VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

(An Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

Project Report on

Smart Sorting: Deep Learning-Powered Metadata Creation for Documentary Video Catalog

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN **COMPUTER ENGINEERING**

By

Pranav Rane 46

Anmol Gyanmote 16

Amisha Chandwani 02

Project Mentor

Prof. Abha Tewari

University of Mumbai

(AY 2023-24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

(An Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

CERTIFICATE

This is to certify that the Mini Project entitled **"Smart Sorting: Deep Learning-Powered Metadata Creation for Documentary Video Catalog"** is a bonafide work of **Pranav Rane(D12B/46), Anmol Gyanmote(D12B/16), Amisha Chandwani (D12A/2)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering" .**

(Prof._____)

Mentor

(Prof._____)

Head of Department

(Prof._____)

Principal

Mini Project Approval

This Mini Project entitled "**Smart Sorting: Deep Learning-Powered Metadata Creation for Documentary Video Catalog"** by **Pranav Rane(D12B/46), Anmol Gyanmote(D12B/16), Amisha Chandwani (D12A/2)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1……

(Internal Examiner Name & Sign)

2……

(External Examiner name & Sign)

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---------- ----------

(Signature) (Signature)

---------- ----------

(Name of student and Roll No.) (Name of student and Roll No.)

(Signature)

(Name of student and Roll No.)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Prof. Abha Tewari** for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to the Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We would also like to express our gratitude to our senior **Mr. Vivek Balani** for providing us major guidance and a helping hand during the implementation of our project and for sharing their knowledge with us.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Index

Abstract

Video is a vital and effective form of data which is capable of providing actual pictures of the scenario at hand. Nowadays this form of data is available in large numbers. Categorizing the videos allows us to easily access them. To incorporate this process manually is a hectic and time consuming task.

Automating the video classification and metadata generation task by utilizing the deep learning methods and computer vision techniques can overcome the issue. Our objective in this paper is to incorporate novel approach in sorting the videos and classifying them into predefined categories and further extracting textual information, carrying out transcription of the video and detecting a number of predefined objects in the same.

For this purpose we have a dataset that includes a large number of clips that are to be sorted into 11 categories by processing them using computer vision and neural networks. Metadata extraction was executed by Natural Language Processing (NLP) and Named Entity Recognition (NER). We believe that this project would be helpful for a number of industries providing them an automated way to store and retrieve their audio visual form of data and reduce the manpower required for the same.

The results for the process have been found after thoroughly examining multiple deep learning approaches and other computational techniques that would be capable of providing us the best result possible.

List of Abbreviations

Sr. No.	Abbreviations	Definitions
1	NER	Named Entity Recognition
2	ISRO	Indian Space Research Organisation
3	CNN	Convolutional Neural Network
4	RNN	Recurrent Neural Network
5	LSTM	Long short-term memory
8	GB	Gigabyte
9	ResNet	Residual Network

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Introduction

Video metadata generation is the process of extracting data from the videos using computer vision techniques. On the other hand video classification is categorizing them as per the genre which they propagate. Essentially, the creation and categorization of video metadata allows users to effortlessly traverse the enormous amount of available video content, facilitating more effective search functions, better content suggestions, and better content management. The accuracy and application of video metadata creation and categorization techniques are guaranteed to transform the way we engage with and find video content as technology advances.

A small amount of the dataset, which contains a variety of clips from ISRO documentaries, will be utilized for training. The aim of this project is to automate the process of creating metadata and classifying videos by utilizing deep learning techniques. Based on features including object detection, speech recognition, text extraction the videos will be categorized according to the genre that ISRO has recommended for the dataset.

The classification work is completed by breaking the project up into multiple parts and utilizing a comprehensive deep learning approach. The first steps involve preprocessing the videos and then using the data produced by the videos to train the deep learning models .The classification is carried out by training the deep learning models such as Convolutional Neural Networks (CNN) 2D, 3D and a combination of CNN & Long short-term memory (LSTM) model, and for speech and text extraction by python libraries . For speech extraction two libraries have been helpful for accurate results that are “movie.py” and “speech recognition”. Additionally, "moviepy easyocr" is used for text extraction.

1.2 Motivation

ISRO works with a huge amount of data, including telemetry and satellite picture data. For making well-informed judgments, having access to structured data is essential. Effective data classification and categorization allows ISRO to speed up decision-making for research initiatives, satellite maintenance, and space missions.In order to handle data more efficiently and free up staff time for other important duties, the project can automate the process of creating metadata.It is possible to apply the deep learning models for categorization and object recognition to find patterns and anomalies in the data collected by ISRO.

The "Smart Sorting: Deep Learning-Powered Metadata Creation for Documentary Video Catalog" project may find use in many different fields. The following are some instances of

how this technology may be modified and applied in other fields:

- Healthcare:Medical Imaging: Deep learning models for object recognition and image classification may be used in medical imaging to help diagnose illnesses and spot abnormalities in X-rays, MRIs, and CT scans.

- Agriculture:Crop monitoring: Drones and satellites can take pictures of fields, and deep learning models can spot problems like pests, nutritional deficits, and illnesses that affect crops.

- Wildlife Conservation: Deep learning may be used to camera trap photos to identify endangered species, assisting with wildlife conservation initiatives.

- Logistics and transport:Traffic Management: Through the study of traffic camera photos, deep learning may be utilized to manage traffic and increase road safety.

1.3 Problem Definition

Video documentaries of various ISRO missions and programs are available. To categorize the all-video programs, generation & verification of huge amount of metadata generation need to be done. With current Deep learning methods-based development in the field of Computer vision and Natural Language Processing this task of video metadata generation is nowadays automated. Given Video programs, process for objects, text, speech recognition, Named Entity recognition etc. to classify the videos in different genres like launch programs, interviews, educational programs, outdoor shots, public shots, traffic etc..

1.4 Lacuna of the existing systems

Low video Resolution: Low resolution videos have a noticeable lack of detail and clarity.

Limitations of dataset: Limitations in datasets can significantly impact the performance and generalizability of machine learning models and data-driven applications.

