

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.

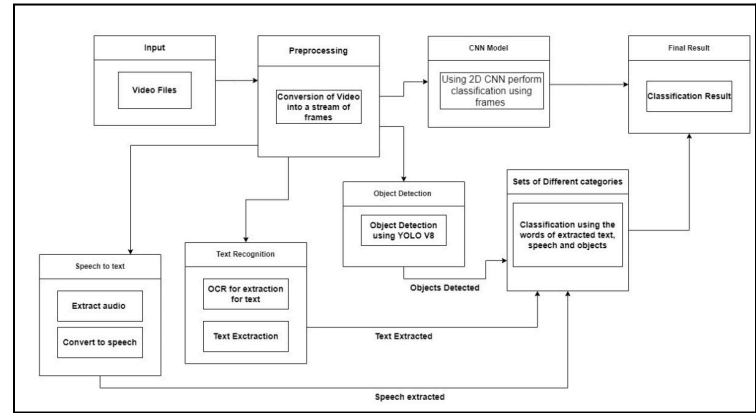


Fig 1: System Architecture

- 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

```
PS C:\Users\ANMOL GYANMOT\Desktop\mini project codes> python C:\Users\ANMOL GYANMOT\Desktop\mini project codes\finalcombine.py
Neither CUDA nor MPS are available - defaulting to CPU
MoviePy - Writing audio in temp_audio.wav
MoviePy - Done.
2024-04-13 10:47:07.124164: I tensorflow/core/platform/default_device_driver.cc:346] Please refer to docs for the list of GPU compiler flags.
To enable the following instructions: SSE SSE2 SSE3 SSE4 SSE4.1 AVX AVX2 FMA FMA4 FMA3 FMA4 FMA5 FMA6 FMA7 FMA8 FMA9 FMA10 FMA11 FMA12 FMA13 FMA14 FMA15 FMA16 FMA17 FMA18 FMA19 FMA20 FMA21 FMA22 FMA23 FMA24 FMA25 FMA26 FMA27 FMA28 FMA29 FMA30 FMA31 FMA32 FMA33 FMA34 FMA35 FMA36 FMA37 FMA38 FMA39 FMA40 FMA41 FMA42 FMA43 FMA44 FMA45 FMA46 FMA47 FMA48 FMA49 FMA50 FMA51 FMA52 FMA53 FMA54 FMA55 FMA56 FMA57 FMA58 FMA59 FMA60 FMA61 FMA62 FMA63 FMA64 FMA65 FMA66 FMA67 FMA68 FMA69 FMA70 FMA71 FMA72 FMA73 FMA74 FMA75 FMA76 FMA77 FMA78 FMA79 FMA80 FMA81 FMA82 FMA83 FMA84 FMA85 FMA86 FMA87 FMA88 FMA89 FMA90 FMA91 FMA92 FMA93 FMA94 FMA95 FMA96 FMA97 FMA98 FMA99 FMA100 FMA101 FMA102 FMA103 FMA104 FMA105 FMA106 FMA107 FMA108 FMA109 FMA110 FMA111 FMA112 FMA113 FMA114 FMA115 FMA116 FMA117 FMA118 FMA119 FMA120 FMA121 FMA122 FMA123 FMA124 FMA125 FMA126 FMA127 FMA128 FMA129 FMA130 FMA131 FMA132 FMA133 FMA134 FMA135 FMA136 FMA137 FMA138 FMA139 FMA140 FMA141 FMA142 FMA143 FMA144 FMA145 FMA146 FMA147 FMA148 FMA149 FMA150 FMA151 FMA152 FMA153 FMA154 FMA155 FMA156 FMA157 FMA158 FMA159 FMA160 FMA161 FMA162 FMA163 FMA164 FMA165 FMA166 FMA167 FMA168 FMA169 FMA170 FMA171 FMA172 FMA173 FMA174 FMA175 FMA176 FMA177 FMA178 FMA179 FMA180 FMA181 FMA182 FMA183 FMA184 FMA185 FMA186 FMA187 FMA188 FMA189 FMA190 FMA191 FMA192 FMA193 FMA194 FMA195 FMA196 FMA197 FMA198 FMA199 FMA200 FMA201 FMA202 FMA203 FMA204 FMA205 FMA206 FMA207 FMA208 FMA209 FMA210 FMA211 FMA212 FMA213 FMA214 FMA215 FMA216 FMA217 