Real Time Video Copy Detection Based on Hadoop

Hardik Parmar Pranav Sangam Sanket Thakur Nilima Patil
Professor

Department of
Computer Engineering
K C College
of Engineering,
Thane (West) – 400603

ABSTRACT

Because of developing enthusiasm for recordings, there are different locales which furnishes with various types of recordings yet it isn't important that each video hold unique substance. Video Copy Detection process comes into picture to separate amongst unique and copy recordings. Video Copy Detection essentially manages discovering likenesses between the substance of two given recordings. Hadoop is a disseminated stage which makes utilization of MapReduce programming model. It has two stages i.e. Mapping and Reducing stage. Shine Sequence calculation alongside TIRI-DCT calculation is executed to conquer the issues in the current framework. OCR is utilized as a part of request to distinguish the duplicated recordings in view of subtitles or some other type of content present in the video. The framegrabber(), which is a JAVA strategy, is utilized to change over the recordings into numerous edges at various time senses.

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 period of 21st century, media has turned into a vital piece of everybody's everyday life. It interfaces the populace with the situations on the planet and educates individuals with things like news, history, excitement and so on which helps for an updated identity in people.

In spite of the fact that there are different sorts of media, video remains in front of all, in different angles. Video is a sort of media which furnishes a person with incredible understanding of learning in different fields. Because of developing enthusiasm for recordings, there are different destinations which furnishes with various types of recordings yet it isn't fundamental that each video hold unique substance. This outcome into recordings with comparable substance as that of the first video and such recordings might be alluded to as copy recordings. Consequently, Video Copy Detection process comes into picture to separate amongst unique and copy recordings.

Video Copy Detection essentially manages discovering likenesses between the substance of two given recordings, subsequently judging whether both of the video is unique or not [5]. This should be possible by ascertaining the hash estimations of the substance exhibit in the recordings by utilizing suitable calculations. Because of billions of recordings introduce in the web, it isn't conceivable to play out the video duplicate location process on a solitary machine approach as it is a riotous procedure. Because of immense measure of figuring present in this procedure, a dispersed processing methodology will get productive outcome when contrasted with that of single machine approach. This is on account of count is conveyed to every PC introduce in the particular dispersing framework. To store and process huge measure of information, Hadoop which is an open

source and Java based programming dialect, assumes an imperative part as it chips away at conveyed condition. In Hadoop, order line utilities are composed utilizing shell content. Hadoop comprises of the Hadoop Common bundle, a MapReduce Engine and the Hadoop Distributed File System (HDFS). In this paper, the Hadoop appropriated stage is utilized to figure the hash estimations of countless and after that coordinating the hash estimation of the given video with the hash estimation of each video show in the HDFS i.e. it tails one-to-numerous cardinality property. To achieve this, two Video Copy Detection calculations are utilized as a part of this unique circumstance. Out of two, one calculation depends on shine succession and other is TIRI-DCT Algorithm [6].

The current arrangement of video duplicate identification needs more data and the more profound examination of the video. It for the most part utilizes PC vision calculations, thus it winds up hard to distinguish the likenesses between the substance of recordings if the data must be put away in a framework with constrained memory.

One of the major methodologies in existing framework which is being utilized is watermarking strategy. In this approach, undetectable flag is included into the recordings. Amid the location, the recordings are changed over into different pictures which have watermark on it, which helps in the recognition procedure. The impediment of this procedure is if the first video isn't watermarked, it winds up difficult to know whether the video is duplicated or not.

Because of this impediment of watermarking, content-based approach came into picture. In this approach, different following highlights are separated from the video and are coordinated with the current highlights exhibit in the database. On the off chance that the highlights coordinate, the video is said to be replicated. Be that as it may, the framework has a tendency to get confounded if two completely unique recordings have to some degree comparative highlights [3].

1.1 Introduction to Hadoop

Hadoop is a disseminated stage which makes utilization of MapReduce programming model. It has two stages i.e. Mapping

and Reducing stage. A definitive utilization of these stages is to store unstructured information as key-esteem matches in HDFS. In mapping, the info is changed over into different key-esteem sets. This transitional key-esteem match is then sent to the Reduce

work where it applies decrease calculations to store it in HDFS in a legitimate key-esteem sets.

