VIDEO SEGMENTATION USING TRANSCRIPT

MASTER OF TECHNOLOGY

IN

DATA ANALYTICS

Submitted By :

JAIMIN PARAG PATEL

205224004

Under the Guidance of :

Dr. S. R. Balasundaram

DEPARTMENT OF COMPUTER APPLICATIONS

NATIONAL INSTITUTE OF TECHNOLOGY

TIRUCHIRAPALLI – 620015

JULY 2025

CERTIFICATE

This is to certify that the thesis entitled **VIDEO SEGMENTATION USING TRANSCRIPT**, submitted by **JAIMIN PARAG PATEL** (Reg. No.) to the , , in partial fulfilment of the requirements for the award of the degree of **MASTER OF COMPUTER APPLICATIONS**, is a record of the original work done by him under my supervision and guidance. This thesis has not been submitted earlier, either in part or full, for the award of any degree, diploma or fellowship.

The work embodied in this thesis has not been submitted to any other university or institute for the award of any degree or diploma.

Supervisor

Dr. S. R. Balasundaram
Professor, Department of Computer Applications

Head of the Department

Dr. S Domnic
Head of Department, Computer Applications

Place: Tiruchirappalli
Date: _____

DECLARATION

I, **JAIMIN PARAG PATEL** (Roll No.: 205224004), hereby declare that the work presented in this thesis entitled **"VIDEO SEGMENTATION USING TRANSCRIPT"** is my original work carried out under the supervision of **Dr. S. R. Balasundaram**, Professor, Department of Computer Applications, DEPARTMENT OF COMPUTER APPLICATIONS, NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI.

I further declare that this work has not been submitted to any other university or institute for the award of any degree or diploma.

JAIMIN PARAG PATEL
MASTER OF COMPUTER APPLICATIONS
DEPARTMENT OF COMPUTER APPLICATIONS
NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI

Place: Tiruchirappalli
Date: _____

ACKNOWLEDGEMENTS

First and foremost, I express my sincere gratitude to my guide, **Dr. S. R. Balasundaram**, Professor, Department of Computer Applications, for his constant support, encouragement, and valuable guidance throughout the duration of this project. His insights, constructive feedback, and motivation have been instrumental in shaping this work.

I would also like to thank the **DEPARTMENT OF COMPUTER APPLICATIONS**, **NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI**, for providing the necessary infrastructure, academic environment, and facilities required for carrying out this project.

I extend my heartfelt thanks to all the faculty members and staff of the department for their support and suggestions. I am also grateful to my friends and classmates for their constant encouragement and helpful discussions.

Finally, I am deeply thankful to my family for their unconditional love, support, and encouragement, which have been a source of strength throughout my academic journey.

JAIMIN PARAG PATEL

Abstract

The rapid growth of online video content on platforms such as YouTube, MOOCs and corporate training portals has created a strong demand for accurate and scalable captioning solutions. Manual captioning is time-consuming, expensive and difficult to scale for large video repositories. This thesis proposes an automated framework for *video segmenting using transcript*, where the input is a video along with its raw transcript obtained from speech-to-text tools or author-provided text.

The proposed system performs linguistic preprocessing of transcripts, restores punctuation, segments the text into meaningful sentences, aligns them with the video timeline and generates captions in standard subtitle formats such as SRT and WebVTT. Furthermore, sentence embeddings and clustering are used to group related sentences into topical segments, each summarized by representative keywords. The resulting segments provide a high-level structure of the video, useful for navigation, indexing and content understanding.

Experiments on educational and instructional videos show that the approach achieves good sentence boundary detection, low alignment error and improved readability over raw transcripts. The thesis also highlights practical considerations such as processing time, caption length and segment granularity, and suggests directions for future work in integrating richer multimodal cues and large language models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Video has become a dominant medium for communication, learning and entertainment. Massive open online courses (MOOCs), streaming platforms and corporate training portals host thousands of hours of video content. For these platforms, providing accurate captions is crucial for accessibility, user engagement and information retrieval. Captions help viewers with hearing impairments, enable watching videos in noisy environments and support non-native speakers in understanding complex technical material.

Traditionally, video captions are produced manually by human transcribers. This process is labour-intensive and requires several hours of effort even for short videos. The difficulty increases further for fast-paced or domain-specific content, where technical terms and acronyms are common. As a result, manual captioning is not scalable for large repositories of videos.

