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

Amisha Chandwani
Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai , India
2021.amisha.chandwani@ves.ac.in

Pranav Rane
Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai , India
2021.pranav.rane@ves.ac.in

Anmol Gyanmote
Department of Computer Engineering
Vivekanand Education Society's Institute of Technology,
Mumbai , India
2021.anmol.gyanmote@ves.ac.in

Mrs. Abha Tewari
Assistant Professor
Department of Computer Engineering
Vivekanand Education Society's Institute of Technology,
Mumbai , India

***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 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.**

**Keywords– Text Extraction, Speech Extraction, Video Classification, CNN, YoloV8, Video Metadata.**

## I. 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.

## II. LITERATURE REVIEW

In [1], 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.

In [2], the author defines ways to develop a novel approach towards the categorization of the video as well as extraction of metadata and provide a description of the video. Foundational technique of NLP for the generation of textual information of the video and the use of LSTM model which is an artificial recurrent neural network (RNN) architecture is used to classify the video. The textual data is obtained from the use of APIs ,web scraping and some other ways which are parsed in the form that is required by NLP for processing and converting them into tokens and applying lemmatization techniques.Once the data is preprocessed it is then input into LSTM model for prediction and once the prediction is done a summary is generated and then a classification result is obtained.

In [3] a detailed description about the various features that can be used for the classification of video into categories is given. The author defines majorly three approaches that are Audio Based Approach, Text Based Approach and Visual Based Features. Audio Based Approach has more usage compared to the Text Based Approach. Audio files have shorter lengths and require less space and also minimizes the utilization of computational resources. There are various parameters that are used in this such as frequency, wavelength,bandwidth,etc.. In Text Based Approach there are two ways in which classification can be carried out. The first is the visual text on the screen using the Optical Character Recognition (OCR) and the second is the transcription of the audio in the video using the Speech Recognition Method. The Visual Based Feature method has features that are Color-based, Shot-based and Object Based.The requirement of space for employing this strategy is more as a single video consists of a number of frames that need to be extracted and stored.

In [4], the content based classification of advertisements is described by the author. The sorting of ad-films helps the companies to reach the correct consumers for their products. A deep learning approach is applied by using the CNN model by passing a series of frames for the classification of the video. The model once trained can be used for the classification of other videos and in the prediction phase.

In[5], the author is discussing various ways to classify the video, such as the traditional hand crafted features on frame level data, but in recent times the most seen model or most used model is CNN as the accuracy of this model is high. Apart from this LSTM and RNN model is also used. In this the model is divided into three main modules Video pre-processing, Feature extraction, Video representation and classification. In Deep feature extraction instead of training CNN from scratch it uses a pre-trained model called efficient-net. Then it undergoes bidirectional LSTM(BiLSTM).It can help any video sharing site either blur/hide any segment with unsettling frames or eliminate the video that contains dangerous clips.

In [6], the author describes and demonstrates CNN that is an extensively used image classifier model of deep learning. Within CNN, videos are treated as a short bag of fixed sized clips. It also describes various time information fusion i.e. the ways in which the information from one frame to other is transferred such as early fusion, late fusion and slow fusion. It also describes multiresolution CNNs which strive to resolve the problem of long training duration and in which input frames are fed into the model

in two separate streams for processing - a context stream that models low-resolution images and a fovea stream that processes high-resolution images. Both the streams have alternating convolution, normalization and pooling layers. Also the author carries out the experimentation of the model on the Sports-1M dataset that consists of 200.000 videos. The results of the experiment are also discussed.

In [7], the author has given a detailed overview about SCORM, data sources are typically represented by plain documents and the context information that goes with them, while metadata needs to be filled in regarding any aspect of document usage, content, or context. The automatic text extraction from video files, in this multi step pre processing technique is used which includes audio extraction, audio splitting, audio segmentation and silence removal. Lastly in performance measure ROUGE is used namely ROUGE N, ROUGE L, ROUGE W, ROUGE S. The results are based on Description and Title field.

In this there is a detailed description about the semi-automatic metadata generation, its type, tools, technique. Categorizing semantic metadata creation into two models, namely ontology-driven semantic tagging and second semantic metadata generation, Nonetheless, the findings suggest that in order to enhance the effectiveness of metadata production in an actual library environment, it will soon be necessary to take advantage of promising findings from experimental research. Growing digital repositories and web resources at a quick pace demonstrate how important automated metadata production is becoming. When working on semi-automatic metadata production, system designers must incorporate the results of experimental research and the specialized knowledge of metadata experts. The study's findings also suggest that additional investigation is required to develop automatic metadata generation for semantic metadata in real-world applications. [8]

In this two different video classification models are considered namely recurrent network based model and Cluster and Aggregate Based Models.Results are on bases of Comparisons of different baselines, Comparing Teacher-Student Network with Uniform k Baseline, Serial Versus Parallel Training of Teacher-Student, Computation time of different models, etc. Specifically, it tested the model on the YouTube-8M dataset and found that the student network—which is computationally less expensive—can reduce computation time by thirty percent while maintaining a performance that is roughly comparable to that of the instructor network. [9]

## III. PROPOSED SYSTEM

In the project,we propose a system in which a video file is provided initially and then it is preprocessed and transformed into a series of uniformly shaped frames. Further the frames are processed for recognition of text, extraction of speech and detection of objects and the results of this are forwarded through a dictionary in which each category has a set of words associated with it . Also they are passed through a 2D-CNN model for frame based classification. Both the results are then taken into considerations as per the weights that define the solution and then a final output is obtained.

