IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-5 E-ISSN: 2347-2693

Lip Localization and Visual Speech Recognition with Optical Flow in Hindi

L.V.S. Raghuveer[1*], Divya Deora[2],

[1]Dept. Electronics and Communication, School of Electronics, Lovely Professional University, Jalandhar, India
[2]Dept. Electronics and Communication, School of Electronics, Lovely Professional University, Jalandhar, India

[*]*Corresponding Author: lvsrv9@gmail.com, Tel.: +91-85006-17758*

Online Available at: www.ijcseonline.org

Received: 17/Apr/2017, Revised: 28/Apr/2017, Accepted: 20/May/2017, Published: 30/May/2017

***Abstract*—** Current era is to make the connection amongst humans and their manufactured accomplices (Computers) and make communication more reliable and easier. One of the real challenges is the utilization of speech recognition. Speech recognition can be improved by visual information of human face. Visual speech recognition (Lip reading) assumes a fundamental part in automatic speech recognition and is an essential stride towards exact and robust speech recognition. In this paper, the technique is developed for visual speech recognition in detail. Optical Flow component is used to extract the feature vector and Artificial Neural Networks (ANN) for training. The effect of variation in velocity of speaking on the execution of the system is minimized by eliminating the zero energy frames and normalizing the number of frames. The efficiency of both approaches (Optical Flow and ANN) is used to evaluate words individually. Considered words are numerical numbers in Indian language (Hindi) from zero to nine, such as ek, do, theen, and so on.

***Keywords*—**Visual speech recognition, lip localization, Optical Flow, ANN, Indian Language

I. INTRODUCTION

Command based systems are helpful as a natural interface for clients to interact and control computers. Such systems give more adaptability than the flexible interfaces, for example, keypad and mouse. Lip reading is utilized to interpret or understand speech without hearing it, a technique especially used by people with hearing problems. The capacity to lip read empowers a man with a hearing impedance to speak with others and to participate in daily activities, most recent advances in the fields of computer vision, pattern recognition, and signal processing have prompted to a developing enthusiasm for automating this testing undertaking of lip reading. Indeed, automating the human ability to lip read, a process referred to as visual speech recognition (VSR), could open the door for other innovative applications [1,2,3,4,5,6].

VSR has obtained a lot of consideration in the recent decade for its potential use in applications, for example, human-computer interaction (HCI), audio-visual speech recognition (AVSR), speaker recognition, sign language recognition, talking heads and video surveillance. Its fundamental point is to understand spoken words by using visual information that is created amid a speech. Hence, VSR manages the visual domain of speech and involves image processing, artificial intelligence, [5] object detection, pattern recognition.

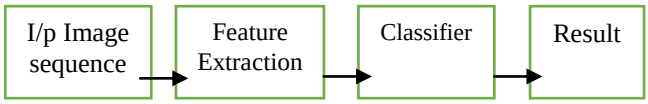

Figure: 1 Visual speech Recognition

Visual features are extracted by visual information from lip moment and recognized to get more precision and strong outcome in speech recognition. Visual Features are like eyes, lip shapes, and mouth region from the conjugative frames and classifier compares the extracted features from the dataset and produce a final result. [1,7,8,9,10]

II. LITERATURE REVIEW

There are two distinctive fundamental approaches to the VSR issue, the visemic (visual phonemes) [10] approach and the word pattern identification, each method has its own qualities and defects. The traditional and most common ways to deal with automatic lip reading are based on visemes. Viseme is the mouth shapes or appearances of mouth dynamics that are required to produce a phoneme in the visual area. However, while using visemes several problems arise in visual speech recognition systems [3], such as the low number of visemes (10 to 14) contrasted with phonemes (45 to 53). Viseme covers a little subspace of the mouth movements of spoken word. [7] These issues add to the terrible execution of the conventional methodologies hence, the visemic approach is something like digitizing the signal of the spoken word, and digitizing causes losing data.

Another approach is considered the whole word as an only the one part. It can give a decent alternative option to the visemic approaches to deal with automatic lip reading. It can be, train the entire of the dictionary words or to train (in any event) the definite ones. This approach can be effective if it is trained in a specific area of words, e.g. numbers, pin codes, cities, etc. The precision of an automatic lip reading system is extremely dependent on exact lip localization and in the robustness of the extracted features. Existing methods for feature extraction can be classified as [3]

1. Geometric based features: Geometric information extracts from the region of mouth like mouth shape Width, Hight, orientation, and area. [6]
2. Appearance-based features: compare conjugative frames of pixel values at same position. mostly apply on mouth region of interest (ROI) such as Optical Flow and PCA etc. [11]
3. Image Transformed approaches: Feature extraction using Image Transform techniques of the mouth region such as DFT, Wavelets, and DCT. [4]

III. METHODOLOGY

A typical lip reading system consists of three major stages: lip localization, feature extraction, and the final step is classification. Fig. 1 demonstrates the significant stages in the proposed lip reading process.

