ISSN 2063-5346

VIDEO PLAGIARISM TO DETECT COPYRIGHT INFRINGEMENT

Aparna Sawant[1], Prithviraj Sangle[2], Riddhi Karandikar[3], Parth Patil[4], Janhavi Pawar[5], Parul Patle[6]

Article History: Received: 01.02.2023 Revised: 07.03.2023 Accepted: 10.04.2023

Abstract

Social-media and the digital world as a whole are expanding quickly today. On the internet, several videos are posted every day. Since everyone has access to the internet, it is simple to download content and re-upload it, which is a severe social issue. Plagiarism occurs when someone steals a video from the internet and publishes it under their name. A technique for feature extraction of video plagiarism detection is suggested in this paper. This method accepts a video as input. We extract frames at a rate of 0.33 per second using the tesseract tool. OCR methods are used to extract textual information from the frames. Then, plagiarism detection software searches Google for these extracted frames.

Keywords—Tesseract, Python, NLP, OCR, Image Processing

Department of Information Technology, Vishwakarma Institute of Technology

[1]aparna.mete20@vit.edu, [2]prithviraj.sangle20@vit.edu, [3]riddhi.karandikar20@vit.edu, [4]parth.patil20@vit.edu, [5]janhavi.pawar20@vit.edu, [6]parul.patle20@vit.edu

DOI:10.31838/ecb/2023.12.s1-B.381

I. INTRODUCTION

In today's digital world, there are millions of videos being uploaded to video-sharing platforms whilst social media-sharing platforms are expanding rapidly. The amount of data on the internet is growing exponentially. Today, the best estimates suggest that at least 2.5 quintillion bytes of data are produced every day, and there is no reliable way determine how much of the data on the internet is genuine. Since the said amount of data is being produced every day, number of these videos uploaded may be plagiarized or reuploaded. Illegally copied content on the Internet and in multimedia technologies is a serious social issue. Because the internet is available to everyone, it is simple to download and re-upload content. Plagiarism can be committed by copying videos from the internet. The issue is determining how to identify plagiarized/copyright-infringing videos on the internet.

II. LITERATURE REVIEW

G. Eason, B. Noble, and I. N. Sneddon in this paper [1], detects copyright infringement on YouTube Videos using YouTube Metadata., April 02, 2013. Today, YouTube is used by millions of individuals worldwide to upload their own videos. Even though it's against YouTube and Copyright regulations, it happens occasionally. In this study, they outline a method for identifying original and illegal YouTube videos. The proposed method makes use of rule-based categorization, and when it is applied to the test data set, it demonstrates that the method is successful and efficient in identifying copyright violations in YouTube videos. They concluded in this experiment that the number of subscribers and hits can be utilized to categorize videos that violate copyright. User profiles can help identify channels that are being infringed, but not the original, genuine channels. The conclusion further suggests that classifying original vs. violated videos is a time-efficient method if irrelevant films are removed from search results. Using two open-source APIs (YouTube and Google Custom Search), we ran experiments using publicly accessible real-world datasets and assessed the overall performance of the suggested approach. The findings (measured in the form of a confusion matrix) show that contextual information about YouTube videos and users can be used to distinguish legitimate and illegal videos with a reasonable degree of accuracy. J. Clerk Maxwell employs a content-based video copy detection method using a discrete wavelet transform in the paper from 2013 (p. Since copyright issues continue to pose difficulties for the multimedia business, video copy detection is a study area that is currently active. Here, a discrete wavelet transform-based method for content-based video copy detection is described. It is possible to extract feature descriptions from video frames using the Daubechies wavelet transform. In this study, the computation needed for similarity search is also decreased. This article presents a discrete wavelet transform-based method for content-based video copy detection. The Daubechies wavelet transform (Db4) is used to divide the video frames into three layers and produce feature descriptors. The amount of computation required for the similarity search is reduced by first locating the related video and then the original portion of the copy. As a result, there is no need to explore the entire feature database. A detection rate of 60% and MAP of 0.38 have been achieved in this study. The method yields outcomes that are superior to those of the competing methods considered in this article. I. S. Jacobs and C. P. Bean, in the paper [3] video copy detection using Spatio-Temporal CNN Features, 2019. We have provided a method for detecting video copies that are based on spatial and temporal CNN features. The CNN features that are recovered from the sparsely