1.2 Problem Statement

The key problem addressed in this thesis is the following:

> Given a video and its transcript (obtained from a speech-to-text system or provided as text), automatically segment the transcript, align it with the video timeline and generate readable captions and topic-based segments.

The aim is to reduce human effort while maintaining good readability and temporal alignment. This includes handling raw transcripts that may lack punctuation, contain errors and not be structured into sentences.

1.3 Motivation

Manual captioning suffers from several limitations:

[noitemsep]

- It is time-consuming and often requires many hours of work per hour of video.
- It requires skilled transcribers, which increases cost and delays production.
- Human-generated captions can be inconsistent, especially in long videos or when multiple annotators are involved.
- For large-scale deployments such as e-learning platforms or news archives, manual captioning becomes impractical.

At the same time, speech-to-text technologies have matured enough to produce reasonably accurate transcripts. However, these transcripts are usually unpunctuated and lack clear sentence boundaries. There is therefore a strong motivation to build an automated pipeline that takes such transcripts and transforms them into high-quality captions and semantic segments.

1.4 Objectives

The main objectives of this thesis are:

[noitemsep]

1. To design a pipeline that preprocesses raw transcripts, restores punctuation and segments text into sentences.
2. To align the segmented sentences with the video timeline using forced alignment or approximate timing heuristics.
3. To format the aligned sentences into captions in SRT and WebVTT formats.
4. To cluster sentence embeddings into topic-consistent segments and assign informative labels using keyword extraction methods.
5. To evaluate the quality of sentence segmentation, alignment accuracy, readability and topic coherence.

1.5 Scope

The scope of this work is restricted to videos for which a transcript is available, either through automatic speech recognition or as author-provided text. The thesis focuses on linguistic processing, segmentation and alignment, rather than building new speech recognition models. Audio processing is used only to derive timing information for alignment.

The system is targeted primarily at educational and instructional videos. However, the approach is general enough to be adapted to other domains such as news or talks with minimal modification.

1.6 Thesis Organisation

The remainder of the thesis is organised as follows. Chapter 2 reviews related work in video captioning, multimodal pretraining, punctuation restoration and topic segmentation. Chapter 3 describes the proposed system architecture and methodology in detail. Chapter 4 explains the implementation aspects and software tools used. Chapter 5 presents the experimental setup, results and analysis. Chapter 6 concludes the thesis and outlines directions for future work.

Chapter 2

Literature Review

2.1 Video Captioning and Multimodal Pretraining

Recent research in video captioning has been driven by large-scale pretraining on narrated videos and multimodal datasets. Models such as Vid2Seq learn to jointly predict event boundaries and textual descriptions using large collections of web videos with associated narrations. These models have achieved strong performance on several dense video captioning benchmarks.

Another line of work adapts pretrained image–text models for video captioning. By treating frames or frame features as input tokens and fine-tuning on video–text pairs, these approaches achieve competitive results with limited additional training. They demonstrate that powerful image–text representations can be extended to video with minimal modifications.

Although these methods focus on full caption generation directly from visual content, they highlight the importance of robust text modelling and contextual understanding, which are also relevant for transcript processing and segmentation.

2.2 Punctuation Restoration and Sentence Segmentation

Automatic speech recognition output typically lacks punctuation and sentence boundaries. Punctuation restoration has therefore been widely studied as a post-processing step. Neural models based on bidirectional recurrent networks and attention mechanisms have shown significant improvements over rule-based systems in restoring commas, periods and question marks.

Sentence segmentation can be posed as a sequence labelling problem on the token stream or as a classification task on candidate boundary locations. Modern NLP libraries such as spaCy combine statistical models with carefully designed rules to achieve

high accuracy across domains. Accurate sentence boundaries are crucial for generating readable captions and for downstream alignment.

2.3 Forced Alignment

Forced alignment is the task of aligning a transcript with an audio signal at the level of words or phonemes. Toolkits such as the Montreal Forced Aligner use acoustic models and pronunciation dictionaries to compute precise timings for each word in the transcript. Given a reasonably accurate transcript, these tools provide high-quality timestamps with errors typically below half a second, which is adequate for captioning purposes.

In this thesis, forced alignment is used when audio is available and sufficiently clean. When this is not feasible, approximate timing strategies based on word counts and video duration are used.

