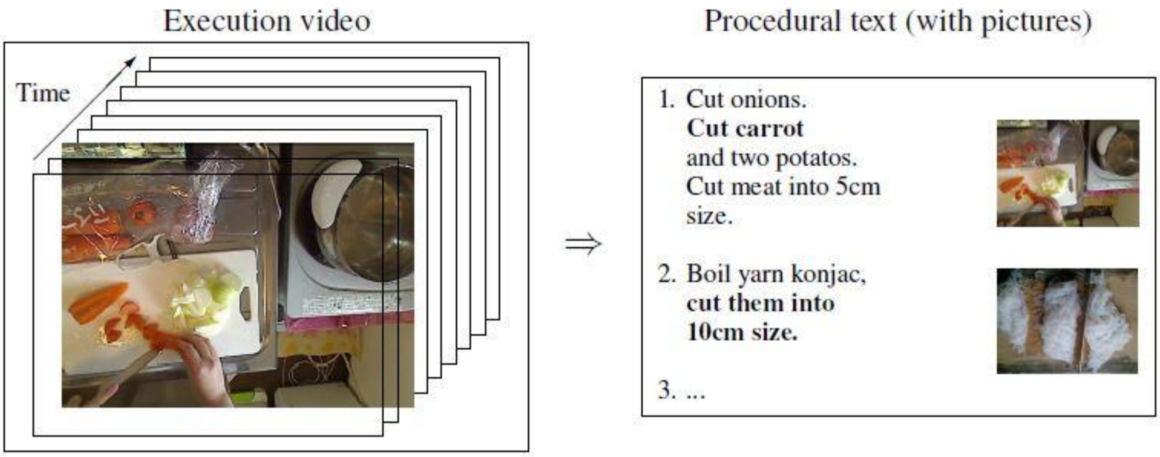
Real-time video to text transcription for visually challenged person

Problem statement

In today’s world, the use of subtitles and text has become important for understanding the video. Transcribing is essential for people who are deaf, visually challenged, who have reading and literacy problems, and can to those who are learning to read. Transcription of videos provide information for individuals who have difficulty understanding the speech and auditory components of the visual. This leads to a valid subject of research in the field of automatic text generation from videos. It provides the users a major benefit of not downloading the subtitles from the internet instead generating them automatically. Eventually the transcription can change to Barile form for visually challenged person.

Background

Various studies have been done to accomplish each module of the work. Hidden Markov Model is used for Speech Recognition for calculating the probability of the occurrence of the words using the acoustic and language model. The main task of such types of methods is generating a procedural text from an execution video. In general, an execution video records a sequence of activities to make or repair something from the beginning to the end.



N

F must

b

A

The input file first goes to the demuxer where the video is separated from the audio. Then, this audio is encoded where the stream is divided into frames and then converted into binary format.

B. Speech Recognition

After the completion of audio extraction, the speech recognition part is carried out by using ant speech recognition method.

C. Subtitle generation

The .srt file generated by the speech recognition method which contains the words (lyrics) spoken in the audio file.

Architecture of the Real-time video to text transcription for visually challenged person is shown in Figure 1.

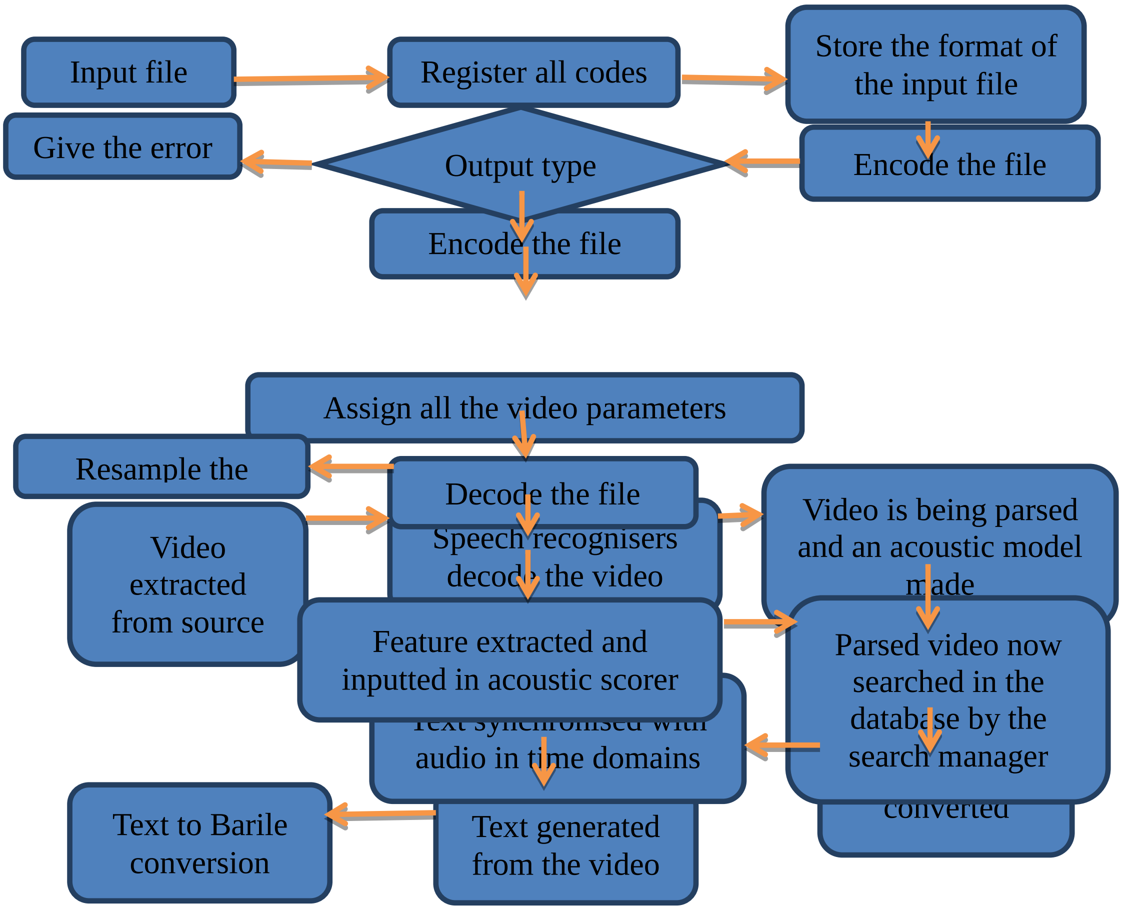

Input file
Register all codes
Store the format of the input file
Give the error
Output type
Encode the file
Encode the file
Assign all the video parameters
Resample the
Decode the file
Video is being parsed and an acoustic model made
Video extracted from source
Speech recognisers decode the video
Feature extracted and inputted in acoustic scorer
Parsed video now searched in the database by the search manager
Text synchronised with audio in time domains
converted
Text to Barile conversion
Text generated from the video

Figure 1: Architecture of the Real-time video to text transcription for visually challenged person.

Experimental design

Test data:

The dataset contains video files in the form of mp3, .mp4, .avi, .au, and .flac supported by FFMPEG standards. Traditionally, such type of algorithm requires annotated video data to learn models.

Evaluation measure:

Measures such as motion features, object detection and tracked in video sequences, flows etc. will be computed. The result can be comparing in term of accuracy with different annotation tool.

Software and hardware requirements:

Python based Deep Learning libraries will be exploited for the development and experimentation of the project. Training will be conducted on NVIDIA GPUs using some speech recognition method.