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

Prof. Abha Tewari

Vivekanand Education Society's Institute of Technology,

An Autonomous Institute affiliated to University of Mumbai

HAMC, Collector's Colony, Chembur,

Mumbai-400074

University of Mumbai (AY 2023-24)

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._Abha Tewari)

(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:

Contents

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 project 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 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.

Acknowledgements

We thank our college Vivekanand Education Society's Institute of Technology for taking an interest in our project and providing assistance when needed throughout our work of information gathering for the project.

We would like to take this opportunity to thank Assistant Professor **Mrs. Abha Tewari** (the project guide) for her kind assistance, insightful counsel, and recommendations while we developed the project description.

We owe a debt of gratitude to our **Principal Dr. (Mrs.) J.M. Nair** and the head of the computer department, **Dr. (Mrs.) Nupur Giri**, for providing us with the precious chance to complete this project.

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.

In addition to gaining information, we hope that our effort will have a constructive impact on society

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
6	VGG	Visual Geometry Group
7	AI/ML	Artificial Intelligence/Machine Learning
8	GB	Gigabyte
9	MRI	Magnetic Resonance Imaging
10	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 Statement & Objectives

Problem Statement

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..

Objectives

1. To classify the videos in the dataset into categories specified.
2. To recognise objects in the videos.
3. To detect speech from the videos.
4. To extract text from the videos if detected any with the help of computer vision.

1.4 Organization of the Report

The report is organized in the following manner:-

Chapter 1: Introduction

Our topic for project is "Smart Sorting: Deep Learning-Powered Metadata Creation for Documentary Video Catalog" it means classification of the videos will be done as per the genre suggested by ISRO in the problems dataset based on the speech recognition, text extraction,object detection,etc.There are different genres such as forests,launchpads,rockets,outdoor ,indoor control room,etc.. We will be using Neural Networks (CNN) for images and Movie.py and OCR for the speech and text.

Chapter 2: Literature Survey

In this we examined a few existing research papers from which we can understand the limitations. Some papers have poor quality of data, some large dataset, less amount of dataset. but everything has one thing in common and that is, the algorithm, the model they have applied.

Chapter 3: Proposed System

The flow of the Architectural framework is Input, Preprocessing, Processing, Output. In this video will be imported then the extraction of the data will be executed (Extracting frames of the video, audio, text). then the extracted data will be trained using deep learning models. The experiment we carried out was done for only 3 classes as there were limitations based on the configurations of hardware and software. The CNN algorithm provided very high accuracy as the classes were limited.

Chapter 2:Literature Review

2.1 Survey of Existing Systems

Title	Author	Summary	Link

Table 1: Literature survey

2.2 Limitation Existing system or Research gap

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.

2.3 Mini Project Contribution

The main contribution of the project is the effective handling and structuring of massive amounts of data. Data becomes easier to access and manage by reducing the time and effort needed for human data curation through the automation of the metadata production and content classification processes. Solid data organization lays the groundwork for better decision-making. The contribution of the project includes:

Data-Driven Insights: Organizations may obtain data-driven insights, spot patterns, and make better decisions by automating classification and the development of metadata.

Timely responses: Quick access to classified data can save lives in industries like emergency response or healthcare.

Impact on Society: By boosting safety, health, and quality of life, better decision-making, faster access to medical information, and effective resource management directly benefit society.

Global Collaboration: The project's results can promote worldwide collaboration in industries like environmental monitoring and space exploration since data can be more readily standardized and shared.

Chapter 3:Proposed System

3.1 Introduction

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 Architectural Framework / Conceptual Design