A. Data set:

Video information was recorded by a normal camera in a room. The camera focused on the face region of the speaker and was kept constant all throughout the recording. Camera position, illumination, and background were kept steady for every speaker. In his experiment, 20 subjects (10 females, 10 males) were utilized. From every speaker recorded 10 numbers were recorded at a sampling rate of 30 frames/sec and each word was recorded twice to generate sufficient variability.

B. Lip Localization

Viola jones algorithm is used to detect face and face parts such as nose, mouth, and eyes. Haar function generates number of features. Threshold value is used to avoid the unnecessary features; this threshold is known as meagre threshold. The merge threshold value is varying for person to person to detect mouth region because of mouth region has dynamic nature, but the nose region is not dynamic so merge threshold is consistent. So easily identify the nose with proper meagre threshold.

Figure: 2 Identifying nose region using viola jones algorithms.

In this paper, a novel approach is proposed to detect mouth region in various brightness conditions and different skin colours. After detecting the nose region using Viola-Jones algorithm [12], easily identify the rectangle ROI of mouth with the help of pixel positions. Viola jones algorithm works with the combination of different algorithms. The steps are showed in figure 3.

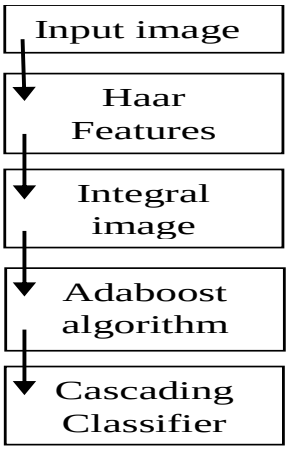

Figure: 3. Flow chart for Viola-Jones algorithm

C. Feature Extraction

Figure (4) represents feature extraction of visual speech recognition. The proposed method consists of automatic temporal segmentation and normalized frames (i.e. started to ending frames) of isolated word. This is obtaining by pair wise comparison (also called template matching) method which extracts the difference in intensity of same pixels in two conjugative frames. The primary step is to convert the colour images into binary images of mouth region frames and compute the mean difference of conjugative frames [1]. The result signal is smoothed using Gaussian filter and linear

interpolation. Utilizing a proper threshold to get required temporal segmentation.

Mean Square Error =

$$\frac{1}{m \cdot n}\sum_{x=I}^{m}\sum_{y=1}^{n}[(I,(x,y_1)-(x_2y_2)I_2)^2 \quad (1)$$

Optical flow [8] [2] is the measure of moving objects (object-tracking) in video data. It measures the spatio-temporal difference between two successive images.

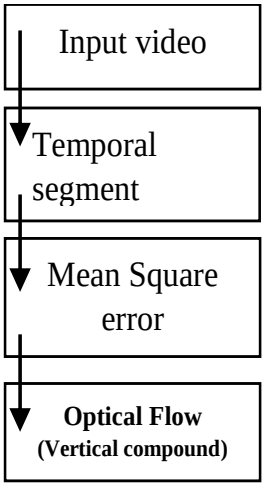

Figure: 4 Feature Extractions

For two-dimensional case consider pixel location and along with time and intensity I (x, y, t) are moved by $\delta x, \delta y, \ in \ \delta t$ in between two conjugative frames.

$$I(x,y,t) = (x+\delta x, y+\delta y, t+\delta t) \quad (2)$$

Simplifying the above equation (2) by Taylor series

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t}+\frac{\partial I}{\partial y}\frac{\delta y}{\delta t}+\frac{\partial I}{\partial t}\frac{\delta t}{\delta t}=0 \quad (3)$$

Let

$$\frac{\delta x}{\delta t}=u \ and \ \frac{\delta y}{\delta t}=v$$

Where x, u and y, v are consistent terms of velocity of optical flow (OP) of $I(x,y,t)$ and $\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \ and \ \frac{\partial I}{\partial t}$ are derivative terms if image derivative terms can also write as into I_x, I_y and I_t substitute these values in equation (3) then

$$I_x u + I_y v = -I_t \quad (4)$$

Equation (4) have two unknown terms (u and v) only, another set of equations is required to find the optical flow. Consider spatial smoothness and brightness consistency essential, Optical flow estimation techniques are another set of equations. Both the brightness and the spatial smoothness constancy are introduced by Horn and Schuck for optical flow estimation. However, the quadratic formulation assumes Gaussian statistics and not robust caused by reflection, motion, and boundaries. Multi resolution technique is used to estimate larger displacement. The optical flow evaluated at a coarser level is utilized to wrap the second image toward the first at the following better level and the flow increment is measure between the first images and wraps the second image. Finally combine the increment flow of all levels

D. Classifier

Artificial Neural Networks (ANN) is used for classification of feature vectors, ANNs are accumulations of little individual interconnected preparing units. Data is passed among these units along interconnections. ANN learn the relation between the output and input.it contains mainly three-layer input layer, hidden layers (No. of hidden layers depend on input and output layers) and an output layer. Connected weights are multiples each input.it simplifies the products and summed, then send through a transfer function to generate a result, and finally, the output is obtained. Training is a most important part; Back-Propagation (BP) is a famous algorithm in ANN.