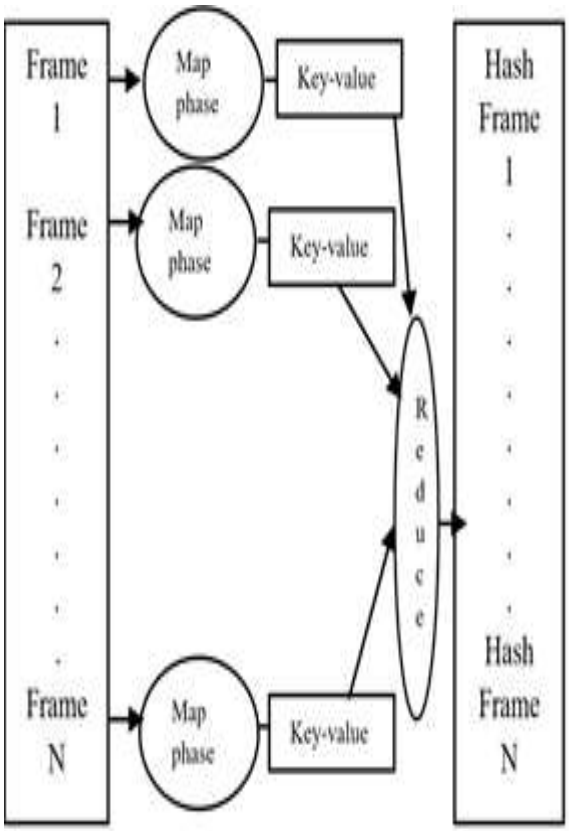


Figure 1: MapReduce Programming Model

2. PROPOSED SYSTEM

In the proposed framework, distinctive techniques to discover the appropriated recordings are utilized. Hadoop stage is favored for this framework since it actualizes MapReduce programming model which gives parallel preparing of enormous volume of information. Shine Sequence Algorithm alongside TIRI-DCT Algorithm is executed to beat the issues in the current framework. Close to these, OCR is utilized as a part of request to identify the replicated recordings in light of subtitles or some other type of content present in the video [4].

3. SYSTEM ARCHITECTURE

:

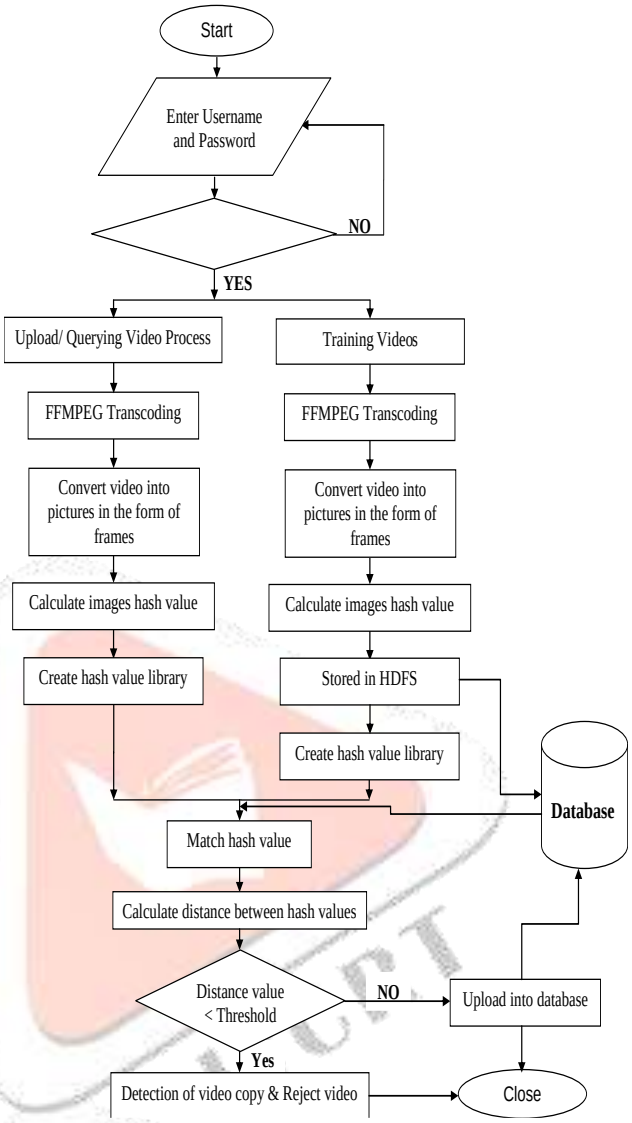


Figure 2 : Flowchart

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 uploads the videos in FLV format

