LIP READING: VISUAL SPEECH RECOGNITION USING LIP READING

Kunal Patil [1], Sandesh Patel [2], Harshad Rathod [3], Ashraf Siddiqui[4]

[1,2,3]*Student, Dept. of Computer Engineering, Universal College of Engineering, Maharashtra, India*
[4]*Professor, Dept. of Computer Engineering, Universal College of Engineering, Maharashtra, India*

Abstract - *When regular sound is unavailable, lip reading is a technique for comprehending speech by visually understanding the motions of the lips, face, and tongue. It also relies on information supplied by the context, linguistic understanding, and any residual hearing. Lip reading is a difficult skill for humans. To anticipate spoken words, it is necessary to have knowledge of the underlying language as well as visual cues. To interpret spoken words, experts must have a particular amount of expertise and awareness of visual expressions. It is now possible to translate lip sequences into meaningful words using deep learning. With visual information, speech recognition in noisy contexts can be improved. Deep learning is a type of machine learning method that employs numerous layers to extract higher-level characteristics from raw input. In image processing, for example, lower layers may recognise boundaries, while higher layers may identify concepts meaningful to humans, such as digits, characters, or faces. This study will look at the advancement of lip identification and recognition technology that outperforms previously offered methods considerably. Approach for automatic word identification and recognition based on varied lip motions. We will demonstrate a method for detecting and recognising human lip expressions.*

Key Words: **Deep learning, lip reading, neural networks, speech recognition, visual speech decoding.**

1. INTRODUCTION

It is common knowledge that humans comprehend speech not just by listening but also by seeing visual clues. Hearing-impaired people can understand human conversation exclusively by visual lip reading, which involves interpreting visual information from a speaker's lips and face. As a result, for decades, researchers have been interested in making lip reading available to electronic speech recognition and processing systems.

In recent years, the job of automated lip reading has gained a lot of academic attention, and significant improvements have been achieved in the domain, with a number of machine learning-based algorithms being deployed. Automated lip reading may be conducted with or without audio aid, and when performed without audio, it is sometimes referred to as visual speech recognition.

2. LITERATURE SURVEY

A neural network-based lip reading system is suggested in this study. The system lacks a language and relies only on visual clues. With only a few number of visemes to recognize as classes, the system is designed to lip read sentences with a wide variety of vocabulary and recognize words that may not have been included in system training. The system was put through its paces on the difficult BBC Lip Reading Sentences 2 (LRS2) benchmark dataset. [1]

Speech identification from visual-only recordings of a speaker's face is possible using a processing pipeline based only on neural networks, producing much higher accuracy than traditional approaches. Feed forward and recurrent neural network layers (specifically, Long Short-Term Memory (LSTM)) are layered to form a single structure that is trained by back-propagating error gradients across all layers. In this study, the performance of a stacked network was experimentally assessed and compared to that of a typical Support Vector Machine classifier using common computer vision characteristics. [2]

They cover the key strategies for audio-visual speech recognition that have been developed during the last two decades in this chapter. They start with the visual feature extraction challenge and then go on to audio-visual fusion. In both cases, they describe some of the strategies used during the Johns Hopkins summer 2000 workshop (Neti et al., 2000). They also explore the issue of audio-visual speaker adaptation, which is critical for establishing systems across databases or building speaker-specific models. They next go over the primary audio-visual corpora that have been utilized in the literature for ASR trials, such as the IBM audio-visual LVCSR database. Following that, they report experimental findings on automated speech reading and audio-visual ASR. They address the challenge of automated recognition of damaged speech as an application of speaker adaptation. Finally, they wrap off the chapter with a review of the present status of audio-visual ASR and what they see as unresolved issues in this field. [3]

This research demonstrates the risk of not employing a variety of speakers in the training and test sets. They provide categorization findings from AVletters 2, a high-definition version of the well-known AVletters database. They show that by carefully selecting features, it is feasible to achieve visual-only lip-reading performance that is very similar to audio-only recognition for single and multi-speaker configurations. [4]

They studied the performance of a machine-based lip-reading system with shape-only parameters as well as comprehensive shape and appearance parameters. They also compared the performance of a machine-based lip-reading system to that of a human lip-reader. The automatic method surpasses human lip-readers, they discovered. Surprisingly, when adding full appearance data to the machine-based approach, there is minimal gain in identification accuracy for relatively basic jobs, but human lip-readers show large improvements in performance. Finally, they tested the effect of 'speaker training' on human lip-reading skill and discovered that even very modest training is enough to increase performance. [5]

