Indian Journal of Science and Technology, Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/98737, August 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Lip Detection and Lip Geometric Feature Extraction using Constrained Local Model for Spoken Language Identification using Visual Speech Recognition

Aparna Brahme[1*] and Umesh Bhadade[2]

[1]Department of Information Technology Engineering, MET's Institute of Engineering, Nashik - 422207, Maharashtra, India; mrsnrkale@gmail.com
[2]Department of IT Engineering, SSBT's College of Engineering and Technology Bhambori, Jalgaon – 425001, Maharashtra, India; umeshbhadade@rediffmail.com

Abstract

Background/Objectives: The aim of our research is to guess the language of spoken utterance by using the cues from visual speech recognition i.e. from movement of lips. The first step towards this task is to detect lips form face image and then to extract various geometric features of lip shape in order to guess the utterance. **Methods/Statistical Analysis:** This paper presents the methodology for detecting lips from face images using constrained local model (CLM) and then extracting the geometric features of lip shape. The two steps involved in lip detection are CLM model building and CLM search. For extracting lip geometric features, twenty feature points are defined on lips and lip height, width, area are defined using these twenty feature points. **Findings:** CLM model is build using images from FGnet Talking face video database and tested using images from FGnet Talking face video database and also using other images. The detection accuracy is more for FGnet images as compare to other images. Feature vector defining the lip shape consists of geometric parameters like height, width and area of inner and outer lip contours. Feature vector is calculated for all test images after detecting lips from face image. So the error in detecting lips leads to the error in feature vector. This indicates the speaker dependency of visual speech recognition systems. **Application/Improvements:** The proposed approach is useful in visual speech recognition for lip detection and feature extraction. Minimizing the speaker dependency and generalizing the approach should be considered for further improvements.

Keywords: CLM, Lip Detection, Language Identification, Visual Speech

1. Introduction

Automatic Language Identification (LID) is the task of recognizing a language of a spoken utterance by a computer**.** Language identification finds many applications in multi-lingual services. An example is the language identification system used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language.

Automatic visual language identification (VLID) is the technology which makes use of visual cues derived from movement of the speech articulators (lip movements) to identify the language of spoken utterance, without using any audio information[1].This technique is useful particularly in noisy environments where audio signal available is very weak or no audio signal is available at all. In our paper[2] an overview of spoken language identification, various language identification cues and basic frame work of visual language identification is discussed. According to the proposed frame work the first task is feature extraction from videos of speech articulators i.e. movements of lips. For this it is necessary to detect Lips from the frames of videos containing face images.

This paper discusses the lip detection using constrained local model. In Section 2 various methods of lip detection are discussed. In Section 3, CLM model build-

*Author for correspondence

ing and searching is discussed. Section 4 gives landmark definition and lip geometric feature definitions. Results are summarized in Section 5. In Section 6 conclusion and future scope from the study of the topic is discussed.

2. Methods of Lip Detection

The various methods for lip detection/localization can be classified as model based and image based techniques.[3,4]

2.1 Model-based Lips Detection Methods

These includes active contour models (Snakes) , Active Shape Models (ASM), Active Appearance Models (AAM).[5-8] In these approaches training set images are used to build lip models and then these models are used to search the lips in new images.

2.1.1 Active Contour Models (Snakes)

Snakes are generally used for shape detection and make use of deformable templates (Splines). Figure1. Shows lip detection using snakes.

The optimal fit between a snake and the detected shape is found by minimizing the energies in the following equation:

$$E_{snake} = {}_0\int^1 (E_{int}\, s(i) + E_{img}\, s(i) + E_{con}\, s(i))di$$

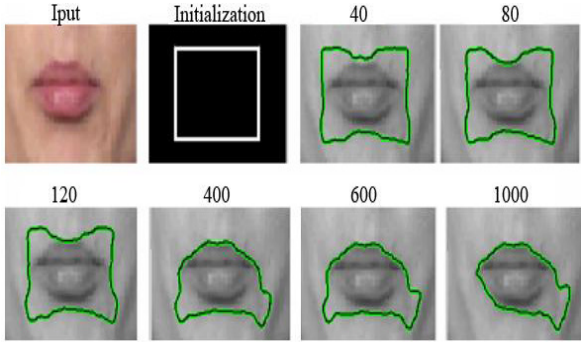

Figure 1. Lip Detection using Snakes.