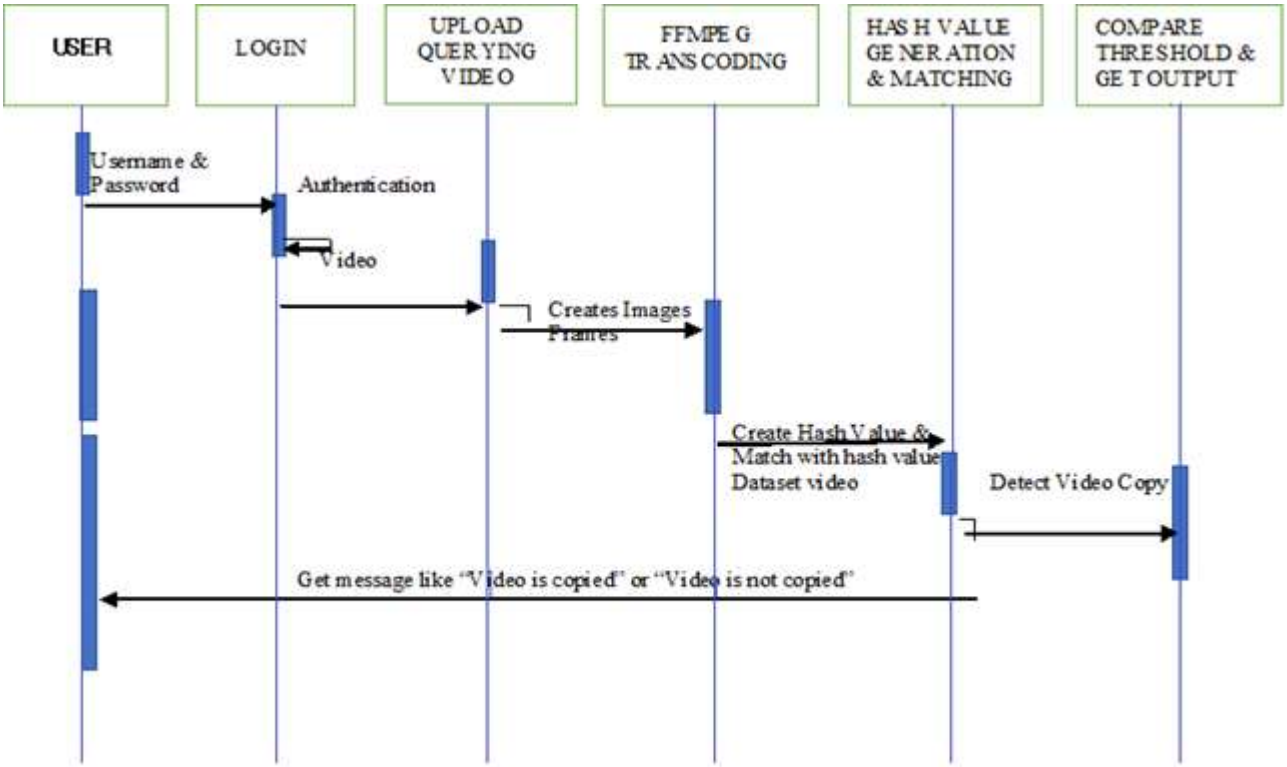


Figure 3: Sequence Diagram of Video Copy Detection

The main strategy for finding the duplicated recordings is by utilizing TIRI-DCT calculation [7]. In this approach, at every moment of time of a specific interim, a casing is drawn. The highlights of that specific casing right then and there of time are separated. The extricated attributes are then contrasted and the qualities of the first recordings show in the database. On the off chance that the highlights of both the recordings are relatively same, at that point the video is said to be replicated. For instance, assume a specific video has four edges. On the off chance that the component of one edge matches with the first video, it is said to be 25 % duplicated video. In the event that two casing matches, it is said to have half inventiveness et cetera.