This project's purpose was to recognize words and sentences uttered by a talking visage, with or without audio. In contrast to prior efforts that concentrated on identifying a restricted number of words or phrases, they approach lip reading as an open-world problem — unrestrained natural language sentences and in-the-wild films. [6]

The creation and recording of a new Czech audio-visual speech database intended for studies on various impaired situations are presented in this study. They focused on adjusting the lighting. As a result, the audio-visual speech database UWB-07-ICAVR was created, which is a great resource for testing algorithms for visual speech parameterization in different lighting. [7]

They suggested a novel visual speech recognition system that makes use of both 2D pictures and depth maps obtained with RGB-D cameras. Their method is based on 3D rigid tracking of the speaker's face to extract image and depth thumbnails reliably. These are then utilized to generate motion and appearance descriptors, which are subsequently fed into an SVM classifier. They created a new dataset (MIRACL-VC) and utilized two other available datasets to evaluate their approach (CUAVE and OuluVS). The assessment findings acquired on these three datasets reveal that their system is very competitive with other recent research in the literature. They also demonstrated that depth data improves the LR system's performance considerably, and that the combination of appearance and motion descriptors improves the SI performance. [8]

In this study, they offer a method for distinguishing a target speaker's speech signal from background noise and other speakers by exploiting visual information from the target speaker's lips. They proved that the deep network can synthesize comprehensible speech from highly noisy audio segments captured in unconstrained 'in the wild' contexts by predicting both the phase and amplitude of the target signal. [9]

They explored how viewers identify languages from visual-only presentations of speech in this paper. Their initial objective was to reproduce and enhance the findings of SotoFaraco et al. (2007) utilizing a different experimental methodology and collection of languages. Overall, the results of Experiment 1 validated previous findings that language identification may be accomplished only through visual cues. Although they were unable to reproduce the impact of linguistic experience (monolingual vs. bilingual) on sensitivity, they did discover an effect on response bias; bilingual individuals displayed a larger bias toward their native language than monolingual speakers. This response bias was greatest among bilingual English speakers, but a similar pattern was observed among Spanish bilinguals, with a greater bias toward Spanish. [10]

They provided a completely automated technique to distinguishing native from non-native speech in English based only on visual cues in this study. Overall, the purpose of this work is to present a fundamental study of visual discrimination between native and nonnative speech, therefore establishing a research topic that can be immensely valuable in biometric applications. They show that important information for distinguishing between native and non-native speech is present in the visual stream, which is likely to boost performance when paired with audio-only approaches, particularly in loud conditions. [11]

3. PROPOSED SYSTEM

One of the most extensively investigated areas in affective computing and human–computer interaction is automatic speech recognition based on lip reading. In this research, we will offer a solution for automated word identification and recognition based on diverse lip motions. We will demonstrate a method for detecting and recognizing human lip expressions. The suggested system's goal is to provide a technique that automatically recognizes and classifies diverse human lip expressions. Automatic speech recognition (ASR) relies heavily on visual speech information, especially when audio is distorted or unavailable. Despite the success of audio-based ASR, visual speech decoding remains a major challenge. Our project's purpose is to recognize words and sentences uttered by a talking visage, with or without audio.

This research focuses on the advancement of lip recognition, which will greatly outperform previously presented systems. A video of a human's lip expression is obtained. A video to frames converter is used to convert the video to frames. As the input image, the average frame is used.

