Agenda

Introduction

Literature Survey

Proposed System

Results and Discussion

Conclusion

References

Introduction

Introduction

Motivation

Objectives &Problem Statement

Introduction

- In the Muted Video To Text Conversion we have created a system that enables users to transcribe the spoken content of a muted video into written text.

- The system utilizes various libraries to accurately capture the frames and alignments in the video and convert it into written text.

- The resulting text can be used to create subtitles or closed captions for videos, making them more accessible to individuals with hearing impairments or those who prefer to read instead of listening.

Motivation

- The motivation behind the development of muted video to text converter is to make video content more accessible and inclusive for everyone.
- This tool helps to ensure that individuals with hearing impairments can access video content and enjoy it on equal terms with those who can hear.
- The motivation behind the muted video to text converter is to enhance accessibility, convenience, and inclusivity in video content.
- It also provides an additional layer of convenience for those who prefer to read instead of listen, allowing them to engage with video content more easily.

Objective

- Providing accurate and reliable transcriptions of the spoken words in a video, to ensure that individuals with hearing impairments can fully understand the content.
- Saving time and effort in the transcription process, by automating the conversion of spoken words into written text.
- Enhancing the overall user experience of video content, by providing an accessible and inclusive viewing experience for all users.

Problem Statement

- With the increasing amount of video content available online, it's becoming common to encounter videos where the audio has been muted or removed. This poses a challenge for individuals who rely on audio cues to access the information presented in the video.
- The current solutions for this problem are limited,time consuming and costly.
- Therefore, the problem statement is to develop a system that can accurately write the text from a muted or audioless video.

Literature Survey

- Literature Survey
- Limitation Existing system

Literature Survey

- In this section,we outline the existing work that has been done in the field. As previously mentioned, most approaches have involved machine learning methods that do not touch on deep learning. It has only been until very recently that deep learning methodshave emerged and produced state-of-the-art results.

- One of the first works to use deep learning in speech recognition was Hinton et al., where neural networkswere used for acoustic processing. Other approaches includelearning multimodal audio-visual representations and learning visualfeatures to then apply to more traditional classifier structures like HMMs .

- Assael et al. created LipNET, a phrase predictor that uses spatiotemporal convolutions and bidirectional GRUs and achieved a 11.4% WER on unseen speakers. Our model is primarily inspired by this work.

Limitation Existing system

- The accuracy of the video with audio to text system can be affected by poor audio quality, such as background noise, low volume, or distorted audio

- systems can struggle to distinguish between homophones and homonyms, which are words that sound alike but have different meanings. This can lead to inaccuracies in the transcription

- Technical issues such as poor audio quality or a low signal-to-noise ratio can also affect the accuracy of systems.

PROPOSED SYSTEM

Introduction

Architecture

Process Design

Details of Hardware & Software

Introduction

- To do so we are going to use OS module to interact with the operating system ,CV2 module for pre-processing and loading the video, TensorFlow for large numerical computations without keeping deep learning in mind.
- Numpy is used for to store arrays. Typing library shows th variable type annotations. From matplotlib, pyplot module is to create 2D graphs and plots by using python scripts and Imageio module is used to create a animated gif of the frames .
- We are going to use CTC (Connectionist temporal classification) as an algorithm which is used to train deep neural network

Architecture

1. Build data loading functions
2. Create Data Pipeline
3. Design the Deep Neural Network
4. Setup Training options and Train
5. Make a Prediction
6. Test on video

