LipSpeak : Enhancing Communication for the Deaf

By

Shravya

NNM22MC091

MASTER OF COMPUTER APPLICATIONS

Under the guidance of

Ms. Prathwini

Assistant Professor, Gd - I

Major Project, Project Progress Evaluation-||
Report Submitted to

DEPARTMENT OF MCA

LipSpeak : Enhancing Communication for the Deaf

By

Shravya

NNM22MC091

MASTER OF COMPUTER APPLICATIONS

Under the guidance of

Ms.Prathwini

Assistant Professor, Gd - I

Major Project, Project Progress Evaluation-|| Report Submitted to

DEPARTMENT OF MCA

NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

April, 2024

NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

*Department of **M.C.A.***

CERTIFICATE

Certified that the Progress Evaluation-|| of the Major Project entitled **Project Title** carried out by **Shravya,** USN **NNM22MC091**, a bonafide student of **NMAM Institute of Technology**, **Nitte** has been carried out satisfactorily. The project work, Progress Evaluation-|| report has been prepared as per the prescribed format.

Name & Signature of Guide(s) **Name & Signature of HOD**

Major Project, Project Progress Evaluation – ||

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

ABSTRACT

This project " LipSpeak : Enhancing Communication for the Deaf " presents a novel approach to speech recognition utilizing computer vision and deep learning techniques for lip reading. The system aims to accurately identify spoken words from a predefined set, catering particularly to individuals with hearing impairments. The methodology involves the creation of a custom dataset comprising around 700 video clips, each labeled with a specific word. These clips undergo rigorous preprocessing steps, including lip segmentation and various image enhancement techniques such as Gaussian blurring, contrast stretching, bilateral filtering, and sharpening. The model architecture incorporates 3D convolutional neural networks to capture spatiotemporal features from the video frames effectively. Training was conducted using TensorFlow and Keras on cloud platforms due to hardware constraints. Moreover, comprehensive evaluation metrics including precision, recall, F1 score, and balanced accuracy attest to the model's effectiveness in word classification. Despite challenges such as data gathering and limited processing power, the developed algorithm shows promise for real-world applications, including aiding communication for individuals with hearing loss and enhancing video surveillance systems.

TABLE OF CONTENTS

1. INTRODUCTION

In a world where communication is fundamental, individuals with hearing impairments face unique challenges in understanding spoken language. With over 1.5 billion people globally experiencing some form of hearing loss, there is an urgent need for innovative solutions to bridge this communication gap. Traditional speech recognition systems often rely on audio cues, rendering them ineffective for individuals who rely primarily on visual cues, such as lip movements, to comprehend speech.This project endeavors to address this pressing need by harnessing the power of computer vision and deep learning to develop a novel speech recognition system tailored specifically for individuals with hearing impairments. By leveraging advancements in technology, particularly in the fields of image processing and neural networks, we aim to create an intuitive and accurate system capable of translating lip movements into spoken words in real-time.

Central to this endeavor is the creation of a robust dataset comprising video clips of individuals speaking predefined words. Through meticulous data collection and annotation, we curate a comprehensive repository that serves as the foundation for training our model. Leveraging state-of-the-art techniques in image processing, we preprocess the data to enhance the clarity and distinctiveness of lip movements, laying the groundwork for accurate recognition.The heart of our approach lies in the design of a sophisticated deep learning architecture, specifically a 3D convolutional neural network (CNN), tailored to capture both spatial and temporal features inherent in video data. Drawing upon libraries such as TensorFlow and Keras, we harness the computational power of cloud resources to train our model, circumventing hardware limitations and ensuring scalability.

Beyond its immediate impact on aiding communication, our system holds promise for broader applications, from enhancing accessibility in everyday interactions to bolstering video surveillance systems where audio data may be inadequate or unavailable. By leveraging cutting-edge technology to address a critical societal need, this project exemplifies the transformative potential of interdisciplinary innovation in improving lives and fostering inclusivity.

2. LITERATURE SURVEY

1. Son Chung, Joon, et al.'s "Lip reading sentences in the wild" enhances visual speech recognition by introducing a large-scale dataset from real-world videos, significantly improving model accuracy in diverse environments. This study, published at the IEEE conference on computer vision and pattern recognition, marks a major step forward in developing practical lip-reading technologies.

