VIDEO TAGGING USING DEEP LEARNING

Major Project

Submitted by:

T.N. Raghunandan (9915103199)
Vidhi Ajay Markhedkar (9915103136)
Shalvi Sanjay Lale (9915103129)

Under the supervision of:
Ms. Varsha Garg

Department of CSE
Jaypee Institute of Information Technology University, Noida

OCTOBER 2018

ACKNOWLEDGEMENT

The completion of any inter-disciplinary project depends upon cooperation, co-ordination and combined efforts of several sources of knowledge. We are grateful to **Ms. Varsha Garg** for her even willingness to give us valuable advice and direction whenever we approached her with a problem. We are thankful to her for providing us with immense guidance for this project.

We are also thankful to Dr. Sanjeev Patel, Mr. Gaurav Nigam and Ms. Himani Bansal for giving their valuable time for evaluating our project. We would also like to thank our College authorities and Head/Dean for giving us the opportunity to pursue our project in this field and helping us to successfully complete this project.

Signature(s) of Students

T.N. Raghunandan (9915103199)

Vidhi Ajay Markhedkar (9915103136)

Shalvi Sanjay Lale (9915103129)

ABSTRACT

Images and videos have become ubiquitous on the internet, In today's digital era, Video Tagging is gaining its importance automatically as people don't want to waste their time in searching for relevant videos which has encouraged the development of algorithms that can analyze their semantic content for various applications, including search and summarization.. In this project, we proposed an automatic video tagging framework which aims to help people so that they can find relevant videos very fast and in an efficient way. Deep Learning model i.e., CNN (Convolution Neural Network) has been used with in-built descriptors instead of separate MPEG-7 Descriptor. Convolutional Neural Networks (CNNs) have been established as a powerful class of models for image recognition problems. Encouraged by the results, we will try to provide an extensive empirical evaluation of CNNs on large-scale video classification.

From a practical standpoint, there are currently no video classification benchmarks that match the scale and variety of existing image datasets because videos are significantly more difficult to collect, annotate and store. We have tried to create a system that will help in automatically tagging and classifying the videos of a data set. These tags will belong to a closed set of tags so that searching and surfing becomes easy for the user. We have tried to find accuracy percentage of all the videos which helps in finding the video according to the users need. The indexing and retrieval of the videos are done by the help of Convolution Neural Network model (CNN). This deep learning model pre-processes and divides the video into a number of frames with a fixed time interval. These frames are then tagged and classified. The frames are also compared with each other to find the accuracy percentage.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

1. IEEE - Institute of Electrical and Electronics Engineers
2. CNN - Convolution Neural Network
3. RNN - Recurrent Neural Network
4. MPEG - Moving Pictures Experts Group
5. ANMRR - Average Normalized Modified Retrieval Rank
6. MPQF - MPEG Query Format

Chapter 1: INTRODUCTION

Video Retrieval in today's world has become as important as Text Retrieval was a decade before because of ease of videomaking tools and simultaneous uploading.

1.1 MOTIVATION

With more and more gadgets in the market, every day a huge number of videos are being recorded and with each and every day, if these videos are not categorized then they will never show up when searched for and will only waste up a lot of memory making the world a cluttered mess of multimedia content. This wastage of memory also results in a loss of performance and resources. Hence due to this popularization of multimedia content on the internet, it has raised a great need of tools which can help in managing, searching, understanding and utilizing this multimedia content with the help of indexing and retrieving the videos or in other words to tag and classify the videos which motivated us to work on this concept. It also drew our interest as it is a new concept which is coming up these days.

1.2 PURPOSE OF PROJECT

Many times when any organization or an individual wishes to find videos of a particular type, finding the perfect video can be a highly time taking task leading to reduction of efficiency. This reduction in efficiency can also reduce the performance of the system. To avoid such situation to no prevent multimedia content a cluttered mess it is our effort is to provide a better algorithm for the above mentioned issue.

1.3 DESCRIPTION OF PROJECT

The project provides a systematic and comparative study of methods and techniques for video tagging using frames which are extracted from videos in a definite interval of time. It also presented a critical analysis of the existing technologies which can be integrated together to achieve the goal of Automatic Video Tagging.

