Project Proposal

Level 4

Automated Step-by-Step Visual Performance Guide Generation from Sinhala Demonstration Videos

Group Name: The TechNovas

Faculty of Information Technology

University of Moratuwa

2024

Project Proposal

Level 4

Automated Step-by-Step Visual Performance Guide Generation from Sinhala Demonstration Videos

Group Name: The TechNovas

Group Members

Index Number	Name
204104H	Kularathna M.D.S.A.
204137K	Nethmini S.A.R.
204041K	Dilshan K.G.A.P.

Supervisor: Dr. L. Ranathunga

Faculty of Information Technology

University of Moratuwa

2024

Table of Contents

1. Introduction

As more demonstration videos become available online across various platforms, users often find it hard to access and understand the information. With so many videos to choose from, it can be overwhelming and frustrating for people to watch an entire video just to see if it contains the specific help they need. This issue is even more noticeable for Sinhala demonstration videos because there are fewer of them available compared to English videos. As a result, people who speak Sinhala may have a harder time finding the information they need. When they do find these videos, they might not have the time to watch the entire video or fully understand the key points. Many Sinhala videos lack structured guides that highlight important actions, ingredients, or instructions. As a result, viewers can miss critical details or waste time re-watching content.

The problem addressed in this project is the absence of performance guides for Sinhala demonstration videos. Visual Performance guides provide viewers with accurate, synchronized summaries that combine text and visual cues. These guides would allow users to quickly understand and follow the most important parts of the demonstration, without needing to watch the entire content.

To solve this, we propose using speech-to-text technology combined with Natural Language Processing to generate accurate transcriptions and extract key information from Sinhala videos [1]. By using image processing techniques, we will obtain key visuals by selecting the best images from specific timestamps [2]. This will make sure that the visuals correspond to the instructions being delivered. Afterwards, we will synchronize both images and text in the correct order. This will create a clear visual performance guide that matches the transcript with the corresponding visuals, making it easier to understand and more engaging.

This solution uses various technologies to improve accessibility for Sinhala video content, saving time for viewers while ensuring they capture the most important details in very little time. It offers a practical and efficient way to extract useful information from demonstration videos.

2. Background & Motivation

The internet is a very familiar place to everyone for sharing knowledge and information on countless topics, including cooking, education, and skills training. It allows people to access demonstrations and resources from anywhere in the world. This makes learning more convenient and accessible for everyone. However, users often face challenges when consuming long-form video content, especially when looking for specific information. For Sinhala-speaking audiences, this problem is made worse by the lack of effective tools that allow easy extraction of key information from video content. Currently, many users either miss important details or spend significant time re-watching segments to understand the critical parts of a demonstration.

Watching videos in Sinhala can be time-consuming, and users often find it difficult to quickly pinpoint key actions, ingredient lists, or step-by-step instructions. This lack of easily accessible performance guides or summaries is a major downside for viewers who want an efficient way to follow demonstrations. In contrast, English-language content has more developed solutions, including advanced video summaries and synchronization tools that do not exist or work well for Sinhala content.

The motivation for this project stems from the growing demand for efficient content summarization and the unique need to cater to Sinhala-speaking audiences who are underserved in this area. By addressing this gap, we aim to provide a solution that not only saves time but also improves user experience. Our team plans to use technologies such as Natural Language Processing, speech-to-text technology, Image processing technology and synchronization technologies, to deliver a solution to this problem. The proposed solution will use these technologies to offer concise, accurate visual performance guides from Sinhala demonstration videos.

3. Problem in Brief

The main issue this project tackles is the absence of effective performance guides for Sinhala demonstrations available online. Specifically, viewers of Sinhala demonstrations do not have access to a tool that can automatically extract key information such as ingredients, steps, and important visuals, and present this information in an accurate, synchronized text-visual format. This gap in content accessibility makes it difficult for Sinhala-speaking users to engage with video content efficiently. This can result in a time-consuming and often frustrating experience.

By developing a solution that generates text and visual summaries using speech-to-text, NLP technologies and image/video processing technologies this project looks to address this communication and accessibility gap for Sinhala video content [1].

4. Aim & Objectives

4.1 Aim

The aim of this project is to develop a system that automatically generates visual performance guides from Sinhala demonstration videos.

4.2 Objectives

To achieve the above aim, our objectives are as follows.