FMA218 FMA219 FMA220 FMA221 FMA222 FMA223 FMA224 FMA225 FMA226 FMA227 FMA228 FMA229 FMA230 FMA231 FMA232 FMA233 FMA234 FMA235 FMA236 FMA237 FMA238 FMA239 FMA240 FMA241 FMA242 FMA243 FMA244 FMA245 FMA246 FMA247 FMA248 FMA249 FMA250 FMA251 FMA252 FMA253 FMA254 FMA255 FMA256 FMA257 FMA258 FMA259 FMA260 FMA261 FMA262 FMA263 FMA264 FMA265 FMA266 FMA267 FMA268 FMA269 FMA270 FMA271 FMA272 FMA273 FMA274 FMA275 FMA276 FMA277 FMA278 FMA279 FMA280 FMA281 FMA282 FMA283 FMA284 FMA285 FMA286 FMA287 FMA288 FMA289 FMA290 FMA291 FMA292 FMA293 FMA294 FMA295 FMA296 FMA297 FMA298 FMA299 FMA300 FMA301 FMA302 FMA303 FMA304 FMA305 FMA306 FMA307 FMA308 FMA309 FMA310 FMA311 FMA312 FMA313 FMA314 FMA315 FMA316 FMA317 FMA318 FMA319 FMA320 FMA321 FMA322 FMA323 FMA324 FMA325 FMA326 FMA327 FMA328 FMA329 FMA330 FMA331 FMA332 FMA333 FMA334 FMA335 FMA336 FMA337 FMA338 FMA339 FMA340 FMA341 FMA342 FMA343 FMA344 FMA345 FMA346 FMA347 FMA348 FMA349 FMA350 FMA351 FMA352 FMA353 FMA354 FMA355 FMA356 FMA357 FMA358 FMA359 FMA360 FMA361 FMA362 FMA363 FMA364 FMA365 FMA366 FMA367 FMA368 FMA369 FMA370 FMA371 FMA372 FMA373 FMA374 FMA375 FMA376 FMA377 FMA378 FMA379 FMA380 FMA381 FMA382 FMA383 FMA384 FMA385 FMA386 FMA387 FMA388 FMA389 FMA390 FMA391 FMA392 FMA393 FMA394 FMA395 FMA396 FMA397 FMA398 FMA399 FMA400 FMA401 FMA402 FMA403 FMA404 FMA405 FMA406 FMA407 FMA408 FMA409 FMA410 FMA411 FMA412 FMA413 FMA414 FMA415 FMA416 FMA417 FMA418 FMA419 FMA420 FMA421 FMA422 FMA423 FMA424 FMA425 FMA426 FMA427 FMA428 FMA429 FMA430 FMA431 FMA432 FMA433 FMA434 FMA435 FMA436 FMA437 FMA438 FMA439 FMA440 FMA441 FMA442 FMA443 FMA444 FMA445 FMA446 FMA447 FMA448 FMA449 FMA450 FMA451 FMA452 FMA453 FMA454 FMA455 FMA456 FMA457 FMA458 FMA459 FMA460 FMA461 FMA462 FMA463 FMA464 FMA465 FMA466 FMA467 FMA468 FMA469 FMA470 FMA471 FMA472 FMA473 FMA474 FMA475 FMA476 FMA477 FMA478 FMA479 FMA480 FMA481 FMA482 FMA483 FMA484 FMA485 FMA486 FMA487 FMA488 FMA489 FMA490 FMA491 FMA492 FMA493 FMA494 FMA495 FMA496 FMA497 FMA498 FMA499 FMA500 FMA501 FMA502 FMA503 FMA504 FMA505 FMA506 FMA507 FMA508 FMA509 FMA510 