2. Published in Image and Vision Computing in 2018, the paper examines various deep learning architectures and methodologies that have significantly improved the performance of lip-reading systems, offering a detailed review of advancements and outlining future research directions in this increasingly important area of study.

3. The study not only compares the effectiveness of different architectures but also introduces an online application designed to demonstrate the practical capabilities of these advanced lip-reading systems. This work highlights significant advancements in model performance, indicating a promising direction for realworld applications of lip-reading technology driven by deep learning.

4. In the study by Nancy A. Neef and Brian A. Iwata titled "The development of generative lipreading skills in deaf persons using cued speech training," the authors investigated how cued speech training can enhance lipreading abilities in deaf individuals. Their methodology involved training participants with cued speech a system where phonemes are represented by hand shapes near the mouth to supplement lip movements and then assessing their ability to generalize these skills to new, untrained words.

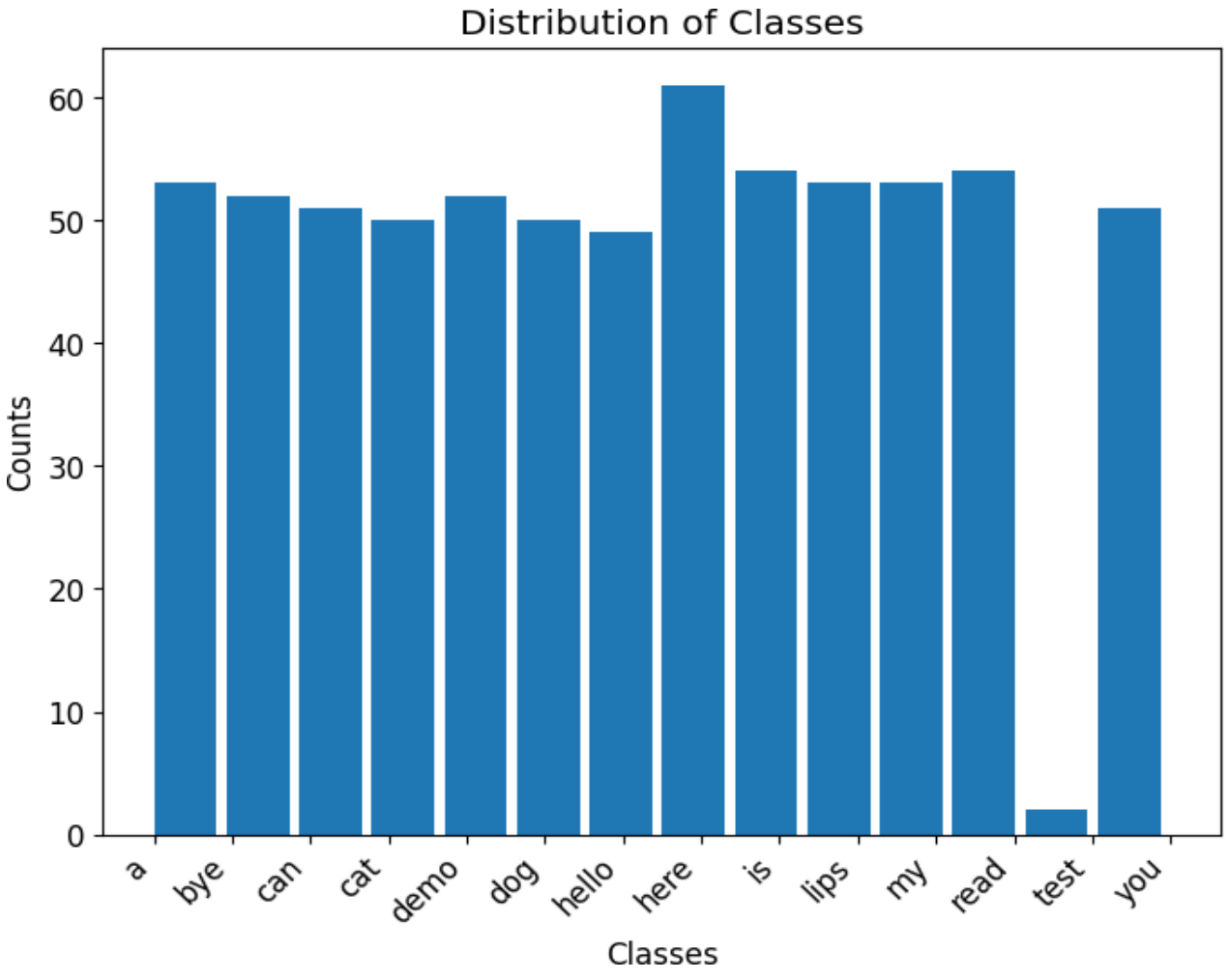
5. In their 2019 study published in Estudos de Psicologia (Campinas), Pinheiro, Rocha-Toffolo, and Vilhena explore how speech and lip reading can aid profoundly deaf users of Brazilian Sign Language (Libras) in enhancing their reading skills. Their research highlights the potential of these visual techniques to improve language comprehension and literacy among the deaf, suggesting new avenues for educational approaches.