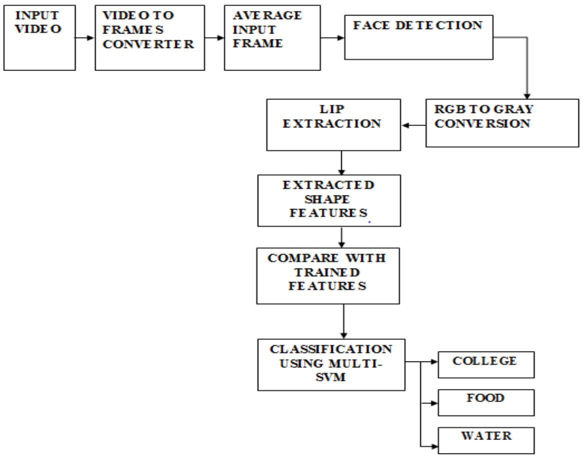

Fig -1: System Architecture

The captured photos include a human's face and a backdrop image. As a result, the captured pictures are initially subjected to pre-processing techniques such as Region Of Interest (ROI) segmentation to eliminate the background. The method of finding the area of a face in a picture is known as face detection. Following the face detection, the face is retrieved using the bounding box approach. The result of the face detection and extraction is an RGB color face image. Using the RGB to grey scale conversion procedure, the RGB face picture is transformed to grey scale. The Lip is the real ROI component. Initially, the face was regarded as the ROI component. The Lip is then recognized and retrieved from the Face using orientation (region) estimation. The lip picture is used to extract the form characteristics. Using the retrieved characteristics, we will categories the various human voice outputs using the Multi-SVM method. As a result, the human's unique lip movement were identified.

4. RESULTS AND DATASET

Fig -2 shows the image we entered after retrieving a number of frames from the video to frame converter model and averaging the entire number of frames.

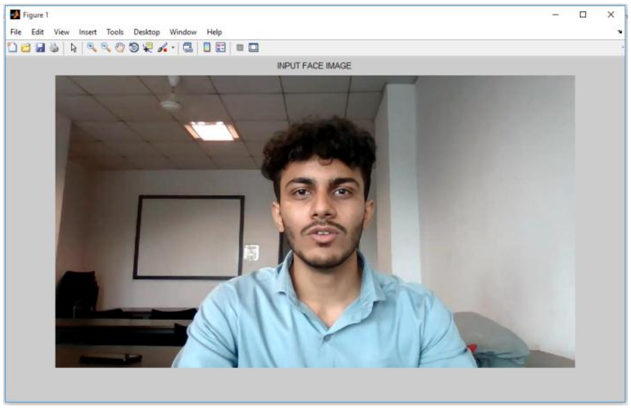

Fig -2: Input image

Fig -3 shows the bounding box approach was used to detect a face in the supplied image.

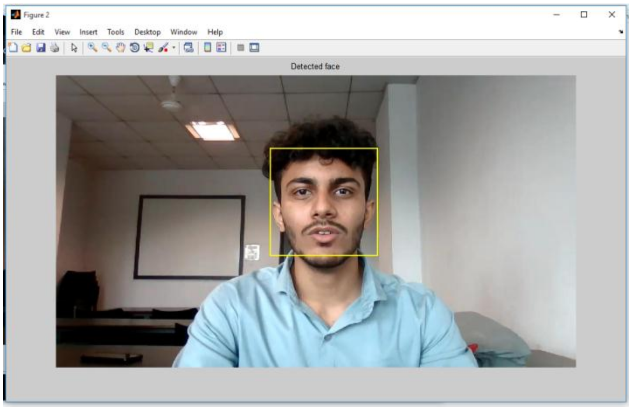

Fig -3: Detected face

Fig -4 shows the retrieved rgb color face from the input image using the bounding box approach.

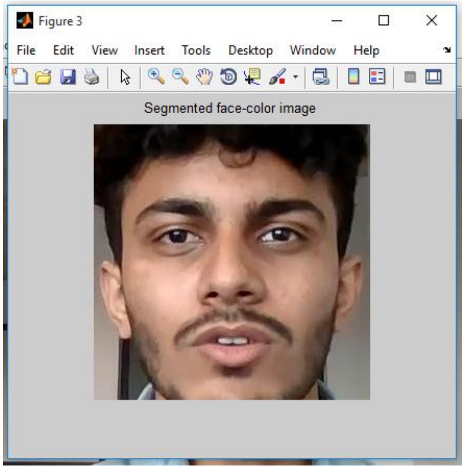
Figure 3
File Edit View Insert Tools Desktop Window Help
Segmented face-color image

Fig -4: Segmented face - color image

Fig -5 shows the extracted rgb color face's gray scale image

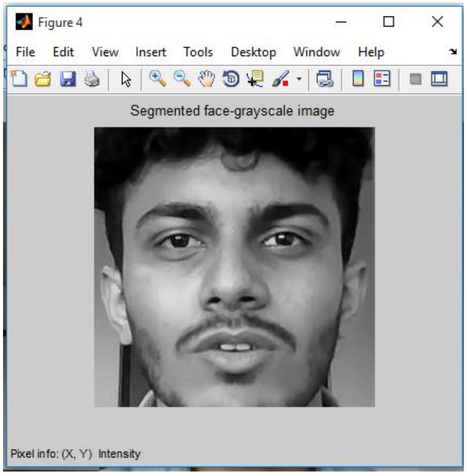
Figure 4
File Edit View Insert Tools Desktop Window Help
Segmented face-grayscale image
Pixel info: (X, Y) Intensity

Fig -5: Segmented face – gray scale image

Fig -6 shows the extracted lip image after conversion to gray scale image of the extracted face.