Limited Accuracy: Existing algorithms for video metadata generation and classification may not always provide accurate results. They can struggle with complex or ambiguous content, leading to misclassification or incomplete metadata.

Data Annotation: Annotating video data for training and evaluation can be time-consuming and expensive. Videos often require frame-level annotations, which can be impractical for large datasets.

Large-Scale Datasets: Collecting and maintaining large-scale video datasets with diverse content is a significant challenge.

1.5 Relevance of the Project

Content Management: With the proliferation of digital media, there's an enormous amount of video content being generated daily. Effective video classification allows for efficient organization and management of this content, making it easier to search, retrieve, and utilize.

Content Moderation: In online platforms and social media, ensuring appropriate content is crucial. Video classification algorithms can help in automatically identifying and flagging inappropriate or sensitive content, facilitating content moderation and ensuring a safer online environment.

Security and Surveillance: Video classification plays a vital role in security and surveillance systems by automatically detecting and recognizing objects, activities, or anomalies in video streams. This capability enhances the effectiveness of surveillance systems for various applications such as public safety, traffic monitoring, and intrusion detection.

Educational Applications: In the realm of education, video classification can be used to categorize educational videos based on subject matter or difficulty level, making it easier for students to find relevant learning resources. Additionally, metadata generation can provide supplementary information such as keywords or summaries, enhancing the educational value of the videos.

Healthcare: In healthcare, video classification can assist in medical image analysis, such as identifying abnormalities in medical scans or monitoring patient movements for rehabilitation purposes. Metadata generation can also provide valuable context for medical videos, aiding in diagnosis and treatment planning.

Chapter 2: Literature Survey

A. Brief Overview of Literature Survey

Title	Author	Summary	Link
"Metadata Extraction and YouTube Video Classification Through Sentiment Analysis"	S. Rangaswamy, S. Ghosh, S. Jha and S. Ramalingam (2016)	It explores the process of classifying video metadata by extracting and analyzing information from videos.This involves categorizing video data into various aspects.	Link
“Efficient Video Classification Using Fewer Frames”	Shweta Bhardwaj, Mukundhan Srinivasan (2019)	It explained a unique approach that uses the notion of distillation to minimize the computing time necessary for video categorization. Training a teacher network, which constructs a video representation using all frames of the movie.	Link
"Video Classification: A Literature Survey"	Pravina Baraiya, Asst. Prof. Disha Sanghani (2018)	The paper explores video classification methods to help users find videos of interest from the plethora available. It discusses three main approaches involving audio, text, and visual features for video classification.	Link
"Automated Metadata Generation for Video Content"	X. Wang, L. Xie, and W. Zhang (2021)	The proposed method involves a multi-task deep learning framework that combines different sources of information, including visual and textual features, to generate metadata for video content.The results demonstrate that the proposed method outperforms existing methods in terms of accuracy, efficiency, and scalability	Link
“Video Captioning Using Deep Learning Approach-A Comprehensive Survey”	Jacob, J., Devassia, V.P. Sharma, H., Saha, A.K., Prasad (2023)	It goes through the most popular neural network variations for feature extraction and language synthesis. ResNet and VGG are common visual feature extractors, according to the report. The most prevalent language model is LSTM	Link

Table 1: Literature survey overview

B. Related Works

2.1 Research Papers Referred

1. **"Metadata Extraction and YouTube Video Classification Through Sentiment Analysis"**

   Abstract-MPEG media have been widely adopted and are very successful in promoting interoperable services that deliver video to consumers on a range of devices. However, media consumption is going beyond the mere playback of a media asset and is geared towards a richer user experience that relies on rich metadata and content description. This paper proposes a technique for extracting and analyzing metadata from a video, followed by decision making related to the video content. The system uses sentiment analysis for such a classification. It is envisaged that the system, when fully developed, is to be applied to determine the existence of illicit multimedia content on the web.

   Inference-In the above mentioned paper the author has proposed a way to classify videos as per the sentiment.The project was developed to determine the existence of illicit multimedia content on the web. Metadata extraction involves a number of stages which can be depicted as an algorithmic structure which involves majorly three sections and the classification can be done on the basis of a dictionary of words each has a sentiment that is associated with . Once the data is obtained from the video it is parsed using NKTL and then a rating is done for the video's sentiment.

2. **“Efficient Video Classification Using Fewer Frames”**

   Abstract-Recent interest in developing compact video classification models with a memory footprint of less than 1 GB has led to the emergence of methods that apply a small weight matrix to all frames in a video, such as recurrent neural networks and cluster-and-aggregate-based approaches like NetVLAD. Despite their compactness, these models still incur significant floating point operations (FLOPs) due to their processing of every frame. This work proposes a novel approach to building compute-efficient video classification models by training a compute-efficient student, which analyzes only a subset of frames, using a compute-heavy teacher model that evaluates all frames. Through extensive evaluation across recurrent, cluster-and-aggregate, and memory-efficient models, it is demonstrated that the proposed student network can achieve a 30% reduction in inference time and

approximately 90% fewer FLOPs with minimal impact on performance, thereby complementing existing research on memory-efficient video classification.

Inference-It explained a unique approach that uses the notion of distillation to minimize the computing time necessary for video categorization. Training a teacher network, which constructs a video representation using all frames of the movie.

3. **"Video Classification: A Literature Survey"**

   Abstract: At present, so many videos are available from many resources. But viewers want videos of their interest. So for users to find a video of interest work has started for video classification. Video Classification literature is presented in this paper. There are mainly three approaches by which the process of video classification can be done. For video classification, features are derived from three different modalities: Audio, Text and Visual. From these features, classification has been done. At last, these different approaches are compared. Advantages and Dis-advantages of each approach/method are described in this paper with appropriate applications.

   Inference:The paper explores video classification methods to help users find videos of interest from the plethora available. It discusses three main approaches involving audio, text, and visual features for video classification.