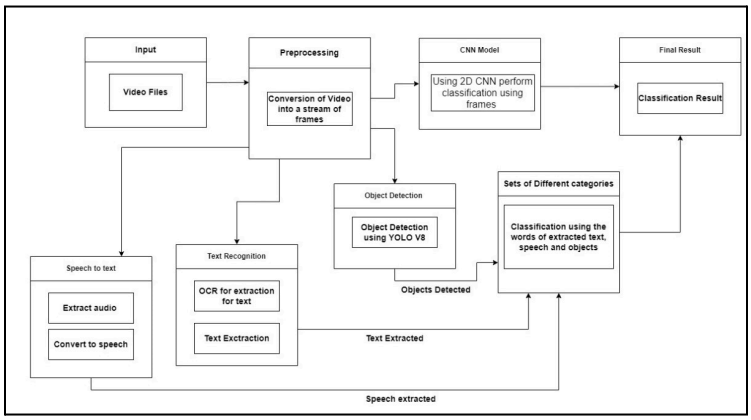


*Fig 1:System Architecture*

- ***Dataset and Preprocessing:*** The dataset has a number of videos that are categorized into 11 different categories. The length of videos is ranging from one second to one minute.The total count of videos is __ .The dataset has videos related to varied aspects of ISRO such as launchpad,rockets,indoor labs,etc.. For the pre-processing task the OpenCV library has been employed which resizes each frame of the video into a constant size image which is further utilized by the deep learning model to proceed with further tasks.The computer vision library first transforms the given videos into an array of frames and then by measuring the dimensions of each converts it into an uniform dimension of __.The preprocessing is carried out as the dataset has videos in different aspect ratio and in order to process the videos it

necessary that we transform them uniformly in such a state that is suitable for the underlined task.

***Video Analysis and Feature Extraction:***

- ***Object Detection****:*Object Detection from the videos was executed using the YOLO V8 model which is a pre-trained object detection model. If the images are not of the same size then resize them into the same size(for example 448). Once images are resized, then for this first various images which had the objects to be detected were collected and then using the LabelImg tool the coordinates of the object for training the model were obtained.After the labeling, the images are passed to train the model and the results are obtained for the detection.

- ***Text Recognition****:*\The text that appears in the frames can provide us utilitarian information that can describe the category to which the video belongs to. This can be achieved using the words detected in the text recognized.This task is carried out by using the Optical Character Recognition (OCR) which provides the textual and numerical information displayed on the screen. It iterates through each frame and captures the text.The accuracy of the text retrieved from a frame is proportional to the quality of the frame.

- ***Speech Recognition****:* Audio that runs in the background can be transcribed to text using the speech-to-text models. Moviepy is a python library which is used to execute this task. The efficiency of this model depends on the quality of the audio in the video.

- ***Classification****:* The features from the videos have been retrieved in the text form. The text form has specific words that can be used in order to classify the video into a set of categories. This is the first stage of classification. It uses a basic set data structure that can be employed to get the categories for which the video at the input belongs to. To prepare these sets the whole dataset is thoroughly examined and words that are frequently appearing in the audio as well as the text displaying on the screen are entered into the set of words for that category. For example, the words such as solar panel, LMV ,appear frequently in the metadata obtained from the videos that belong to the category of Satellites, so these words would be collectively mentioned in a set named Satellites. This is carried out for all other categorical labels. As a single word would belong to multiple categories the output would be a list of labels for which the video belongs to.

In the second phase of the classification process the video is passed on to frame wise classification which is performed by the 2D-CNN deep learning model. For the framewise classification we had considered two more models that can perform the classification process so that the accuracy can be verified. CNN-LSTM, 2D-CNN and 3D-CNN were considered for it. But as most of the videos were intended to have a short length considering the temporal aspect would result in an adverse effect on the classification result. The 3D-CNN and CNN-LSTM models which consider the temporal aspect of the video proved to be less proficient towards the classification task. But the 2D-CNN model which considers only the spatial aspect of the video was more efficient towards the classification and gave an accuracy of 95.30%. Fitting this model to a new video out of the dataset mostly provided accurate results of classification.