FMA511 FMA512 FMA513 FMA514 FMA515 FMA516 FMA517 FMA518 FMA519 FMA520 FMA521 FMA522 FMA523 FMA524 FMA525 FMA526 FMA527 FMA528 FMA529 FMA530 FMA531 FMA532 FMA533 FMA534 FMA535 FMA536 FMA537 FMA538 FMA539 FMA540 FMA541 FMA542 FMA543 FMA544 FMA545 FMA546 FMA547 FMA548 FMA549 FMA550 FMA551 FMA552 FMA553 FMA554 FMA555 FMA556 FMA557 FMA558 FMA559 FMA560 FMA561 FMA562 FMA563 FMA564 FMA565 FMA566 FMA567 FMA568 FMA569 FMA570 FMA571 FMA572 FMA573 FMA574 FMA575 FMA576 FMA577 FMA578 FMA579 FMA580 FMA581 FMA582 FMA583 FMA584 FMA585 FMA586 FMA587 FMA588 FMA589 FMA590 FMA591 FMA592 FMA593 FMA594 FMA595 FMA596 FMA597 FMA598 FMA599 FMA600 FMA601 FMA602 FMA603 FMA604 FMA605 FMA606 FMA607 FMA608 FMA609 FMA610 FMA611 FMA612 FMA613 FMA614 FMA615 FMA616 FMA617 FMA618 FMA619 FMA620 FMA621 FMA622 FMA623 FMA624 FMA625 FMA626 FMA627 FMA628 FMA629 FMA630 FMA631 FMA632 FMA633 FMA634 FMA635 FMA636 FMA637 FMA638 FMA639 FMA640 FMA641 FMA642 FMA643 FMA644 FMA645 FMA646 FMA647 FMA648 FMA649 FMA650 FMA651 FMA652 FMA653 FMA654 FMA655 FMA656 FMA657 FMA658 FMA659 FMA660 FMA661 FMA662 FMA663 FMA664 FMA665 FMA666 FMA667 FMA668 FMA669 FMA670 FMA671 FMA672 FMA673 FMA674 FMA675 FMA676 FMA677 FMA678 FMA679 FMA680 FMA681 FMA682 FMA683 FMA684 FMA685 FMA686 FMA687 FMA688 FMA689 FMA690 FMA691 FMA692 FMA693 FMA694 FMA695 FMA696 FMA697 FMA698 FMA699 FMA700 FMA701 FMA702 FMA703 FMA704 FMA705 FMA706 FMA707 FMA708 FMA709 FMA710 FMA711 FMA712 FMA713 FMA714 FMA715 FMA716 FMA717 FMA718 FMA719 FMA720 FMA721 FMA722 FMA723 FMA724 FMA725 FMA726 FMA727 FMA728 FMA729 FMA730 FMA731 FMA732 FMA733 FMA734 FMA735 FMA736 FMA737 FMA738 FMA739 FMA740 FMA741 FMA742 FMA743 FMA744 FMA745 FMA746 FMA747 FMA748 FMA749 FMA750 FMA751 FMA752 FMA753 FMA754 FMA755 FMA756 FMA757 FMA758 FMA759 FMA760 FMA761 FMA762 FMA763 FMA764 FMA765 FMA766 FMA767 FMA768 FMA769 FMA770 FMA771 FMA772 FMA773 FMA774 FMA775 FMA776 FMA777 FMA778 FMA779 FMA780 FMA781 FMA782 FMA783 FMA784 FMA785 FMA786 FMA787 FMA788 FMA789 FMA790 FMA791 FMA792 FMA793 FMA794 FMA795 FMA796 FMA797 FMA798 FMA799 FMA800 FMA801 FMA802 FMA803 