IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 6, PP 70-74

Content-Based Real Time Video Copy Detection Using Hadoop

Pramodini Kamble[1], Priyanka Shejawal[2], Tejaswi Surve[3],
Kushal Yadav[4], Prof. Bhushan Patil[5]
[1,2,3,4]*(UG student, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai India)*
[5]*(Professor, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai India)*

Abstract : *Due to emerging interest in videos along with the evolution of new technology and facilities, today the amount of videos on the internet is increasing enormously. But these videos do not hold original content and are simply revised versions of the original videos. The purpose of the Video Content Copy Detection is to detect the copied videos and find the similarities among them. The currently available video copy detection algorithms have certain limitations, such as it requires more time for searching and gives inaccurate output. TIRI-DCT algorithm along with Brightness Sequence algorithm is implemented to overcome the problems in the existing system. For the significant improvement in the processing of videos we use Hadoop platform in distributed file system and for vital practical significance in video tracking and real time video content retrieval.*

Keywords: *Hadoop, brightness sequence, cluster – based similarity search, inverted file-based similarity search and TIRI - DCT*

I. Introduction

From the last ten years, the large number of video contents produced, stored, distributed, and broadcasted digitally and also has grown extensively. Video is a kind of media which provides an individual with great insight of knowledge in various fields.

Due to the propagation of videos has made accessibility of video contents much easier and cheaper. There are many sources, e.g., the illegal distribution of copyright movies via file sharing services on the Internet.

The massive capacity of these services makes the tracing of video content into a very hard problem for video professionals. Now a day increasing problems facing professionals those are copyrights infringement and data piracy. The problems associated with such problems digital videos require an efficient method for protecting, managing and indexing video contents. According to statics 27% redundant videos are duplicated on the most popular version of a video in the search results from video search engines. Therefore, an effective and efficient method for video copy detection has become more and more important. Also, users are often frustrated when they need to spend their valuable time to find videos of interest; they have to go through number of duplicated or nearly duplicated videos that are streamed over the internet before arriving at an interesting one. It increasing burden on the network like bandwidth, storage space, traffic and users time can be wasted.

Hence, Video Copy Detection process comes to differentiate between original and duplicate videos. Video copy detection is useful to find out same contents between the videos, i.e., to criticize whether the two videos are from the original content. This can be achieved by calculating the hash values of the contents present in the videos. Hadoop has advantages of high fault tolerance, high throughput, easy scalability and etc.

Video Copy detection is divided into two parts First, the features of the video are extracted and a hash library (Hadoop distributed platform is applied to calculate the hash values) of the videos is formed; secondly, the features of the querying video are extracted and a hash value of the querying video is formed and then it is compared with the hash values of the training videos, so that whether the query is a copy video or not can be determined.

As there are so many videos are present in the internet, it is very difficult to execute the video copy detection process on only single machine approach because it is very furious process. Even complex calculations are existing in this process so for that distributed approach will bring the significant result comparatively that of the single machine approach.

1.1 Introduction to Hadoop Platform

Hadoop was introduced by the Apache Foundation. It is tool that provides infrastructure needed for distributed storage and distributed processing. A MapReduce process is called a job. It includes MapReduce programs, it breaks data processing into two phases Map phase and Reduce Phase. Hadoop Mapper generates large chunk of intermediate data that is passed to Hadoop Reducer for further processing. Now Reducer involves finding distinct elements from massive data string. Sorting and grouping at the reducer side happens at the basis of the key from the output of the mapper. Hadoop has advantages of high fault tolerance, high throughput, easy scalability and etc.

HDFS is the distributed file storage system of Hadoop. It consists of a NameNode and several DataNode. NameNode is responsible for managing file system, and DataNode is responsible for storing files. Through setting the number of backup files, the file storage can be made more reliable. Pig is a high-level query language built on MapReduce, which can simplify the development process of MapReduce. Hive is a data warehouse tools on Hadoop, which can provide SQL queries. HBase is a distributed database, which is based on the column storage model. ZooKeeper provides coordination services in distributed systems on Hadoop.

II. Literature Review

We have studied that if there is no watermark present in the original video then it is impossible to detect the duplicate video based on watermarking techniques [7].

This paper introduced an approach for fast and parallel video processing on MapReduce-based clusters such as Apache Hadoop. By using clusters, the approach is able to handle large-scale of video data and the running time can be significantly reduced [5].

In this paper, the efficiency of Hadoop platform is also focused which has high processing speed and also makes use of HBase which is optimum for storing and retrieval of data. We also studied about optimization of both the techniques i.e. TIRI-DCT and Brightness Sequence Algorithm, their correctness, speed and efficiency are analysed. [2].

We studied about the TIRI DCT algorithm which is based on temporally informative representative Images. Temporally Informative Representative Images (TIRI) - Discrete Cosine Transform (DCT) Extracts compact content- based signatures from Special images constructed from the video [4].

