IJARIIT
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
(Volume2, Issue5)
Available online at: www.Ijariit.com

An Improved Visual Recognition of Letters of English Language Using Lip Reading Technique

Ishu Garg[#1]
Research Scholar
Punjabi University of Information and Technology
ishugarg1002@gmail.com

Amandeep Verma[#2]
Assistant Professor
Punjabi University of Information and Technology
vaman71@gmail.com

Abstract: Two phenomena have determined the emergence of a new research field. First, the drop of costs involved in Electronically collecting and storing interpretation of the world has brought the need for sophisticated technique to handle the outcome data collection. The mapping from acoustic to visual information is the focus of this part of the thesis. The challenge is to produce adequately precise movements, in order to convey useful information to the listener in a real-time system with low latency.

Keywords— Lip-reading, image, machine learning, ANN.

I. INTRODUCTION

With the rapid development in computing technologies, the demand for enhanced elasticity and futuristic capabilities in human-computer interfaces (HCI) is increasing. Speech identification systems appear as a new generation of HCI that utilizes the natural ability of users to communicate with computers through voice commands. Speech identification can be usually viewed as a sample matching issue to find an assumes for spoken utterance given the speech signals.

Lip-reading is not affected by the noise. The visual information is used to improve the speech recognition. The audio features have main contribution and play an important role in speech however in some cases it is difficult to extract useful information from the audio.

Lip-reading is defined as the procedure of recognizing utterances by observing the speaker's lip movement. The term 'speechreading' more enlarge the meaning of lip-reading to cover for facial movement at time of speech (Campbell et al., 1998). Kaplan et al. (1999) defines speechreading as "the ability to understand a speaker's thoughts by watching the movements of the face and body and by using information provided by the situation and the language". Facial movements in the mouth region are known to include the most important amount of visual cues functional for speechreading.

Speech can be divided into sound segments known as phonemes. Phonemes are the smallest structural units of spoken language that differentiate meanings of words. Phonemes can be mostly classified into vowels and consonants depending on the relative sonority of the sounds.

Lip-reading is a visual proficiency based on the ability to recognize rapid lip movements. Speechreading relies on visual proficiency, visual discrimination and visual memory. Visual aptitude is defined as the capability to focus speedily and to be visually considerate to the speaker's face for long periods of time. Visual discrimination is the ability to differentiate the subtle differences in the speech articulators (lips, tongue and jaw) movements. Visual memory is associated with the capability to remember the visual sample of speech movements.

Computer-based lip-reading is a relatively new area and research in this field has only started two decades ago. The prime task of speech reading systems is to procedure the mouth images to attain meaningful features that can be used to correctly recognize utterances. The input videos are recorded from video capturing devices such as cameras. The video captured can be full face video or mouth video from various viewpoints, e.g., frontal view and profile view.

In the last decades, two phenomena have determined the emergence of a new research field. First, the drop of costs involved in electronically collecting and storing interpretation of the world has brought the need for sophisticated technique to handle the outcome data collections.
The mapping from acoustic to visual information is the focus of this part of the thesis. The challenge is to produce adequately precise movements, in order to convey useful information to the listener in a real-time system with low latency.

The real-time requirement arises from the intention to run the system on home computers with no extraordinary processing power. This limits the amount of calculations that can be performed per time unit without overloading the machine, and translates into a constraint on the complexity of the techniques employed.

Active Appearance Models (AAMs) are frequently used to represent shapes and textures in a compact form. The term AAM is most frequently used to define a composite model which consists of: the 2D shape — initially extracted from the image as x and y coordinates and to be represented by the Point Distribution.

In speech recognition, the language model is used to determine the most likely word sequence and can be used to discriminate between similar sounds in different contexts. In the majority of cases, a language model is used to model the probabilities of whole-word sequences. However, they can also be applied to sub-word units such as phonemes (phonotactics). The sparsity of available training data can be a problem when building a language model.

As the value of n in n-gram increases, the structure of the language is better defined because of the larger context, but the training procedure also requires a larger training set. Smoothing techniques are applied to a language model to remove some probability from higher frequency sequences and re-distribute probability mass across to other lower probability sequences. The term smoothing refers to the technique which attempts to adjust the probability distribution to be more uniform, increasing the much smaller probabilities and decreasing the high probabilities. The main objective in speech recognition is to find the word sequence, W, that maximizes P r(W|A) and A represents the acoustic signal.