1. **Automated Speech-to-Text Conversion and Content Structuring:** Develop an automated system to extract and convert Sinhala audio from demonstration videos into text [1]. This obtained text will be refined by removing background noise, correcting grammatical and punctuation mistakes, and organizing the content into meaningful sections for easier instruction extraction

2. **Automated Key Point Extraction and Instruction Generation**: Implement a machine learning model to automatically identify and extract key points from transcribed Sinhala audio while associating them with relevant video timestamps [3]. Simultaneously, generate clear, step-by-step written instructions based on the key points.

3. **Automated Visual Highlighting and Screenshot Extraction**: Extract screenshots at key video timestamps and enhance these visuals with zoom effects and other highlights to improve clarity and focus on important steps [2].

4. **Generate a Document by Synchronizing Instructions and Visuals:** Create a visual performance guide by synchronizing the selected visuals with the written instructions, ensuring a cohesive and engaging presentation of the information.

5. **Integration of Machine Learning for Enhanced Fine-Tuning:** Use machine learning techniques to continuously improve the accuracy and coherence of the text, screenshot selection, and overall output for generating performance guides from Sinhala videos.

5. Proposed Solution

In this project, we propose a solution to generate visual performance guides from Sinhala demonstration videos, by using speech-to-text technology, Natural Language Processing, Image Processing and synchronization tools. The system will transcribe, summarize and obtain the steps mentioned in the demonstration video. Afterwards, it will obtain the correct screenshots from the video which correspond to the steps that we have obtained earlier. Finally, we will synchronize the text and images we have derived in order to create a step-by-step visual performance guide as the end product. This solution will focus on demonstration videos in the cooking and technology-related domains. Videos considered will have a maximum duration of 10 minutes.

Technology Adapted:
The key technologies involved in this solution include:

- **Speech-to-Text Technology:** To convert the audio content of Sinhala demonstration videos into text. We will use open- source Speech-to-Text APIs that support the Sinhala language [1].

- **Natural Language Processing:** NLP techniques will be applied to the transcribed text to extract important information, such as instructions to do a certain task. This will help in identifying the most relevant parts of the transcript.

- **Image/Video Processing Technology**: Image/Video processing techniques will be used to analyze video frames and identify key visuals that correspond to the extracted instructions. This will enhance the overall clarity and effectiveness of the performance guide [2].

- **Document Generation Tools:** Different libraries will be used to generate PDF documents. Other synchronization technologies will be used for synchronizing the visuals with the instructions.

Nature of the Solution:

Input:
The input will be Sinhala demonstrations and potentially other instructional videos in Sinhala.

Process:

1. The system will first transcribe the video using speech to text technology.
2. The transcription will be cleaned and processed to correct any grammatical and language errors.
3. NLP will identify and extract key sections, such as instructions, and relevant processes.
4. The video will be analyzed to mark visual timestamps that correspond to the key sections.
5. The key sections will be summarized and written as instructions.
6. The required screenshots will be obtained from the video using the timestamps recorded. Zoom and other effects will be used for enhancing clarity.
7. Finally, a performance guide that combines text and visuals will be generated. The guide will provide synchronized instructions with video cues.

Output:
The output will be a performance guide that includes both textual summaries (instructions) and highlighted visual cues from the video.
The guide will be downloadable or accessible online, with an option to generate a printable version.

Users:
The primary users will be Sinhala-speaking audiences, particularly those looking for accurate guides to Sinhala demonstrations. This includes home cooks, learners, and content creators looking to improve the accessibility of their content.

Feasibility of Implementation:

Technical Feasibility: Existing technologies like Open-Source Speech-to-Text APIs and NLP tools support Sinhala, making the transcription and extraction processes feasible. The combination of NLP, image processing and video synchronization can be integrated to produce a solution.

Resource Feasibility: The necessary technologies, including APIs and video editing tools, are accessible. The system will be developed, using NLP, STT, Image and Image processing to ensure accurate results.

Team Capability: Our team is studying on how to implement machine learning, NLP, and video processing, which are important for developing and implementing the system. Additionally, familiarity with Sinhala content ensures that we understand the unique requirements of the target audience.

