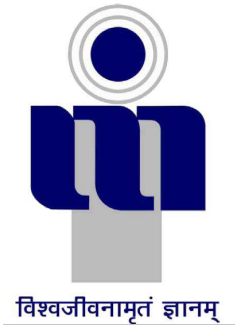
Detecting Original and duplicate YouTube videos

A project report submitted in partial fulfillment of the requirements for M.Tech. Project

M.Tech.

by

Prashant Gupta (2010IPG-72)



ABV INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR-474 010

2014

ABSTRACT

YouTube is one of the largest video sharing website on the Internet. Several music and record companies, artists and bands have official channels on YouTube to promote and monetize their music videos. YouTube consists of huge amount of copyright violated content including music videos despite the fact that they have defined several policies and implemented measures to combat copyright violations of content. This work is a method to detect copyright violated videos by studying YouTube textual/linguistic metadata, for example: title and description of video, comments posted on video.i.e..mining video as well as uploader meta-data. We propose a multi-step approach consisting of mining user profile data, idea to use google images for searching most popular publicized frames, analyzing the popularity of the uploader and video, comparison of near duplicate image detection to predict the category (original or copyright-violated) of the video. Currently two solution approaches are proposed with slight modifications in the initial stages of proposed methodology.

Keywords: YouTube Copyright Infringement Detection, Social Media Analytics, Mining User Generated Content, Information Retrieval

ACKNOWLEDGEMENTS

I am highly indebted to Dr. Joydip Dhar and obliged for giving me the autonomy of functioning and experimenting with ideas. I would like to take the opportunity to express profound gratitude to him not only for his academic guidance but also for his interest in this project and constant support coupled with confidence boosting and motivating sessions which proved very fruitful and were instrumental in infusing self assurance and trust within me. The nurturing and blossoming of the present work was mainly due to his valuable guidance, suggestions, astute judgement, constructive criticism and an eye for perfection. My mentor always answered myriad of doubts with smiling graciousness and prodigious patience, never letting me feel that I am novice by always lending an ear to my views, appreciating and improving them and by giving me a free hand in the project. Its only because of his overwhelming interest and helpful attitude, the present work has attained the stage it has.

Finally, I am grateful to all my friends and colleagues, whose constant encouragement served to renew our spirit, refocus our attention and energy and helped us in carrying out this work.

(Prashant Gupta)

Contents

1 INTRODUCTION **5**

2 MOTIVATION **5**

3 LITERATURE REVIEW **8**

4 OBJECTIVE **10**

5 METHODOLOGY **10**

6 EXPECTED DELIVERABLES **14**

List of Tables

1 Literature review-I 8
2 Literature review-II 9

List of Figures

1 Loss of revenue 6
2 First Approach 11
3 Second Approach 12

1 INTRODUCTION

Today, YouTube is the largest user-driven video content provider in the world. it has become a major platform for disseminating multimedia information. YouTube is a key international platform for socially-enabled media diffusion. According to public statistics, more than 48 hours of video content is uploaded every minute and 3 billion views are generated every day. To complement the content broadcast/consume experience, YouTube connects seamlessly with major online social networks such as Facebook, Twitter, and Google+. In fact, 12 million users have linked their YouTube account with at least one such OSN for auto-sharing, and more than 150 years of YouTube content is watched on Facebook every day.

One of the major problems encountered by YouTube is uploading of copy- right infringement videos. Despite continuous attempts by YouTube to counter copyright infringement problem, the problem of copyright infringement still persists on a large scale and is a serious issue. Piracy in YouTube is found in television shows, music videos and movies. Copyright infringement costs the music industry and the government millions of dollars every year. As per the Economic Times statistics, the eventual increase in piracy on YouTube has affected the market value of right owners and over the past decade music and movie industry has faced a loss of 5000 crore rupees in terms of revenue.

Therefore, if we could develop a robust model, that could successfully detect copyright infringement videos, it could serve as a boon against the piracy of copyright infringement videos.