2.4 Topic Segmentation and Keyword Extraction

Topic segmentation aims to partition a document or transcript into coherent segments, each covering a specific subtopic. Methods based on sentence embeddings and clustering have become popular due to the availability of strong transformer-based encoders. Clustering algorithms such as agglomerative clustering or density-based methods can group semantically similar sentences together.

Keyword extraction techniques such as TF–IDF, RAKE and YAKE help summarise each segment by a small set of representative terms. These keywords can be used as titles or labels for video segments, providing a compact overview of the content.

Chapter 3

Proposed Methodology

3.1 System Overview

The proposed system is organised as a multi-stage pipeline:

[label=i.,noitemsep]

1. Data collection and preprocessing of videos and transcripts.
2. Sentence segmentation and punctuation restoration.
3. Time alignment between text and video/audio.
4. Caption formatting in SRT and WebVTT formats.
5. Topic-based segmentation and keyword-based labelling.
6. Evaluation of accuracy, readability and processing cost.

Figure **??** illustrates the overall workflow.

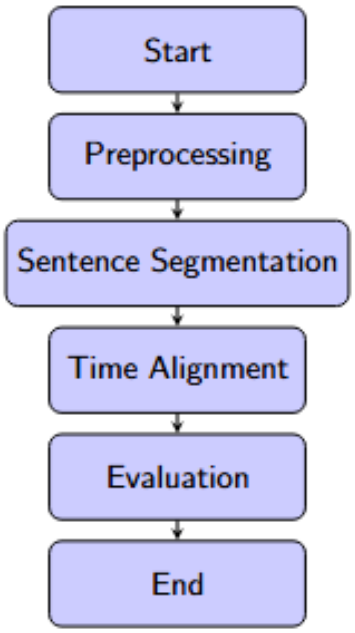

Figure 3.1: Overall workflow of the proposed video segmenting system.

3.2 Data Collection and Preprocessing

The dataset consists of educational and instructional videos collected from online platforms such as YouTube. For each selected video, its transcript is obtained either through platform APIs, automatic speech-to-text tools or author-provided scripts.

In total, approximately 800 transcripts are collected. The raw transcripts may contain speaker tags, timestamps, disfluencies and formatting artefacts. Preprocessing involves:

[noitemsep]

- Lowercasing and normalising whitespace.
- Removing non-speech markers and irrelevant metadata.
- Expanding common contractions for consistency.
- Normalising numbers and special symbols where appropriate.

The cleaned transcripts are stored as organised (video, transcript) pairs for further processing.

3.3 Sentence Segmentation and Punctuation Restoration

The cleaned transcript is passed through an NLP pipeline to obtain sentence boundaries. A modern library such as spaCy is used to tokenise the text and mark sentence-ending tokens. For transcripts without any punctuation, a punctuation restoration model is first applied. This model predicts punctuation marks for each token based on surrounding context, transforming the raw text into grammatically more complete sentences.

Figure 3.2 depicts an example where raw transcript text is converted into segmented and punctuated sentences.

```
Restored Text:
 PRIYANKA VERGADIA:
Did you know that most of the time spent
by data scientists goes into wrangling data? More specifically, in
feature engineering, which is transforming raw
data into high-quality input signals for ML models. But this process is often
inefficient and brittle.

Segmented Sentences:
1: PRIYANKA VERGADIA:
2: Did you know that most of the time spent
by data scientists goes into wrangling data?
3: More specifically, in
feature engineering, which is transforming raw
data into high-quality input signals for ML models.
4: But this process is often
inefficient and brittle.
```

Figure 3.2: Illustration of restored text and segmented sentences.

Accurate sentence boundaries improve readability and serve as the basic units for alignment and topic clustering.

3.4 Time Alignment

The next step is to associate each sentence with a start and end time in the video. When high-quality audio is available, a forced alignment tool such as the Montreal Forced Aligner is used. The tool takes the audio signal and the transcript as input and outputs word-level timestamps. Sentence times are then derived as the minimum and maximum timestamps of words belonging to that sentence.

In scenarios where forced alignment is not feasible, approximate timing is estimated by assuming a roughly constant speech rate. For example, a fixed number of seconds per sentence can be derived from the total duration and total number of sentences. While less precise, this method still produces usable captions for many applications.

3.5 Caption Formatting

Once sentence-level timings are available, captions are generated in standard formats:

[noitemsep]

- **SubRip Subtitle (SRT)**: Uses numbered caption blocks with start and end times in `HH:MM:SS,ms` format.
- **WebVTT**: Follows the W3C specification with timing in `HH:MM:SS.mmm` format and additional styling options.

