2001. In their approach, a new method for transformation of the RGB colour space and a chromatic map was applied to separate out the lips and facial skin. They argued that their proposed approach is able to provide robust lip detection under variable lighting conditions. Later on Eveno et al introduced a novel method where the pseudo-hue [12] was applied for lip segmentation that has been embedded in an active contour framework. They applied the proposed algorithm for visual speech recognition and the results show significant improvement in terms of accuracy in lip modelling.

Another method for mouth segmentation has been proposed by Liew in 2003 [13]. In their approach, they used new transformation method to convert given colour image into the CIE-Lab colour space and CIE-LUV colour space, and then they calculated a lip membership map using the fuzzy clustering algorithm. The ROI around the mouth can be identified from the face area after application of morphological filtering on given image.

Guan [14] has given a new approach based on Discrete Hartley Transform (DHT) to improve the contrast between lip region and the other regions of face. In their paper, they presented method of extraction of lips by applying wavelet multi-scale edge detection across the C3 component of the DHT. Their method takes into account both the colour information and the geometric characteristic of image.

3. VISUAL FEATURE EXTRACTION METHOD.

3.1. Proposed Method

Feature extraction process is preceded by a number of pre-processing steps to be done as shown in Figure 1. This involves face detection followed by ROI extraction. Then, the lips of the speaker are tracked in consecutive frames of recorded video. Following these steps, and given an informative set of features, the visual front-end module can proceed with feature extraction.

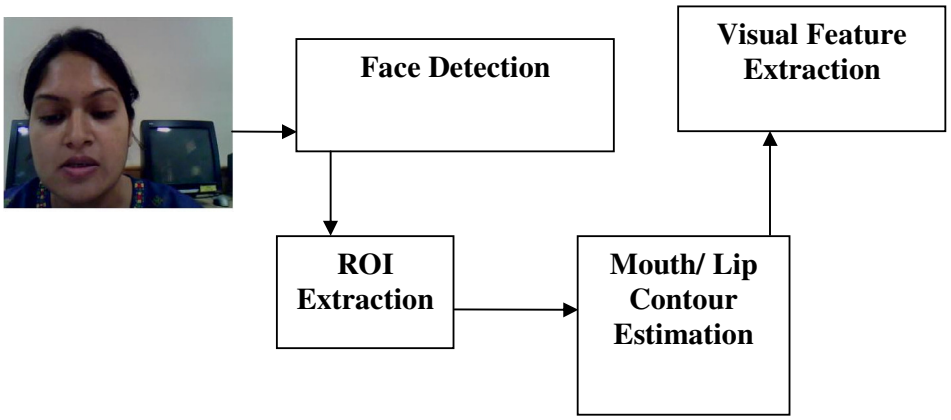

Figure 1. Visual Feature Extraction Process

3.2. Dataset and Implementation

A viseme is the visual appearance of a phoneme or visual unit of speech in spoken language. However, currently there is no standard viseme table available to be used by all researchers. Due to this fact, a viseme table for this study was based on the work performed by Lee and Sook in 2002 [15, 16]. This was used in a table which effectively mapped all the possible phonemes to viseme. They have shown the typical 48 phonemes used in English language including silence and short pauses, grouped into their 14 viseme classes.

Table 1: List of Phonemes & Viseme Used for Digits One to Five

Word	Phoneme	Viseme
One	w ah n/w an n/v ao n	v ao n
Two	t uw/d uw	t uw
Three	th r iy/dh iy	th r iy
Four	f ao r	f ao r
Five	f ay v	f ay v
Six	s ih k s	s ih k s
Seven	s eh v ax n	s eh v ax n
Eight	ey t	ey t
Nine	n ay n	n ay n
Ten	t eh n	t eh n

We have recorded videos for 12 different viseme classes as listed in Table1. Table1 shows list of phonemes and corresponding viseme for digits one to ten. We extracted mouth parameters for each viseme class for 6 different speakers. Videos recorded for different speakers include female and male speakers' videos. Videos are recorded in normal conditions using webcam. No artificial marker like lipstick is used for detecting lips. Videos are recorded in avi format with 30fps.