In this paper, we evaluate the performance of Inverted File based search & Cluster based search algorithm and compare search method. It can be seen that proposed Cluster based approach is faster than that the inverted file search method. Thus, the cluster-based algorithm is selected as the search engine for our copy detection system to obtain the secure version of fingerprinting algorithm. It greatly speeds up the search process as well as improves the retrieval accuracy. It also maintains a good performance for different attacks on video signals, frame loss, changes in brightness/contrast, rotation, spatial/temporal shift, including noise addition. [3].

We learned about fingerprinting technology and application of video copy detection system in managing copyrighted content, video tracking, identifying. This paper has done considerations, such as robustness, compactness and discriminability for finger printing methods. It also discusses fingerprint matching technologies complexities [8].

A simple exhaustive search method which has a complexity of, where is the number of the fingerprints in the database. Search methods inverted file-Based Similarity Search is modified version of search algorithm that was proposed [3].

III. Content Based Similarities

It is another method for identification of images and video clips. It is based on idea of "the media itself is the watermark," i.e., the media like video, audio, image contains. enough unique information for detecting copies. Signature is extracted from the trained media and stored in a database. The same procedure of signature extraction is done for query media and compared to the trained media signature which is already extracted and stored in a database to determine if the test stream contains a copy of the query media. The signatures which are extracted are called as "Fingerprint".

3.1 Fingerprinting

Fingerprints is used to summarize a video signal. These signature features vectors that uniquely characterize the specific signal. Video fingerprinting is a technique that can be used for content-based copy detection. The major task is to detect whether a particular part of the video is based on the same original video in a database of reference video.

3.1.1 Temporal Fingerprints

This technique extracted from video sequence characteristics all the time. It always performs good with huge video sequences. But due to insufficient discriminant temporal information, for short video clips it does not work well because short video clips contain a large amount of share of online video databases, temporal fingerprints alone do not suit online applications.

3.1.2 Spatial Fingerprints

This algorithm is useful for conversion of a video image into YUV colour space among these the chrominance components (U, V) are discarded and the luminance (Y) component is kept. Spatial fingerprints are unique value which derived from each frame or from a key frame. Spatial fingerprints can be classified into global and local fingerprints. Global fingerprints look globally to check global properties of key frame. (e.g. Image histograms) However local fingerprints use to check local information. (e.g. Edges, corners)

3.1.3 Spatio-Temporal Fingerprints

One disadvantage of spatial fingerprints is that they are not able to capture the video's temporal data, which is an important differentiating factor.

IV. Proposed System

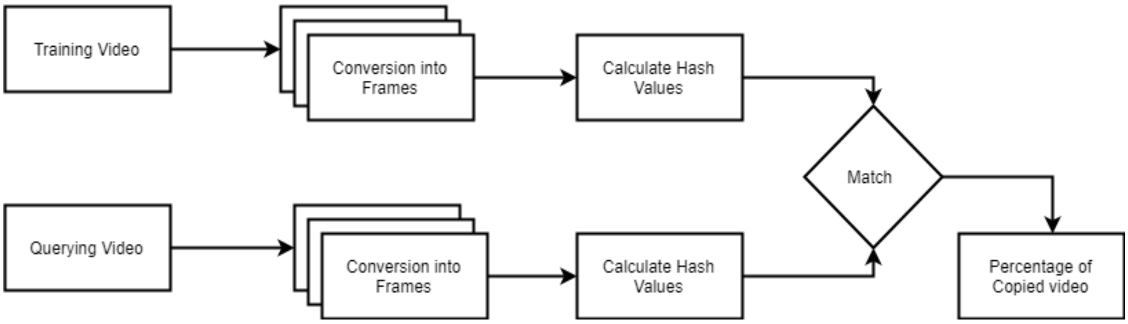

Fig 4.1: Block Diagram of Proposed System

The Training video i.e., the original video is first converted into the number of frames. After the creation of number of frames and then hash value for the frames is calculated and stored in HDFS. Same process for querying video i.e., for the duplicate video is done and stored in the HDFS. These two videos are then compared with each other and if the % of content matched is 100 then the video is said to be fully copied video else the video is said to be partially copied. The comparison between these two videos is done with the help of two algorithms known as TIRI-DCT and Brightness Sequence Algorithms.

V. Algorithms

We have studied following algorithms related to video copy detection which are based on content-based copy detection approach.

5.1 Existing Algorithm

5.1.1 TIRI DCT

Temporally informative representative images-discrete cosine transforms (TIRI-DCT) is based on spatial and temporal information of a video sequence. Representative images are generated based on the weighted average of the frames. This sequence contains the temporal as well as spatial information. The frames are then divided into the blocks and then the horizontal and vertical features are being extracted from these blocks. The feature vector is created based on the concatenation of the values of the features extracted. Features are compared to threshold (median of feature value) and then binary sequence fingerprint is generated. In order to determine whether a query video is an attacked version of a video in a database or not, it's fingerprint is first extracted. The fingerprint database (previously created from the videos in the video database) is then searched for the closest fingerprint to the extracted query fingerprint [3].