2.2 Comparison with the existing system

The proposed system for classifying ISRO video documentaries and generating metadata stands out in several key aspects. It offers tailored customization for ISRO-specific content, leveraging advanced deep learning techniques for higher accuracy and automation compared to existing systems. Integration with ISRO data management systems ensures seamless access and sharing of metadata. While the proposed system may require initial investment in resources, its scalability and efficiency surpass those of existing systems, which may lack the specialized capabilities needed for ISRO video analysis. Ultimately, the proposed system promises enhanced accuracy, automation, and user experience, addressing specific needs while paving the way for efficient management and utilization of ISRO video content.

Chapter 3: Requirement Gathering for the Proposed System

3.1 Introduction to requirement gathering

The sorting of videos and the generating metadata is a crucial task in today's world. The amount of data that is being handled by the industries in making predictions is huge and withdrawing metadata from those data is very important.If the data is in the form audio visuals it’s a tedious task taking into consideration the size of the data.Taking this problem into consideration we to develop a software based solution which tries to provide an efficient way to carry out the task. The proposed solution has various stages.The process begins from collecting the data that is the clips of documentaries. Further extracting the frames and resizing them and imposing techniques for detecting objects, extraction of text from the videos in turn generating metadata. This metadata is utilized in training the models of deep learning. Using neural networks the classification of the clips of documentaries from the data set are classified into various categories.There are a number of steps involved in this process.They are as follows:

- Data Collection and Preprocessing: Gather and prepare a diverse dataset of ISRO-related video documentaries.
- Object Detection and Text Extraction: Use object detection models such as YOLOv8 to identify objects and OCR to extract on-screen text.
- Speech Recognition and NER: Transcribe audio using speech recognition and apply Named Entity Recognition to identify key entities.
- Genre Classification Model: Train a deep learning model that fuses information from object detection, text, speech, and NER to predict video genres.
- Training Data and Model Training: Label a subset of the data, train the model, and validate its performance.

3.2 Functional Requirements

- **Video Classification**: The system should be able to classify video documentaries of ISRO missions and programs into different genres such as launch programs, interviews, educational programs, outdoor shots, public shots, and traffic.
- **Object Detection**: It should detect and recognize objects, such as spacecraft, satellites, launch vehicles, and scientific instruments, within the video content.

- **Text Recognition**: The system should extract and recognize text present within the video, including on-screen captions, subtitles, and textual information displayed during interviews or presentations.

- **Speech Recognition**: It should transcribe spoken content within the video, including dialogue, narration, and commentary, to facilitate further analysis and indexing.

3.3 Non-Functional Requirements

- **Accuracy**: The system should achieve high accuracy in video classification, object detection, text recognition, speech recognition, and named entity recognition to ensure reliable metadata generation.

- **Scalability**: It should be scalable to handle a large volume of video content efficiently, accommodating future growth in the dataset and user base.

- **Performance**: The system should be capable of processing video content in real-time or near-real-time, minimizing processing delays and latency.

- **Robustness**: It should be resilient to variations in video quality, lighting conditions, background noise, accents, and languages, ensuring consistent performance across diverse scenarios.

3.4.Hardware, Software Requirements

Type of Requirement	**Description**
Hardware Requirements	1. Processor:A good processor that will be efficiently utilized in training the model. 2. Storage: Storage is required to store the training and testing data. 3. Memory: A memory of 16GB would be sufficient for the tasks. 4. Internet connection: It would be necessary to install various libraries used for the project.

Software Requirements	1. Python: It is the programming language used for machine learning and deep learning models. 2. Deep Learning Frameworks: You'll need deep learning frameworks such as TensorFlow or PyTorch to build and train your models. 3. Computer Vision Libraries: For object detection and image processing, you might use libraries like OpenCV. 4. Natural Language Processing Libraries: Libraries like NLTK, spaCy can be used for text processing and NER (Name Entity Recognition). 5. Speech Recognition Libraries: Libraries like SpeechRecognition or Google's Speech-to-Text API can be used for transcribing audio. 6. Data Annotation Tools: Tools like labeling for image annotation and various text annotation platforms can help with creating labeled datasets.

Table 2: Hardware and Software Requirements

3.5 Constraints

- **Data Availability**: The availability of a diverse and representative dataset of video documentaries covering ISRO missions and programs may be limited, posing a constraint on the training and evaluation of machine learning models.

- **Computational Resources**: The computational resources required for training deep learning models and processing large volumes of video data may be limited, impacting the scalability and performance of the system.

- **Time Constraints**: There may be time constraints for completing the project, especially if there are deadlines associated with the delivery of the system or the availability of ISRO video content for analysis.

- **Budget**: Budgetary constraints may limit the acquisition of necessary hardware, software, and data resources, as well as the hiring of skilled personnel for development and implementation.

- **Technological Limitations**: The effectiveness and accuracy of machine learning models for tasks such as object detection, text recognition, and speech recognition may be constrained by the current state of technology and available algorithms.

- **Regulatory Compliance**: Compliance with legal and regulatory requirements, such as data privacy regulations and copyright laws, may impose constraints on the collection, processing, and sharing of video content and associated metadata.
- **Integration Challenges:** Integrating the developed system with existing ISRO data management systems or platforms may pose technical challenges, such as interoperability issues or data format compatibility.

Chapter 4: Proposed Design

4.1 Block diagram of the system

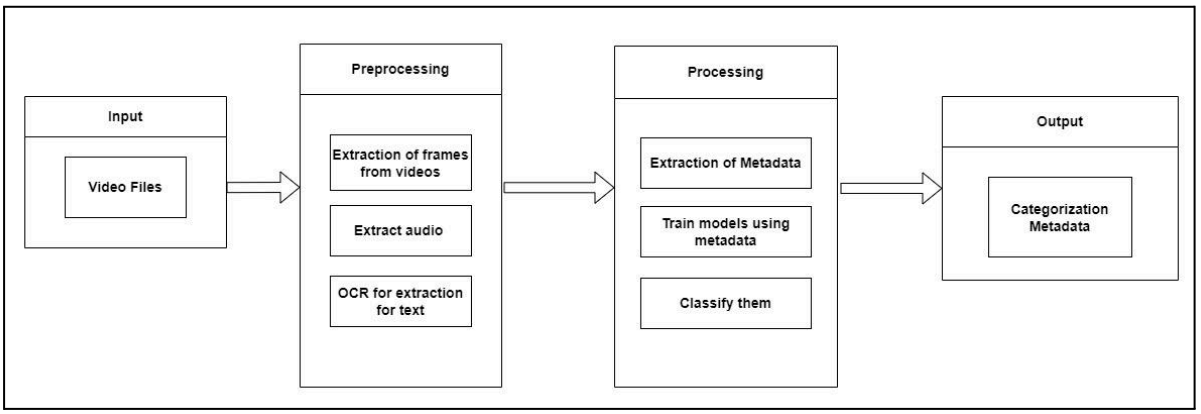


Fig 1: Block diagram of the project

Input:

This is the initial data that your system receives. In the context of your block diagram, it likely refers to the multimedia content you're analyzing, such as a video file. This content serves as the raw material for your processing pipeline.

