International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

Automatic System for Lip Reading

Sahana B Mestha[1], Sahana V[2], Siddiqa Akram[3], Shruthi B[4]

Student, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India[1,2,3]

Assistant Professor, Department of Information Science & Engineering, Atria Institute of Technology, Bangalore, India [4]

Abstract -The proposed visible speech popularity method has used the idea of deep learning. A CNN shall detect the movement of lips and judge the words spoken. This trained CNN detects the words that have been spoken in the video and displayed in the text format. The CNN additionally is based on records furnished via way of means of the context, expertise of the language, and any residual hearing. The aim is to verify whether the use of artificial intelligence methods, namely Deep Neural Network, is a suitable candidate for solving this problem. In the practical part, the focus is on presenting the results in terms of the accuracy of the trained neural network on test data.

***Key Words*: Artificial Intelligence, visual motion, CNN**

1. INTRODUCTION

A machine that can read lip movement has great practicality in numerous applications such as: automated lip-reading of speakers with damaged vocal tracts, biometric person identification, multi-talker simultaneous speech decoding, silent movie processing and improvement of audio-visual speech recognition in general. The advancements in machine learning made automated lip-reading possible. However, many tries that hired conventional probabilistic fashions did now no longer acquire the predicted results. Most of these lip-reading methods were exclusively used to enchant the performance of audio-visual speech recognition systems in case of low-quality audio data.

Speech popularity in popular changed into revolutionized with the aid of using deep getting to know and the supply of huge datasets for education the deep neural networks. Lip-reading is an inherently supervised problem in machine learning and more specifically a classification task. Most existing deep visual recognition systems have approached lip-reading as a word classification task or a character sequence prediction problem. In the first case, a lip-reading network receives a video where a single word is spoken and predicts a word label from the vocabulary of the dataset. In the second one case, the enter video can also additionally include a complete sentence (more than one words) and a deep neural community outputs a series of characters, that's the anticipated textual content given the enter sentences. This type of network performs classification in the character level.

Obtaining inspiration from existing deep lip-reading networks, the proposed system aims to describe the process of designing, implementing and training two different deep lip-reading network. And finally evaluating their predictive performance. First, a NN that performs word classification. Then, a second neural network that decodes a sequence of characters from the input video sample. However, instead of training the second network on videos with full sentences, a single word video is used which resembles, the simple neural network

2. LITERATURE SURVEY

Lip Reading is one of the major problems that many researchers are trying to solve. The inspiration is taken from the audiovisual and visual speech recognition systems where ResNet blocks are used for the audio as well as video aspect of word-level prediction.

CTC models coupled with LSTMs is another approach to this problem where character level recognition is done by performing predicting labels frame-wise.

Another approach that is greatly followed these days is using sequence-to-sequence models which is an encoder decoder approach using LSTM blocks. This approach's shortcoming is that it fails in longer input sequences. To overcome this, the attention mechanism is implemented, mixing both the audio and video input sequences.
Another approach is the use of Convolutional Auto Encoders for feature extraction and coupling it with LSTMs for decoding the hidden vectors as explained in this paper.

3. REQUIREMENTS

Non-functional requirements, as the name suggests, are those requirements that are not directly concerned with the specific function delivered by the applications. They may relate to emergent properties such as reliability, response time and performance. Alternatively, they may define constraints on the application interface. Many non-functional requirements relate to the system as a whole rather than to individual application feature. This method they're frequently essential than the character useful requirements. These are the non-functional requirements listed:

3.1 Hardware Requirements

- System : Intel Core i7 9750H
- Speed : 4.5 GHz
- Hard Disk : 20 GB
- Monitor : LED/LCD Display
- RAM : 8 GB
- Keyboard : Standard Windows keyboard
- Mouse : Optical mouse
- GPU : NVIDIA GeForce GTX 1650

3.2 Software Requirements

- Operating System : Ubuntu/Windows 10 Home
- Platform : Python
- Frontend : Python interface
- Tool-kit : CUDA 10.0 and cuDNN
- Packages :TensorFlow(1.0),Keras(2.0), OpenCV

4. THE MODEL

4.1 Data Flow Diagram

Designing with the Data Flow Diagrams involves creating a model of the system. The entities and attributes are a version, of the states of the system. Processes version the guidelines of a System. The stimuli and reaction are modelled through Data Flows. All of those fashions are mixed into one photo version referred to as a Data Flow Diagram.