FMA804 FMA805 FMA806 FMA807 FMA808 FMA809 FMA810 FMA811 FMA812 FMA813 FMA814 FMA815 FMA816 FMA817 FMA818 FMA819 FMA820 FMA821 FMA822 FMA823 FMA824 FMA825 FMA826 FMA827 FMA828 FMA829 FMA830 FMA831 FMA832 FMA833 FMA834 FMA835 FMA836 FMA837 FMA838 FMA839 FMA840 FMA841 FMA842 FMA843 FMA844 FMA845 FMA846 FMA847 FMA848 FMA849 FMA850 FMA851 FMA852 FMA853 FMA854 FMA855 FMA856 FMA857 FMA858 FMA859 FMA860 FMA861 FMA862 FMA863 FMA864 FMA865 FMA866 FMA867 FMA868 FMA869 FMA870 FMA871 FMA872 FMA873 FMA874 FMA875 FMA876 FMA877 FMA878 FMA879 FMA880 FMA881 FMA882 FMA883 FMA884 FMA885 FMA886 FMA887 FMA888 FMA889 FMA890 FMA891 FMA892 FMA893 FMA894 FMA895 FMA896 FMA897 FMA898 FMA899 FMA900 FMA901 FMA902 FMA903 FMA904 FMA905 FMA906 FMA907 FMA908 FMA909 FMA910 FMA911 FMA912 FMA913 FMA914 FMA915 FMA916 FMA917 FMA918 FMA919 FMA920 FMA921 FMA922 FMA923 FMA924 FMA925 FMA926 FMA927 FMA928 FMA929 FMA930 FMA931 FMA932 FMA933 FMA934 FMA935 FMA936 FMA937 FMA938 FMA939 FMA940 FMA941 FMA942 FMA943 FMA944 FMA945 FMA946 FMA947 FMA948 FMA949 FMA950 FMA951 FMA952 FMA953 FMA954 FMA955 FMA956 FMA957 FMA958 FMA959 FMA960 FMA961 FMA962 FMA963 FMA964 FMA965 FMA966 FMA967 FMA968 FMA969 FMA970 FMA971 FMA972 FMA973 FMA974 FMA975 FMA976 FMA977 FMA978 FMA979 FMA980 FMA981 FMA982 FMA983 FMA984 FMA985 FMA986 FMA987 FMA988 FMA989 FMA990 FMA991 FMA992 FMA993 FMA994 FMA995 FMA996 FMA997 FMA998 FMA999 FMA1000 FMA1001 FMA1002 FMA1003 FMA1004 FMA1005 FMA1006 FMA1007 FMA1008 FMA1009 FMA1010 FMA1011 FMA1012 FMA1013 FMA1014 FMA1015 FMA1016 FMA1017 FMA1018 FMA1019 FMA1020 FMA1021 FMA1022 FMA1023 FMA1024 FMA1025 FMA1026 FMA1027 FMA1028 FMA1029 FMA1030 FMA1031 FMA1032 FMA1033 FMA1034 FMA1035 FMA1036 FMA1037 FMA1038 FMA1039 FMA1040 FMA1041 FMA1042 FMA1043 FMA1044 FMA1045 FMA1046 FMA1047 FMA1048 FMA1049 FMA1050 FMA1051 FMA1052 FMA1053 FMA1054 FMA1055 FMA1056 FMA1057 FMA1058 FMA1059 FMA1060 FMA1061 FMA1062 FMA1063 FMA1064 FMA1065 FMA1066 FMA1067 FMA1068 FMA1069 FMA1070 FMA1071 FMA1072 FMA1073 FMA1074 FMA1075 FMA1076 FMA1077 FMA1078 FMA1079 FMA1080 FMA1081 FMA1082 FMA1083 FMA1084 