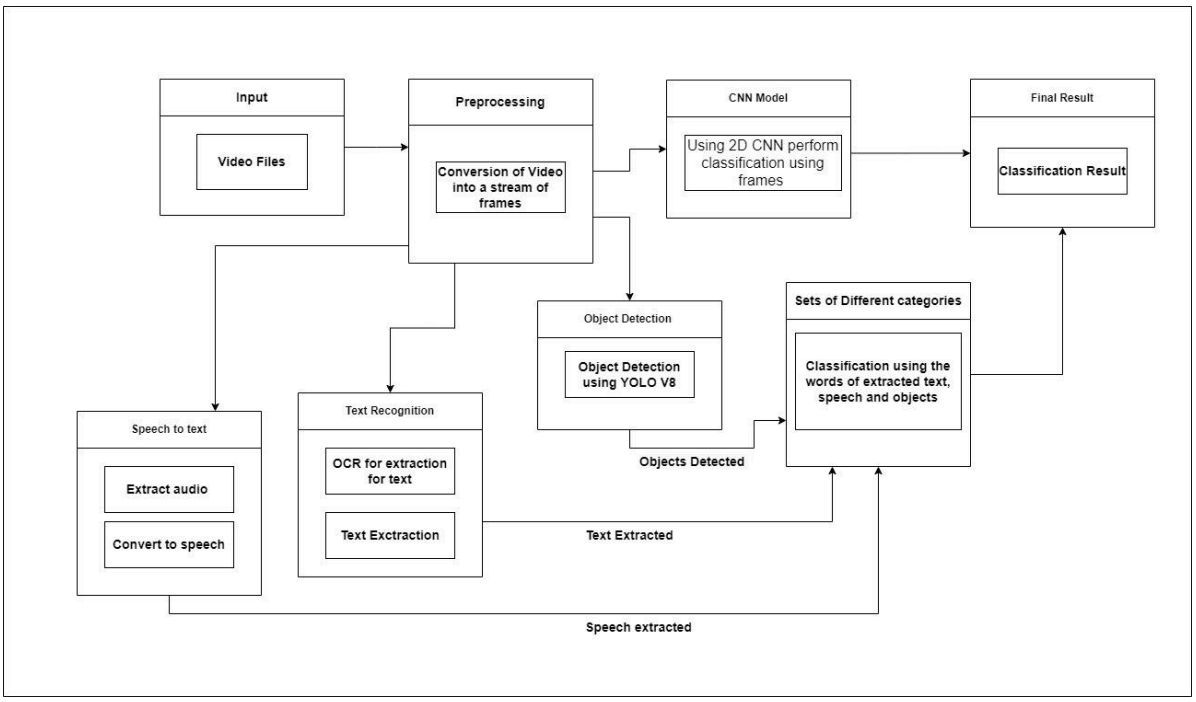

Figure 1:Input, processing and output diagram of the project

3.3 Algorithm and Process Design

The algorithms that are incorporated in the following project involve a number of deep learning models.The process of video metadata generation and classification is a complex task that requires a number of stages initiating from pre-processing to categorization.These steps are discussed below:

1. Data Pre-processing: The videos that are provided to the deep learning models cannot be processed directly consumed by the model for classification or metadata generation.These videos have to be pre-processed by extraction of frames from the videos and converting them to a uniform size by the help of computer vision. These frames can be provided as input to the models of deep learning.
2. Metadata extraction: This process is responsible for extracting the data that the video possesses. The audio extraction,object recognition and various other features from the video would be extracted from the input video.
3. Training of Deep Learning Models:The deep learning models are trained by using the pre-processed dataset.The effective training of models is essential for accurate classification of videos.
4. Processing:The metadata extracted is processed and the results of the classification are obtained.The classification model's accuracy depends on the type of data and the amount of data that is provided for the training of the model.

5. Parameterization:The parameters of the models can be fine tuned in order to get maximum accuracy from the model.
6. Output: In the output we get the label for the video which is provided to model.

3.4 Methodology Applied

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.

3.5 Hardware & Software Specifications

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

3.6 Experiment and Results for Validation and Verification

For the initial experiment we tried to implement the CNN model for the classification task.The experiment we carried out was done for only 3 classes as there were limitations

based on the configurations of hardware and software. Based on the experiment we received the following result. The CNN algorithm provided very high accuracy as the classes were limited.

Results

```
Epoch 1/10
94/94 [==============================] - 17s 172ms/step - loss: 0.3160 - accuracy: 0.8804 - val_loss: 0.0170 - val_accuracy: 0.9987
Epoch 2/10
94/94 [==============================] - 17s 182ms/step - loss: 0.0219 - accuracy: 0.9963 - val_loss: 0.0018 - val_accuracy: 1.0000
Epoch 3/10
94/94 [==============================] - 17s 180ms/step - loss: 0.0095 - accuracy: 0.9983 - val_loss: 0.0031 - val_accuracy: 0.9987
Epoch 4/10
94/94 [==============================] - 16s 170ms/step - loss: 0.0064 - accuracy: 0.9993 - val_loss: 0.0044 - val_accuracy: 0.9987
Epoch 5/10
94/94 [==============================] - 16s 166ms/step - loss: 0.0034 - accuracy: 0.9983 - val_loss: 0.0029 - val_accuracy: 0.9987
Epoch 6/10
94/94 [==============================] - 16s 165ms/step - loss: 0.0037 - accuracy: 0.9997 - val_loss: 0.0035 - val_accuracy: 0.9987
Epoch 7/10
94/94 [==============================] - 16s 168ms/step - loss: 0.0136 - accuracy: 0.9960 - val_loss: 0.0037 - val_accuracy: 0.9987
Epoch 8/10
94/94 [==============================] - 16s 169ms/step - loss: 0.0020 - accuracy: 0.9997 - val_loss: 0.0059 - val_accuracy: 0.9987
Epoch 9/10
94/94 [==============================] - 15s 164ms/step - loss: 0.0056 - accuracy: 0.9987 - val_loss: 0.0113 - val_accuracy: 0.9987
Epoch 10/10
94/94 [==============================] - 16s 169ms/step - loss: 0.0033 - accuracy: 0.9990 - val_loss: 0.0020 - val_accuracy: 0.9987
24/24 [==============================] - 1s 37ms/step - loss: 0.0020 - accuracy: 0.9987
Test loss: 0.002027408219873905
Test accuracy: 0.9986613392829895
24/24 [==============================] - 1s 36ms/step
```

