International Scientific Journal of Engineering and Management
ISSN: 2583-6129
Volume: 03 Issue: 07 | July – 2024
DOI: 10.55041/ISJEM02092
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Audio/Video Transcriber Using NMT in Different Languages

Dr Sayyada Fahmeeda Sultana[1], Arusa Konain[2], Bhavana Reddy[3], Neelambika Kolar[4]

[#1]*Department of Computer Science, Visveswaraya Technological University PDA Kalaburagi, Karnataka, IndiaEmail: arusakonain09@gmail.com*

[*2]*Department of Computer Science, Visveswaraya Technological University PDA Kalaburagi, Karnataka, IndiaEmail: bhavanareddy250@gmail.com*

[#3]*Department of Computer Science, Visveswaraya Technological University PDA Kalaburagi, Karnataka, IndiaEmail: neelambikaskolar5167 @gmail.com*

***Abstract*— Multimedia data is represented as electronic signals that can be recorded, processed, and reproduced. The detection and extraction of scene and caption text from unconstrained, educational video is an important research problem in the context of content-based retrieval. The project presents a reliable system for detecting, localizing, extracting, tracking and binarizing text from unconstrained, educational video. The features of speech differ with each language, even while communicating in the same language, the pace and the dialect varies with each person. Speech recognition which is an inter disciplinary field of computational linguistics aids in developing technologies that empowers the recognition and translation of speech into text. The project proposed to transcribe and translate educational audio/videos in different regional languages. Audio files are transcribed into text using Natural Language processing (NLP) techniques like Flask, Speech recognition, Pytest, Gunicorn. The translation is performed using Neural Machine Translation (NMT). As education videos contain presentations which include important point those important points are extracted by using Optical Character Recognition (OCR). The project focuses on NPTEL videos, educational and news video. Neural Machine Translation (NMT) for translating text into different languages i.e. the neural network is trained on vast amounts of multilingual text data. The objective is to achieve time efficiency and accuracy in transcription.**

***Keywords*— NLP (Natural Language Processing), OCR (Optical Character Recognition), Flask, Speech Recognition, NMT (Neural Machine Translation).**

I. INTRODUCTION

The essential component for advancement in the current sector is communication. Not just on a corporate level, but also on a personal level, it is crucial to convey information to the appropriate person and in the appropriate way. In today's technologically advanced society, communication via phone calls, emails, text messages, and other channels has become essential. Now a days the online mode or work in online has the major part in the educational departments, jobs, and much more. The project focus is on NPTEL, educational and news video. Many applications that function as a mediator and aid in efficiently transmitting messages in the form of text or audio signals over miles of networks have emerged in order to fulfil the objective of effective communication between two parties without obstacles.

Natural Language Processing (NLP) stands at the forefront of artificial intelligence, focusing on the interaction between computers and human language. This interdisciplinary field combines elements of linguistics, computer science, and cognitive psychology to enable machines to understand, interpret, and generate human-like language. NLP empowers computers to process and analyze vast amounts of natural language data, encompassing written, spoken, or textual communication. One of the primary goals of NLP is to bridge the gap between human communication and computer understanding.

The NLP techniques include Flask, Speech Recognition, Pytest, Gunicorn functions. Building a audio to text conversion application can be streamlined using Flask, Speech Recognition, Pytest, and Gunicorn. Flask, a lightweight web framework for Python, facilitates the development of web applications and APIs, providing the foundation for handling HTTP requests and responses. The Speech Recognition library in Python simplifies the process of capturing and converting spoken language into text by supporting multiple speech engines and APIs, making it easy to integrate speech-to-text functionality into the Flask application. To ensure the reliability and accuracy of the application, Pytest offers a powerful testing framework that allows developers to write and execute tests, ensuring each component of the application performs as expected. Finally, Gunicorn, a Python WSGI HTTP server, is employed to serve the Flask application, providing robust, production-ready deployment with the ability to handle multiple requests simultaneously, thereby ensuring scalability and performance. Together, these tools create a seamless, efficient pipeline for developing, testing, and deploying a speech-to-text conversion service.