sampled video frames have a high degree of compactness and are capable of describing the temporal and spatial properties of films, which results in a high degree of discriminability. The results of the experiments demonstrate that the suggested spatiotemporal CNN features may be used to detect video copies with great performance in both efficiency and effectiveness. Learning Spatial-Temporal Features for Video Copy Detection by the Combination of CNN and RNN is described by K. Elissa in her article [4]. Based on spatial and temporal CNN features, we have presented a technique for identifying video duplicates. With a high degree of compactness and the ability to describe the temporal and spatial characteristics of films, the CNN features that are reconstructed from the sparsely sampled video frames have a high degree of discriminability. The findings of the experiments show that it is possible to detect video copies with excellent performance in both efficiency and effectiveness using the proposed spatiotemporal CNN features. [5]

Detecting video copies by quickly looking for inverted files. They recommend an efficient searching technique using inverted files for a fingerprint-based video duplicate detection system, where fingerprint matching only requires simple table lookup and word-counting operations. Because the video fragment matching is based on matched fingerprint counting and fingerprint order, the recommended searching method avoids the computationally intensive Hamming distance calculation. In order to enhance similarity matching with pixel correlation, we introduced the rash method, which creates fingerprints using the local mean and subregion. This novel fingerprint-generating technique reduced the search time of the proposed fast searching method while keeping the low computational cost and dependability of the aHash method. The proposed fast searching algorithm and fingerprint generation technique were evaluated and compared with other state-of-the-art algorithms and methods on an experimental video copy detection system using the TRECVID 2011 and the VCDB datasets. The suggested fast searching algorithm and the rHash method provide better accuracy and searching speed for video copy detection when compared to the conventional inverted file-based searching approaches. There are no high-level features acquired by rHash. The existing approach can therefore only manage copied videos with limited spatial similarity. High-level video fingerprint techniques can, however, be incorporated into the proposed system to improve it. As a result, the video copy detection system will be able to identify cloned videos that share comparable abstract meaning in addition to similar spatial similarities.

III. METHODOLOGY

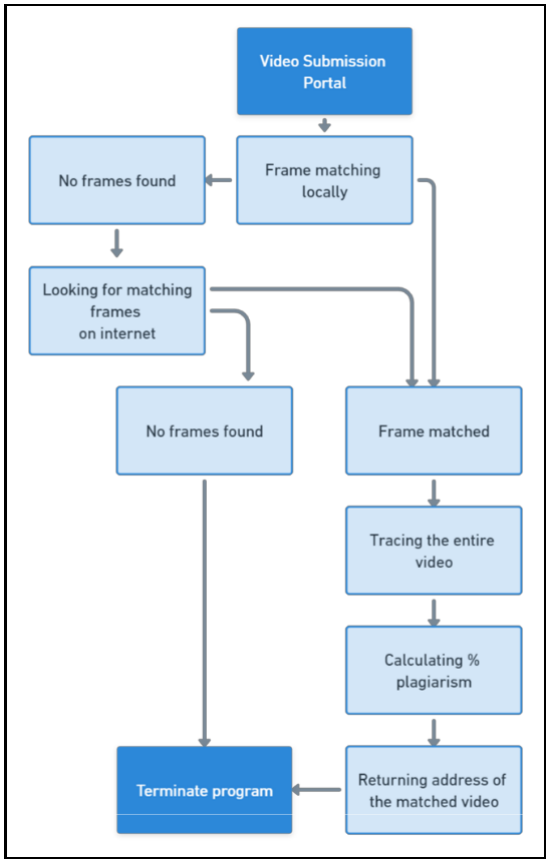

Fig. 1. Proposed Diagram

The complete process of creating our Model is covered in this section. The dataset is chosen, the data is pre-processed, the model is designed, trained, and its performance is evaluated using the inferences derived from the study of related literature.