Figure 2: CNN Model Result

```
Classification Report:
              precision    recall  f1-score   support

   IndoorLab       1.00      1.00      1.00       217
   Animation       1.00      1.00      1.00       361
    Graphics       1.00      1.00      1.00       169

    accuracy                           1.00       747
   macro avg       1.00      1.00      1.00       747
weighted avg       1.00      1.00      1.00       747
```

Figure 3: Classification Report

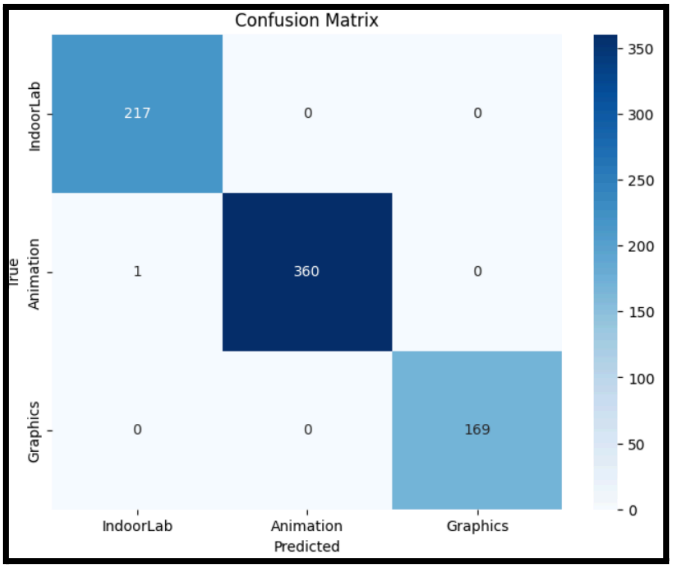

Figure 4: Confusion Matrix

3.7 Result Analysis and Discussion

Result:

The CNN model has given the accuracy of the

Discussion:
This is a very high accuracy for the sorting of videos. Once the number of entities in the classification is increased the accuracy of the model would degrade from the current value.

In Order to increase the accuracy we need to fine tune the model.The quality and quantity of dataset that is utilized by the model in order to carry out the training part would play a very crucial role as if the data is of good quality then the model will be trained at a better level.

3.8 Conclusion and Future work.

In conclusion, the automated video metadata generation and genre classification project presents a comprehensive solution for enhancing the accessibility and organization of video documentaries related to ISRO missions and programs.

By leveraging the power of computer vision and natural language processing, this project aims to streamline the process of categorizing videos into different genres, such as launch programs, interviews, educational content, and more.

However, it's important to recognize that the success of the project hinges on diligent data collection, through model training, and careful user interaction design.

In essence, the automated video metadata generation and genre classification project stands as a testament to the capabilities of modern technology in automating complex tasks, enhancing user experiences, and contributing to the effective management of valuable video content.

Future Work

1. **Models Implementation:** Implementing pre-trained models for the video classification task.Using CNN and RNN architectures for classification task.
2. **Features for metadata generation**:This involves the implementation of various computer vision libraries for extraction of metadata.
3. **Training and Testing**: Once the selection of the model as well as the features we need to train and test the dataset on the model for getting the highest accuracy of that model.
4. **Parameterizing the model**:In this we have to train and test the model for various

parameters of the model and manipulate those in order to obtain an efficient accuracy.Also storing the results for various parameter values in a form of spreadsheet.

References

Journal Paper,
A Deep Learning Approach for Video Metadata Generation and Classification Dr. T. Raghunadha Reddy1 , P. Sreekari2 , J. Nikhil Kumar Reddy3 , and V. Jyothsna4 1Associate Professor, Department of CSE, Matrusri Engineering College, Hyderabad 2, 3, 4 Student, Department of CSE, Matrusri Engineering College, Hyderabad

Journal Paper,
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.

Journal Paper,
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.

Journal Paper,
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.

Journal Paper,
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.

Journal Paper,
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.

Journal Paper,
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.

Journal Paper,
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.