Hybrid based approach is used for Lip feature extraction. Dynamic Time Wrapping method and linear interpolation is used for classification. Guoying Zhao et. al. (Zhao et al. 2009)In this paper, we propose an approach for lip reading, i.e. visual speech recognition, which could improve the human-computer interaction and understanding especially in noisy environments or for listeners with hearing impairments. A new appearance feature representation based on spatiotemporal local binary patterns is proposed, taking into account the motion of mouth region and time order in pronunciation. The local binary pattern (LBP) operator is a gray-scale invariant texture primitive statistic, which has shown excellent performance in the classification of various kinds of textures. For each pixel in an image, a binary code is produced by thresholding its neighbourhood with the value of the centre pixel A method for temporal texture recognition using spatiotemporal Local Binary Patterns extracted from Three Orthogonal Planes (LBP-TOP) was proposed .

With this approach the ordinary LBP for static images was extended to the spatiotemporal domain. If we do not consider the time order, the two phrases “see- you” and “you-see” would generate almost the same features. Spatiotemporal multi-resolution descriptors are introduced. Multi-resolution features can provide more information and improve the analysis of dynamic events and feature selection is done using AdaBoost to select more important slices (principal appearance and motion). A Support Vector Machine (SVM) classifier is utilized for recognition.

II. LITERATURE REVIEW

M. Subrahmanyam et al[1]: A new algorithm meant for content based image retrieval (CBIR) and object tracking applications is presented in this paper. The local region of image is represented by local maximum edge binary patterns (LMEBP), which are appraised by taking into deliberation the magnitude of local difference among the center pixel and its neighbours. This LMEBP differs from the existing LBP in an approach that it extracts the information based on allocation of edges in an image. Additional, the efficiency of our algorithm is confirmed by combining it with Gabor transform. Four experiments have been carried out for proving the worth of our algorithm. Out of which three are meant for CBIR and one for object tracking. It is more mentioned that the database considered for first three experiments are Brodatz texture database (DB1), MIT VisTex database (DB2), rotated Brodatz database (DB3) and the fourth acquires three observations. The outcomes after being investigated show an important development in terms of their appraisal measures as compared to LBP and other available transform domain method.

Yong-Ki Kim et al[2], investigate the discrimination of various features extracted from lip image data. A total of 90 pieces of data were collected as five subjects's uttered six isolated words three times. The results of speech recognition through two different feature generation methods showed mean recognition rates of up to 60%. Although the grid-based feature extraction method yielded higher recognition rates for certain isolated words, the highest recognition rate was found to be the coordinate-based feature of the combined vector of width/height ratio of the outer lip and the height of inner lip.

Lai Pai Mei [3], presents interpretation of alphabets by images of lips movement for native language (Mandarin) by means of building a system that can detect visual speech recognition and Matlab as the platform for face feature detection, mouth segmentation, pre-processing, thresholding and recognition. The aim of this project is to understand the overall process and components involved in visual speech recognition. In addition, it is done to understand and develop the suitable input file for neural network pattern recognition. Image acquisition is implemented whereby a video consists of lips movement of a person who utters alphabets or speech can be captured using camera. The video is then separated into frames and the best frame resembling the uttered speech are chosen and implemented with Matlab coding for face detection, face feature detection, lips detection, lips feature point. The images obtained are then pre-processed and segmented with aid of Photoshop CS4. Next, the preprossed image is then implemented with Matlab coding to convert the coloured image to grayscale image. Subsequently, the grayscale image is converted into binary image and then adjusted to different levels which are 100, 130, 160 and 190 from the range of 0-255 and these acts as input files which will be fed into neural network to produce the content in text for Mandarin.

Salah Werda et al [4], need for an automatic lip-reading system is ever increasing. Infact, today, extraction and reliable analysis of facial movements make up an important part in many multimedia systems such as videoconference, low communication systems, lip-reading systems. In addition, visual information is imperative among people with special needs. We can imagine, for example, a dependent person ordering a machine with an easy lip movement or by a simple syllable pronunciation. Moreover, people with hearing problems compensate for their special needs by lip-reading as well as listening to the person with whom they are talking. We present in this paper a new approach to automatically localize lip feature points in a speaker's face and to carry out a spatial-temporal tracking of these points. The extracted visual information is then classified in order to recognize the uttered viseme (visual phoneme). We have developed our Automatic Lip Feature Extraction prototype (ALiFE). Experiments revealed that our system recognizes 72.73% of French Vowels uttered by multiple speakers (female and male) under natural conditions.

Bregler et al. [5], used a modular time-delay neural network (TDNN) which consists of an input layer, a hidden layer, and a phone state layer. The network was trained by back-propagation. They have described another connectionist approach for combining acoustic and visual information into hybrid Multi layer Perceptron MLP/HMM speech recognition system. Given the audio-visual data MLP is trained to estimate the posterior probabilities of the phonemes. The likelihoods are obtained from the posterior probabilities and used as the emission probabilities for the HMM.