3. Project Progress

Our project on Computer Vision Lip Reading has made substantial advancements towards its objective of developing a speech recognition system tailored for individuals with hearing impairments.

DATASET:

As this is a supervised learning project, I needed a dataset to train my model on. Since a suitable dataset was not available for this problem, I took the initiative to create my own dataset by collecting approximately 700 video clips of words being spoken. Each video clip was manually labeled with a word from a predefined set and in the end, I had around 3 gigabytes worth of total data. My dataset contains data for 13 different words, and to give an overview of the dataset's distribution, you can view the histogram chart on the left showing the number of video clips for each word. The x-axis represents the words in the set, and the y-axis represents the number of video clips. The chart confirms that the dataset has a roughly equal number of video clips for each word, with some minor variation.



The dataset was subsequently divided into training and testing sets, with 80% of the data allocated for training purposes and the remaining portion reserved for testing. The resulting sizes of the training and testing datasets were (544, 22, 80, 112, 3) and (137, 22, 80, 112, 3) respectively. Each video clip comprises 22 frames and possesses dimensions of 112x80 for width and height respectively. Below, I have provided several examples of the training data.

Libraries:

For training my models, I relied on a selection of essential libraries. TensorFlow and Keras served as the primary frameworks for model development and implementation. In addition, I utilized the image processing capabilities of OpenCV and PIL (Python Imaging Library) for data collection and preprocessing tasks. To handle data manipulation and preparation, numpy and scikit-learn were indispensable tools. For visualizing key insights and performance metrics, matplotlib and seaborn were employed, providing clear and insightful graphical representations.

Lip Segmentation:

Since the goal of this project aims to translate lip movements into words, I want to segment out just the lips and surrounding area. I created constant values for width and height, and after segmenting out the lips, I would add padding to the segmented image so that the final frame matches the predefined constant dimensions.

Gaussian Blurring:

Gaussian Blurring is a useful image pre-processing technique that passes a Gaussian filter through an image to blur edges and reduce contrast. By softening sharp curves previously present in the image, my model will have an easier time performing calculations in the hidden layers. Ultimately, Gaussian Blurring will reduce train time and overfitting.

Contrast Stretching:

Contrast Stretching is an image enhancement technique used to improve the contrast of an image by stretching the range of intensity values. By increasing the contrast between the darker and lighter pixels in an image, the details become more visible, which can help my models perform better. Contrast stretching is particularly helpful in this case since it enhances the contrast between different regions in the image, making the lips more distinguishable and prominent.

Bilateral Filtering:

Bilateral Filtering is a non-linear image filtering technique that smooths the image while preserving edges. This technique uses both the spatial and intensity distances between pixels to achieve this. By reducing noise in the image, Bilateral Filtering can improve image quality, which can improve the performance of my models. Bilateral Filtering can also help to remove small objects or noise in the image, which is particularly useful in this situation.

Sharpening:

Sharpening is an image enhancement technique that accentuates the edges of objects in an image, making them appear more defined and distinct. This technique can help to increase the contrast between objects and their surroundings, making it easier for my models to detect and recognize them. By improving the edges in an image, sharpening can improve the accuracy of object recognition tasks.

4. CURRENT PROGRESS

The current progress of the lip reading project involves two primary stages: data collection and prediction.