Chapter 2: LITERATURE SURVEY

Research Paper [1] represents the results to evaluate the effectiveness of MPEG7 Color descriptors in Visual Surveillance retrieval problems. Color descriptors from the MPEG7 standard are used, including Dominant Color, Color Layout, Color Structure and Scalable Color. Experiments are presented that compare the performance of these, and also compare automatic and manual techniques to examine the sensitivity of the retrieval rate on segmentation accuracy. In addition, results are presented on innovative methods to combine the output from different descriptors, and also different components of the observed people. The evaluation measure used is the ANMRR, a standard in Content-Based Retrieval experiments.

In the research paper [2] they conducted an empirical evaluation of MPEG-7 visual part of experimentation model (XM) color descriptors in a challenging problem of content-based retrieval of semantic image categories. The performance of the four color descriptors provided in the current XM reference implementation, Color Layout, Color Structure, Dominant Color and Scalable Color, is compared to that of HSV autocorrelogram, which has done well in recent empirical studies. Experimental results show that Color Structure provides best retrieval accuracy, whereas the computationally most expensive descriptor, Dominant Color, is worst in this problem. Considering the difficulty of the problem, Color Structure achieves a decent 35% average precision in retrieving the first 10% of the images from the semantic category of the query image. Dominant Color, which is the most expensive XM Descriptor in computational terms, provides the worst retrieval performance.
(Auto)correlograms capture both global occurrence statistics and local spatial organization of colors by a simple spatial constraint. Future work includes validation whether performance could be improved by a more extensive, possibly color independent, spatial constraint. Also, the de facto interpretations of (auto)correlograms are feature vectors, which largely for computational reasons are compared with L1 norm. However, (auto)correlograms are essentially histograms, and better performance could be achieved with information theoretic similarity measures, albeit at higher computational cost.

In the above research papers we see that earlier this classification of videos was done manually with the help of MGEP-7 descriptor. In the coming research papers we see how this classification is done automatically with the help of deep learning, which is then applied in our project.

In the research paper [3], it is told that GPU-enabled hardware is needed for the first time for indexing a single pass of feature vector extraction and storing into the database automatically. They have shown in this work that feature vector fv belongs to $\mathscr{R}$^1024 extracted by CNN [6] contains enough semantic information for segmenting raw video into shots with 0.92 precision; retrieving video shots by keywords with 0.84 precision; retrieving videos by sample video clip with 0.86 precision and retrieving videos by online learning with 0.64 precision.
They have shown in this work that all query types i.e., QueryByMedia, QueryByFreeText, SpatialQuery and TemporalQuery can be implemented using the semantic features extracted from video by deep learning algorithms, namely by convolutional neural networks. Their contribution is presenting a video indexing and retrieval architecture based on unified semantic features and capable to implement MPQF query interface and sharing the results of real-world testing.

In Research paper [4], they have reviewed two lines of research aiming to stimulate the comprehension of videos with deep learning: video classification and video captioning.
While video classification concentrates on automatically labelling video clips based on their semantic contents like human actions or complex events, video captioning attempts to generate a complete and natural sentence, enriching video classification's single label to capture the most informative dynamics in videos. They have reviewed basic deep learning modules like CNN(Convolution Neural Network), RNN(Recurrent Neural Network) that have been widely adopted in the literature for video analysis. Long short-term memory (LSTM) is reviewed which overcome cons of RNN like —vanishing and exploding gradients. It is a variant that was designed to store and access information in a long time sequence. For Video Captioning, they have tried Supervised Deep Learning and Unsupervised Deep Learning on different datasets.