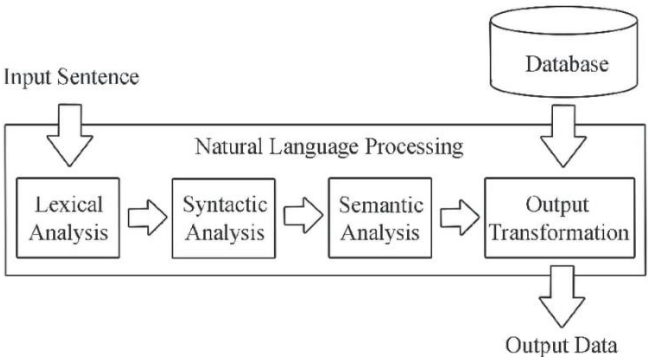

Figure 1: NLP (Natural Language Processing) Workflow

International Scientific Journal of Engineering and Management
ISSN: 2583-6129
Volume: 03 Issue: 07 | July – 2024
DOI: 10.55041/ISJEM02092
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

NLP workflow consists of 4 major steps:

1. Lexical Analysis

The process of splitting a sentence into words or small units called lexical analysis. In order to identify the meaning of it and its relationship to the entire sentence. Lexical analysis, also known as tokenization, is a fundamental stage in natural language processing (NLP) workflows.

2. Syntactic Analysis

The process of identifying the relationship between the different words and phrases within a sentence, standardizing their structure. Syntactic analysis, or parsing, is a critical stage in natural language processing (NLP) that involves analysing the grammatical structure of sentences.

3. Semantic Analysis

The process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings. Semantic analysis, also known as semantic parsing, is a crucial stage in natural language processing (NLP) that focuses on understanding the meaning conveyed by a text

4. Output Transformation

The process of generating an output based on the semantic analysis of the text or speech which fits the target of the application. Output transformation is a crucial stage in the natural language processing (NLP) workflow where the processed and analyzed data is converted into a desired format or response

II. RELATED WORK

Recent Advances in Augmented Reality the paper explores the latest developments in augmented reality technology, highlighting its applications and advancements. [1] Spatial Augmented Reality Merging Real and Virtual Worlds (2020) The paper delves into the integration of physical and virtual environments, presenting techniques for spatial augmented reality. [2] Beyond Attention: Exploring Transformer Variants in NMT (2019) The investigates various transformer models in neural machine translation beyond traditional attention mechanisms.[3] Multilingual Neural Machine Translation: State-of-the-Art (2018) provides an overview of the latest advancements in multilingual neural machine translation techniques, showcasing the current state-of-the-art methods. [4] Textless Speech-to-Speech Translation on Real Data" (2018) presents research on speech-to-speech translation without the need for text intermediaries, focusing on real-world data applications. [5] Spectral Subtraction Based on an Adaptive Filter for Noise Reduction (2017) introduces a method for noise reduction in speech signals using adaptive filtering based on spectral subtraction. [6] End-to-End Multilingual Speech Recognition with Transformer (2015) proposes an end-to-end approach for multilingual speech recognition using transformer models, offering improved accuracy and efficiency. [7] Shot detection in video sequences using entropy-based metrics (2017) presents a method for shot detection in video sequences using entropy-based metrics, enhancing the automation of video analysis tasks.

III. PROPOSED SYSTEM

The proposed system for an Audio/Video to text transcriber project is designed to efficiently transcribe spoken language from audio sources into text. The system would likely consist of several key components, including audio input processing, speech recognition, neural machine translation, natural language processing for improving accuracy, and a user-friendly interface. The input audio would be processed, and speech recognition software would convert it into text, which could then be further refined using language processing techniques to enhance accuracy and readability. Users would interact with the system through a user-friendly interface, allowing them to input audio or video files or live speech for transcription. The proposed audio/video to text converter system is designed as a versatile and accurate solution for transcribing content in multiple languages. This comprehensive system incorporates advanced features such as language-agnostic preprocessing techniques for both audio and video inputs, ensuring adaptability to diverse linguistic contexts. The core of the system lies in a modular and multilingual Optical Character Recognition (OCR) engine, equipped with dynamic language identification capabilities for real-time language switching. Context-aware language models enhance transcription accuracy, while user-friendly interfaces allow for customization and manual corrections. The system prioritizes security and privacy, complying with data protection regulations, and integrates continuous learning mechanisms for improved accuracy over time. With scalable architecture, performance optimization, and thorough documentation, the proposed system aims to provide a seamless and efficient experience for converting audio and video content into text across a spectrum of languages.