Data waft diagram (DFD) is a graphical illustration of the "waft" of information via an records machine, modeling its system aspects. Often, they're a initial step used to create a top level view of the machine that can later be elaborated. DFDs also can be used for the visualization of information processing (dependent design).

In the proposed system, the flow of data occurs as it is shown in the figure 3.1. In the first phase the data is represented as raw video as it is fed to the preprocessing stage and then it is divided into frames before passing through the face recognition stage. Then the face data is systematically passed to the cropping stage, where the ROI is cropped for feature extraction.

The necessary features are extracted and the video is normalized to maintain uniformity. Finally, normalized video data is the one which emerges as text output after having been passed through the decoding phase.

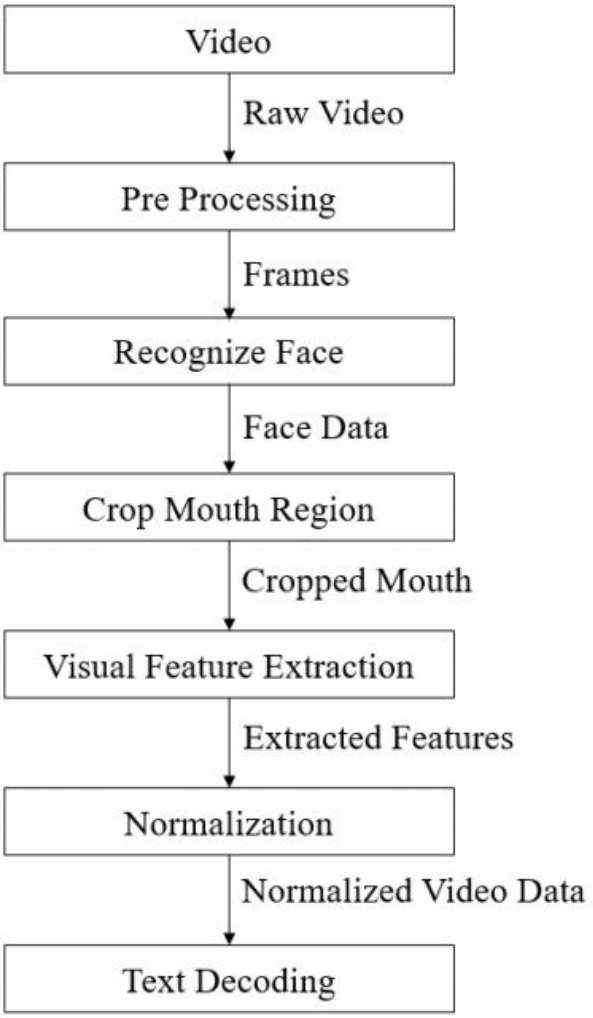

Figure 4.1 : Data Flow Diagram

4.2 Implementation

4.2.1 Face Detection using Viola Jones Algorithm

The fundamental belongings of this set of rules is that education is slow, however detection is fast. This set of rules makes use of Haar foundation function filters, so it does now no longer use multiplications. The performance of the Viola Jones set of rules may be substantially elevated through first producing the fundamental photograph The fundamental photograph permits integrals for the Haar extractors to be calculated through including most effective 4 numbers.

$$II(y,x) = \sum_{p=0}^{y} \sum_{q=0}^{x} Y(p,q)$$

4.2.2 Convolution Neural Network

In deep mastering, a convolutional neural network (CNN, or ConvNet) most commonly done to analyzing seen imagery. CNNs are regularized variations of multilayer perceptron. Multilayer perceptron normally refers to completely related networks, that is, every neuron in a single layer is attached to all neurons with inside the subsequent layers. The "completely-connectedness" of those networks lead them to liable to overfitting records. Typical methods of regularization consist of including a few shape of importance size of weights to the loss function. However, CNNs take a special method closer to regularization: they take benefit of the hierarchical sample in records and collect extra complicated styles the usage of smaller and less difficult styles. Therefore, on the size of connectedness and complexity, CNNs are at the decrease extreme. A convolutional neural network (CNN, or ConvNet) is one of the maximum famous algorithms for deep mastering with snap shots and video. Like different neural networks, a CNN consists of an enter layer, an output layer, and plenty of hidden layers in between.