In this section, we will discuss quick summary of how exactly lip features are extracted from the video of speech. Entire video is converted into frames. RGB colour scheme of the image is not suitable for immediate processing as it contains a lot of mixed information about lightness etc. Another colour scheme should be used. Very convenient colour scheme is YCbCr as it separates luminance, blue chrominance and red chrominance. This method is convenient as the mouth region in image contains high red and low blue components in comparison with other face regions.

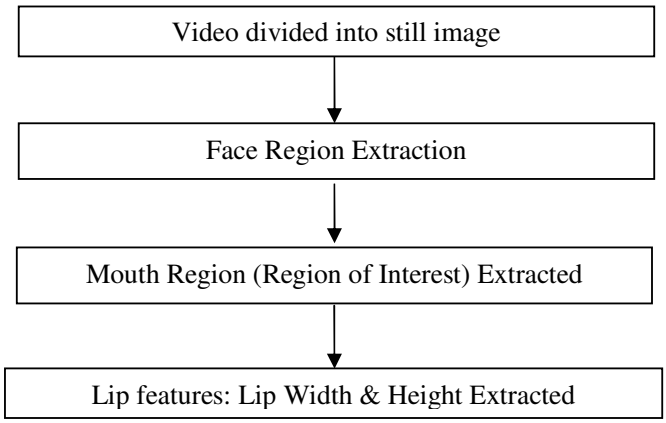

Figure 2: Overview of the Proposed Algorithm for Lip Feature Extraction

The mouth region has a greater value of red chromaticity (Cr) than of blue chromaticity (C_b). The value of the chromaticity of the Cr component is increased by using its square value. On the other hand, the region of the mouth has a good response to the ratio C_b/C_r.

Lip Feature Extraction Algorithm steps are described in figure 2. First, Video divided into frames (frame rate 30 frames per second). On each frame all operations described in algorithm are performed. We have used YCBCR colour space. Based on luminance and chromaticity information extracted from frame we are extracting mouth region surrounding Region of interest i.e. Lips. The ratio Cb/Cr is used to extract face region and eliminate background from input video frame. From the cropped face image we extracted lip area based on H and S value by using HSV colour model. On cropped lip image edge detection is applied first. Then to eliminate extra info morphological operations are applied to get final lip parameters. We have used these features to estimate mouth region with information such as height and width for each viseme.

3.2. Results

Four video sequences were taken from different subjects, each one having at most than 3X30 frames. Those sequences were analysed using the proposed lip segmentation algorithm. The

figure 3 shows examples of segmentation results for each subject. The best results were obtained for the pale skinned, non-bearded subject. We have used s and h value of HSV colour space to identify lip region based on which all viseme are extracted. Sometimes due to poor illumination we are not get good results only based on H and S value so we also try-out combining results of contrast stretched and log applied images to get exactly mouth region.

Using Lip Feature Extraction Algorithm we extracted Viseme for 10 different viseme classes. Figure 3 shows results of viseme V for four speakers. We tested our algorithm for many frames for each speaker. For testing purpose for each speaker we selected two videos and extracted frames to apply algorithm on it. On an average we are getting 85% result for all speakers.

Figure 3: Results viseme 'v'

4. CONCLUSIONS

After the investigation done on existing methods, we have implemented a lip localization based visual feature extraction method which gives good accuracy for our database. Our algorithm is suitable for offline applications which can be extended to work in real time applications. Another improvement to the program can be achieved by trying it for available datasets. We have planned to use results obtained to use for lip-reading application as future enhancement.

ACKNOWLEDGEMENTS

We would like to thank everyone in my friends, staff members and students who helped me in recording videos.

REFERENCES

[1] Tsuhan Chen and R. R, Rao, "Audio-visual interaction in multimedia communication," in Proc. Of IEEE Intern. Con$ on Acoustics, Speech and Signal Processing (ICASSP'97), vol. 1, 1997, pp. 179-182.