Preprocessing:

- Extraction of Frames from Video: Involves breaking down the video file into individual frames. This step is crucial for analyzing the visual content of the video. Each frame represents a single image, and by extracting frames, you enable further analysis on a frame-by-frame basis.
- Extraction of Audio: This step involves separating the audio track from the video. Audio analysis can include tasks like speech recognition, sound classification, or sentiment analysis.
- OCR for Text Extraction: OCR stands for Optical Character Recognition. This process involves extracting text from images or frames. In the context of your system, OCR might be used to extract text from any textual elements present within the video frames.

Processing:

- Metadata Extraction: Metadata refers to additional information about the content. In the case of a video, metadata might include details like the date of creation, location, camera settings, etc. Extracting metadata can provide valuable context for further analysis or organization of the content.

- Train: This likely refers to the training of machine learning models. Training involves feeding labeled data into a machine learning algorithm to allow it to learn patterns and make predictions or classifications.
- Classify: Classification is a machine learning task where the system categorizes data points into predefined classes or categories. In your context, classification might involve categorizing video frames, audio segments, or text snippets based on their content or characteristics.

Output:

This is the final result or output generated by your system after processing the input data. The output could include various forms of analysis results, such as labeled frames, transcribed text, classified audio segments, or any other relevant information extracted during the preprocessing and processing stages.

