Real Time Video Copy Detection using Hadoop

Abrum Jeysudha
Department of Information Technology
S.I.E.S. Graduate School of Technology,
Nerul, Navi-Mumbai - 400 706

Lavanya Muthukutty
Department of Information Technology
S.I.E.S. Graduate School of Technology,
Nerul, Navi-Mumbai - 400 706

Ashwati Krishnan
Department of Information Technology
S.I.E.S. Graduate School of Technology,
Nerul, Navi-Mumbai - 400 706

Samit Shivadekar
Professor
Department of Information Technology
S.I.E.S. Graduate School of Technology,
Nerul, Navi-Mumbai - 400 706

ABSTRACT

Due to emerging interest in videos, there are various sites which provides with different kinds of videos but it is not necessary that every video hold original content. Video Copy Detection process comes into picture to differentiate between original and duplicate videos. Video Copy Detection basically deals with finding out similarities between the content of two given videos. Hadoop is a distributed platform which makes use of MapReduce programming model. It has two phases i.e. Mapping and Reducing phase. Brightness Sequence algorithm along with TIRI-DCT algorithm is implemented to overcome the problems in the existing system. OCR is used in order to detect the copied videos based on subtitles or any other form of text present in the video. The framegrabber(), which is a JAVA method, is used to convert the videos into multiple frames at different time instincts.

Keywords

Video copy, TIRI-DCT, Brightness sequence, OCR, training video, querying video, Hadoop, MapReduce, hash, plagiarism, HDFS, FFMPEG, frames, copied video.

1. INTRODUCTION

In this era of 21st century, media has become an important part of everyone's day-to-day life. It connects the population with the scenarios in the world and informs people with things like news, history, entertainment etc. which helps for an upgraded personality in humans.

Though there are various kinds of media, video stands ahead of all, in various aspects. Video is a kind of media which provides an individual with great insight of knowledge in various fields. Due to emerging interest in videos, there are various sites which provides with different kinds of videos but it is not necessary that every video hold original content. This result into videos with similar content as that of the original video and such videos may be referred to as duplicate videos. Hence, Video Copy Detection process comes into picture to differentiate between original and duplicate videos.

Video Copy Detection basically deals with finding out similarities between the content of two given videos, hence judging whether either of the video is original or not [5]. This can be done by calculating the hash values of the content present in the videos by using appropriate algorithms. Due to billions of videos present in the internet, it is not possible to perform the video copy detection process on a single machine approach as it is a hectic process. Due to huge amount of calculation present in this process, a distributed computing approach will fetch efficient result as compared to that of single machine approach. This is because calculation is distributed to each computer present in the respective distributing system. To store and process large amount of data, Hadoop which is an open source and Java based programming language, plays an important role as it works on distributed environment. In Hadoop, command line utilities are written using shell script. Hadoop consists of the Hadoop Common package, a MapReduce Engine and the Hadoop Distributed File System (HDFS). In this paper, the Hadoop distributed platform is used to calculate the hash values of a large number of videos and then matching the hash value of the given video with the hash value of every video present in the HDFS i.e. it follows one-to-many cardinality property. In order to accomplish this, two Video Copy Detection algorithms are used in this context. Out of two, one algorithm is based on brightness sequence and other is TIRI-DCT Algorithm [6].

The existing system of video copy detection needs more information and the deeper analysis of the video. It mainly uses computer vision algorithms, hence it becomes difficult to detect the similarities between the content of videos if the information has to be stored in a system with limited memory.

One of the major approaches in existing system which is being used is watermarking technique. In this approach, invisible signal is added into the videos. During the detection, the videos are converted into various images which have watermark on it, which helps in the detection process. The limitation of this technique is if the original video is not watermarked, it becomes impossible to know whether the video is copied or not.

Due to this limitation of watermarking, content-based approach came into picture. In this approach, various tracing features are extracted from the video and are matched with the existing features present in the database. If the features match, the video is said to be copied. But, the system tends to get confused if two entirely different videos have somewhat similar features [3].

1.1 Introduction to Hadoop

Hadoop is a distributed platform which makes use of MapReduce programming model. It has two phases i.e. Mapping and Reducing phase. The ultimate use of these phases is to store unstructured data in the form of key-value pairs in HDFS. In mapping, the input is converted into various key-value pairs. This intermediate key-value pair is then sent to the Reduce function where it applies reduce algorithms to store it in HDFS in a logical key-value pairs.