According to Ahmad B. A. Hassanat, [4] snakes have reported good result but some of the problems using Snakes are as follows:

Initial position far away from the lip edges may results in fitting to the wrong feature, such as the nose or the chin. Also they are affected by facial hair (moustache and beard).As snakes do not bend easily, it becomes difficult to locate sharp curves e.g. corners of the mouth. Sometimes it may take long time or more iteration for tuning of its parameters.

2.1.2 Active Shape Models

ASM are statistical models of the shapes of objects.[9-11] The shape of the object is defined by set of labeled landmark points as shown in Figure.2.

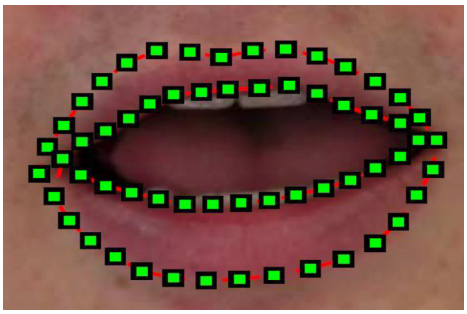

Figure 2. ASM landmark point.

Each landmark point is represented by its x and y coordinates X= {(x1, y1), (x2, y2), …., (xn,yn)}. A training set of a land-marked object in images, is used for building statistical shape model using principal component analysis (PCA). Eigen values and Eigen vectors of a covariance matrix describes the modes of variation in the shape of an object from mean shape.

2.1.3 Active Appearance Models (AAM)

An active appearance model describes the variation in gray level of the object along with the shape information. T.F. Cootes presented AAM algorithm and its performance in[7]. In ASM models of the image texture around the landmark points is used, whereas in AAM, model of the image texture of the entire region is used. AAM iteratively adjust to minimize the difference between the synthesized model image and the target image. In [9] I. Matthews et.al compared some approaches of visual feature extraction for lip-reading using a principal component analysis of shape or both shape and appearance.

2.2 Image-based Lips Detection Methods

These are based on the fact that there is difference between the color of lips and color of skin and hence also referred as color based methods[12, 13].The lip and skin colors are discriminated using a transformation based on RGB color space, hue of HSV color system or red and blue components in YCbCr color system.

2.2.1 RGB Approach

Red (R), green (G) and blue (B) are primary colors. Any color can be produced by variable combinations of these primary colors. The primary colors can be added to produce the secondary colors of light magenta, cyan, and yellow. Skin and lip pixels have different components in RGB space. The red is dominant in both skin and lips. Skin has more green components than blue and appears more yellow than lips. Red, green and blue components are linearly combined to transform the image in RGB space and then filtered using a high pass filter. Then both the images are converted to binary image. The largest area in the binary image is recognized as the lips.

2.2.2 HSV Approach

In HSV color system, 'H' represents Hue i.e. dominant color perceived by someone. 'S' represents saturation, the amount of white light mixed with hue and 'V' represents brightness or intensity.

There are specific equations to convert from RGB to HSV and vice versa. Hue can be used for discrimination between lip and face as hue value of the face pixels is more than that of the lip pixels.

2.2.3 YCbCr Approach

This color space is used in digital videos, where Y is the luminous component; Cb and Cr are the blue difference and red-difference chroma components respectively. Lips are more red than faces and have high Cr and low Cb values. YCbCr approach is based on this fact. So Cr component is maximized using a specific equation and Cb is minimized. After using edge detection as a mask to remove any unnecessary information, the image output of the equation is threshold, and the largest area is found to be the mouth area.

Model-based lip-detection methods needs a significant amount of processing time, and training time which makes them difficult to be applied for the online systems, or to be applied on low resources machines, like the PDAs. The implementation of ASM and AAM is always difficult to run in real time.AAM is slower than ASM. On the other hand image based methods though computationally efficient do not give satisfactory performance on images with weak color contrast. The drawback of color based approach is that it is Vulnerable to variations in light conditions.