4.2 Modular design of the system

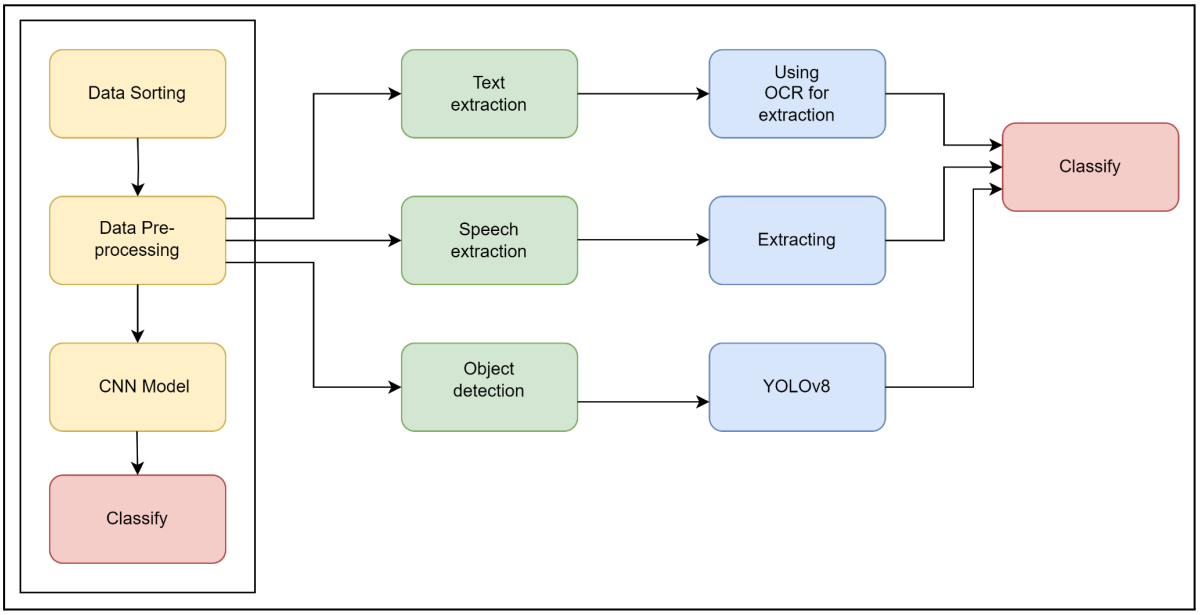


Fig 2 : Modular design

4.3 Detailed Design

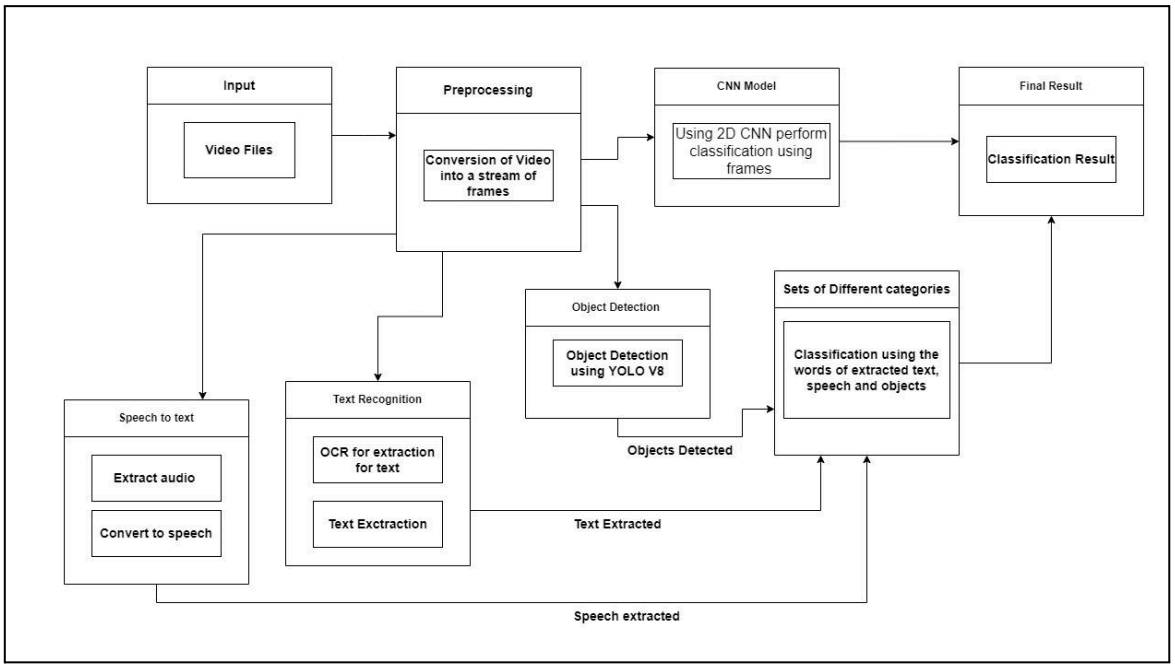


Fig 3 : Detailed design of the project

Dataset and Preprocessing:

The dataset comprises videos categorized into 11 different categories related to ISRO, with lengths ranging from one second to one minute. OpenCV is used for preprocessing, resizing each frame to a uniform dimension for efficient processing. Object Detection employs the YOLO V8 model to detect objects in resized images, ensuring uniformity for accurate detection. Text Recognition extracts utilitarian information from frames using Optical Character Recognition (OCR), capturing textual and numerical data displayed on the screen. Speech Recognition transcribes audio to text using the Moviepy library, with efficiency dependent on audio quality.

Classification:

Text-based features from videos are used for initial classification, where specific words indicate categories. Sets of frequently appearing words in audio and text are used to classify videos. Frame-wise classification is performed using a 2D-CNN deep learning model, achieving 95% accuracy. Other models considered, like CNN-LSTM and 3D-CNN, were less proficient due to their consideration of temporal aspects, which is less relevant for shorter videos. Results from both classification phases are aggregated to obtain the final classification label for the video.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

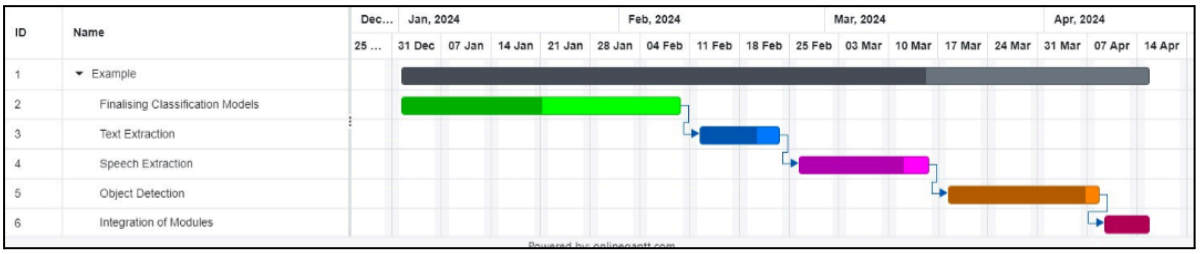


Fig 4 : Gantt Chart

Chapter 5: Implementation of the Proposed System

5.1. Methodology employed for development

Data collection: The dataset of video files required for classification and metadata generation purpose are to be gathered.

Data pre-processing: Storing the data in an appropriate way such that it can be suitable for analysis.

Feature extraction: Analyzing the pre-processed videos for extracting characteristics of the video with the help of deep learning models that are Convolutional neural networks (CNN) and Long short-term memory networks(LSTM) models.

Metadata generation: Obtaining the data that the video has using various techniques of computer vision.

Classifying videos: Using deep learning techniques to categorize video data into predefined labels.

Scalability: Training and testing deep learning models for larger datasets.

Hyper Parameterizing: Manipulating the parameters of deep learning techniques to increase the accuracy of prediction.

User Interface: Try to develop an application that would allow you to generate metadata and classify the videos.

5.2 Algorithms and flowcharts for the respective modules developed

- **Data Ingestion Module:**
  Algorithm:
  Obtain video files from specified sources (e.g., directories, databases).
  Iterate through each video file.
  Read video files and extract frames at regular intervals.
  Store extracted frames in a suitable data structure (e.g., arrays, lists).
- **Preprocessing Module:**
  Algorithm:
  Resize each frame to a uniform dimension using OpenCV.
  Apply any necessary preprocessing techniques (e.g., normalization, denoising).

- **Feature Extraction Module:**

  Algorithm:

  Use the YOLOv8 model for object detection to extract visual features from frames.

  Utilize OCR for text recognition to extract textual features from frames.

  Apply speech-to-text models for speech recognition to extract audio features from videos.

- **Classification Module:**

  Algorithm:

  Extract features (visual, textual, audio) from preprocessed frames.

  Perform initial classification based on extracted features using predefined categories.

  Apply 2D CNN model for frame-wise classification.

  Aggregate results from initial and frame-wise classification to obtain final classification labels.

- Flowchart:

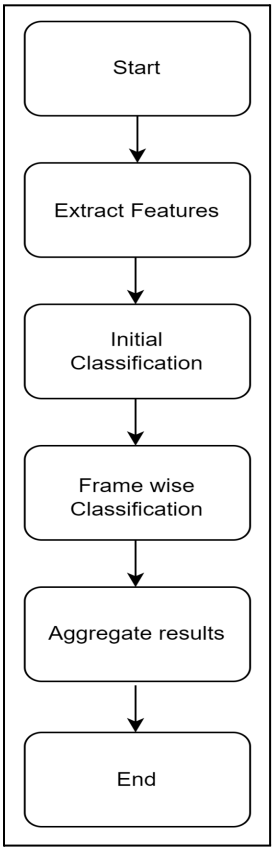


Fig 5 : Flow Chart

5.3 Datasets source and utilization

Datasets:

- For CNN, Text extraction and Speech extraction the data was from ISRO website with their problem statement.
- For Object Detection the data was from various sites (e.g. google, pinterest edge,etc).
- Dataset was utilized for extraction, classification and detection.