FMA1085 FMA1086 FMA1087 FMA1088 FMA1089 FMA1090 FMA1091 FMA1092 FMA1093 FMA1094 FMA1095 FMA1096 FMA1097 FMA1098 FMA1099 FMA1100 FMA1101 FMA1102 FMA1103 FMA1104 FMA1105 FMA1106 FMA1107 FMA1108 FMA1109 FMA1110 FMA1111 FMA1112 FMA1113 FMA1114 FMA1115 FMA1116 FMA1117 FMA1118 FMA1119 FMA1120 FMA1121 FMA1122 FMA1123 FMA1124 FMA1125 FMA1126 FMA1127 FMA1128 FMA1129 FMA1130 FMA1131 FMA1132 FMA1133 FMA1134 FMA1135 FMA1136 FMA1137 FMA1138 FMA1139 FMA1140 FMA1141 FMA1142 FMA1143 FMA1144 FMA1145 FMA1146 FMA1147 FMA1148 FMA1149 FMA1150 FMA1151 FMA1152 FMA1153 FMA1154 FMA1155 FMA1156 FMA1157 FMA1158 FMA1159 FMA1160 FMA1161 FMA1162 FMA1163 FMA1164 FMA1165 FMA1166 FMA1167 FMA1168 FMA1169 FMA1170 FMA1171 FMA1172 FMA1173 FMA1174 FMA1175 FMA1176 FMA1177 FMA1178 FMA1179 FMA1180 FMA1181 FMA1182 FMA1183 FMA1184 FMA1185 FMA1186 FMA1187 FMA1188 FMA1189 FMA1190 FMA1191 FMA1192 FMA1193 FMA1194 FMA1195 FMA1196 FMA1197 FMA1198 FMA1199 FMA1200 FMA1201 FMA1202 FMA1203 FMA1204 FMA1205 FMA1206 FMA1207 FMA1208 FMA1209 FMA1210 FMA1211 FMA1212 FMA1213 FMA1214 FMA1215 FMA1216 FMA1217 FMA1218 FMA1219 FMA1220 FMA1221 FMA1222 FMA1223 FMA1224 FMA1225 FMA1226 FMA1227 FMA1228 FMA1229 FMA1230 FMA1231 FMA1232 FMA1233 FMA1234 FMA1235 FMA1236 FMA1237 FMA1238 FMA1239 FMA1240 FMA1241 FMA1242 FMA1243 FMA1244 FMA1245 FMA1246 FMA1247 FMA1248 FMA1249 FMA1250 FMA1251 FMA1252 FMA1253 FMA1254 FMA1255 FMA1256 FMA1257 FMA1258 FMA1259 FMA1260 FMA1261 FMA1262 FMA1263 FMA1264 FMA1265 FMA1266 FMA1267 FMA1268 FMA1269 FMA1270 FMA1271 FMA1272 FMA1273 FMA1274 FMA1275 FMA1276 FMA1277 FMA1278 FMA1279 FMA1280 FMA1281 FMA1282 FMA1283 FMA1284 FMA1285 FMA1286 FMA1287 FMA1288 FMA1289 FMA1290 FMA1291 FMA1292 FMA1293 FMA1294 FMA1295 FMA1296 FMA1297 FMA1298 FMA1299 FMA1300 FMA1301 FMA1302 FMA1303 FMA1304 FMA1305 FMA1306 FMA1307 FMA1308 FMA1309 FMA1310 FMA1311 FMA1312 FMA1313 FMA1314 FMA1315 FMA1316 FMA1317 FMA1318 FMA1319 FMA1320 FMA1321 FMA1322 FMA1323 FMA1324 FMA1325 FMA1326 FMA1327 FMA1328 FMA1329 FMA1330 FMA1331 FMA1332 FMA1333 FMA1334 FMA1335 FMA1336 FMA1337 FMA1338 FMA1339 FMA1340 FMA1341 