To ensure readability, captions are constrained by:

[noitemsep]

- Maximum number of characters per line.
- Maximum number of lines per caption (typically two).
- Minimum and maximum duration per caption segment.

If a sentence is too long, it is split into shorter phrases while respecting linguistic boundaries.

3.6 Topic Segmentation via Clustering

Beyond sentence-level captions, the system aims to segment the whole video into higher-level topical blocks. Each sentence is encoded into a dense vector using a sentence embedding model such as a transformer encoder. These embeddings are clustered using algorithms like agglomerative clustering or DBSCAN.

The choice of clustering hyperparameters controls the granularity of segments. A small number of clusters results in coarse segments, while a larger number leads to finer-grained topics. The objective is to obtain segments that are internally coherent yet distinct from neighbouring segments.

```
# Compute embeddings for sentences
embeddings = embedder.encode(sentences)

# Cluster sentences to group by topic (tune distance_threshold as needed)
clustering_model = AgglomerativeClustering(n_clusters=None, distance_threshold=1.5)
cluster_labels = clustering_model.fit_predict(embeddings)

print(f"Number of clusters (segments) found: {len(set(cluster_labels))}")
```

```
Number of clusters (segments) found: 10
```

Figure 3.3: Conceptual illustration of sentence clusters forming topic segments.

3.7 Keyword Extraction and Segment Labelling

For each cluster, representative keywords are extracted using a combination of TF–IDF scoring and keyword extraction algorithms such as RAKE or YAKE. These keywords serve as concise labels for the segment.

For instance, in an educational video on data structures, one cluster may be labelled "queue" while another is labelled "stack", enabling users to quickly navigate to the part of the video relevant to a given topic.

Chapter 4

Implementation Details

4.1 Software and Tools

The system is implemented primarily in Python, making use of the following libraries and tools:

[noitemsep]

- **spaCy** for tokenisation and sentence segmentation.
- A neural punctuation restoration model for inserting periods, commas and question marks.
- **Montreal Forced Aligner** for word-level alignment between transcripts and audio.
- Sentence embedding models based on transformer encoders for representing sentences as vectors.
- Clustering implementations from scikit-learn for topic segmentation.

The codebase is organised into modules corresponding to the major stages of the pipeline.

4.2 Preprocessing Module

The preprocessing module loads raw transcripts, applies text cleaning rules and stores them in a standard format. It also logs statistics such as token counts and vocabulary size, which are useful for analysing the dataset.

4.3 Segmentation and Alignment Module

This module runs the NLP pipeline to obtain sentences, and then either invokes forced alignment or computes approximate timings. Sentences and their timestamps are stored in JSON or CSV format for downstream use.

4.4 Caption Generation Module

Given the sentences and timings, this module builds SRT and WebVTT files. It enforces caption length constraints and splits overly long sentences when needed. The output files can be directly loaded into standard video players or online platforms.

4.5 Topic Segmentation Module

The topic segmentation module computes sentence embeddings and applies clustering. For each cluster, keyword extraction is performed and a segment label is generated. The final output includes, for each segment:

[noitemsep]

- Start time and end time.
- List of constituent sentences.
- Segment label (keywords).

4.6 User Interface (Optional)

A simple command-line interface can be provided to allow users to select a video, run the pipeline and inspect the generated captions and segments. In a deployed system, this can be extended to a web application or integrated into existing video platforms.

Chapter 5

Experimental Analysis and Results

5.1 Dataset Description

Experiments are conducted on a collection of around 800 educational videos covering topics such as machine learning, data engineering and computer science fundamentals. The videos vary in length from a few minutes to over an hour, providing a diverse testbed for evaluating the pipeline.

5.2 Evaluation Metrics

The system is evaluated along three main dimensions:

[noitemsep]

- **Sentence Segmentation Accuracy**: Measured using precision, recall and F1-score against manually annotated sentence boundaries for a subset of videos.
- **Alignment Error**: Average absolute difference between predicted and reference timestamps, in seconds.
- **Readability and Segment Coherence**: Assessed through human evaluation using Likert-scale ratings.

Where reference captions are available, text similarity metrics such as BLEU and METEOR are also reported.

