Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

Lip Reading to Text using Artificial Intelligence

Dr. Mamatha G[1]
Head of the Department
Information Science & Engineering
Nagarjuna college of Engineering & Technology
Bangalore, India

Bharath Roshan B R[2]
Student
Information Science & Engineering
Nagarjuna college of Engineering & Technology
Bangalore, India

Vasudha S R[3]
Student
Information Science & Engineering
Nagarjuna college of Engineering and technology
Bangalore, India

***Abstract* - An application that uses the camera of a smartphone to detect the lip movements of the person and convert the movements into text that can be understood by the hearing-impaired person. This application uses LRW dataset to visualise every movement of the lips. After the visualisation is completed, the captured movements are then converted into the form the person can understand easily.**

Keywords – LSTM, RNN, LRW, Deep Learning

INTRODUCTION

Lip perusing permits you to "tune in" to a speaker by viewing the speaker's face to make sense of their discourse designs, developments, signals and demeanours. Frequently called "a third ear," lip perusing goes past just perusing the lips of a speaker to decode singular words.

Figuring out how to lip read includes creating and rehearsing certain abilities that can make the procedure a lot simpler and progressively successful. These include:

- Learning to utilize the signals gave by the developments of the speaker's mouth, teeth and tongue
- Reading and assessing the data gave by outward appearances, non-verbal communication and motions related to the words being said
- Using vision to help with tuning in
- Using earlier information to fill in the holes that can happen in comprehension since it is difficult to peruse each word said.
- Curiously, it is simpler to peruse longer words and entire sentences than shorter words.

DATASET

Huge scale datasets have progressively demonstrated their central significance in a few research fields, particularly for early advancement in some rising themes. Right now, centre around the issue of visual discourse acknowledgment, otherwise called lipreading, which has gotten expanding enthusiasm for late years. We present a normally circulated enormous scale benchmark for lip perusing in the wild, named LRW-1000, which contains 1,000 classes with 718,018 examples from in excess of 2,000 individual speakers. Each class relates to the syllables of a Mandarin word made out of one or a few Chinese characters. Apparently, it is right now the biggest word-level lipreading dataset and furthermore the main open enormous scale Mandarin lipreading dataset. This dataset targets covering a "characteristic" changeability over various discourse modes and imaging conditions to join difficulties experienced in functional applications. It has demonstrated a huge variety right now a few viewpoints, remembering the quantity of tests for each class, video goals, lighting conditions, and speakers' characteristics, for example, present, age, sexual orientation, and make-up. Other than giving a definite depiction of the dataset and its assortment pipeline, we assess a few ordinary well-known lipreading strategies and play out an intensive investigation of the outcomes from a few perspectives. The outcomes show the consistency and difficulties of our dataset, which may open up some new encouraging bearings for future work.

BACKGROUND

AI techniques have greatly affected social advancement in late years, which advanced the fast improvement of man-made brainpower innovation and tackled numerous down to earth issues. Programmed lip-perusing innovation is one of the significant segments of human–PC cooperation innovation and computer-generated reality (VR) innovation. It assumes an imperative job in human language correspondence and visual observation. Particularly in loud conditions or VR situations, visual signs can expel repetitive data, supplement discourse information, increase the multi-modular info measurement of vivid interaction, reduce the time and remaining task at hand of human on learning lip language and lip development, and improve programmed discourse acknowledgment capacity. It improves the genuine experience of vivid VR.

ALGORITHM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture utilized in the field of profound learning. Dissimilar to standard feedforward neural systems, LSTM has criticism associations. It cannot just process single information focuses, (for example, pictures), yet in addition whole arrangements of information.

A typical LSTM unit is made out of a cell, an information door, a yield entryway and an overlook door. The cell recalls esteems over discretionary time interims and the three entryways direct the progression of data into and out of the phone.

LSTM systems are appropriate to ordering, preparing and making expectations dependent on time arrangement information, since there can be slacks of obscure span between significant occasions in a period arrangement. LSTMs were created to manage the detonating and disappearing inclination issues that can be experienced when preparing customary RNNs.