5.1.2 Brightness Sequence

We must calculate the brightness of each frames of querying video and at the same time of instance we must keep comparing it with the brightness of training video i.e., original one. If we get the brightness value is almost similar then we can consider that it is a copied video [1].

5.2 Proposed Algorithm

The above two algorithms i.e., TIRI-DCT and brightness sequence would run parallel to get the efficient output:

5.2.1 Hash value extraction:

For a video sequence X = x (1), x (2), x (3) …… x(n), N is the frame number of the videos. The hash sequence corresponding to Video X is H = (f(x1), f(x2) …... f(x{n})), f(x) = (a (1), a (2), a (3) …. a(m)) is the hash extracting function, M is the number of hash bits extracted from a single frame.

Hash value extraction algorithm is:
1) The video frame x(i) is divided into w*h areas
2) Calculate the average brightness of each block
3) Sort the brightness of each block
4) Now I = I +1, repeat step1, end the process until I = N.

5.2.2 Hash value matching:

For video sequences X = x (1), x (2), x (3), …… x(n), and Y = y (1), y (2), y (3), …… y(n), the hash sequences corresponding to videos X and Y are H(x) = [f(x1), f(x2), …. f(x{n}) and H(y) = [f(y1), f(y2) …. f(y{n}) respectively, wherein f(x) = [a (1), a (2), a (3) …. A(m)] and f(y) = [b (1), b (2), b (3) …. B(m)].

The distance between the two video sequences is calculated i.e., D (X, Y). When the distance D (X, Y) =T, the two videos are regarded as the same video. T is the predetermined threshold.

VI. Searching Technique

6.1 Inverted file-based similarity search:

Step 1: Firstly, the fingerprints which are in the form of binary unique values are divided into n words each of same bits.
Step 2: The representation of the word position is done by the horizontal length of the table.
Step 3: The representation of the possible values of the words is done by the vertical length of the table.
Step 4: For entry in column corresponding to the value assigned to the word an Index is added to each word of the fingerprints
Step 5: The hamming distance is calculated for the fingerprints in the database and the query fingerprints.
Step 6: A threshold value is fixed and then compared with the calculated distance. If calculated distance is less than the threshold value the query video is matched and is declared as matching.
Step 7: Else the results will be videos not matched

6.2 Cluster based similarity search:

- This is another algorithm for implementing binary fingerprints.
- Cluster based similarity search is used to make use of the clustering to decrease the number of queries that are analyzed within the database.
- If only one cluster has all the fingerprints, then those fingerprints in the database will be clustered into non-overlapping groups.
- To perform the same, for each cluster a centroid is chosen and named as cluster head. The cluster will be assigned a fingerprint if it is closest to this cluster's head.
- The cluster head closest to the query is found to determine if the query fingerprint matches a fingerprint in the database.
- The fingerprints which belong to this cluster are searched to find the match i.e., the one which has the minimum Hamming distance (of less than a certain threshold) from the query. If match is not found, then the cluster which is the next closest to the query is examined.

VII. Conclusion

Proposed system would remove the burden on the internet which has the multiple set of revised copy of original videos. Our system will make an ease for the user to get the video of interest very quick instead of going through numerous videos. The system works on Hadoop platform using both the techniques TIRI-DCT and Brightness sequence which focus on high processing speed and optimize the retrieval and storage of data.

References

[1] Jing Li, Xuquan Lian, Qiang Wu and Jiande Sun "*Real time Video Copy Detection Based on Hadoop," Sixth International Conference on Information Science and Technology Dalian, China; May 6-8, 2016.*

[2] Abrum Jeysudha, Samit Shivadekar, "Real Time Video Copy Detection using Hadoop", International Journal of Computer Application 2017.

[3] Ms. Laxmi Gupta, Prof. M.B Limkar, "*Search Methods for Fast Matching of Video Fingerprints within a Large Database", by SSRG International Journal of Electronics and Communication Engineering (SSRG- IJECE) – volume1 issue 3 May 2014.*

[4] Vaishali V. Sarbhukan, Prof. V.B.Gaikwad ,"*Video Fingerprinting Extraction Using TIRI-DCT" , International Journal of Engineering Research and Development 2013*

[5] Hanlin Tan, Lidong Chen," *An Approach For Fast And Parallel Video Processing On Apache Hadoop Clusters" IEEE International Conference on Multimedia and Expo 2014.*

[6] M. M. Esmaeili, M. Fatourechi, and R. K. Ward. ”*A robust and fast video copy detection system using content-based fingerprinting," IEEE Transactions on Information Forensics and Security, vol. 6, no. l, pp. 213-226, 2011.*

[7] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, *Video copy detection: A comparative study, in Proc. Conf. Image Video Retrieval (CIVR), 2007.*

[8] Jian Lu, "*Video fingerprinting for copy identification: from research to industry applications", Proceedings of SPIE - Media Forensics and Security XI, Vol. 7254, January 2009.*