Nankaku and Tokuda [6], use the continuous density HMM method. The ordinary approach in the conventional normalization is to provide a criterion independently of the HMM, and to apply normalization before learning. In their approach, normalization by the ML (Maximum Likelihood) criterion is considered. Normalized training is proposed such a way that the normalization processes for elements such as the position, size, inclination, mean brightness, and contrast of the lips are integrated with the training of the model.

Potamianos and Matthews [7], proved that image approaches allowed better performances in terms of recognition rates than image-based approaches in different conditions. On the other hand for Even the most promising methods in labial segmentation, are those model-based approaches because they are based on lip models. He proceeds by detecting the full contour of lips, but it is necessary to doubt on the interest of the extraction of the labial contours in their entirety for the recognition stage. According to speech specialists the pertinent features of verbal communication expression are: the heights, widths and inter-labial surface. From this interpretation we notice that it will be judicious to opt for an extraction method of these features based on the detection and the tracking of some "Points Of Interest" (POI) sufficient to characterize labial movements. Therefore, the problem of labial segmentation is to detect some POI on the lips and to track them throughout the speech sequence.

Ayaz A. Shaikh et al [8], the proposed technique is based on the use of directional motion history images (DMHIs), which is an extension of the popular optical flow method for object tracking. Zernike moments of each DMHI are computed in order to perform the classification. The technique incorporates automatic temporal segmentation of isolated utterances. The segmentation of isolated utterance is achieved using pair-wise pixel comparison. Support vector machine is used for classification and the results are based on leave-one-out paradigm. Experimental results show that the proposed technique achieves better performance in visemes recognition than others reported in literature.

A lip-reading method has been proposed which can recognize the five widely used English vowels in a word. First, the images of speaker's utterance of different words collected and the images are converted into gray scale images. As this work use only visual features, main concentration is on lip movement and lip area of the images are automatically cropped as a region of interest.. In previous work of lip-reading, the key-points are selected manually but this work proposed an approach to select the key-points automatically. To train the network total 45 images are used, 9 images for each class and in testing phase total 135 images are used, 45 images for each class. The recognition accuracy achieved 87.57% with neural network and 97.7% with SVM Previously some work of lip reading done in using the HMM but this work is first attempt in English to distinguish the vowels using Euclidian method.

III. **Conclusions**

A lip-reading method has been proposed which can recognize the five widely used English vowels in a word. First, the images of speaker's utterance of different words collected and the images are converted into gray scale images. As this work use only visual features, main concentration is on lip movement and lip area of the images are automatically cropped as a region of interest. In previous work of lip-reading, the key-points are selected manually but this work proposed an approach to select the key-points automatically. To train the network total 45 images are used, 9 images for each class and in testing phase total 135 images are used, 45 images for each class. The recognition accuracy achieved 87.57% with neural network and 97.7% with SVM Previously, some work of lip reading done in using the HMM but this work is first attempt in English to distinguish the vowels using Euclidian method.

References

[1] Subrahmanyam, Murala, R. P. Maheshwari, and R. Balasubramanian. "Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking." *Signal Processing* 92.6 (2012): 1467-1479.

[2] Kim, Yong-Ki, Jong Gwan Lim, and Mi-Hye Kim. "Comparison of Lip Image Feature Extraction Methods for Improvement of Isolated Word Recognition Rate." Vol.107, pp.57-61 (2015).

[3]Lai Pei Mei, 'INTERPRETATION OF ALPHABETS BY IMAGES OF LIPS MOVEMENT FOR NATIVE LANGUAGE JUNE 2014

[4] Werda, Salah, Walid Mahdi, and Abdelmajid Ben Hamadou. "Lip localization and viseme classification for visual speech recognition." *arXiv preprint arXiv:1301.4558* (2013).

[5] C. Bregler and Y. Konig. Eigenlips for robust "speech recognition. In Proc. IEEE Int. Conf. on Acoust.", Speech, and Signal Processing, pages 669-672, Adelaide, 1994.

[6] Y. Nankaku, K. Tokuda. "Normalized Training for HMM-Based Visual Speech Recognition". Electronics and Communications in Japan, Part 3, Vol. 89, No. 11, 2006.

[7] G. Potamianos, H. P. Graft et E. Gosatto. An Image transform approach For HM based automatic lipreading". Proc, ICIP, Volume III, pages 173-177, Chicago, IL, USA Octb 1998.

[8] Shaikh, Ayaz A., Dinesh K. Kumar, and Jayavardhana Gubbi. "Automatic visual speech segmentation and recognition using directional motion history images and Zernike moments." *The Visual Computer* 29.10 (2013): 969-982.

[9] Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimed. 11(7), 1254–1265 (2009)

[10] Matthews, I., Cootes, T., Bangham, J., Cox, S., Harvey, R.: Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. 24(2), 198–213 (2002)