2 MOTIVATION

It was not the global recession but piracy that did the Indian film industry in 2008. While nearly every sector increased the number of pink slips, the film and TV industry created nearly 1.8 million jobs and contributed $6.2 billion to the Indian economy in 2008. A new report "Economic Contribution of Indian Film and Television Industry" by PricewaterhouseCoopers said the film industry was in fact significantly impacted by online piracy. In 2008, piracy cost the Indian film industry $959 million and about 571,000 jobs 1. Speaking on the issue of piracy, Motion Picture Association of America (MPAA) chairman Dan Glickman said, "Free is great and everybody likes things for free". In a civilised society, we need to pay for products and services. I believe that as an industry, if we offer people products at reasonable prices in a hassle free manner, people will not steal. A study undertaken by the Motion Picture Distributors Association

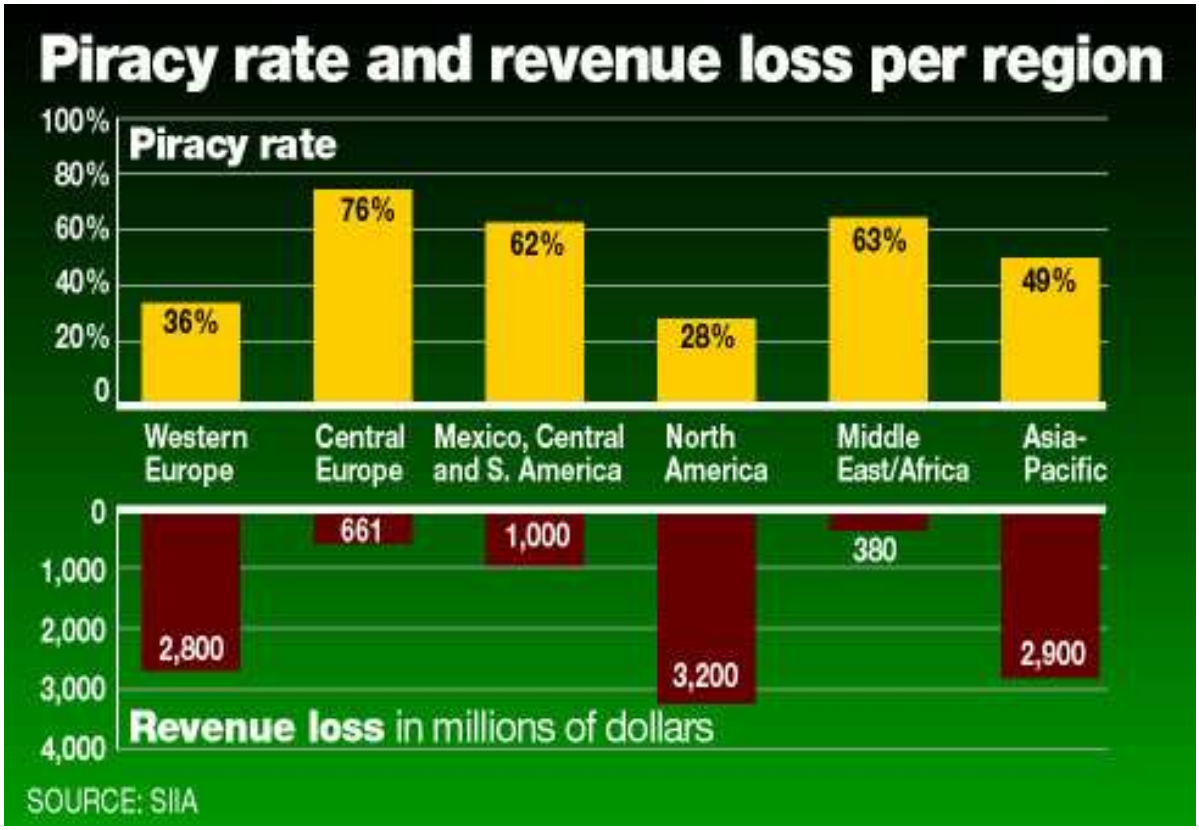


Figure 1: Loss of revenue

has put India among the top ten countries in the world where online piracy is highest. Research has shown that online piracy of film and TV network is mainly through file sharing networks.

As per the article by Times of India, Over 200 sites blocked in India after Sony's piracy complaint on 7th July 2014. In it's complaint Sony said "various websites are indulging in hosting,streaming, providing access to, etc of infringing its exclusive rights and broadcast and re-production rights." And that "the acts of infringement is not only causing Sony loss of substantial revenues but will also take away the legitimate revenue to the government through service tax etc which are payable on the subscription fee payable by the named and unnamed defendants if they conduct their business illegitimately."

Pike (2007) describes some obsolete video hosting websites which were shut down due to the sharing of copyright content. For example, Vishal Bharadwaj's 'Kaminey' was downloaded a record number of times (estimated at 350,000 times) in India and abroad. The situation in equally bad for regional language films with 88% of Telugu and 80% of Tamil films being downloaded from the internet, the report said. The menace of copyright theft jeopardises a movie's ability to make money – if at all. This

affects the level of investment available for new films and the ability to create new jobs for workers throughout the country. He said the main mode of piracy was use of camcorders in theatres and the industry needed government support against use of the device in screening locations. Despite the piracy issues, the Indian film and television industry is poised to get bigger.