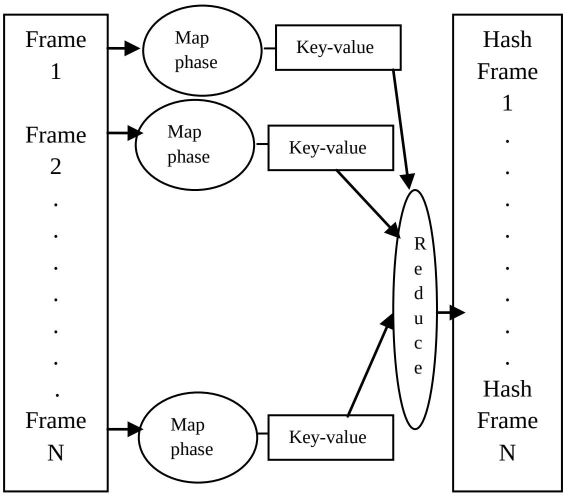

Figure 1: MapReduce Programming Model

2. PROPOSED SYSTEM

In the proposed system, different methods to find out the plagiarised videos are used. Hadoop platform is preferred for this system because it implements MapReduce programming model which provides parallel processing of huge volume of data. Brightness Sequence Algorithm along with TIRI-DCT Algorithm is implemented to overcome the problems in the existing system. Beside these, OCR is used in order to detect the copied videos based on subtitles or any other form of text present in the video [4].

3. SYSTEM ARCHITECTURE

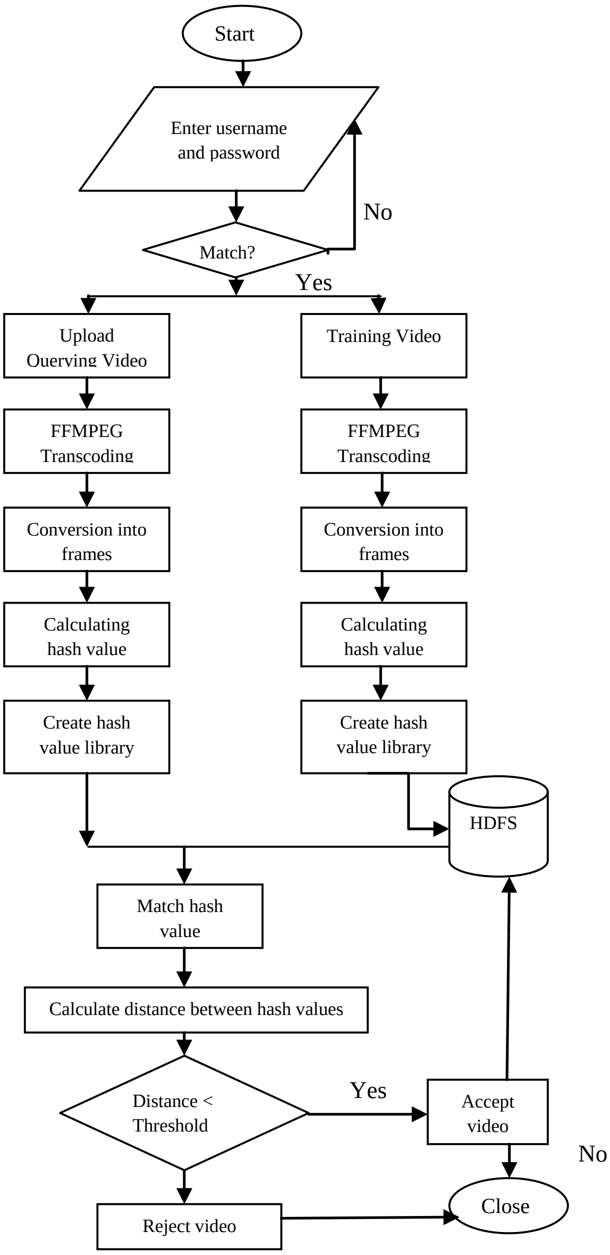

Figure 2: Flowchart

The admin can login into the system with given username and password stored in the database. If the password matches, the admin is authorized and given access to the system.

The original videos, which are also known as training videos, are transcoded using FFMPEG. Since Hadoop doesn't support FLV format videos, it has to be converted to different format if in case the admin upload the videos in FLV format.