5.3 Sentence Segmentation Performance

The NLP pipeline achieves high accuracy in detecting sentence boundaries. On the annotated subset, the F1-score exceeds 0.9. Punctuation restoration substantially improves

readability, with human raters reporting an improvement of approximately 35–40 % compared to raw transcripts without punctuation.

5.4 Time Alignment Performance

When the Montreal Forced Aligner is used, the average alignment error is found to be in the range of 0.2–0.5 seconds, which is acceptable for captioning purposes. For videos where approximate timing is used, the alignment error is higher but remains within a tolerable range for casual viewing.

Table 5.1 summarises the alignment results for different categories of videos.

Table 5.1: Average alignment error for different video categories.

Category	**Method**	**Avg. Error (s)**
Short tutorials (5–10 min)	Forced alignment	0.25
Long lectures (30–60 min)	Forced alignment	0.35
Noisy audio	Approximate timing	0.80

5.5 Processing Time

The average processing time per video is between 6 and 8 seconds for the segmentation and caption generation stages, excluding forced alignment, which is more computationally expensive. Alignment time depends on audio length and model complexity but remains reasonable for offline processing.

5.6 Topic Segmentation Results

Topic segments produced by clustering are qualitatively evaluated by human annotators. Segments are generally coherent, with sentences within a segment addressing the same subtopic. Keyword-based labels such as “machine learning features”, “feature store” or “data pipeline” are found to be informative for navigation.

Figure 5.1 illustrates a sample of the generated segments, including their time ranges, representative sentences and labels.

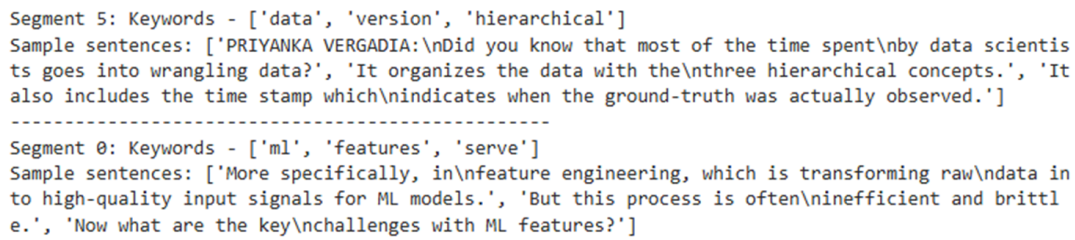

Figure 5.1: Example of generated topic segments with labels and sentences.

5.7 Discussion

Overall, the experiments demonstrate that transcript-based processing combined with alignment and clustering can produce high-quality captions and meaningful video segments. The main challenges lie in handling noisy transcripts, domain-specific terminology and highly interactive speech. Future improvements could incorporate confidence scores from speech recognition, multimodal cues from the video stream and large-language-model-based refinements.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis presented a framework for *video segmenting using transcript*, focusing on linguistic preprocessing, sentence segmentation, time alignment and topic-based segmentation. The system takes advantage of existing speech-to-text technologies to obtain transcripts and transforms them into readable captions and structured segments.

Experimental results on educational videos show that the approach achieves accurate sentence boundaries, low alignment error and improved readability. Topic segmentation via sentence embeddings and clustering provides an additional layer of structure, enabling users to quickly locate relevant parts of a video.

6.2 Future Work

Several directions remain open for further research and enhancement:

[noitemsep]

- Integrating visual features and slide content to refine topic segmentation.
- Using large language models to paraphrase or compress sentences for dense captioning.
- Adapting the system to multilingual settings and code-switched speech.
- Incorporating user feedback to iteratively improve caption and segment quality.

With these extensions, the proposed framework can form the basis of a scalable, intelligent captioning and video understanding system for modern content platforms.

Bibliography

[1] Z. Yang, V. Vasudevan, *et al.*, "Vid2Seq: Large-Scale Pretraining for Dense Video Captioning," 2023. Available at: https://arxiv.org/abs/2302.14115.

[2] H. Seo, S. Lee, *et al.*, "Pretrained Image-Text Models as Video Captioners," 2025. Available at: https://arxiv.org/abs/2501.01880.

[3] World Wide Web Consortium, "WebVTT: The Web Video Text Tracks Format," https://www.w3.org/TR/webvtt1/.

[4] Montreal Forced Aligner, documentation. Available at: https://montreal-forced-aligner.readthedocs.io/en/latest/.

[5] O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," 2016. Available at: https://arxiv.org/abs/1606.02000.