3. CLM : Constrained Local Model

Given a face image detecting specific points on face, such as eye, mouth, Lips, Nose is a difficult task. A real person use models of eyes, nose etc to search in local region around where the corresponding item might appear and also use knowledge of shape of a face to constrain the search.Figure.3 illustrates this concept.

To search individual items such as nose, eyes etc, their local models are used and the search is constrained using the shape model. Hence the name is Constrained Local Models (CLM).A Tutorial by Xiaoguang Yan[14] describes constrained Local Model for Face Alignment in detail. CLM implementation consists of two steps, first is CLM model building from training images and second is CLM search in new images.

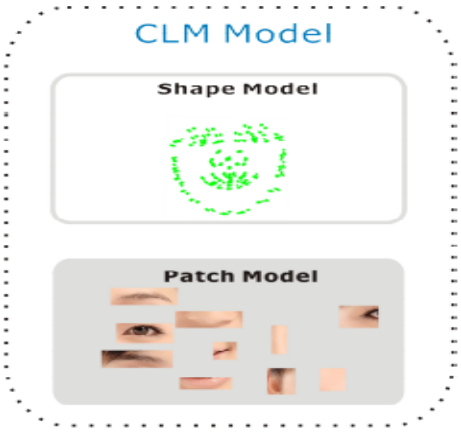

Figure 3. CLM shape and patch model.

3.1 CLM Model-Building

Model-building is a training phase.CLM model contains two models, shape model and patch model. **Shape model** is build using PCA and describes variation in shape of feature points. Procrustes analysis is used to remove translation, scale and rotation as a preprocessing step before applying PCA. Image texture around each feature point is described by **Patch model.**

3.2 CLM Search Process

CLM model build can be used to find the position of nose, eyes and mouth in a given face image. This is called CLM search process.The steps in the search process are as follows

1. Make initial guess of feature point position.

2. For each feature point, obtain SVM response image by searching in the local region of feature point using SVM.

3. Fit each response image with a quadratic function.

4. Find best feature point position by optimizing quadratic functions and shape constraints.

5. Repeat step 2-4 until converges.

4. Lip Geometric Feature Extraction

Landmarks and their definitions:

After detecting lips from the face image next step in automatic lip reading is feature extraction to capture speech related movements from frames of video. Typical features are height, width and area of inner and outer lips. In order to compute these parameters landmarks points are defined on the lip[15]. The lip model used consists of 20 feature points on inner and outer lip contour as shown in Figure. 4

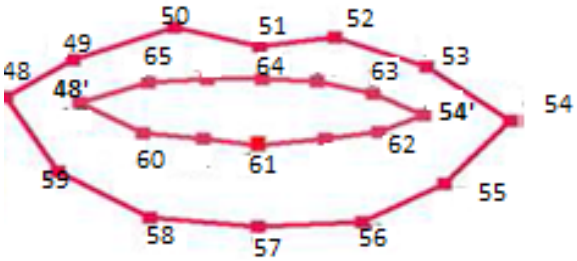

Figure 4. Model of Lip.

4.1 Outer Mouth Contour

The points on the outer mouth contour are defined as follows:

- Point 48 is left corner of the mouth.
- Point 54 is right corner of the mouth.
- Point 51 is at middle of upper lip where philtrum meets the upper lip.
- Points 50 and 52 are placed at the foot of the philtral columns.
- Points 56, 57 and 58 are placed on lower lip corresponding to upper lip points 52, 51 and 50, respectively.
- Points 49, 53, 55, and 59 are placed such that entire lip area is covered. These points are placed at equal distances from their neighboring points.

4.2 Inner Mouth Contour

Inner mouth counter should describe the opening of mouth accurately. The points on the inner lip contour are similar to the ones on the outer lip contour.

- Point 48' is left corner of inner lip and coincides with point 48.
- Point 54' is right corner of inner lip and coincides with point 54.
- Points 61 and 64 correspond to points 51 and 57 below philtrum. All 4 points lie on the same line.