Chapter 6: Testing of the Proposed System

6.1. Introduction to testing

Testing is an indispensable phase in the development lifecycle of any system or software. It serves as a systematic and methodical approach to verify that the system meets specified requirements, functions correctly, and delivers expected results. Through testing, developers can identify defects, errors, or inconsistencies in the system's behavior and address them before deployment. Testing encompasses various techniques and methodologies tailored to different stages of development, including unit testing, integration testing, system testing, and acceptance testing. By rigorously testing each component and aspect of the system, from individual modules to the system as a whole, developers can ensure its reliability, stability, and performance, ultimately enhancing user satisfaction and trust in the product.

6.2. Types of tests Considered

Input Testing: Testing begins with the examination of the input data. This involves verifying that the system can correctly ingest various types of multimedia content, such as video files, ensuring compatibility and proper handling of different formats. Input testing may include scenarios with different resolutions, frame rates, audio codecs, and text encoding to ensure robustness across diverse inputs.

Preprocessing Testing:

- **Frame Extraction Testing**: Preprocessing testing involves validating the accuracy and efficiency of frame extraction from videos. This includes confirming that all frames are extracted without loss or distortion and that the timing and sequencing of frames are preserved correctly.

- **Audio Extraction Testing:** For audio extraction, testing ensures that the system accurately separates audio tracks from videos, maintaining audio quality and synchronization with visual content.

- **OCR Testing**: Optical Character Recognition (OCR) testing assesses the system's ability to accurately extract text from video frames, verifying the correctness of text recognition and ensuring that it can handle various fonts, sizes, orientations, and backgrounds.

Processing Testing:

- **Metadata Extraction Testing:** Metadata extraction testing focuses on validating the correctness and completeness of metadata extracted from the multimedia content. This involves confirming that relevant metadata such as timestamps, location data, camera settings, and any other pertinent information are extracted accurately.

- **Training Testing:** In the training phase, testing involves evaluating the performance of machine learning models. This includes assessing model accuracy, precision, recall, and other relevant metrics using validation datasets to ensure that the models generalize well to unseen data.

- **Classification Testing:** Testing classification involves evaluating the accuracy of class predictions made by the system. This includes testing the system's ability to correctly categorize video frames, audio segments, or text snippets based on the trained models' classifications.

- **Output Testing:** Output testing ensures that the system generates the expected results accurately and efficiently. This includes verifying the correctness of labeled frames, transcribed text, classified audio segments, and any other output produced by the system. Output testing also assesses the system's performance in handling large volumes of data and its scalability to accommodate increasing workloads.

6.3 Various test case scenarios considered

- **Positive Classification Test**: Test the system's ability to correctly classify videos into their respective categories. Provide videos from each category in the dataset and verify that the system assigns the correct label to each video.

- **Negative Classification Test**: Test the system's ability to handle videos that do not belong to any predefined category. Provide videos outside the scope of the dataset and ensure that the system correctly identifies them as unclassified or unknown.

- **Performance Test**: Evaluate the system's performance in terms of classification accuracy, precision, recall, and F1 score. Use a separate test dataset with ground truth labels to measure the system's performance metrics.

- **Temporal Test**: Assess the system's performance on videos with temporal dynamics, such as changes in activity over time or sequences of events. Test models like CNN-LSTM or 3D-CNN on videos with temporal dependencies to evaluate their effectiveness.

6.4. Inference drawn from the test cases

- The 2D CNN model achieved a high accuracy of 95%, indicating its effectiveness in classifying videos in the dataset.
- Other models, such as CNN-LSTM or 3D-CNN, did not perform as well, possibly due to the small size of the videos in the dataset.
- The CNN model demonstrated the ability to classify videos not present in the dataset for training, suggesting its generalization capability and robustness.

Chapter 7: Results and Discussion

7.1. Screenshots of User Interface (UI) for the respective module

The output of the process involves the testing of the whole process on a video that is being obtained from another source and is not present in the dataset.

1. **CNN**

```
Predicted Label according to CNN: Satellite
Confidence: 0.9999994
```

Fig 6 :Output of CNN model

This was the result obtained for the CNN model.

2. **Text Extraction**

Fig 7 :Frame of a video

The above picture shows the frame from a video and the result of text extraction for the following video is as follows.

```
WARNING:easyocr.easyocr:Neither CUDA nor MPS are available - defaulting to CPU. Note: This module is much
WARNING:easyocr.easyocr:Downloading detection model, please wait. This may take several minutes depending
Progress: |                                        | 100.0% CompleteWARNING:easyocr.easyocr:Down
Progress: |                                        | 100.0% Complete22 of the 29 Nano Satellites
and one Micro Satellite are from the USA:
22 of the 29 Nano Satellites
and one Micro Satellite are from the USA:
22 of the 29 Nano Satellites
and one Micro Satellite are ftom the USA:
22 ofithe 29 Nano Satellites
and one Micro Satellite are from the USA:
22 of the 29 Nano Satellites
and one Micro Satellite are from the USA:
22 of the 29 Nano Satellites
and one Micro Satellite are from the USA:
22 of the 29 Nano Satellites
and one Micro Satellite are from the USA:
22 of the 29 Nano Satellites
and one Micro Satellite are from the USA:
```

Fig 8 :Text captured through the frame

3. **Speech Extraction**

```
MoviePy - Writing audio in temp_audio.wav
                                                                MoviePy - Done.
Extracted speech:  22 of the 29 Nano satellites and 1 micro satellite are from the USA
```

Fig 9 :Speech that is being transcripted

4. **Object detection**

Fig 10 : Object detected

5. **Integrated process**

```
Extracted Speech: communication satellite GSAT 19
Predicted Label according to Text,Speech and CNN: IndoorLab
Confidence: 0.9950689
Extracted Text: successfully launched Indias high
throughput communication satellite GSAT-19
Shd
successfully launched Indias high
throughput communication satellite GSAT-19
ISrd
successfulljy launched Indials high
throughput communication satellite GSAT419
Shd
successfully launched Indials high
throughput communication satellite GSAT419

Extracted Speech: communication satellite GSAT 19
Predicted Label according to Text,Speech and CNN: IndoorLab
```