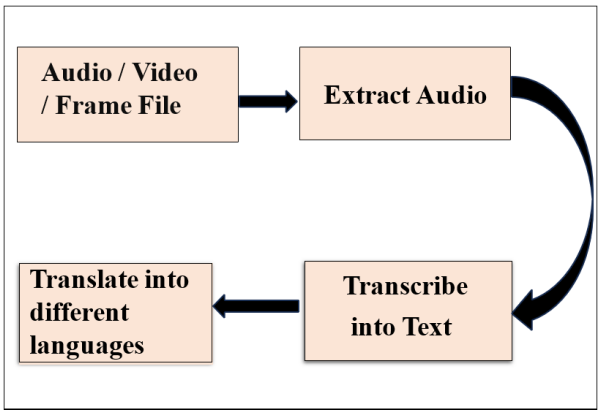

Figure 2: Block Diagram of the Proposed System

Figure 2 shows the block diagram that involves the steps that the system goes through to transcription of the text. For a perfect transcription, the system has to first take a source document or file that is in the form of .wav file type as an input and when we click on transcribe button it displays the clean text to the user on the user interface in the web page.

International Scientific Journal of Engineering and Management
ISSN: 2583-6129
Volume: 03 Issue: 07 | July – 2024
DOI: 10.55041/ISJEM02092
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

VI. METHODOLOGY

The encoder transforms the audio into an acoustic embedding. The model is made up of an encoder and a decoder. First, the model's encoder uses deep recurrent neural networks to extract features from the audio at multiple levels of abstraction, called an acoustic embedding. An acoustic embedding is multidimensional representation of sound that can be used for classification tasks. The decoder predicts the next character using the acoustic embedding and the previously predicted characters to form the final transcription.

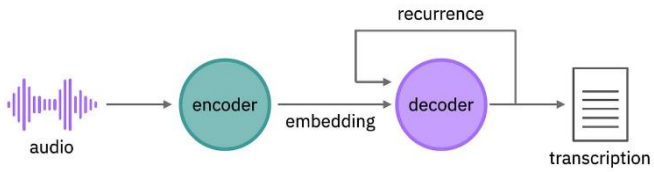

Figure 3: Real Time Transcription

This project uses popular API's like: -

Flask: Is a lightweight web framework for Python, widely used for building web applications.

Speech Recognition: The application uses advanced speech recognition algorithms to accurately transcribe spoken words.

Pytest: Is a testing framework for Python that makes it easy to write simple unit tests as well as complex functional tests.

Gunicorn: HTTP server for running Python web applications. It's commonly used to deploy web applications built with frameworks like Flask.

The technique's used for converting video, frames to text includes as follows:

Optical Character Recognition (OCR): OCR is a technology that recognizes and extracts text from images or frames. In video processing, each frame is treated as an image, and OCR is applied to recognize and extract text from these frames.

Speech-to-Text (STT): This technique involves converting spoken words in a video into text and analyse audio content and transcribe it into a textual format.

Subtitle Extraction (FFmpeg): Many videos include subtitles or closed captions. Extracting text from these subtitles is a straightforward method for video-to-text conversion. Tools like FFmpeg or specialized subtitle extraction libraries can be used to isolate and utilize the textual information already present in the video.