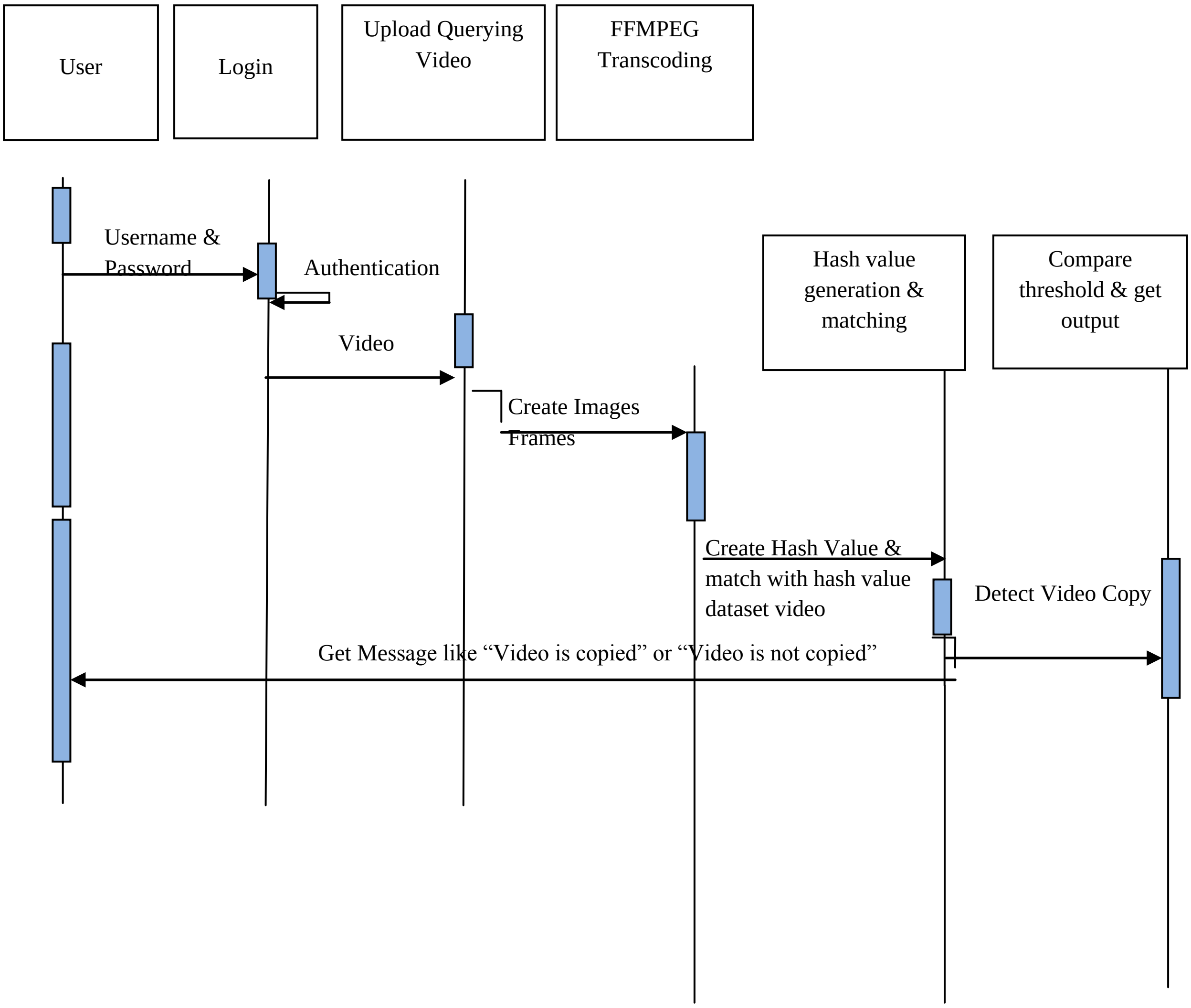
User
Login
Upload Querying Video
FFMPEG Transcoding
Username & Password
Authentication
Hash value generation & matching
Compare threshold & get output
Video
Create Images Frames
Create Hash Value & match with hash value dataset video
Detect Video Copy
Get Message like "Video is copied" or "Video is not copied"

Figure 3: Sequence Diagram of Video Copy Detection

The first method of finding the copied videos is by using TIRI-DCT algorithm [7]. In this approach, at each instant of time of a particular interval, a frame is drawn. The features of that particular frame at that instant of time are extracted. The extracted characteristics are then compared with the characteristics of the original videos present in the database. If the features of both the videos are almost same, then the video is said to be copied. For example, suppose a particular video has four frames. If the feature of one frame matches with the original video, it is said to be 25 % copied video. If two frame matches, it is said to have 50% originality and so on.

The second method is Brightness Sequence Algorithm. In this approach, the brightness of each frame is calculated of the video whose originality is to be found. This brightness values are compared with the values of the brightness of the original video at the same instant of time. If the brightness of both the videos is almost same, then it is said to be copied.

$$R_{x,y} = \sum^{J}_{k-1} 0.65^{k} * l_{k,x,y} \ldots\ldots\ldots [1]$$

where $R_{x,y}$ = Frame

$l_{k,x,y}$ = Brightness value of the kth frame at the location (x,y)

This algorithm is implemented on the frames extracted from the video in hand. The hash values of the same is calculated and then the distance between the hash values of querying video as well as the training video is determined which is then compared against a predefined threshold value. If the distance is greater than the predefined threshold value, then the two videos are considered as different videos else they are regarded as same video, hence the copied one.