Fig 11 : Output of the integrated process

7.2. Performance Evaluation measures

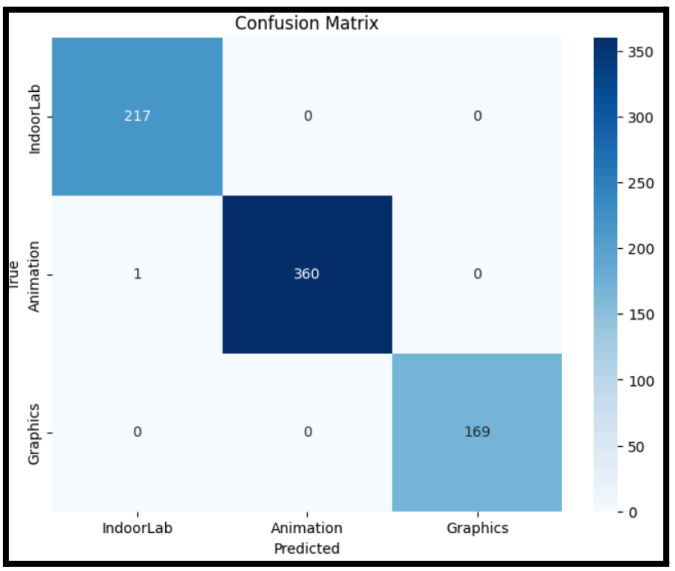
- Accuracy: Measures the proportion of correctly classified videos out of the total number of videos. It provides an overall assessment of the classification model's effectiveness.
- Precision: Indicates the proportion of correctly classified positive cases (true positives) out of all videos classified as positive. It focuses on the accuracy of positive predictions.

- Recall (Sensitivity): Measures the proportion of correctly classified positive cases (true positives) out of all actual positive cases. It assesses the model's ability to identify all relevant instances.
- F1 Score: Harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when there is an imbalance between positive and negative classes.
- Confusion Matrix: A table that summarizes the performance of a classification model by comparing actual and predicted classifications. It provides insights into true positives, true negatives, false positives, and false negatives.
- Time and Resource Consumption: Evaluation of the computational resources and time required for model training, inference, and evaluation. It helps assess the efficiency and scalability of the classification system.

7.3. Input Parameters / Features considered

- Visual Features: These include features extracted from video frames using techniques like object detection, which identify objects, scenes, or patterns within the video content.
- Textual Features: Features extracted from text appearing in the video, obtained through techniques like Optical Character Recognition (OCR). This includes textual information displayed on-screen, such as captions, subtitles, or other textual overlays.
- Audio Features: Features extracted from audio content, such as speech recognition results or audio characteristics like pitch, amplitude, and frequency. These features provide insights into the audio content and can be used for classification.
- Temporal Features: Features related to the temporal dynamics of the video content, such as the sequence of frames, changes over time, or patterns of activity. These features capture the temporal context and structure of the video and can be utilized by models like CNN-LSTM or 3D-CNN.
- Metadata Features: Additional metadata associated with the video content, such as timestamps, file format, resolution, duration, or other relevant information. Metadata features provide contextual information about the video and can aid in classification and analysis.

7.4. Graphical and statistical output



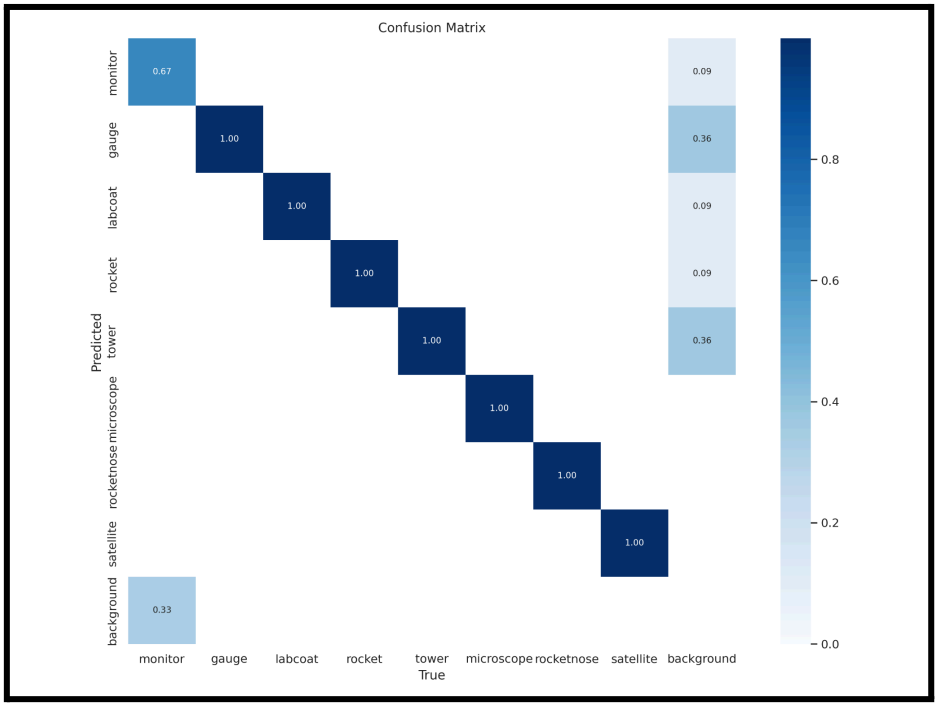


Fig 12 : Graphical Outputs

7.5. Comparison of results with existing systems

- Scope and Coverage: The proposed system aims to classify a wider range of ISRO-related video content with enhanced metadata generation, compared to existing systems which may have limited coverage or focus on specific categories.

- User Interface: The development of an intuitive user interface sets the proposed system apart, providing a seamless experience for users to upload videos, view

metadata, and access categorized content, which may be lacking in some existing systems.