1. Data Collection

To begin with, the data collection phase uses a video feed to capture and analyze lip movements in real-time. The process initiates by setting up a video capture device, typically a webcam, and initializing essential libraries such as OpenCV for image processing, Dlib for face detection and landmark identification, and various Python libraries for data handling. A frontal face detector from Dlib is employed to locate the face in each frame. Subsequently, a shape predictor is used to identify specific facial landmarks, particularly around the mouth region.

As the user speaks, the system tracks the distance between the upper and lower lips. If this distance exceeds a certain threshold, it identifies that the user is talking and records the corresponding frames. These frames are processed and stored, ensuring they fit within a specific size by adding padding and applying various image enhancement techniques such as contrast stretching and Gaussian blur. The data collection continues until a complete set of frames representing a word is gathered. Each collected word is stored with its associated frames, and the system also creates a video from these frames for future reference.

2. Prediction

In the prediction phase, the collected data is used to train a deep learning model designed to recognize spoken words based on lip movements. The model architecture includes several convolutional layers for feature extraction, followed by dense layers for classification. The input to the model is a sequence of frames representing a spoken word.

The system captures new video frames, processes them similarly to the data collection phase, and then feeds them into the trained model for prediction. The model outputs a probability distribution over a predefined set of words, identifying the word with the highest probability as the predicted output. To ensure accuracy, the system checks if the predicted word has been previously spoken in the current session, and if so, it adjusts the prediction accordingly.

Overall, the project has made significant strides in both capturing and processing lip movement data and in developing a robust predictive model to identify spoken words based solely on visual input. The combination of real-time video processing, advanced image enhancement techniques, and a sophisticated deep learning model forms the backbone of this innovative lip reading system.

5. References

1. Son Chung, Joon, et al. "Lip reading sentences in the wild." Proceedings of the IEEE conference on computer vision and pattern recognition.

2. Fernandez-Lopez, Adriana, and Federico M. Sukno. "Survey on automatic lip-reading in the era of deep learning." Image and Vision Computing 78 (2018): 53-72.

3. Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "Deep lip reading: a comparison of models and an online application." arXiv preprint arXiv:1806.06053 (2018).

4. Alegria, Jesus, Brigitte L. Charlier, and Sven Mattys. "The role of lip-reading and cued speech in the processing of phonological information in French-educated deaf children." European journal of cognitive psychology 11.4 (1999): 451-472.

5. Pinheiro, Ângela Maria Vieira, Andreia Chagas Rocha-Toffolo, and Douglas de Araújo Vilhena. "Reading strategies for the profoundly deaf Libras users: Benefits of speech and lip reading for strengthening linguistic skills." Estudos de Psicologia (Campinas) 37 (2019): e190003.

6. Conclusion

In conclusion, this project stands at the forefront of innovation, epitomizing the fusion of computer vision and deep learning to pioneer an algorithm capable of translating intricate lip movements into coherent spoken words. Leveraging the power of transfer learning, I meticulously curated a dataset comprising 700 video clips, each meticulously processed through a spectrum of image enhancement techniques, including lip segmentation, Gaussian blurring, contrast stretching, bilateral filtering, and sharpening. Central to the project's success were the indispensable contributions of TensorFlow, Keras, OpenCV, PIL, numpy, and scikit-learn, serving as pillars for data preparation and model training. The resulting model architecture, characterized by convolutional and dense layers, achieved remarkable training accuracy of 95.7% and validation accuracy of 98.5%, positioning it as a beacon of efficacy and reliability. Moreover, the model's lightweight nature holds promise for seamless integration into real-time applications, paving the way for its deployment in live settings to facilitate communication for individuals with hearing impairments. Beyond its technical prowess, this project underscores the profound impact of interdisciplinary collaboration, ushering in a new era of accessibility and inclusivity through transformative technological advancements.

In addition to its technical achievements, this project resonates profoundly as a testament to the transformative potential of technology in fostering inclusivity and accessibility. By bridging the gap between computer vision and deep learning, we have not only created a groundbreaking algorithm but also opened doors to enhanced communication for diverse communities. As we look ahead, the real-world implications of this work extend far beyond its immediate scope, promising to empower individuals with hearing impairments and catalyze societal change. With continued innovation and collaboration, we can harness the full potential of technology to build a more inclusive world where communication barriers are dismantled, and everyone's voice is heard.