Once both the phases of classification were completed all the results were displayed and the classification label of the video was obtained.

## IV. OUTPUT

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

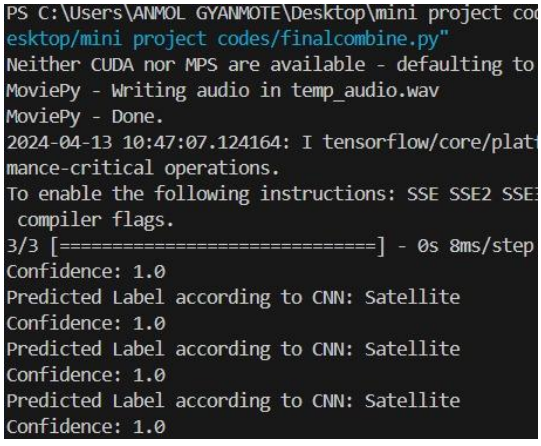


*Fig 2:Frame of a video*

The 2D-CNN model was trained on a large video dataset for each category and the results of the prediction for a video are displayed with the confidence for the model.

2. Text Extraction


*Fig 3: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. A particular text is printed multiple times as it captures the text on the screen after a regular interval and prints it if there are any changes with the previously captured text.

```
WARNING:easyocr.easyocr:Neither CUDA nor MPS are available - defaulting to CPU. Note: This module is much f
WARNING:easyocr.easyocr:Downloading detection model, please wait. This may take several minutes depending u
Progress: |                                        | 100.0% CompleteWARNING:easyocr.easyocr:Downl
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 4: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 5:Speech that is being transcripted from the video*

The figure above shows the audio that is being played in the background of the video was extracted and transformed into text using the Movie.py library and Google speech recognition services.

4. Object detection

```
Ultralytics YOLOv8.0.20 🚀 Python-3.10.12 torch-2.2.1+cu121 CUDA:0 (Tesla T4, 15102MiB)
Model summary (fused): 218 layers, 25844392 parameters, 0 gradients, 78.7 GFLOPs
val: Scanning /content/drive/MyDrive/miniproj Dataset Object new/val/labels.cache... 16 images, 0 ba
                 Class     Images  Instances      Box(P          R      mAP50  mAP50-95):   0% 0/1 [
  self.pid = os.fork()
                 Class     Images  Instances      Box(P          R      mAP50  mAP50-95): 100% 1/1 [
                   all         16         31      0.813      0.895      0.896      0.636
               monitor         16          3      0.655      0.667      0.562      0.328
                 gauge         16          4      0.541       0.75      0.688      0.378
               labcoat         16          8      0.856      0.743      0.939      0.731
                rocket         16          4      0.793          1      0.995       0.69
                 tower         16          6       0.82          1      0.995      0.617
            microscope         16          2      0.961          1      0.995      0.747
            rocketnose         16          2      0.917          1      0.995      0.846
             satellite         16          2      0.961          1      0.995      0.747
Speed: 0.3ms pre-process, 22.3ms inference, 0.0ms loss, 34.5ms post-process per image
```

*Fig 6:Results of the YOLOv8 training*

The figure above shows the YOLOv8 model trained for 9 custom object classes and the results of the model.The size of the YOLOv8 model considered for the above training is medium.


*Fig 7:Result of the trained YOLOv8 model*

The figure above shows the result of the validation of the YOLOv8 model which was trained earlier.The model could correctly predict the rocket,monitor,lab coat ,gauge, towers.The integration of this object detection model is one of the future scope of the project as there were few problems faced during the integration regarding the dependencies.

5. Integrated process

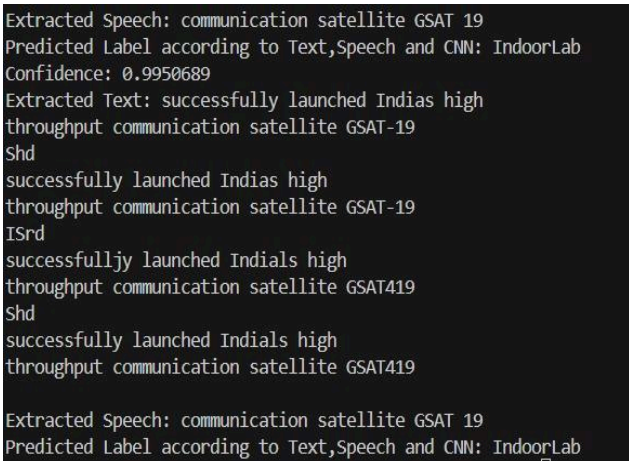


*Fig 8:Prediction of the video category by integration of CNN,text extracted and speech detected.*

A video was provided in the input and the speech in the video,text displayed on the video and the frames of the video were detected and the category of the video is predicted using the data.

## V. 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.

## IV. REFERENCE


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.