In the research paper [5], they have proposed the construction of a YouTube Recommender Network (YRN) and a recommender system derived from it. The YRN was created from the data collected from the YouTube website using their API. The data was composed of a number of videos and comments on each video. The YRN is undirected and weighted. The nodes represent the videos, whereas the edges are established between two nodes if there is a user who commented on both of them. The edge weight represents the number of times two nodes are associated with comments from different users in both of them. After the YRN was generated, they observed scale-free and small-world properties in the network with a number of communities. It was demonstrated that the distribution of tags inside communities is diverse and follows a power law. The weight of an edge predicts the strength of the relation between the two nodes it connects. A utility value was also introduced which represents the importance

of a node. The higher the utility value, the more important the node is. Finally, a way to build a recommender system derived from our YRN was proposed in which firstly the videos are recommended to users from the highest utility value to the lowest. Secondly, when a user is watching a video, other nodes connected to it were recommended.

Similarly, the research paper [6] has introduced ViTS, an industrial Video Tagging System which generates tags based on information crawled from the Internet and learns relations between concepts. The core of the system is a knowledge base that is constantly updated to capture the dynamics of the indexed concepts. ViTS was tested on a subset of videos from the YouTube- 8M dataset. The tags generated by ViTS were highly graded by human users exposed to a visual summary of the video and its metadata. The accuracy of 80.87% is comparable to the inter-annotator agreement of (non-expert) humans in the task of semantic annotation. This high quality, combined with its capability of capturing not-only visual concepts, shows the capability of ViTS as a rich video indexing system. Moreover, experiment results on Youtube-8M are publicly available. The presented tagging system shows how contextual data is a powerful source of information when indexing web videos. Exploiting the relations between concepts allows generating a rich set of tags with a light computation, desirable when addressing a web-scale indexing. However, content-based techniques could also extend our content based tags. Our future work will address exploiting these tags as weak labels for computer vision and audio processing deep models, which have been shown impressive recognition performances in the recent years.

Research Paper [7] studies the exploration of user searching behavior through click-through data, which is largely available and freely accessible by search engines, for learning video relationship and applying the relationship for the economic way of annotating online videos. They have demonstrated that, by a simple approach using co-click statistics, promising results were obtained in contrast to feature-based similarity measurement.
A new method based on polynomial semantic indexing is proposed to learn a latent space for alleviating the sparsity problem of click-through data. The proposed approaches are then applied for three major tasks in tagging: tag assignment, ranking, and enrichment. On a bipartite graph constructed from click-through data with over 15 million queries and 20 million video URL clicks, they have shown that annotation can be performed for free with competitive performance and minimum computing resource.

In the research paper [8] they have explained what is tagging and geo-tagging. It mentioned that there are still a number of videos that are untagged and the need to do so in order to improve

the performance of multimedia retrieval. The paper is based on the three techniques of media eval 2010 benchmark initiative which are discussed later in the paper. For auto-tagging in a video, two approaches were used, first is called extraction which tags videos on the bases of its metadata. The second approach is called assignment in which tags are assigned from a fixed data set. Auto spoken audio tagging was also applied but was not possible due to noise in the voice. It was only possible on speech on telephones. For geo-tagging spatial distribution was used in which location is represented as grid cells. The three techniques of media eval 2010 benchmark initiative are then discussed which are: tagging task professional which needs human activity for assigning tags from a closed tags using mean average precision (MAP) technique, tagging task wild wild web which includes tagging by users in an online video community using MAP and metadata, finally is placing task which requires participants to automatically assign geo-ordinates to the video.

Automatic video tagging was also done in research paper [9] which tells how video content on the World Wide Web continues to expand and it is considerable to annotate videos for effective and accurate search and mining properly. The paper says that while the idea of annotating imagery using keywords is simple and well known, it facilitates basic for annotating videos with natural keywords automatically to enhance search is a significant arising problem with the excessive potential to improve the quality of video search. The paper then discusses the benefits of levelling large scale video datasets for automated annotation also presents fresh challenges and requires methods specialized for scalability and efficiency.