The second strategy is Brightness Sequence Algorithm. In this approach, the shine of each edge is ascertained of the video whose inventiveness is to be found. This shine esteems are contrasted and the estimations of the brilliance of the first video at a similar moment of time. On the off chance that the splendor of both the recordings is relatively same, at that point it is said to be duplicated.

$$Rx,y = \sum Jk\text{-}1\ 0.65k * lk, x, y \ldots [1]$$

where Rx,y = Frame

lk, x, y = Brightness estimation of the kth outline at the area (x,y)

This calculation is executed on the casings separated from the video close by. The hash estimations of the same is figured and

afterward the separation between the hash benefits of questioning video as

well as the preparation video is resolved which is then looked at against a predefined limit esteem. On the off chance that the separation is more prominent than the predefined edge esteem, at that point the two recordings are considered as various recordings else they are viewed as same video, consequently the duplicated one.

In the accompanying framework, OCR (Optical Character Recognition) is utilized to recover the characters from the casings. The character recovered is then masterminded in some proper consecutive request and analyzed. The characters can be subtitles display in the recordings.

4. IMPLEMENTATION

In usage stage, similitudes and dissimilarities between the recordings are discovered utilizing Hadoop stage and by making utilization of two stages, i.e. Guide stage and Reduce stage. The two particular calculations which are being utilized are TIRI-DCT and Brightness Sequence Algorithm. The fundamental point of these calculations is to change over the video into various casings, extricating their highlights and after that at last looking at both questioning and also preparing video. The preparation video is as of now introduce in the HBase, while the questioning video is taken as contribution from the client.

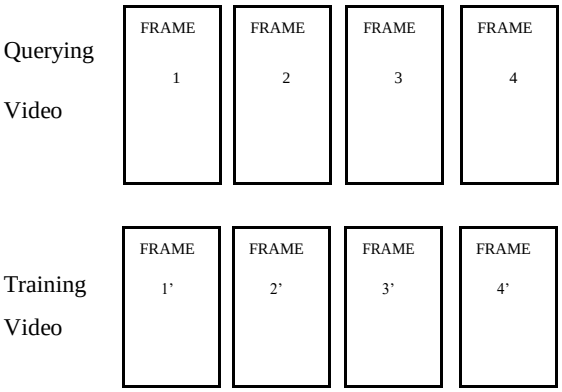


Figure 4: Comparison of Training and Querying Videos

In figure 4, consider that four casings are drawn from the video. On the off chance that FRAME 1 = FRAME 1', FRAME 2 = FRAME 2', FRAME 3 = FRAME 3' and FRAME 4 = FRAME 4', at that point questioning recordings is said to be 100% replicated.

On the off chance that FRAME 1 ≠ FRAME 1', FRAME 2 = FRAME 2', FRAME 3 = FRAME 3' and FRAME 4 = FRAME 4', at that point questioning recordings is said to be 75% replicated i.e. on the off chance that any of the one edge out of four matches with the preparation video, at that point it is said to be 75% duplicated.

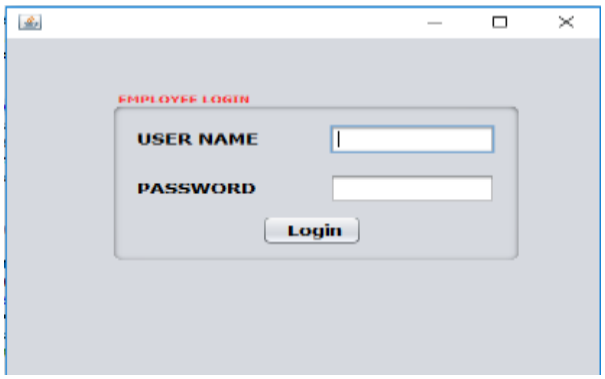
In the event that FRAME 1 ≠ FRAME 1', FRAME 2 ≠ FRAME 2', FRAME 3 = FRAME 3' and FRAME 4 = FRAME 4', at that point questioning recordings is said to be half duplicated i.e. in the event that any of the two casings out of four matches with the preparation video, at that point it is said to be half replicated.

In the event that FRAME 1 ≠ FRAME 1', FRAME 2 ≠ FRAME 2', FRAME 3 ≠ FRAME 3' and FRAME 4 = FRAME 4', at that point questioning recordings is said to be 25% duplicated i.e. on the off chance that any of the three edges out of four matches with the preparation video, at that point it is said to be 25% replicated.