TEXT RECOGNITION AND EXTRACTION ALGORITHM

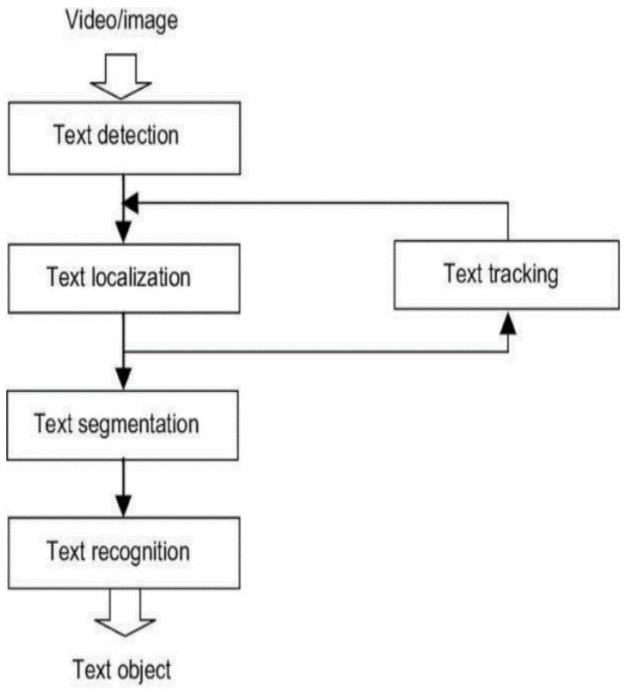

Figure 4: Text Detection Process

• **Text Region Recognition**: The MSER (Maximally Stable External Region) algorithm is used to detect candidate text region from the given image.

• **Removal of Non-Text Region**: The MSER may also detect non text regions. Stroke width is used to discriminate between text and non-text regions.

• **Merge Text Regions for Final Detection**: All the detection results are composed of individual text characters. To use this result for recognition task, the individual text characters must be merged into words.

• **Recognize Detected Text Using OCR**: After detecting the text regions, the OCR (optical character recognition) method is used to recognize the text.

V. OBJECTIVES

• Transcribe multiple languages of Audio/Video recordings or file to English text.

• The goal is to translate text into multiple languages English, Hindi, Kannada, Telugu).

• The main objective is to achieve time efficiency and accuracy in transcription.

• The main focus is on the frames with text in NPTEL videos, it transcribes English text into different languages.

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 07 | July – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02092

V. RESULTS AND DISCUSSIONS

The main module developed in the system. The description of the module is done below.

- Home Page
- Audio Conversion
- Video Conversion

Home Page

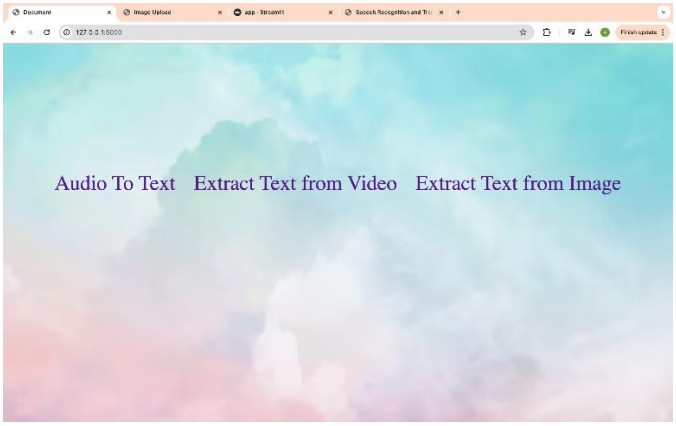

Figure 5: Home Page

The Figure 5 shows web page that contains Audio to text, Image to text, Video to text. The home page serves as the main interface for users interacting with the web application. The home page is designed to be user- friendly and accessible, allowing users to easily navigate through the application. The Navigation Bar likely contains links to Audio to Text, Frames to Text, and Video to Text.

Audio Conversion

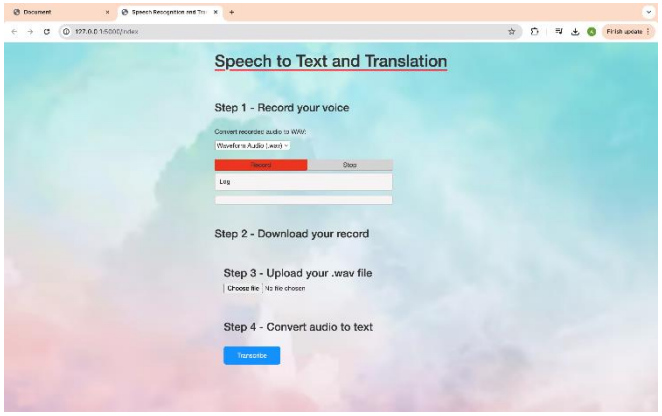

Figure 6: Audio to text Conversion

The Figure 6 shows web page that takes the live recording of the user of any duration. Now, download the recorded file. The file converts into the .wav file type. Click on Transcribe button. The system displays result of the clean text. your personnel computer as input and then takes the text from the file. Or else the system directly chooses audio file from the computer and transcribe the text.