The methodology of the project consists of four steps –

1. Data Extraction
2. Data Classification
3. Data Manipulation and Searching
4. Video Plagiarism Detection

a) Data Extraction The process of gathering or retrieving various kinds of data from various sources, many of which may be erratically organized or entirely unstructured, is known as data extraction. Data extraction makes it possible to combine, process, and refine data so that it can be stored in one place and changed.

For detection purposes three types of data have been extracted:

1) Image data: It is the frames extracted at the frequency of 3 frames/sec from a video input.

2) Textual data: Any type of text that might be present in the image data.

3) Captioned data: The entire captions of the video are drawn out by converting video to audio, then audio to text. All three types of data are stored in a CSV file for further manipulation and searching.

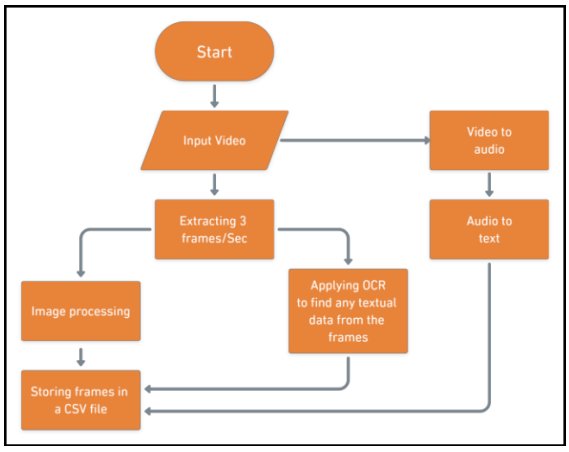

Fig. 2. Method of Data Extraction

b) Data Classification A general definition of data classification states that it is the process of grouping data into suitable categories so that it can be used and safeguarded more efficiently. The classification process essentially makes it easier to locate and retrieve data. Data classification is crucial for risk management, compliance, and data protection. Data classification entails categorizing information to make it simple to track and search for. Additionally, it gets rid of duplicate copies of the data, which can lower storage and backup costs while accelerating search times.

Data classification consists of –

- Processed frame Images: Image size is reduced for space efficiency whilst retrieving the quality of the image.

- OCR Textual Data: Might Contain relevant information about the source of the video such as website name, creator name, video information, platform name, movie, series, etc.

- Subtitles from video to text: This string data will be containing information regarding the video.

c) Data Manipulation and Searching The entire searching and matching of the videos for piracy detection are done using two different methods. Searching and finding plagiarism locally on the model. Searching for pirated videos globally on the internet.

1) Local Approach

Preparing Training and Testing Data - For preparing the training data and testing data we extract the frames from the videos in the train and test dataset. Because we can't feed the entire video to the sequence model. After the data is prepared the entire data is stored in train.csv and test.csv files We then feed the videos to the network to resize the frames as the frames are not of the same size so we resize them and store all the frames in the numpy array.

a) Feature Extraction

The feature extraction process is completed after the videos are resized and saved in a numpy array. Deep learning and machine learning feature extraction. The process of turning raw data into numerical features that can be processed while keeping the information in the original data collection is known as feature extraction. Compared to using machine learning on the raw data directly, it produces superior results.

b) Label Encoding

The labels, are encoded for the movies' similar and unidentified data. Label encoding is the process of transforming labels into a numeric shape so that computers can read them. The operation of those labels can then be better determined by machine learning techniques. It is a crucial supervised learning preprocessing stage for the structured dataset.

c) Sequence model

To analyze how much percentage of frames of videos are copied and how much is original we use sequence modeling. We used the CNN model keeping the epoch value to 30 and calculated the test accuracy. To calculate the percentage of unknown and similar data we matched the frames of all train datasets with the test dataset.

d) Data Source

There are two types of datasets we created to train and test datasets. The test dataset consists of videos which are original videos and the training dataset consists of all the plagiarized or copied videos.