.

4.2.3Feature Detection Layers

These layers perform one of three types of operations on the data: convolution, pooling, or rectified linear unit (ReLU).

• Convolution places the enter photographs via a fixed of convolutional filters, every of which turns on positive capabilities from the photographs.

• Pooling simplifies the output by performing nonlinear down sampling, reducing the number of parameters that the network needs to learn about.

• Rectified linear unit (ReLU) permits for quicker and extra powerful education through mapping terrible values to 0 and keeping nice values.

These 3 operations are repeated over tens or masses of layers, with every layer mastering to hit upon special features.

Classification Layers

After feature detection, the architecture of a CNN shifts to classification. The next to-last layer is a fully connected layer (FC) that outputs a vector of K dimensions where K is the number of classes that the network will be able to predict. This vector includes the chances for every elegance of any photo being classified. The very last layer of the CNN structure makes use of a SoftMax feature to offer the type output

SoftMax

The SoftMax characteristic is regularly hired in the very last layer of neural network-primarily based totally classifiers. For a Multilayer perceptron classifier, the SoftMax characteristic transforms the MLP output, that is a vector to a brand new vector.

ReLU Layer

This step will increase the non-linear residences of the general community. Different activation capabilities can be additionally used, which include the hyperbolic tangent and the sigmoid function. However, the ReLU is favored in numerous networks, because it enables triumph over the hassle of vanishing gradients, that's a not unusual place difficulty in neural networks. During training, whilst the back-propagation set of rules is used to propagate the mistake from the closing layers of the community to the front layers, the gradient with inside the first layers takes small values near zero. This outcomes in sluggish updates with inside the weights of the primary layers.

5. RESULTS

The proposed system has been trained within GRID CORPUS dataset. The system shows variable accuracy between 70-80 % on the test dataset. The Accuracy achieved is depicted while comparing the kernel sizes. It is evident that while increasing the kernel size of CNN from 3X3X3 to 5X5X5 the accuracy increases significantly subject to number of epochs.

6. CONCLUSION

It can be concluded that the proposed system is a unique model which uses artificial intelligence to predict the text from a video sample. It can be confidently said that with minimal requirements the model designed reaches an accuracy oddly close to two-fold higher than that of a human lip reader. As of now the technology progression of this system is extended as far as on small scale devices which may only be in the prototype state and an improvement in performance of the lip-reading system would pave the path for different applications to be integrated with the uses of the system.

REFERENCES

[1] "VSR: https://youtube.com/playlist?list=PLXkuFIFnXUAPIrXKgtIpctv2NuSo7 xw3k"

[2] "T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in Interspeech, 2017."

[3] "X. Zhang, C. C. Broun, R. M. Mersereau and M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces," EURASIP Journal on Advances in Signal Processing, 2002, vol. 2002, no. 11, pp. 1228-1247"

[4] "Joon Son Chung and Andrew Zisserman. Learning to lip read words by watching videos. Computer Vision and Image Understanding, pages 1– 10, 2018."

[5] "S. Yang et al., "LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019, pp. 1-8, doi: 10.1109/FG.2019.8756582"

[6] "P. Sujatha and M. R. Krishnan, "Lip feature extraction for visual speech recognition using Hidden Markov Model," 2012 International Conference on Computing, Communication and Applications, Dindigul, Tamilnadu, 2012, pp. 1-5."

[7] "A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional ´ LSTM networks for improved phoneme classification and recognition," in International Conference on Artificial Neural Networks. Springer, 2005, pp. 799–804."

[8] "N. Eveno, A. Caplier, and P.Y. Coulon, "Accurate and Quasi-Automatic Lip Tracking, "Ieee Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 5, 2004, pp. 706-715."

[9] "Chung, Junyoung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." ArXiv abs/1412.3555 (2014): n. pag."