2.1 INTEGRATED SUMMARY OF LITERATURE SURVEY

Research paper name	Author	Summary
"Annotation for Free: Video Tagging by Mining User Search Behavior" , IEEE 04-September 2017	Ting Yao, Tao Mei , Chong-Wah Ngo, Shipeng Li	Method based on polynomial semantic indexing is proposed to learn a latent space for alleviating the sparsity problem of click-through data. The proposed approaches are then applied for three major tasks in tagging: tag assignment, ranking, and enrichment. They have shown on a bipartite graph that annotation can be performed for free with competitive performance and minimum computing resource.
"Evaluation of MPEG7 color descriptors for visual surveillance retrieval." IEEE , 2005	James Annesley, James Orwell, John-Paul Renno	It evaluates the effectiveness of MPEG7 Colour descriptors in Visual Surveillance retrieval problems. Colour descriptors from the MPEG7 standard are used, including Dominant Colour, Colour Layout, Colour Structure and Scalable Colour. Experiments compare the performance of these, and also compare automatic and manual techniques to examine the sensitivity of the retrieval rate on segmentation accuracy.
"Deep learning based semantic video indexing and retrieval. " Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2016.	Anna Podlesnaya, Sergey Podlesnyy	They have shown in this work that all query types i.e., QueryByMedia, QueryByFreeText, SpatialQuery and Temporal Query can be implemented using the semantic features extracted from video by deep learning algorithms, namely by convolutional neural networks.
"Deep learning for video classification and	Wu, Zuxuan, et al.	They have reviewed two lines of research aiming to stimulate the comprehension of

captioning." arXiv preprint arXiv:1609.06782 (2016).		videos with deep learning: video classification and video captioning.Basic deep learning modules like CNN(Convolution Neural Network), RNN(Recurrent Neural Network) and Long short-term memory (LSTM) are also reviewed.
A recommender system for youtube based on its network of reviewers." Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010.	Qin, Song, Ronaldo Menezes, and Marius Silaghi	They have proposed the construction of a YouTube Recommender Network (YRN). It was created from the data collected from the YouTube website using their API. A recommender system derived from YRN was proposed in which firstly the videos are recommended to users from the highest utility value to the lowest. Secondly, when a user is watching a video, other nodes connected to it were recommended.
Video tagging system from massive web multimedia collections." Proceedings of the 5th Workshop on Web-scale Vision and Social Media (VSM). IEEE Press, 2017.	Fernández, Delia, et al. "ViTS:	This has introduced ViTS, an industrial Video Tagging System which generates tags based on information crawled from the Internet and learns relations between concepts. The core of the system is a knowledge base that is constantly updated to capture the dynamics of the indexed concepts. The accuracy obtained is 80.87%.
"Automatic tagging and geotagging in video collections and communities." Proceedings of the 1st ACM international conference on multimedia retrieval. ACM, 2011.	Larson, Martha, et al.	In this, they have explained what is tagging and geo-tagging. The paper is based on the three techniques of media eval 2010 benchmark initiative: tagging task professional which needs human activity for assigning tags from a closed tags using mean average precision (MAP) technique, tagging task wild wild web which includes tagging by users in an online video community using

		MAP and metadata, finally is placing task which requires participants to automatically assign geo-ordinates to the video.
"Semantic video search by automatic video annotation using TensorFlow." Manufacturing & Industrial Engineering Symposium (MIES). IEEE, 2016.	Ashangani, Kithmi, et al.	Automatic video tagging was also done in research paper [8] which tells how video content on the World Wide Web continues to expand and it is considerable to annotate videos for effective and accurate search and mining properly. It facilitates basic for annotating videos with natural keywords automatically to enhance the quality of video search.
Ojala, Timo, Markus Aittola, and Esa Matinmikko. "Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories." Pattern Recognition, 2002. Proceedings. 16th International Conference on. Vol. 2. IEEE, 2002.	Timo Ojala, Markus Aittola and Esa Matinmikko	The performance of the four color descriptors:- Color Layout, Color Structure, Dominant Color and Scalable Color, Experimental results show that Color Structure provides best retrieval accuracy, whereas the computationally most expensive descriptor, Dominant Color, is worst in this problem.