6. Resource Requirements

The following resources are required to develop the system for obtaining visual performance guides from Sinhala demonstration videos:

1. Hardware Resources:
    - Laptop: HP Omen
        - 16 GB RAM
        - 2.6 GHz
        - 500 GB SSD storage
    - Graphics card: NVIDIA GeForce GTX 1660 Ti
    - External storage: 5TB Seagate hard disk

2. Software Resources:
    - Operating System: Windows 11
    - IDE/Code Editors: Visual Studio Code, PyCharm
    - Programming Languages: Python
    - Libraries/Frameworks:
        - Machine Learning Libraries
        - Natural Language Toolkit
        - Image/Video Processing Libraries
    - Database

3. APIs & Cloud Services:
    - Speech-to-Text API (with Sinhala language support)
    - Cloud Storage

4. Internet Connectivity: High-speed internet connection.

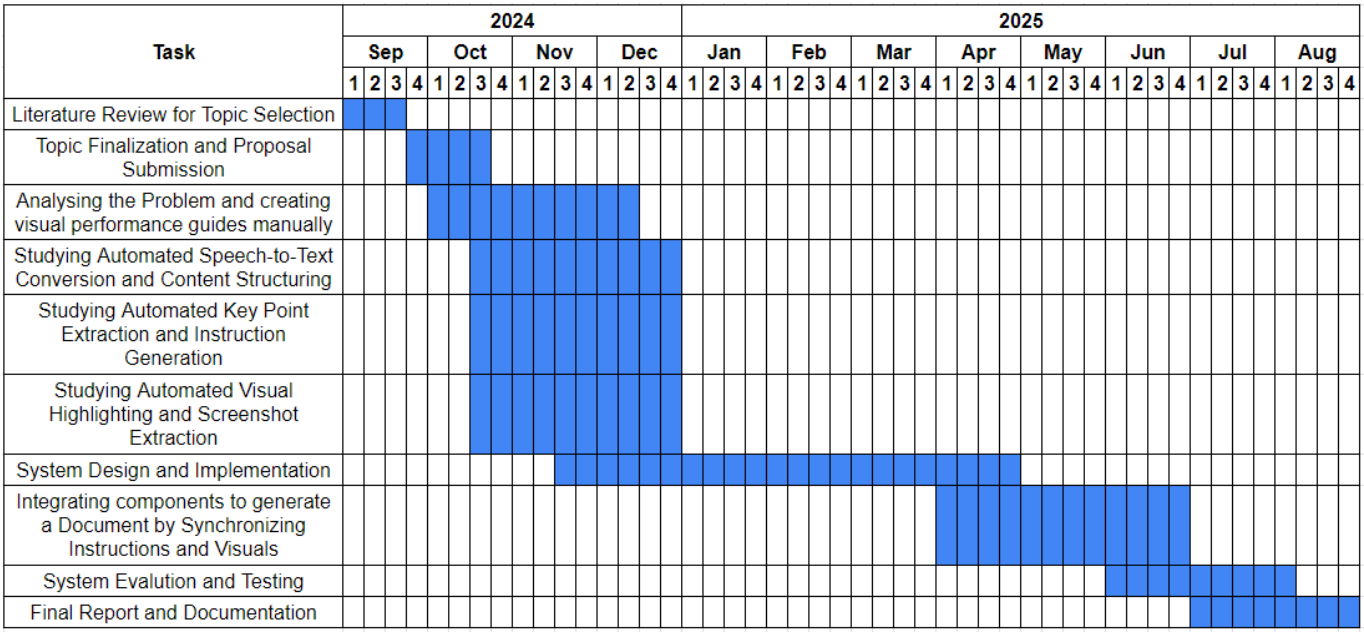
5. Other Tools:
    - Version Control: Git/GitHub for project management and code versioning
    - Image/Video Editing Software
    - Documentation Tools

References

[1] "PR6941-48.pdf." Accessed: Oct. 16, 2024. [Online]. Available: http://viduketha.nsf.gov.lk:8585/slsipr/PR6941/PR6941-48.pdf

[2] B. Kang, "A Review on Image & Video Processing," *Int. J. Multimed. Ubiquitous Eng.*, vol. 2, May 2007.

[3] M. A. Jahan and K. Wijesekara, "Automated text summarization of Sinhala online articles," 2023.


Task
2024
2025
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
Literature Review for Topic Selection
Topic Finalization and Proposal Submission
Analysing the Problem and creating visual performance guides manually
Studying Automated Speech-to-Text Conversion and Content Structuring
Studying Automated Key Point Extraction and Instruction Generation
Studying Automated Visual Highlighting and Screenshot Extraction
System Design and Implementation
Integrating components to generate a Document by Synchronizing Instructions and Visuals
System Evalution and Testing
Final Report and Documentation