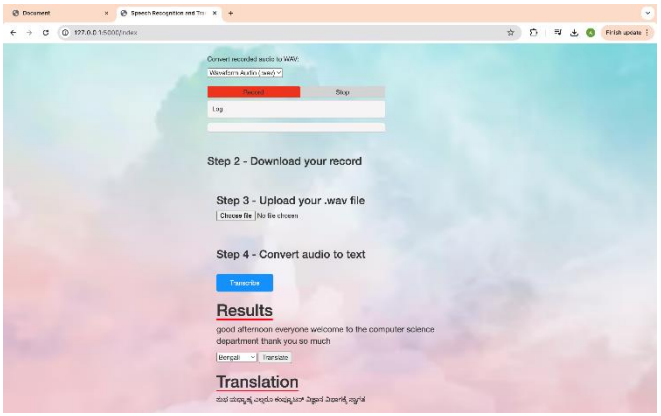

Figure 7: Audio To Text, Translation

In Figure 7, the page is specifically designed for processing audio files. There is a File Upload Button where users can upload audio files (e.g., in WAV or MP3 format) to the system. Once the audio file is uploaded, the system uses Speech Recognition, Flask, Pytest technology to convert the spoken words into text. After transcription, users can select target languages for translation. The system employs Neural Machine Translation (NMT) models to translate the transcribed text into the selected language. The page displays the original transcribed text and the translated version.

Video Conversion

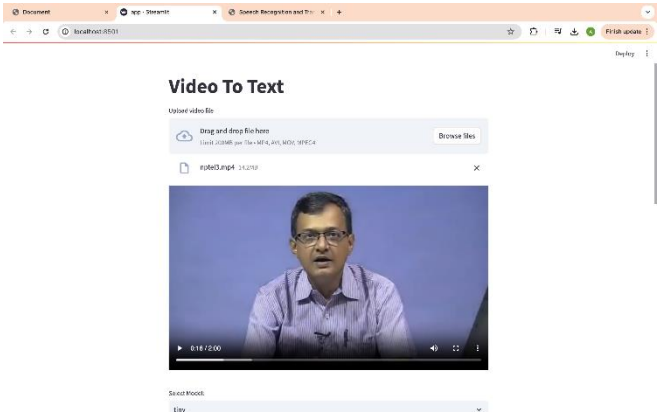

Figure 8: Video to text Conversion

The Figure 8 shows web page that contains video to text. The system contains choose browser button to select the file i.e. educational video and a submit button and then the translation button that converts in different languages.

International Scientific Journal of Engineering and Management
ISSN: 2583-6129
Volume: 03 Issue: 07 | July – 2024
DOI: 10.55041/ISJEM02092
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

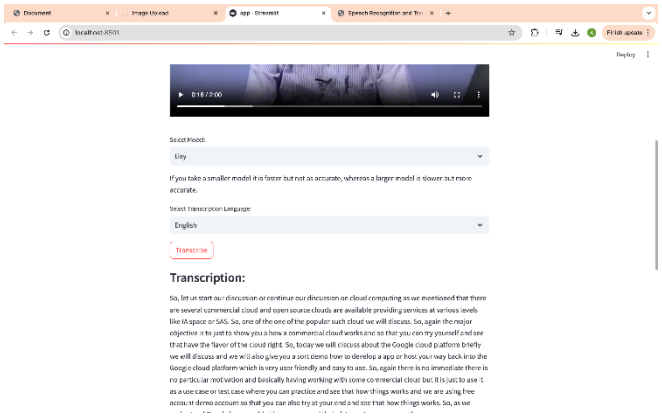

Figure 9: Video to text to different languages

In Figure 9, Users can translate the combined text into multiple languages using the Neural Machine Translation (NMT) models. The final output, including transcribed and translated text.