Table 2.1 Integrated summary of literature survey

2.2 CURRENT PROBLEMS

Given the substantial amounts of videos generated at an astounding speed every hour and every day, it remains a challenging open problem how to derive better video representations with new methodologies, the abundant interactions of objects and their evolution over time with limited supervisory signals to facilitate video content understanding (i.e. , the recognition of human activities and events as well as the generation of free-form and open-vocabulary sentences for describing videos)

In doing Video Tagging using Mining User Search behavior, there is a problem with partial visual near duplicates, which frequently happen in Web videos. The visual features can be exploited together with click through and document features for a more comprehensive manner of characterizing visual similarities. Similarly, in Deep Learning Methodology like CNN, performance of sample-based video retrieval should be enhanced. Several approaches should be explored for lowering the feature vector dimensionality in order to search in log time scale, e.g. random projections and compact binary descriptors.

While using MPEG-7 Descriptor, again some problem arises like it fails to adequately address certain issues familiar to Visual Surveillance researchers. For example, the rank ordering method cannot in itself provide evidence that a given query example does not appear in a dataset: there will always be one element of the data set most similar to the example. Similarly, probabilistic estimates of identity, for incorporation with other uncertain cues, are not easily deduced from the rank method. One challenge is the unification of retrieval metrics across the research communities. Also, Multi-camera retrieval rate should be enhanced by specification of a pre-processing method which is lacking in color constancy. Color Structure achieves a decent 35% average precision in retrieving the first 10% of the images from the semantic category of the query image. Dominant Color, which is the most expensive XM Descriptor in computational terms, provides the worst retrieval performance. This is understandable, since the descriptor is targeted for presenting local features such as regions or objects, not complete images.

2.3 PROBLEM STATEMENT

To efficient video retrieval based on Tagging using Deep Learning.
The feature vector extracted by CNN gives 0.92 precision for segmenting raw video into shots, 0.84 precision for retrieving video shots by keywords, 0.86 precision for retrieving videos by sample video clip and 0.64 precision for retrieving videos by online learning. [3]

In this project we will try to increase this performance and efficiency which are obtained as above.

2.4 OVERVIEW OF PROPOSED SOLUTION

As per given Problem Statement above, Accuracy of video tagging is not up to the mark. We need to improve performance of our system.

So, to increase efficiency as well as accuracy of our system, we need to modify our technique.

We can improve accuracy by following methods:

1. Extract features with a CNN, pass the sequence to a separate RNN.

2. Extract features from each frame with a CNN and pass the sequence to an MLP (Multilayer Perceptron).

2.5 TASK DIVISION

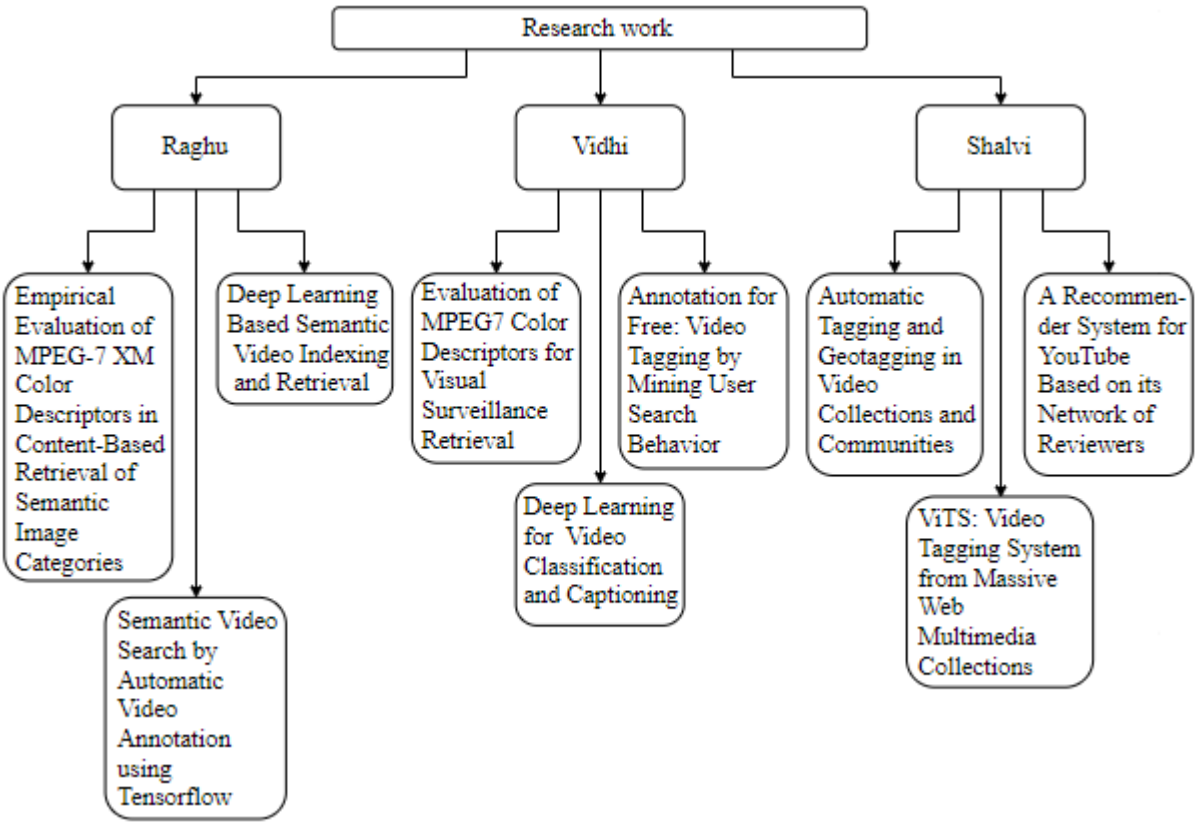

Fig. 2.1 Task division

Chapter 3: REQUIREMENT ANALYSIS

3.1 SOFTWARE REQUIREMENTS

- Windows 7 and above , or Ubuntu or Centos
- Python Idle, Anaconda

3.2 HARDWARE COMPONENTS

- Processor – i3
- Hard Disk – 5 GB
- Memory – 1GB RAM

3.3 FUNCTIONAL REQUIREMENTS

- The user should have information of the libraries used.
- The user should be familiar with python.
- The user should be familiar with Anaconda.

3.4 NON-FUNCTIONAL REQUIREMENTS

- The program should not have any reliability issues. The program will be thoroughly tested.
- The program should run on any software mentioned above.
- The program should be able to classify videos.

3.5 USER REQUIREMENTS

- User should be able to work in python.
- User should be able to work in Anaconda.

Chapter 4: ANALYSIS, DESIGN AND MODELING

4.1 OVERALL ARCHITECTURE

The previous sections presented a critical analysis of the existing technologies which can be integrated together to achieve the goal of Automatic Video Tagging. In this section, new design and architectural proposals will be made for the target system.

Our dataset had been created using a large number of videos which can be taken from "www.youtube.com". Then each video was divided into frames in a fixed interval of time. Large no of frames will be taken as a dataset and appropriate deep learning Algorithm i.e., CNN (Convolution Neural Network) will be applied. After this step, we will apply feature vector function to get the feature vector of each frame. This function includes pre-processing: image re-scaling into 256x256 BGR, selecting single central crop 224x224 and applying the CNN calculation. The function returns an output of the last average pooling layer of the network which has the dimension 1024. As each frame is continuous, so to avoid this we will use a sampling period. So, Frames will be retrieved in an interval of Sampling Period.

After that, we will find the distance between two frames i.e., Previous frame feature vector and current frame feature vector. Later, Filtering operation will be performed. We will use simple low-pass filter e.g. convolution of 4-window of last distance values with vector [0.1, 0.1, 0.1, 0.99]. Then, we will check if the filtered value of vector distance exceeds a threshold value, and add frame number to its relevant class if it will exceed. If the threshold value is maximum it means both frames are different, so we will create another class. Frames containing same object notation will lie in a class and tagging will be applied according to its class.