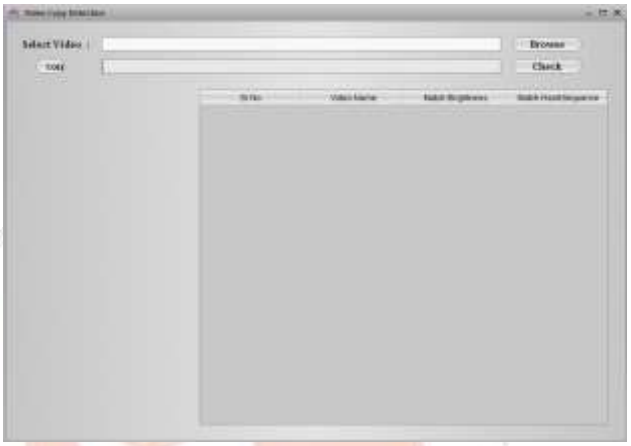
On the off chance that FRAME 1 ≠ FRAME 1', FRAME 2 ≠ FRAME 2', FRAME 3 ≠ FRAME 3' and FRAME 4 ≠ FRAME 4, at that point questioning recordings is said to be 0% replicated consequently, it isn't a duplicated video.

5. SCREENSHOTS

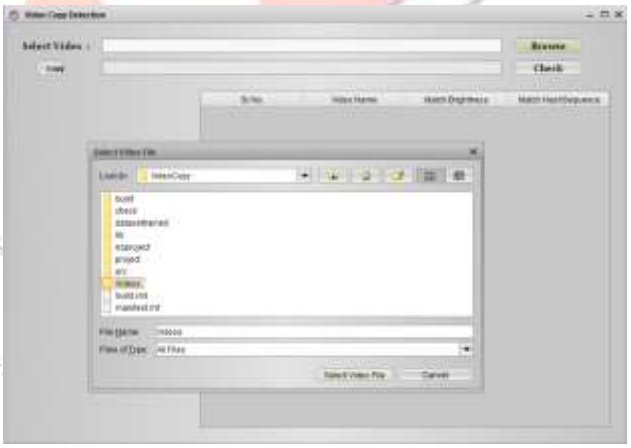
Screenshots 1: Registration Page



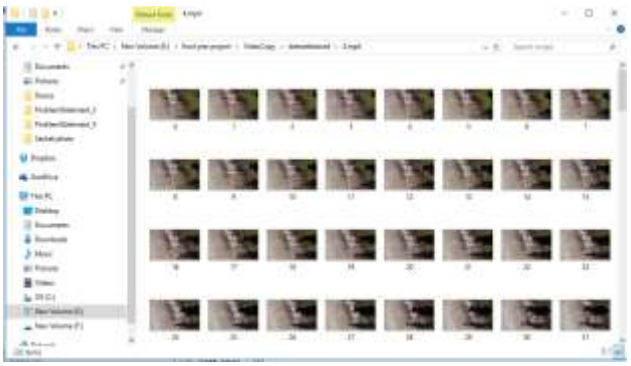
Screenshot 2: Login page



Screenshot 3: Video Copy Detection



Screenshots 4: Select Videos



Screenshots 5: Video Frames

6. FUTURE SCOPE

I. Furthermore, for video sequence matching, we propose a graph-based video sequence matching method. It skillfully converts the video sequence matching result to a matching result graph. Thus, detecting the copy video becomes finding the longest path in the matching result graph.

II. Further improving the scalability of the system. We will investigate other local visual features, including SURF, which take less time to compute. Global visual features, including color histogram, edge and texture information, may also be used as a pre-processing step to speed up the local visual feature search.

III. Incorporating the audio information. Videos normally contain both audio and video streams. The audio information can be used to further improve the detection accuracy and reduce the computational complexity

7. CONCLUSION

In this venture, two video duplicate location strategies, the strategy in light of shine arrangement and the technique in view of TIRI-DCT are executed and the reviews and precisions of the two techniques with various video numbers and diverse edges are broke down. The calculations are actualized on Hadoop disseminated registering stage and the efficiencies are thought about in various video sums and diverse guide sums. The further work is to look into more video duplicate discovery calculations and influence examinations between them; to improve the handling of the calculations in Hadoop stage; make great utilization of HBase circulated database to secure the extraction and recovery of video hash; stretch out the calculations to the duplicate location for picture and content.

8. REFERENCES

[1] Jing Li, Xuquan Lian, Qiang Wu and Jiande Sun "Realtime 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