[2] X. Zhang and R. M. Mersereau, M. Clements, C. C. Broun, "Visual Speech Feature Extraction for Improved Speech Recognition," in Proc. IEEE Intern. Con$ on Acoustics, Speech, and Signal Processing (ICASSP'O2), vol. 2, 2002, pp. 1993-1996.

[3] D. Lo, R. A. Goubran, J. Gammal, G. Thompson, and D. Schulz, "Robust Joint Audio-Video Localization in Video Conferencing using Reliability Information," in Proc. 20th IEEE Instrumentation and Measurement Technology Conf: (IMTC'O3), vol. 2, May 2003, pp. 141-418.

[4] K. L. Sum, W. H. Lau, S. H. Leung, A. W. C. Liew, and K. W. Tse, "A New Optimization Procedure for Extracting the Point-based Lip Contour Using Active Shape Model," in Proc. Of IEEE Intern. Con. on Acoustics, Speech and Signal Processing (ICASSP 'Ol), vol. 3, May 2001, pp. 1485-1488.

[5] A. W. C Liew, S.H. Leung, and W.H. Lau, "Lip Contour Extraction Using a Deformable Model," in Proc. of IEEE Intern. Con. on Image Processing (ICIP'OO), vol. 2, Sept. 2000, pp. 255-258.

[6] R. Kaucic and A. Blake, "Accurate, Real-time, Unadomed Lip Tracking," in Proc. of 6th International Conference on Computer Vision, 1998, pp. 370-375.

[7] T. Coianiz, L. Torresani and B. Caprile, "2D Deformable Models for Visual Speech Analysis", Proceedings of Springer, Speech reading by Humans and Machines, D.G. stork & M. E. Hennecke Eds., NY, 1996.

[8] X. Zhang and R.M. Mersereau, "Lip Features extraction Towards an Automatic Speech-reading System", Proceedings of ICIP00, Wa07.05, Vancouver, Canada,2000.

[9] M. Heckmann, F. Berthommier and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition", Proceedings of EURASIP J. Appl. Signal, vol. 2002, pp. 1260-1273, Nov. 2002.

[10] J. Chaloupka, "Automatic Lips Reading for Audio-Visual Speech Processing and Recognition", Proceedings of Of ICSLP, pp. 2505-2508, Jeju Island, Korea, Oct. 2004.

[11] N. Eveno, A. Caplier, P.Y. Coulon. "A new color transformation for lips segmentation", Proceedings of IEEE Fourth Workshop on Multimedia Signal, pp. 3-8, Cannes, France, 2001.

[12] N. Eveno, A. Caplier and P. Coulon, "Accurate and Quasi-Automatic Lip Tracking", Proceedings of IEEE Trans. Circuits Syst. Video Techn. 14(5): 706-715., 2004.

[13] A.W.C. Liew, S.H. Leung, and W.H. Lau, "Segmentation of color lip images by spatial fuzzy clustering" Proceedings of IEEE Trans. Fuzzy Syst. vol. 11 no. 4, pp. 542-549, Aug. 2003.

[14] Y. P. Guan, "Automatic Extraction of Lip Based on Wavelet Edge Detection", Proceedings of the Eighth international Symposium on Symbolic and Numeric Algorithms, Scientific Computing SYNASC. IEEE Computer Society, pp. 125-132. Sept. 2006.

[15] S. Lee and D. Yook, "Viseme recognition experiment using context dependent Hidden Markov Models", Proceedings of Third Intl. Conf. on Intelligent Data Engineering and Automated learning, Vol. 2412 pp. 557-561, 2002.

[16] N. Dave, N. M. Patel. "Phoneme and Viseme based Approach for Lip Synchronization.", International Journal of Signal Processing, Image Processing and Pattern Recognition. 7(3), pp. 385-394, 2014.

Authors

Namrata A Dave
Assistant Professor,
Computer Engineering Department,
G H Patel College of Engg. & Tech.
GUJARAT, INDIA.