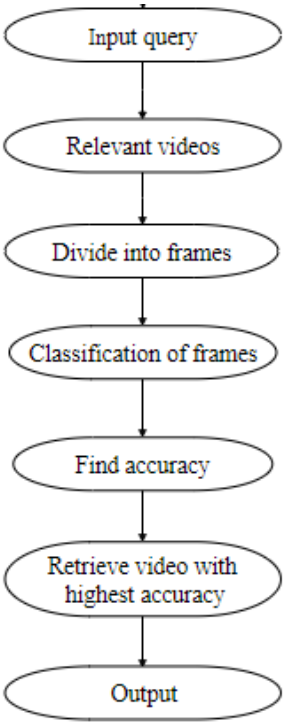

Fig. 4.1 Flow chart

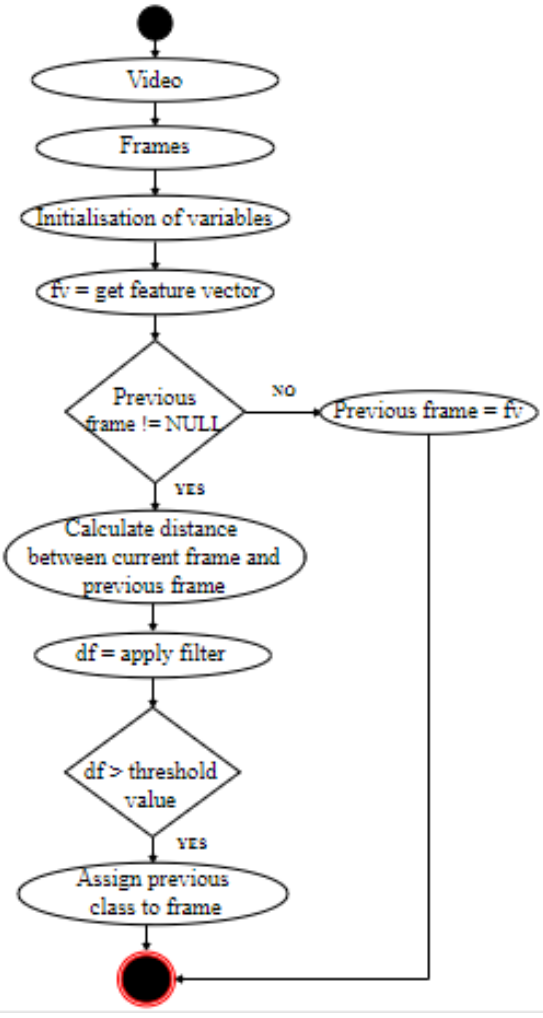

Fig. 4.2 Activity diagram

4.2 PROPOSED ALGORITHM

```
1:      prev_fv <= Null; K <= {1}; InitFilter();
2:      for i=1 to N with steps S
3:          fv <= GetFeatureVector(fi)
4:          if prev_fv is not Null
5:              d <= Distance(fv, prev_fv)
6:              df <= Filter(d)
7:              if df > T
8:                  k << i
9:              end if
10:         end if
11:         prev_fv <= fv
12:     end for
```

4.3 RISK ANALYSIS AND MITIGATION PLAN

Risk ID	Description of Risk	Risk Area	Probability (P)	Impact (I)	RE = PxI	Risk Selected for Mitigation on	Mitigation on plan	Contingency plan
1	It requires system with effective GPU	Hardware	3	3	9	N	N	Use AWS
2	Large dataset is required	Algorithm	1	5	5	Y	Y	Not planned yet

Table 4.3 Risk Analysis and Mitigation Plan

4.4 TEST PLAN

In this process, we have to find the accuracy of the image. Firstly, we will tag the video manually and save the answer in a variable. Now, we will apply deep learning on the video for tagging, find out the solution and finally save it in another variable. Finally, we will compare both the solutions, if both of them matches then it is accurate and if it doesn't match then it is an error.