- Integration of YOLOv8: Integrating YOLOv8 for object detection enhances the proposed system's capabilities, allowing for detailed object information in metadata generation, a feature that may be absent in many existing systems.
- Scalability and Performance: The optimization of the proposed system for scalability and performance ensures efficient handling of larger datasets and volumes of video content, addressing potential limitations of existing systems in managing increased data loads.
- Advanced Feature Extraction: Utilizing advanced feature extraction techniques, such as deep learning-based methods, sets the proposed system apart in capturing richer metadata from video content compared to traditional feature extraction methods used in existing systems.

7.6. Inference drawn

The comparison highlights several key areas where the proposed system excels compared to existing systems. By offering a wider scope, intuitive user interface, advanced object detection integration, scalability, and continuous improvement mechanisms, the proposed system provides a comprehensive solution for video classification and metadata generation in the domain of ISRO missions and programs. These enhancements ensure better coverage, usability, accuracy, and adaptability, setting the proposed system apart as a more advanced and effective tool for analyzing ISRO-related video content.

Chapter 8: Conclusion

8.1 Limitations

- **Data Quality and Quantity:** The performance of your CNN model heavily relies on the quality and quantity of data available for training. Limited or low-quality data could result in a model that fails to generalize well to new inputs.

- **Variability in Text and Speech:** Natural language and speech can be highly variable due to factors like accents, dialects, background noise, and vocabulary diversity. Your models may struggle to accurately extract information from sources with significant variability.

- **Computational Resources:** Training and deploying deep learning models, especially CNNs, can be computationally expensive. Limited computational resources may restrict the size and complexity of your models or increase the time required for training and inference.

- **Model Interpretability**: CNNs, particularly when applied to text and speech data, can be challenging to interpret. Understanding how the model makes decisions and identifying potential biases or errors may be difficult, limiting your ability to diagnose and address performance issues.

- **Domain Specificity:** Models trained on one domain may not generalize well to others. If your project operates in a specific domain (e.g., healthcare, finance), the performance of your models may be limited when applied to new or unseen data from different domains.

- **Speech Recognition Accuracy:** Speech-to-text conversion accuracy can vary depending on factors like background noise, speaker accent, and speech rate. Errors introduced during the speech extraction phase may propagate to downstream tasks, impacting the overall performance of your system.

- **Integration Challenges:** Integrating multiple components (text extraction, speech extraction, and CNN model) into a cohesive system can be complex. Ensuring seamless communication and compatibility between these components while maintaining system performance can present challenges.

- **Scalability:** As the volume of data or user interactions increases, scalability becomes a concern. Your system may face performance bottlenecks or resource constraints when dealing with large-scale data processing or user requests.

- **Ethical and Legal Considerations:** Utilizing speech and text data raises ethical considerations regarding user privacy, consent, and data security. Additionally, depending on your application and the data sources involved, you may need to navigate legal regulations such as GDPR or HIPAA compliance.

- **Maintenance and Updates:** Machine learning models require ongoing maintenance and updates to remain effective over time. Changes in data distribution, user behavior, or technological advancements may necessitate periodic retraining or fine-tuning of your models.

8.2 Conclusion

Through this project we have successfully completed the classification of videos as well as tested the videos that do not belong to the dataset for classification along with the generation of metadata for the videos. In the future, our major task for taking this project to the next level of functionality is to develop a UI with which an user can efficiently get the classification as well as the metadata from the video accurately as well as be able to classify a number of videos in a folder simultaneously. This would save the time for the user to want the video and classify it after viewing it by automation of the whole process.

The system can be used by industries which deal with a lot of video files and can leverage the machine learning technique to reduce the time required for sorting of video documentaries based on a set of predefined categories.

8.3 Future Scope

- **Category Expansion**: Add more categories to cover a wider range of ISRO-related video content, such as space exploration, satellite technologies, and astronaut training.
- **User Interface Development**: Create an intuitive interface for easy video upload, metadata viewing, and categorized video access.
- **YOLOv8 Integration**: Integrate YOLOv8 for improved object detection, enhancing metadata generation with detailed object information.
- **Enhanced Metadata Generation**: Utilize advanced feature extraction techniques to capture richer video metadata.
- **Scalability and Performance Optimization**: Optimize the system for handling larger datasets and volumes of video content efficiently.
- **Integration with External Systems**: Integrate with external data sources or platforms for additional information and analysis.
- **Continuous Improvement**: Implement mechanisms for ongoing system updates, model retraining, and user feedback for continuous enhancement.

References

1. Rangaswamy, Shanta & Ghosh, Shubham & Jha, Srishti & Ramalingam, Soodamani. (2016). Metadata extraction and classification of YouTube videos using sentiment analysis. 1-2. 10.1109/CCST.2016.7815692.

2. A Deep Learning Approach for Video Metadata Generation and Classification Dr. T. Raghunadha Reddy , P. Sreekari , J. Nikhil Kumar Reddy , and V. Jyothsna Associate Professor, Department of CSE, Matrusri Engineering College, Hyderabad Student, Department of CSE, Matrusri Engineering College, Hyderabad

3. Baraiya, Pravina, and Disha Sanghani. "Video Classification: A Literature Survey." *International Journal on Recent and Innovation Trends in Computing and Communication* 6.3: 01-05.

4. Konapure, R. C., and L. M. R. J. Lobo. "Video content-based advertisement recommendation system using classification technique of machine learning." *Journal of Physics: Conference Series*. Vol. 1854. No. 1. IOP Publishing, 2021.

5. Yousaf, Kanwal, and Tabassam Nawaz. "A deep learning-based approach for inappropriate content detection and classification of youtube videos." *IEEE Access* 10 (2022): 16283-16298.

6. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.

7. Maratea, Antonio, Alfredo Petrosino, and Mario Manzo. "Generation of description metadata for video files." *Proceedings of the 14th International Conference on Computer Systems and Technologies*. 2013.

8. Park, Jung-ran, and Caimei Lu. "Application of semi-automatic metadata generation in libraries: Types, tools, and techniques." *Library & Information Science Research* 31.4 (2009): 225-231.

9. S. Bhardwaj, M. Srinivasan and M. Khapra, "Efficient Video Classification Using Fewer Frames," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019 pp. 354-363.