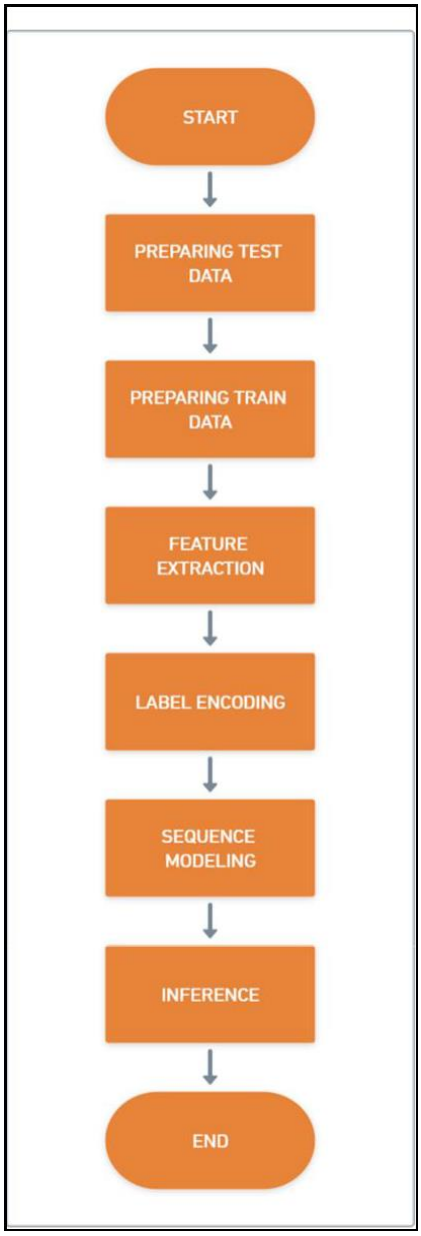

Fig. 3. Data Classification- Local Approach

2) Global Approach

The fight against digital piracy has been a continuous battle for content creators, copyright holders, and distribution companies. With the widespread use of the Internet and the ease with which digital content can be shared and distributed, the threat of digital piracy has become increasingly prevalent. To combat this, different tools and techniques are used to search the Internet for pirated video of the original content.

a) Using Google Image Search

One of the most common techniques used in the search for pirated videos is using Google Image Search. This technique involves creating a custom image search engine that utilizes Google's image search capabilities to search for processed video frames on the Internet. The search engine is designed to match image-to-image or image-to-video. The URLs of the top ten results of the search are then returned. The idea behind this technique is that the processed video frames, or individual frames taken from the video, are unique enough to match against other instances of the same frame found on the Internet. This method can be useful in identifying pirated

videos that have been posted online, as the same frames from the original video will appear in the pirated version.

b) Searching Textual data

The utilization of textual data in the search for pirated videos is a simple and straightforward method. By utilizing the information obtained from the images captured in the video, such as the title of the video or the name of the creator, this method can be highly effective in locating pirated videos. The retention of the original title in pirated videos makes searching for the title a quick and efficient way to locate such videos. It's important to note, however, that not all pirated videos will retain the original title, so this technique may not be as effective in these cases. Despite this, the use of textual data remains valuable as it provides the opportunity to uncover the video's title and make it easier to find. In essence, this type of search leverages the power of the extracted textual data to help identify pirated videos on the Internet.

c) Applying NLP to Textual data

The use of natural language processing (NLP) in the hunt for pirated videos is a more sophisticated method that has grown in favor in recent years. This involves using NLP tools, such as NLTK, to extract meta tags, titles, and nouns from the textual data associated with the original video. The extracted information is then stored as separate strings, which can be later searched on the web using a Python script. The purpose of using NLP in the search for pirated videos is to gain a more accurate and relevant result by extracting nouns, verbs, and adjectives from the textual data. This method can be more effective than traditional text-based searching methods, as it takes into account the context of the textual data and the relationships between different words and phrases.