FMA1342 FMA1343 FMA1344 FMA1345 FMA1346 FMA1347 FMA1348 FMA1349 FMA1350 FMA1351 FMA1352 FMA1353 FMA1354 FMA1355 FMA1356 FMA1357 FMA1358 FMA1359 FMA1360 FMA1361 FMA1362 FMA1363 FMA1364 FMA1365 FMA1366 FMA1367 FMA1368 FMA1369 FMA1370 FMA1371 FMA1372 FMA1373 FMA1374 FMA1375 FMA1376 FMA1377 FMA1378 FMA1379 FMA1380 FMA1381 FMA1382 FMA1383 FMA1384 FMA1385 FMA1386 FMA1387 FMA1388 FMA1389 FMA1390 FMA1391 FMA1392 FMA1393 FMA1394 FMA1395 FMA1396 FMA1397 FMA1398 FMA1399 FMA1400 FMA1401 FMA1402 FMA1403 FMA1404 FMA1405 FMA1406 FMA1407 FMA1408 FMA1409 FMA1410 FMA1411 FMA1412 FMA1413 FMA1414 FMA1415 FMA1416 FMA1417 FMA1418 FMA1419 FMA1420 FMA1421 FMA1422 FMA1423 FMA1424 FMA1425 FMA1426 FMA1427 FMA1428 FMA1429 FMA1430 FMA1431 FMA1432 FMA1433 FMA1434 FMA1435 FMA1436 FMA1437 FMA1438 FMA1439 FMA1440 FMA1441 FMA1442 FMA1443 FMA1444 FMA1445 FMA1446 FMA1447 FMA1448 FMA1449 FMA1450 FMA1451 FMA1452 FMA1453 FMA1454 FMA1455 FMA1456 FMA1457 FMA1458 FMA1459 FMA1460 FMA1461 FMA1462 FMA1463 FMA1464 FMA1465 FMA1466 FMA1467 FMA1468 FMA1469 FMA1470 FMA1471 FMA1472 FMA1473 FMA1474 FMA1475 FMA1476 FMA1477 FMA1478 FMA1479 FMA1480 FMA1481 FMA1482 FMA1483 FMA1484 FMA1485 FMA1486 FMA1487 FMA1488 FMA1489 FMA1490 FMA1491 FMA1492 FMA1493 FMA1494 FMA1495 FMA1496 FMA1497 FMA1498 FMA1499 FMA1500 FMA1501 FMA1502 FMA1503 FMA1504 FMA1505 FMA1506 FMA1507 FMA1508 FMA1509 FMA1510 FMA1511 FMA1512 FMA1513 FMA1514 FMA1515 FMA1516 FMA1517 FMA1518 FMA1519 FMA1520 FMA1521 FMA1522 FMA1523 FMA1524 FMA1525 FMA1526 FMA1527 FMA1528 FMA1529 FMA1530 FMA1531 FMA1532 FMA1533 FMA1534 FMA1535 FMA1536 FMA1537 FMA1538 FMA1539 FMA1540 FMA1541 FMA1542 FMA1543 FMA1544 FMA1545 FMA1546 FMA1547 FMA1548 FMA1549 FMA1550 FMA1551 FMA1552 FMA1553 FMA1554 FMA1555 FMA1556 FMA1557 FMA1558 FMA1559 FMA1560 FMA1561 FMA1562 FMA1563 FMA1564 FMA1565 FMA1566 FMA1567 FMA1568 FMA1569 FMA1570 FMA1571 FMA1572 FMA1573 FMA1574 FMA1575 FMA1576 FMA1577 FMA1578 FMA1579 FMA1580 FMA1581 FMA1582 FMA1583 FMA1584 FMA1585 FMA1586 FMA1587 FMA1588 FMA1589 FMA1590 FMA1591 FMA1592 FMA1593 FMA1594 FMA1595 FMA1596 FMA1597 