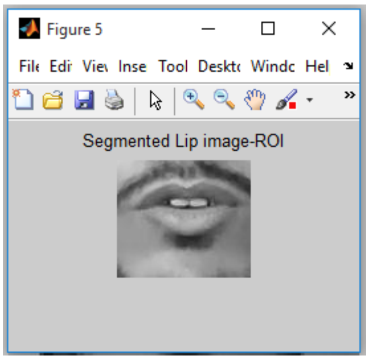

Fig -6: Segmented lip image

Fig -7 shows the end result of lip reading of the supplied video.

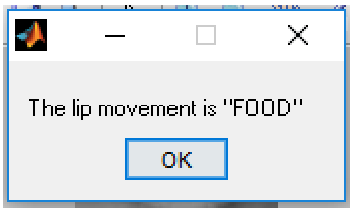
The lip movement is "FOOD"

Fig -7: Detected word

We contributed the dataset utilized in this system. We were able to get 75 percent efficiency by utilizing our own dataset. Which is fairly decent given that the dataset utilized was provided by us.

5. CONCLUSIONS

The ability of recognizing what is being said only from visual information is an impressive skill but a difficult task for the novice. Machine lip reading is a special type of automatic speech recognition (ASR) which transcribes human speech by visually interpreting the movement of related face regions including lips, face, and tongue. Recently, deep neural network based lip reading methods show great potential and have exceeded the accuracy of experienced human lip readers in some benchmark datasets. However, lip reading is still far from being solved, and existing methods tend to have high error rates on the wild data.

This research looks on the advancement of lip identification and recognition. We describe a method for detecting and distinguishing diverse human lip expressions. A video of a human's lip expression is obtained. A video to frames converter is used to convert the video to frames. As the input image, the average frame is used. The bounding-box method is used to detect the face. Based on orientation estimate, the ROI section (lip) is extracted. Shape characteristics are used to extract feature values. The varied human lip expressions were further categorized using the SVM algorithm based on the selected characteristics and the measured area attributes of the lip.

In this research, we manually provided an input and received an output. In our future research, we intend to create this project in real time, which means that the input will be real-time video delivered from the camera and the output will be real-time output.

REFERENCES

[1] Fenghour, S., Chen, D., Guo, K. and Xiao, P., 2020. Lip reading sentences using deep learning with only visual cues. IEEE Access, 8, pp.215516-215530.

[2] Wand, M., Koutník, J. and Schmidhuber, J., 2016, March. Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6115-6119). IEEE.

[3] Potamianos, G., Neti, C., Luettin, J. and Matthews, I., 2004. Audio-visual automatic speech recognition: An overview. Issues in visual and audio-visual speech processing, 22, p.23.

[4] Cox, S.J., Harvey, R.W., Lan, Y., Newman, J.L. and Theobald, B.J., 2008, September. The challenge of multispeaker lip-reading. In AVSP (pp. 179-184).

[5] Hilder, S., Harvey, R.W. and Theobald, B.J., 2009, September. Comparison of human and machine-based lip-reading. In AVSP (pp. 86-89).

[6] Chung, J.S., Senior, A., Vinyals, O. and Zisserman, A., 2017, July. Lip reading sentences in the wild. In 2017 IEEE conference on computer vision and pattern recognition (CVPR) (pp. 3444-3453). IEEE.

[7] Trojanová, J., Hrúz, M., Campr, P. and Železný, M., 2008, May. Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).

[8] Rekik, A., Ben-Hamadou, A. and Mahdi, W., 2014, October. A new visual speech recognition approach for RGB-D cameras. In International conference image analysis and recognition (pp. 21-28). Springer, Cham

[9] Afouras, T., Chung, J.S. and Zisserman, A., 2018. The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121.

[10] Ronquest, R.E., Levi, S.V. and Pisoni, D.B., 2010. Language identification from visual-only speech signals. Attention, Perception, & Psychophysics, 72(6), pp.1601-1613.

[11] Georgakis, C., Petridis, S. and Pantic, M., 2015. Discrimination between native and non-native speech using visual features only. IEEE transactions on cybernetics, 46(12), pp.2758-2771.