3 LITERATURE REVIEW

Table 1: Literature review-I

Kim (2007)	YouTube itself uses some techniques to avoid and detect the copyright violated videos. Copyright owners can use a system called Content ID to easily identify and manage their content on YouTube. Videos uploaded to YouTube are scanned against a database of files that have been submitted to us by content owners. Copyright owners get to decide what happens when content in a video on YouTube matches a work they own. When this happens, the video gets a Content ID claim. YouTube also allows us to Submit a copyright infringement notification. If one believes that his copyright-protected work was posted on YouTube without authorization, he/she may submit a copyright infringement notification.
Agrawal and Sureka (2013)	This work describes the discriminatory features that can be used to distinguish between the original & the duplicate videos. These include examining the user profile based upon its number of subscribers, the views on the uploaded videos, total number of videos uploaded by the user and the textual comments on the video.
Zhang and Chang (2004)	defines image near-duplicate(IND) as a pair of images in which one is close to the exact duplicate of the other, but differs slightly due to variations of capturing conditions (camera, camera parameter, view angle, etc), acquisition times, rendering conditions, or editing operations. In this paper detecting IND is for for Linking Multimedia Content. Detecting IND can be used in a variety of applications, for example, linking multimedia content, identifying copyright infringement, managing photo albums, etc. Broadcast channels often report and track a news story for a few of days or months. News videos on the same event or topic often contain Near-Duplicate frames. Finding Image Near-Duplicates in News videos is therefore very useful for linking news videos from different channels and reported in different days.

Table 2: Literature review-II

Potthast and Becker (2010)	Typically, submitted comments are published immediately on the same page, so that new visitors can get an idea of the opinions of previous visitors.Popular multimedia items, such as videos and images, frequently get up to thousands of comments, which is too much to be read in reasonable time. i.e., visitors read, if at all, only the newest comments and hence get an incomplete and possibly misleading picture of the overall opinion. So, introduces OPINIONCLOUD, a technology to summarize and visualize opinions that are expressed in the form of Web comments. Also we need to pre-process all the comments of the video, such that there must not be any misleading picture of the overall opinion. Moreover, python script can also be used to retrieve all the comments.
Sureka (2011)	gives an insight to Mine User Comment Activity for detecting forum spammers in YouTube. The proposed technique is based on mining comment activity log of a user and extracting patterns (such as presence of exactly same comment across multiple unrelated videos, time interval between subsequent comments) indicating spam behavior. Rigourous testing is performed on the crawled data from YouTube which demonstrates that the proposed method is effective for the task of comment spammer detection. This could be the additional work based on online multimedia content in my project.
Nahm and Mooney (2002)	This paper gives insights about the basic measures for text retrieval, so that relationship between the set of relevant documents and the set of retrieved documents can be established. Various text retrieval and text indexing techniques are being discussed.

This section describes the literature survey of the in the area of copyright infringement web video detection, duplicate and near-duplicate video identification. In addition to the application of multimedia and image processing on the video content meta-data based features for the task of copyright infringement detection are also studied in this project.

So overall, the scope of work in the multimedia content especially YouTube has much far reaching consequences on the online community as well as the service providers. Whether the research is to detect copyright infringement videos or detect spammers or detect viral videos or summarize the video, YouTube has a vast database to work upon and also offers the huge diverse database in terms of content.

4 OBJECTIVE

- To conduct a study on the extent of copyright infringement in YouTube and investigate solutions to automatically detect original and copyright infringement videos from the search results of a given user query.
- To investigate the effectiveness of contextual features as discriminatory attributes for the task of original and copy-right infringement video detection.
- To detect near duplicate image detection.

5 METHODOLOGY

To achieve the objective mentioned, work is divided into four phases.

Phase 1:

- Literature Review
- Devise different ways to detect copyright infringement videos.

Phase 2:

- Study YouTube API
- Information retrieval from Video Metadata

Phase 3:

- devise an appropriate method for near duplicate image detection.
- extract video metadata

Phase 4:

- Integrating above modules
- Comparing & Testing both approaches.

fig 2 represents one way to detect copyright infringement videos on YouTube. This method basically depends upon the efficient retrieval of user comments on YouTube video. Starting with YouTube API, comments are fetched and then needs to be processed properly. On a YouTube video, comments may be of several types For example:

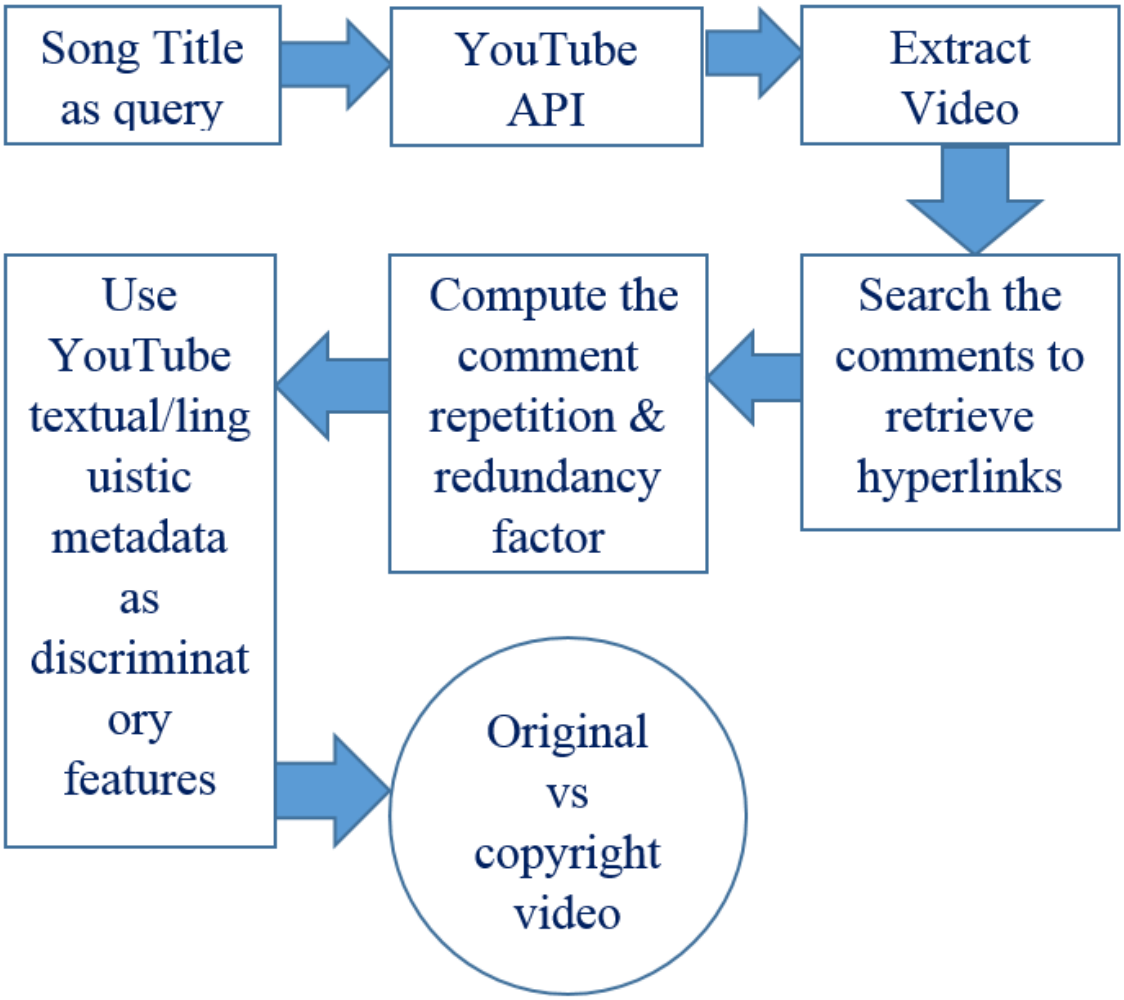


Figure 2: First Approach

spam and promotional comments, pornographic content, hate and extremism promoting comments, harassment and insulting comments and copyright violated comments. Type of comment to be focused on depends on our objective. Here video response as comments or copyright violated comments is our main objective. As mentioned in Sureka (2011) comment repetition and redundancy factor will increase the probability to detect copyright violated comments. Now, based upon the Near Duplicate image detection as proposed by **?** redirected video frames in video responses will be compared with original video frames.

Given a query, firstly top k search results are being retrieved.Thereafter three parameters are used to classify the videos

- Number of subscribers
- user profile

- username popularity

We can first find out the original videos with above method. And then use the most popular publicized frames(may be multiple) of that video to search on google image. This method therefore retrieves the list of duplicate videos in google image search results.

After getting these duplicate videos, we can once again check frame by frame and reach a level of satisfaction(yes retrieved videos were "exact" or "almost" duplicate copy of original video).