Fig. 4. Text Extraction using NLP

The results of the three procedures are used to assess whether or not a video is violating any copyrights. The local and global models' appropriate outcomes are examined for plagiarized content. Different video hosting services, like YouTube, Netflix, and Amazon Prime, have their own copyright policies and only permit a certain proportion of plagiarism when utilizing their content. The video frames from the test dataset and the training dataset are included. As a result, every frame from the training dataset is compared to the training dataset's frames. The number of frames that match in each dataset is counted in order to determine how much of the movie is plagiarized.

The dataset is divided into two parts:

1) Testing data

2) Training data

The percentage of the video plagiarized can be found out by: *Plagiarism = (Frames Matched / Total No of frames)*100*

By copying the actual film, training and validation were conducted for videos. The duplicate video was then combined with a different, original video.

The original video and the bigger videos' frames were taken. The bigger video was then tested by utilizing CNN to compare the smaller and larger videos. A simple percentage calculation was used to quantify the proportion of plagiarism after instances of matching were recorded

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

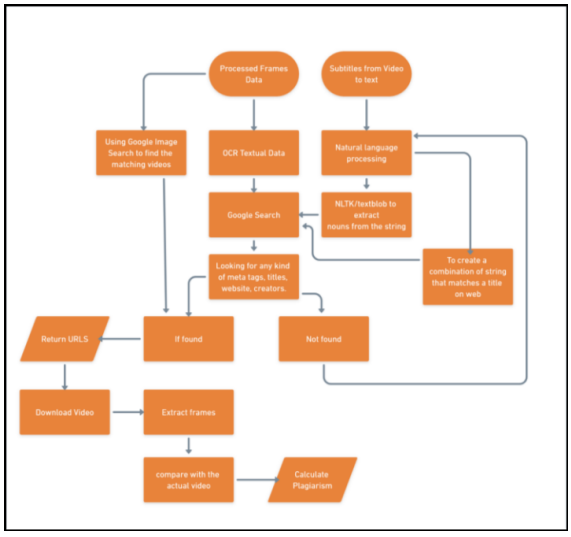

Fig. 5. Method to find copyright video

d) Video Plagiarism Detection

IV. RESULTS

```
Test video path: dataset/test/2.mp4
1/1 [==============================] - 2s 2s/step
  SIMILAR: 51.50%
  [UNK]: 48.50%
```

Fig. 6. Results

It was found after substantial model training under supervised learning utilising the global technique.

We reached some findings when the frames coincided. Only 48.50% of the video content was judged to be original, with 51.50% of it being copied.

Since the ratio of the two videos' lengths was 1:1, the projected similarity was 60%, however only 51.5% of the original test video's content was preserved.

The accuracy of the plagiarism-calculating model was 85.83%.The accuracy of the

model will undoubtedly rise to over 90% with the addition of more data.

V. CONCLUSION

After utilizing numerous techniques to locate a film that violated copyrights, this research was able to locally identify plagiarism in the video. Because using the custom search engine requires substantial training, using web scraping for this purpose is currently ineffective.

Copyright detection and content theft are significant social problems that call for practical solutions. We encourage additional study and project extension.

VI. SCOPE OF RESEARCH

The global issue of content piracy, which results in significant losses for the authors, may be solved with the successful completion of this project. Finding and blacklisting websites and authors who violate copyrights is necessary to stop copyright infringement. This issue can be solved effectively by using cutting-edge techniques like web scraping, natural language processing, and a skilled search engine.

VII. FUTURE SCOPE

To create a public use web gateway for this video plagiarism detection system. Use more accurate picture search engines as well, as the ones currently in use are not adequately trained to provide results that are similar to the input image frames. to extract the audio-to-text data, arrange the text words into a random string, and then use it to search the internet for video titles or the name and URL of the video content creator. By doing this, we can get better plagiarism-checking results.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil.Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.