PROPOSED SYSTEM

The Proposed system consists of a mobile application that uses the camera to capture the lip movements and use LSTM algorithm to convert it to text and this uses LRW dataset to use the features that was captured by the camera to convert to text.

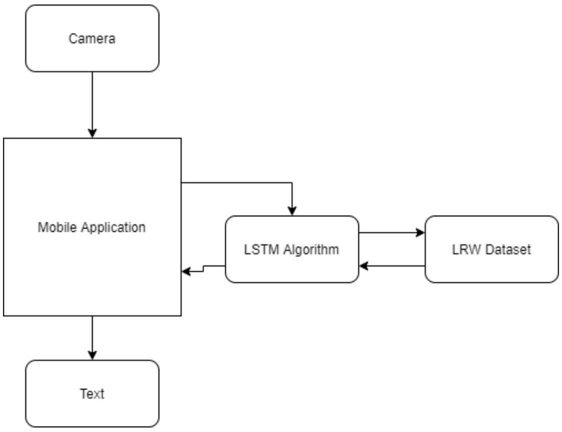

Fig 1 – System Architecture

Past examinations have demonstrated that both CNN and RNN models can accomplish better lip-perusing acknowledgment execution alone. We have discovered that the half and half system of consideration-based CNN-LSTM can additionally improve execution. The arrangement-based consideration instrument can be applied to assignments identified with time-arrangement PC vision and help the model in concentrating on some succession data of the video. Thinking about the impact of light, point and lucidity of the information pictures, we utilize a superior nature of the camera, and the proposed model is prepared with RGB pictures. The piece of CNN is improved by utilizing the model dependent on VGG19 and it does exclude the last two completely associated layers and we keep preparing the model dependent on pre-preparing parameters of ImageNet. The structure graph of VGG19 is appeared in Figure 5, and the contribution of VGG19 is 224 × 244 pixel RGB picture. In this way, the yield of CNN is 4096 × 10 and the consideration component is acquainted with the LSTM system to weight keyframes. From that point, the system expands two completely associated layers and a SoftMax layer for grouping.

CONCLUSION

Right now, neural system design of CNN and consideration-based LSTM is proposed for lip-perusing acknowledgment frameworks. Right off the bat, CNN (VGG19) separated visual highlights from the mouth ROI. At that point, we utilized the consideration-based LSTM to get familiar with the grouping loads and succession data between the edge level highlights. At long last, the grouping was accomplished by utilizing two completely associated layers and a SoftMax layer. The test dataset was worked by us autonomously and it comprised of three guys and three females. American English way to express numbers from zero to nine, and each advanced articulation were separated into autonomous video cuts, every free speaker was most certainly not prepared in proficient elocution. The test results show that contrasted and the general CNN-RNN model, the proposed design can adequately foresee words from the arrangement of lip area pictures without anyone else dataset, and the exactness of the proposed model is 88.2% in the test dataset which is 3.3% higher than the general CNN-RNN. In future research, we will prepare the lip-perusing acknowledgment model on datasets of continuous communicate recordings, including video tests from news communicates and true situations to investigate our proposed approach for speaker-free video discourse acknowledgment framework.

REFERENCES

[1] Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. Comput. Vis. Image Underst. 2007, 108, 116–134.

[2] Loomis, J.M.; Blascovich, J.J.; Beall, A.C. Immersive virtual environment technology as a basic research tool in psychology. Behav. Res. Methods Instrum. Comput. 1999, 31, 557–564. [CrossRef] [PubMed] Appl. Sci. 2019, 9, 1599 11 of 12

[3] Hassanat, A.B. Visual passwords using automatic lip reading. arXiv, 2014; arXiv:1409.0924.

[4] Thanda, A.; Venkatesan, S.M. Multi-task learning of deep neural networks for audio visual automatic speech recognition.

[5] Biswas, A.; Sahu, P.K.; Chandra, M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. Int. J. Speech Technol. 2016, 19, 159–171.

[6] Scanlon, P.; Reilly, R. Feature analysis for automatic speechreading. In Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564), Cannes, France, 3–5 October 2001;