VI. CONCLUSION

Audio and video transcription in multiple languages is an indispensable tool for the society, offering a bridge across linguistic barriers. The process involves converting spoken content from audio or video files into written text, thereby facilitating accessibility, comprehension, and searchability. This service plays a pivotal role in sectors such as media, entertainment, education, legal proceedings, healthcare, and more, enabling wider audience reach and inclusivity. Furthermore, transcription in different languages enhances content localization, catering to diverse global audiences and fostering cross-cultural communication. It serves as a foundation for subtitling, translation, content analysis, and data mining, contributing significantly to research, language preservation, and information dissemination. The efficacy of transcription depends on accuracy, linguistic nuances, and technological advancements in speech recognition an language processing tools.

Future scope:

The future scope for audio-video transcription in multiple languages is promising and expansive. With the growing digital landscape and global connectivity, the demand for accurate, real-time transcription services across languages is expected to soar. The incorporation of artificial intelligence and machine learning algorithms into transcription services is poised to make these processes more adaptable, scalable, and cost effective.

References:

[1] "Recent Advances in Augmented Reality‖ by Hua, H. Gao published in the year 2021.
[2] "Spatial Augmented Reality: Merging Real and Virtual Worlds‖ by Bimber, O, Raskar, p published in the year 2020.
[3] "Beyond Attention: Exploring Transformer Variants in NMT‖ by Azuma, R.T, Baillot.Y, Behringer, R published in the year 2019.
[4] "Multilingual Neural Machine Translation: State-of-the-Art‖ by D. Krishnan, J. Salomon, and P. Antoni published in the year 2018.
[5] —Textless Speech-to-Speech Translation on Real Data‖ by Ann Lee, Hongyu Gong, PaulAmbroise Duquenne published in the year 2018.
[6] —Spectral Subtraction Based on an Adaptive Filter for Noise Reduction‖ by Daniel S. Park, William Chan, Yu Zhang, et al published in the year 2017.
[7] —End-to-End Multilingual Speech Recognition with Transformer‖ by Xiaoyi Zhang, Fuming Fang, Lei Xie, et al published in the year 2015.
[8] "Shot detection in video sequences using entropy-based metrics‖ by Cerenkov, Z Greece
Nikou, C. Pitas, I published in the year 2017.
[10] "Enhancing Video Captioning Through Transfer Learning in Speech Recognition" by Garcia, R published in the year 2023.
[9] —SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition by Daniel S. Park, William Chan, Yu Zhang, et al published in the year 2014.
[11] —Extraction of Unconstrained Caption Text from General-Purpose Video‖ by D.J. Crandall Master's thesis, The Pennsylvania State University, University Park, USA, published in the year May 2015.
[12] —Image Sequence Analysis for Object Detection and Segmentation‖ by T. Gandhi. PhD thesis, The Pennsylvania State University, University Park, published in the year 2014.
[13] —Character String Extraction by Multi-stage Relaxation‖ by H. Hase, T. Shinokawa, M. Yoneda, H. Sakai, and H. Maruyama In International Conference on Document Analysis and Recognition, published in the year 2013.
[14] —Automatic Text Location in Images and Video Frames‖ by A.K. Jain and B. Yu published in the year 1998.
[15] —Robust Multifont OCR System from Gray Level Images‖ by F. LeBourgeois in Inter- national Conference on Document Analysis and Recognition, published in the year 1997.
[16] —Automatic Text Recognition for Video Indexing‖ by R. Lienhart and F. Stuber in Proceedings of the ACM International Multimedia Conference & Exhibition, published in the year 1996.
[17] —Extraction of special effects caption text events from digital video, ‖ by David Crandall, Sameer Antani, Rangachar Kasturi, International Journal on Document Analysis and Recognition, published in the year 2019.
[18] —A new approach for video text detection, ‖ by Min Cai, Jiqiang Song, and Michael R. Lyu, IEEE International Conference on Image Processing, published in the year 2002.
[19] —Arvind, Mohamed Rafi "Text Extraction from Images Using Connected Component Method" Journal of Artificial Intelligence Research & Advances Volume in 2019.
[20] Lifang Gu "Text Detection and Extraction in MPEG Video Sequences" In Proceedings of the International Workshop on Content-Based Multimedia Indexing, 2015.