FMA1598 FMA1599 FMA1600 FMA1601 FMA1602 FMA1603 FMA1604 FMA1605 FMA1606 FMA1607 FMA1608 FMA1609 FMA1610 FMA1611 FMA1612 FMA1613 FMA1614 FMA1615 FMA1616 FMA1617 FMA1618 FMA1619 FMA1620 FMA1621 FMA1622 FMA1623 FMA1624 FMA1625 FMA1626 FMA1627 FMA1628 FMA1629 FMA1630 FMA1631 FMA1632 FMA1633 FMA1634 FMA1635 FMA1636 FMA1637 FMA1638 FMA1639 FMA1640 FMA1641 FMA1642 FMA1643 FMA1644 FMA1645 FMA1646 FMA1647 FMA1648 FMA1649 FMA1650 FMA1651 FMA1652 FMA1653 FMA1654 FMA1655 FMA1656 FMA1657 FMA1658 FMA1659 FMA1660 FMA1661 FMA1662 FMA1663 FMA1664 FMA1665 FMA1666 FMA1667 FMA1668 FMA1669 FMA1670 FMA1671 FMA1672 FMA1673 FMA1674 FMA1675 FMA1676 FMA1677 FMA1678 FMA1679 FMA1680 FMA1681 FMA1682 FMA1683 FMA1684 FMA1685 FMA1686 FMA1687 FMA1688 FMA1689 FMA1690 FMA1691 FMA1692 FMA1693 FMA1694 FMA1695 FMA1696 FMA1697 FMA1698 FMA1699 FMA1700 FMA1701 FMA1702 FMA1703 FMA1704 FMA1705 FMA1706 FMA1707 FMA1708 FMA1709 FMA1710 FMA1711 FMA1712 FMA1713 FMA1714 FMA1715 FMA1716 FMA1717 FMA1718 FMA1719 FMA1720 FMA1721 FMA1722 FMA1723 FMA1724 FMA1725 FMA1726 FMA1727 FMA1728 FMA1729 FMA1730 FMA1731 FMA1732 FMA1733 FMA1734 FMA1735 FMA1736 FMA1737 FMA1738 FMA1739 FMA1740 FMA1741 FMA1742 FMA1743 FMA1744 FMA1745 FMA1746 FMA1747 FMA1748 FMA1749 FMA1750 FMA1751 FMA1752 FMA1753 FMA1754 FMA1755 FMA1756 FMA1757 FMA1758 FMA1759 FMA1760 FMA1761 FMA1762 FMA1763 FMA1764 FMA1765 FMA1766 FMA1767 FMA1768 FMA1769 FMA1770 FMA1771 FMA1772 FMA1773 FMA1774 FMA1775 FMA1776 FMA1777 FMA1778 FMA1779 FMA1780 FMA1781 FMA1782 FMA1783 FMA1784 FMA1785 FMA1786 FMA1787 FMA1788 FMA1789 FMA1790 FMA1791 FMA1792 FMA1793 FMA1794 FMA1795 FMA1796 FMA1797 FMA1798 FMA1799 FMA1800 FMA1801 FMA1802 FMA1803 FMA1804 FMA1805 FMA1806 FMA1807 FMA1808 FMA1809 FMA1810 FMA1811 FMA1812 FMA1813 FMA1814 FMA1815 FMA1816 FMA1817 FMA1818 FMA1819 FMA1820 FMA1821 FMA1822 FMA1823 FMA1824 FMA1825 FMA1826 FMA1827 FMA1828 FMA1829 FMA1830 FMA1831 FMA1832 FMA1833 FMA1834 FMA1835 FMA1836 FMA1837 FMA1838 FMA1839 FMA1840 FMA1841 FMA1842 FMA1843 FMA1844 FMA1845 FMA1846 FMA1847 FMA1848 FMA1849 FMA1850 FMA1851 FMA1852 FMA1853 FMA1854 