4.5 IMPLEMENTATION

```
import cv2
import numpy as np
import os

# Playing video from file:
cap = cv2.VideoCapture('Toy_car.mov')

try:
    if not os.path.exists('datall'):
        os.makedirs('datall')
except OSError:
    print ('Error: Creating directory of data')

currentFrame = 0
c=0
while(c<100):
    ret, frame = cap.read()
    if currentFrame%2==0:
        # Saves image of the current frame in jpg file
        name = './datall/frame' + str(currentFrame) + '.jpg'
        print ('Creating...' + name)
        cv2.imwrite(name, frame)
    # To stop duplicate images
    currentFrame += 1
    c=c+1

cap.release()
cv2.destroyAllWindows()
```

Fig. 4.3 Code of dividing a video into frames in a fixed time interval

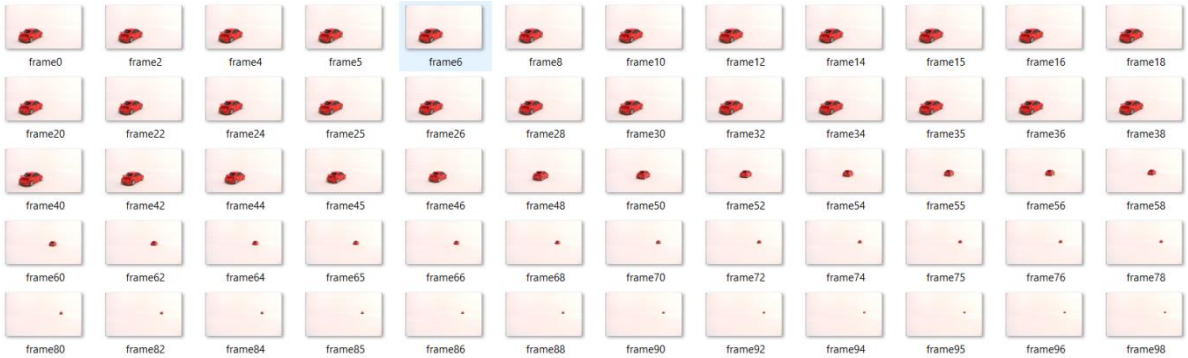

Fig. 4.4 Set of frames obtained

Fig. 4.5 Closer view of frame6

Fig. 4.6 Closer view of frame46

Fig. 4.7 Closer view of frame68

Fig. 4.8 Closer view of frame90

CONCLUSION AND FUTURE SCOPE

To efficient retrieval of video based on video tagging, we had studied different research papers based on video tagging as well as methods to classify videos for an appropriate tagging using Deep Learning. We tried to divide videos into no. of frames with a fixed time interval.

In future, we will try to develop an efficient system with Deep Learning methodologies to overcome drawbacks of previously researched work/project like accuracy, memory wastage etc. Also, we will try to implement an algorithm which has not been used till now.

APPENDIX

REFERENCES

[1] Annesley, James, James Orwell, and J-P. Renno. "Evaluation of MPEG7 color descriptors for visual surveillance retrieval." Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005

[2] Ojala, Timo, Markus Aittola, and Esa Matinmikko. "Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories." Pattern Recognition, 2002. Proceedings. 16th International Conference on. Vol. 2. IEEE, 2002.

[3] Podlesnaya, Anna, and Sergey Podlesnyy. "Deep learning based semantic video indexing and retrieval." Proceedings of SAI Intelligent Systems Conference. Springer, Cham, 2016.

[4] Wu, Zuxuan, et al. "Deep learning for video classification and captioning." arXiv preprint arXiv:1609.06782 (2016).

[5] Qin, Song, Ronaldo Menezes, and Marius Silaghi. "A recommender system for youtube based on its network of reviewers." Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010.

[6] Fernández, Delia, et al. "ViTS: Video tagging system from massive web multimedia collections." Proceedings of the 5th Workshop on Web-scale Vision and Social Media (VSM). IEEE Press, 2017.

[7] Ting Yao †, Tao Mei ‡, Chong-Wah Ngo †, Shipeng Li Annotation for Free: Video Tagging by Mining User Search Behavior, IEEE 04-September 2017

[8] Larson, Martha, et al. "Automatic tagging and geotagging in video collections and communities." Proceedings of the 1st ACM international conference on multimedia retrieval. ACM, 2011.

[9] Ashangani, Kithmi, et al. "Semantic video search by automatic video annotation using TensorFlow." Manufacturing & Industrial Engineering Symposium (MIES). IEEE, 2016.