Process Design

1. Input Layer
2. Feature Extraction Layer
3. Encoding Layer
4. Machine Learning Algorithm
5. Output Layer

Details of Hardware & Software

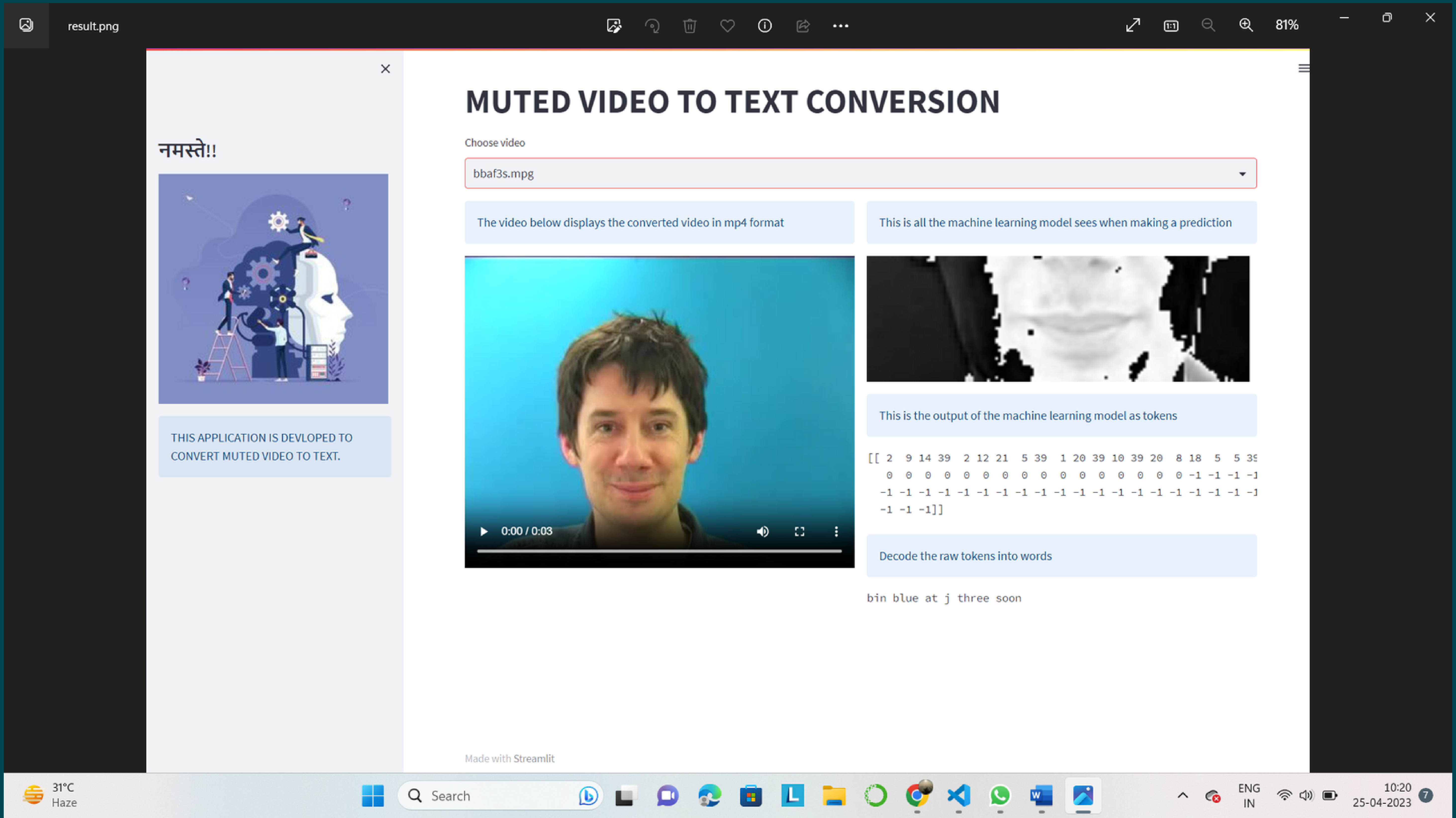
Hardware:

- RAM- 8GB DDR4 or above
- Storage- 1GB ROM
- Graphics- nvidia RTX 3050 or above

Software:

- OS- Windows/Linux
- python IDLE

Result



नमस्ते!!

THIS APPLICATION IS DEVLOPED TO CONVERT MUTED VIDEO TO TEXT.

MUTED VIDEO TO TEXT CONVERSION

Choose video

bbaf3s.mpg

The video below displays the converted video in mp4 format

This is all the machine learning model sees when making a prediction

This is the output of the machine learning model as tokens

```
[[ 2  9 14 39  2 12 21  5 39  1 20 39 10 39 20  8 18  5  5 39
   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1 -1 -1 -1
  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
  -1 -1 -1]]
```

Decode the raw tokens into words

```
bin blue at j three soon
```

Made with Streamlit

Conclusion

- **The result of a muted video to text converter is a transcript of the video in written form. The accuracy of the transcript depends on the quality of the transcription tools used and the quality of the video being transcribed.**
- **It can make the video content more accessible to people who are deaf or hard of hearing, or those who prefer to read rather than watch a video.**
- **Also save time for people who want to quickly scan through the content of the video.**

References

- https://arxiv.org/pdf/1001.2267
- https://www.mdpi.com/1424-8220/23/4/2284
- https://ijettjournal.org/assets/year/2016/volume-37/number-6/IJETT-V37P254.pdf

THANK YOU!