In the following system, OCR (Optical Character Recognition) is used to retrieve the characters from the frames. The character retrieved is then arranged in some appropriate sequential order and compared. The characters can be subtitles present in the videos.

4. IMPLEMENTATION

In implementation phase, similarities and dissimilarities between the videos are found out using Hadoop platform and by making use of two phases, i.e. Map phase and Reduce phase. The two distinct algorithms which are being used are TIRI-DCT and Brightness Sequence Algorithm. The main aim of these algorithms is to convert the video into multiple frames, extracting their features and then finally comparing both querying as well as training video. The training video is already present in the HBase, while the querying video is taken as input from the user.

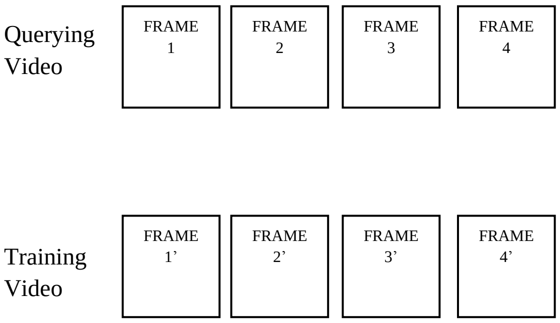

Figure 4: Comparison of Training and Querying Videos

In figure 4, consider that four frames are drawn from the video.

If FRAME 1 = FRAME 1', FRAME 2 = FRAME 2', FRAME 3 = FRAME 3' and FRAME 4 = FRAME 4', then querying videos is said to be 100% copied.

If FRAME 1 ≠ FRAME 1', FRAME 2 = FRAME 2', FRAME 3 = FRAME 3' and FRAME 4 = FRAME 4', then querying videos is said to be 75% copied i.e. if any of the one frame out of four matches with the training video, then it is said to be 75% copied.

If FRAME 1 ≠ FRAME 1', FRAME 2 ≠ FRAME 2', FRAME 3 = FRAME 3' and FRAME 4 = FRAME 4', then querying videos is said to be 50% copied i.e. if any of the two frames out of four matches with the training video, then it is said to be 50% copied.

If FRAME 1 ≠ FRAME 1', FRAME 2 ≠ FRAME 2', FRAME 3 ≠ FRAME 3' and FRAME 4 = FRAME 4', then querying videos is said to be 25% copied i.e. if any of the three frames out of four matches with the training video, then it is said to be 25% copied.

If FRAME 1 ≠ FRAME 1', FRAME 2 ≠ FRAME 2', FRAME 3 ≠ FRAME 3' and FRAME 4 ≠ FRAME 4', then querying videos is said to be 0% copied hence, it isn't a copied video.

5. FUTURE WORK

1. Time-feature graph can be plotted to show the features of the frame with respect to time. This can be done for both training and querying videos. If the graphs coincide then it is said to be copied video [2].

2. Instead of using visuals, frequencies of audio of both the videos can be used with moving time. If the frequencies are somewhat similar, it can be declared as copied video [8].

3. Histograms, edge and texture information can also be plotted to detect whether the video is original or not.

6. CONCLUSIONS

In this paper, optimization of both the techniques i.e. TIRI-DCT and Brightness Sequence Algorithm, their correctness, speed and efficiency are analyzed. 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.

7. 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] Chih-Yi Chiu, Cheng-Chih Yang and Chu-Song Chen "Efficient and Effective Video Copy Detection Based on Spatiotemporal Analysis," Ninth IEEE International Symposium on Multimedia 2007.

[3] Mani Malek Esmaeili, Mehrdad Fatourechi, and Rabab Kreidieh Ward "A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting," IEEE Transactions on Information Forensics and Security, VOL. 6, NO. 1, March 2011.

[4] Datong Chen*, Jean-Marc Odobez and Herv/e Bourlard "Text detection and recognition in images and video frames," D. Chen et al. / Pattern Recognition 37 (2004) 595 – 608.

[5] Shikui Wei, Yao Zhao, Ce Zhu, Changsheng Xu and Zhenfeng Zhu "Frame Fusion for Video Copy Detection," IEEE Transactions on Circuit and System for Video Technology, VOL. 21, NO. 1, January 2011.

[6] Nan Nan and Guizhong Liu "Video Copy Detection Based on Path Merging and Query Content Prediction," IEEE Transactions on Circuit and System for Video Technology, VOL. 25, NO. 10, October 2015.

[7] Suman Elizabeth Daniel and Binu A "An Exploration based on Multifarious Video Copy Detection Strategies," Proc. of Int. Conf. on Advances in Recent Technologies in Communication and Computing.

[8] Lezi Wang, Yuan Dong, Hongliang Bai, Jiwei Zhang , Chong Huang and Wei Liu "Contended-based large scale Web Audio Copy Detection," 2012 IEEE International Conference on Multimedia and Expo.