FMA1855 FMA1856 FMA1857 FMA1858 FMA1859 FMA1860 FMA1861 FMA1862 FMA1863 FMA1864 FMA1865 FMA1866 FMA1867 FMA1868 FMA1869 FMA1870 FMA1871 FMA1872 FMA1873 FMA1874 FMA1875 FMA1876 FMA1877 FMA1878 FMA1879 FMA1880 FMA1881 FMA1882 FMA1883 FMA1884 FMA1885 FMA1886 FMA1887 FMA1888 FMA1889 FMA1890 FMA1891 FMA1892 FMA1893 FMA1894 FMA1895 FMA1896 FMA1897 FMA1898 FMA1899 FMA1900 FMA1901 FMA1902 FMA1903 FMA1904 FMA1905 FMA1906 FMA1907 FMA1908 FMA1909 FMA1910 FMA1911 FMA1912 FMA1913 FMA1914 FMA1915 FMA1916 FMA1917 FMA1918 FMA1919 FMA1920 FMA1921 FMA1922 FMA1923 FMA1924 FMA1925 FMA1926 FMA1927 FMA1928 FMA1929 FMA1930 FMA1931 FMA1932 FMA1933 FMA1934 FMA1935 FMA1936 FMA1937 FMA1938 FMA1939 FMA1940 FMA1941 FMA1942 FMA1943 FMA1944 FMA1945 FMA1946 FMA1947 FMA1948 FMA1949 FMA1950 FMA1951 FMA1952 FMA1953 FMA1954 FMA1955 FMA1956 FMA1957 FMA1958 FMA1959 FMA1960 FMA1961 FMA1962 FMA1963 FMA1964 FMA1965 FMA1966 FMA1967 FMA1968 FMA1969 FMA1970 FMA1971 FMA1972 FMA1973 FMA1974 FMA1975 FMA1976 FMA1977 FMA1978 FMA1979 FMA1980 FMA1981 FMA1982 FMA1983 FMA1984 FMA1985 FMA1986 FMA1987 FMA1988 FMA1989 FMA1990 FMA1991 FMA1992 FMA1993 FMA1994 FMA1995 FMA1996 FMA1997 FMA1998 FMA1999 FMA2000 FMA2001 FMA2002 FMA2003 FMA2004 FMA2005 FMA2006 FMA2007 FMA2008 FMA2009 FMA2010 FMA2011 FMA2012 FMA2013 FMA2014 FMA2015 FMA2016 FMA2017 FMA2018 FMA2019 FMA2020 FMA2021 FMA2022 FMA2023 FMA2024 FMA2025 FMA2026 FMA2027 FMA2028 FMA2029 FMA2030 FMA2031 FMA2032 FMA2033 FMA2034 FMA2035 FMA2036 FMA2037 FMA2038 F
```

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.

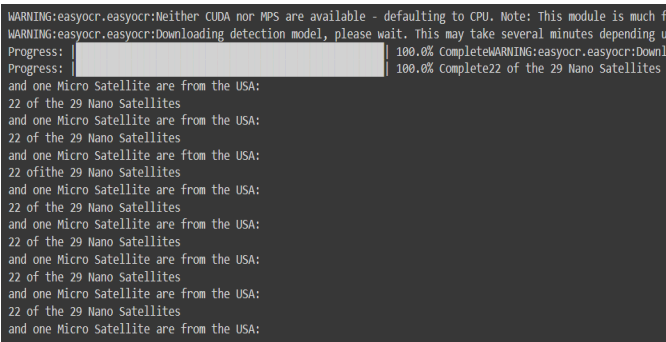


Fig 4:Text captured through the frame

3. Speech Extraction

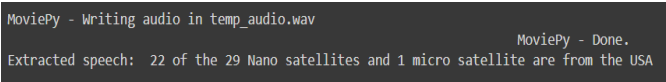


Fig 5:Speech that is being transcribed from the video

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
self.pid = os.fork()
Class      Images  Instances  Box(P      R      mAP50  mAP50-95): 0% 0/1
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

```
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 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.