The remaining pairs of points 62 and 63, 60 and 65 are placed at equal distances so as to cover entire inner lip area.

4.3 Geometric Feature Definition

Shape of the mouth is described using following geometric features:

- **mouthht** is height of outer lip contour. It is the Euclidian distance between points 51 and 57
- **mouthwd** is width of outer lip contour. It is the Euclidian distance between points 48 and 54.
- **outliparea** is defined as the area of mouth covered by the outer lip contour.
- **inlipht** is height of inner lip contour .It is Euclidian distance between the points 61and 64.
- **inlipwd** is width of inner lip contour. It is Euclidian distance between the left corner (48') and right corner (54') of the inner lip contour.
- **inliparea** is the area of mouth covered by inner lip contour.
- Parameters inlipht, inlipwd and inliparea indicates the opening of the mouth.

5. Results

Implementation of Constrained Local Model for Face Alignment is modified for detection of lips and lip geometric feature extraction. The annotations for modeling the lips are used from FGnet Talking face video database. It is tested using images from FGnet Talking face video database and also using other images. Figure 5 shows the initial position during search and Figure 6 shows the position of lips after convergence for FGnet database images. Figure 7 shows initial position and Figure 8 shows position after convergence for image other than FGnet dataset images. The geometric features of the lips as explained in section 4 are summarized in Table 1 and Table 2.

Figure 5. Initial Position (FGnet dataset image).

Figure 6. After Convergence (FGnet dataset image).

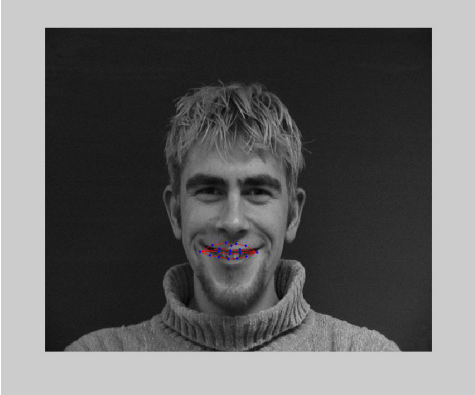

Figure 7. Initial Position (Other than FGnet dataset image).

Figure 8. After Convergence (Other than FGnet dataset image).

6. Conclusion and Future Scope

Visual speech recognition is the process of automating the human ability to lip read. Visual speech is transcribed using Visem i.e. visual phoneme, is the visual appearance of speech articulators (Lips) involved in speech. So the first step towards spoken language identification using VSR is the detection of lips. Further step is feature extraction to guess the utterance by recognizing the corresponding visem. Visem recognition uses the geometric shape information of lips such as height, width, area etc. To calculate these geometric features it is necessary to define feature points on lips.

Various methods of lip detections are discussed in section 2 with their merits and demerits. Color based methods though simple requires some method to define feature points on detected lips for the purpose of finding geometric shape parameters. On the other hand Active shape models make use of labeled landmark points to build statistical shape model. As each landmark point defines X and Y coordinates of the feature points, geometric shape parameters can be defined in terms of these coordinates.

Constrained local model uses Shape Model which is build using landmark points that defines the shape of lips and patch model that defines texture of image around each feature point i.e. patch. So we get the feature points on lips while detection which will be further useful in extracting features like height, width, opening of lips etc. to be used for visual feature extraction.

Section 5 shows the results obtained. From Figures 5-8 it is observed that lips detected after convergence is more accurate for images from FGnet dataset as compare to other images. This is because the model is trained using FGnet dataset. Feature vector is calculated for all test images after detecting lips from face image. So the error in detecting lips leads to the error in feature vector. This indicates the speaker dependency of the performance of visual speech recognition systems. So in order to generalize this approach there is need to develop training dataset which incorporate images of lips of different speakers with different poses. Developing this dataset requires manually labeling the landmark points on each image frame. So this research further concentrates on developing such dataset.

The further step in this research of spoken language identification using VSR is to track the movement of lips by observing the difference between various feature points defined on lips and thereby guessing the possible utterance.