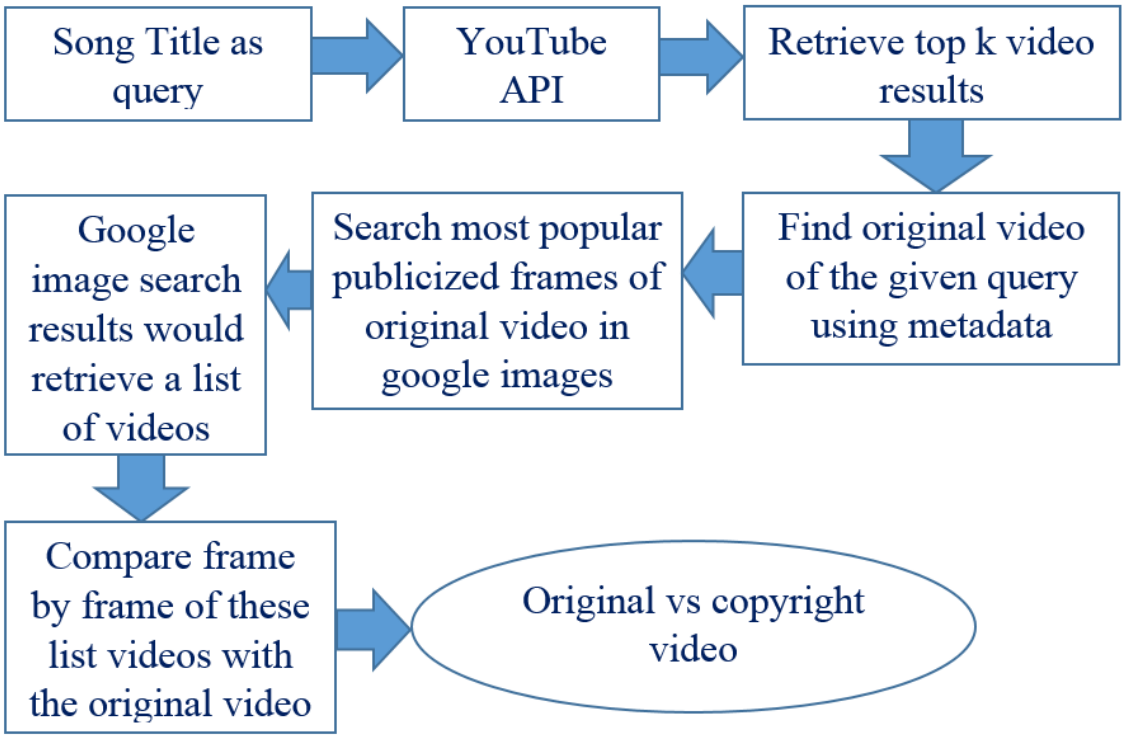


Figure 3: Second Approach

At the end the problem pertains to finding Near-Duplicate Image Detection. At the moment I have a system that does histogram analysis between two images, but this is a very expensive operation and seems too overkill. Optimally I am looking for a algorithm that would give each image a score (for example a integer score, such as the RGB Average) and I can just sort by that score. Identical Scores or scores next to each other are possible duplicates.

0299393
0599483

```
0499994 <- possible dupe
0499999 <- possible dupe
1002039
4995994
6004994
```

Later it is found that there has been a lot of research on image searching and similarity measures. It's not an easy problem. In general, a single integer value won't be enough to determine if images are very similar. This method will have a high false-positive rate.

6 EXPECTED DELIVERABLES

We have gone through the different aspects of the currently proposed project, which includes types of comments on the video, scalability related to a particular video, meta-data of the video & the frame by frame comparison of the near duplicate image detection.

So, from our work an efficient technique to detect original and duplicate videos is expected.

References

[1] Agrawal, S. and Sureka, A.: 2013, Copyright infringement detection of music videos on youtube by mining video and uploader meta-data, *Big Data Analytics*, Springer, pp. 48–67.

[2] Kim, E. C.: 2007, Youtube: Testing the safe harbors of digital copyright law, *S. Cal. Interdisc. LJ* **17**, 139.

[3] Nahm, U. Y. and Mooney, R. J.: 2002, Text mining with information extraction, *AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, Vol. 1.

[4] Pike, G. H.: 2007, Google, youtube, copyright, and privacy, *Information Today* **24**(4), 15.

[5] Potthast, M. and Becker, S.: 2010, Opinion summarization of web comments, *Advances in Information Retrieval*, Springer, pp. 668–669.

[6] Sureka, A.: 2011, Mining user comment activity for detecting forum spammers in youtube, *arXiv preprint arXiv:1103.5044* .

[7] Zhang, D.-Q. and Chang, S.-F.: 2004, Detecting image near-duplicate by stochastic attributed relational graph matching with learning, *Proceedings of the 12th annual ACM international conference on Multimedia*, ACM, pp. 877–884.