IV. RESULTS

This section Deals about outcomes of the experiments. Considered as input 10 Male and 10 Female persons, each person speaks 10 numerical numbers in Hindi such as shoony, eak, do, theen, chaar, panj, Chee, saat, aath, nau. Totally 200 samples are considered (20persons*10numbers). Results are classified in gender wise and numerical number wise.

Figure: 5 Input Frames

Figure: 6 Lip Localization from input video

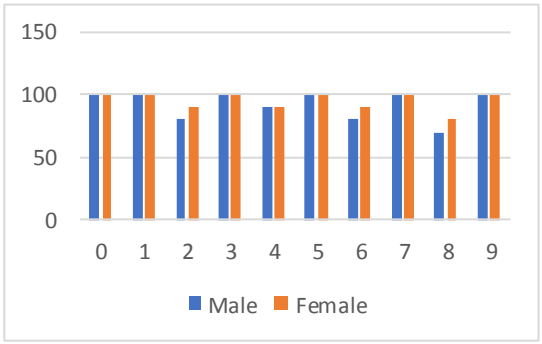

Figure: 7 Result analysis gender wise and word wise

Propose method has been 93.5% (Average) accuracy. Both genders shown different accuracy in result. Male has been shown 92% and female gender shown 95%. Figure: 6 represent each word and both genders results.

V. CONCLUSION

This paper reports visual speech recognition based on lip moments catching using Optical flow and ANN. The test comes about show that the inner and intra-subject speed of speech difference can be overcome through normalization utilizing by MSE and linear interpolation, and speech recognition recognized by using an optical flow of horizontal component. Feature extraction is mainly depending on using fixed size non-overlapped in the column. Extracted feature vectors are classified using ANN. The results indicate that the reported technique can produce very high success rates. Average accuracy of 93.5 % has been obtained. Such a system may be connected to drive computerized machinery in noisy places and can be used for the rehabilitation of hairless people.

REFERENCES

[1] Bor-Shing Lin, Yu-Hsien Yao, Ching-Feng Liu, Ching-Feng Lien, Bor-Shyh Lin, "*Development of Novel Lip-reading Recognition Algorithm*", IEEE Access, vol. 5, no.1 , pp. 794-801, 2017.

[2] Jun Shiraishi, Takeshi Saitoh, "*Optical Flow based Lip Reading using Non-Rectangular ROI and Head Motion Reduction*", 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, pp. 1-6, 2015.

[3] Ahmad B. A. Hassanat, "*Visual Passwords Using Automatic Lip Reading*", IJSBAR, Vol.13, No.1, pp.218-231, 2013.

[4] SS. Morade, "*Suprava Patnaik Lip Reading Using DWT and LSDA"*, IEEE International Advance Computing Conference, India, pp.1000-1003, 2014.

[5] WR. Butt, L. Lombardi, "*A Survey of Automatic Lip Reading Approaches*", Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, pp. 299-302, 2013.

[6] SS. Morade, B.S. Patnaik, "*Automatic Lip Tracking and Extraction of Lip Geometric Features for Lip Reading*", International Journal of Machine Learning and Computing, Vol. 3, No. 2, pp.23-30, 2013

[7] JI. Newman, SJ. Cox, *"language identification using visual features*", IEEE transactions on audio speech and language processing, vol. 20, no.7, pp. 1936-1947, 2012.

[8] AA. Shaikh, DK. Kumar, WC. Yau, M. Z. Che Azemin, J. Gubbi, "*Lip Reading using Optical Flow and Support Vector Machines*", 3rd International Congress on Image and Signal Processing (CISP2010), India, pp.327-330, 2010.

[9] W.C. Yau, D.K. Kumar, S.P. Arjunan, "*Visual speech recognition using dynamic features and support vector machines*, International Journal of Image and Graphics, vol.8, Issue.3, pp. 419-437, 2008.

[10] Salah Werda, Walid Mahdi, Abdel Majid, "*Lip Localization and Viseme Classification for Visual Speech Recognition"*, International Journal of Computing & Information Sciences, Vol.5, No.1, pp.67-75, 2007.

[11] X. Hong, H. Yao, Y. Wan, R. Chen, "*A PCA based visual DCT feature extraction method for lip-reading*", in Proc. Int. Conf. Intell Inf. Hiding Multimedia Signal Process, CA, pp. 321-326, 2006.

[12] T. Chen, R.R. Rao, "*Audio-Visual Integration in Multimodal Communication*", Special Issue on Multimedia Signal Processing, IEEE Proceedings, vol. 86, Issue.5, pp. 837-852, 1998.

Authors Profile

Mr. L V S Raghuveer pursed Master of Technology from Lovely Professional University, India in 2016. He is currently and currently working as Assistant Professor in Department of Electronics and Communication in JNTU, His main research work focuses on Image processing, Signal processing. He has 1 years of teaching experience and 1 years of Research Experience.

Mis. Divya Deora pursed Master of Technology from Lovely Professional University, India in year 2013. She is currently working as Assistant Professor in Department of Electronics and Communication, Lovely Professional University, India since 2013. Her main research work focuses on image processing, meachin learning algorithems. She has 4 years of teaching experience and 3 years of Research Experience.