Table 1. Results with FGnet dataset images

Test File	Mouth ht	mouthwd	inlipht	Inlip wd	Outlip area	Inlip area
franck_02599.jpg	30.32	102.99	11.04	102.99	2292.34	688.77
franck_02814.jpg	15.94	103.06	3.07	103.06	1591.84	229.08
franck_02542.jpg	24.49	107.03	6.19	107.03	2044.17	489.13
franck_02301.jpg	23.65	101.37	6.28	101.37	1852.63	431.00
franck_02239.jpg	26.08	106.60	7.04	106.60	2033.50	461.97
franck_04053.jpg	25.63	106.74	7.26	106.74	2113.81	550.22
franck_04138.jpg	24.97	105.82	6.91	105.82	2467.84	712.76
franck_04194.jpg	20.60	115.53	5.38	115.53	1930.58	382.13
franck_04277.jpg	23.32	102.84	6.71	102.84	1981.57	549.01

Table 2. Results with other images

Test File	Mouth ht	mouthwd	inlipht	Inlip wd	Outlip area	Inlip area
01-1m.jpg	35.78	98.34	16.73	98.33	2264.42	729.93
01-2m.jpg	22.15	100.99	6.09	100.99	1717.21	485.28
07-3m.jpg	22.87	85.13	6.39	85.13	1484.71	353.14
05-2m.jpg	20.32	92.99	6.26	92.99	1519.22	430.63
02-3m.jpg	18.07	80.54	4.29	80.54	1131.60	240.30
08-2f.jpg	16.83	87.15	3.63	87.15	1067.38	181.54
19-2m.jpg	21.65	85.75	8.07	85.75	1439.88	371.47

7. References

1. Jacob L. Newman and Stephen J. Cox. Language Identification Using Visual Features. IEEE Transactions on audio, speech, and language processing. 2012 Sep; 20(7):1936–7.
2. Kale N, Bhadade US. An overview of spoken language Identification using Visual Cues from Speech. Cyber Time International Journal of Technology and Management. 2014 Apr; 7(2):219–25
3. Hassanat ABA. Visual Speech Recognition, Speech and Language Technologies. IvoIpsic editor. ISBN:978-953-307-322-4,InTech. Available from: http://www.intechopen.com/books/speechandlanguagetechnologies/visual-speech-recognition
4. Hassanat ABA, Jassim S. Color-based Lip Localization Method. Proceedings of SPIE- The International Society for Optical Engineering. 2010 Apr; 7708.
5. Cootes TF, Hill A, Taylor CJ, Haslam J. The use of active shape models for locating structures in medical images. J Image Vis Comput. 1994; 12(6):355–66
6. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models their training and application. J Comput Vis Image Underst. 1995; 61(1):38–59
7. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. Proc European Conf Comput Vis. 1998 Jun. p. 484–98.
8. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour model. Int J Comput Vis. 1987; 1:321–31.
9. Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R. Extraction of visual features for lip-reading. IEEE Trans Pattern Anal Mach Intell. 2002; 24(2):198–213
10. Cootes TF, Taylor CJ. Active Shape Models - Smart Snakes. Proc. British Machine Vision Conference, Springer-Verlag. 1992. p. 266–75.
11. Mehrotra H, Agrawal G, Srivastava MC. Automatic Lip Contour Tracking and Visual Character Recognition for Computerized Lip Reading. International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering. 2009 Apr; 3(4):664–73.
12. Eveno N, Caplier A, Coulon PY. New color transformation for lips segmentation. Proc IEEE 4th Workshop Multimedia Signal Proc, France. 2001; 3–8.
13. Wark T, Sridharan S, Chandran V. An approach to statistical lip modeling for speaker identification via Chromatic Feature Extraction. Proc 4th Intl Conf Pattern Recognition. Brisbane, Australia. 1998. p. 123–5.
14. Yan X. Constrained Local Model for Face Alignment, a Tutorial. Available from: http://sites.google.com./site/xgyanhome/home/projects/clm-implementation.Date Accessed: 28/08/2015
15. Chitu AG, Rothkrantz LJM. Visual Speech Recognition Automatic System for Lip Reading of Dutch. Journal of Information Technologies and Control. 2009; 3:2–9.