International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

A Systematic Review of Near Duplicate Video Retrieval Techniques

Dhanashree A. Phalke[1]
Dr. Sunita Jahirabadkar[2]
[1]Research Scholar,
Department of Technology,
SPPU, Pune
[2]Department of Computer Engineering,
Cummins College of Engineering for Women,
sunita.jahirabadkar@cumminscollege.in,
a_dhanashree@yahoo.com ,

May 27, 2018

Abstract

The tremendous growth of videos shared on Internet has been observed in last few years. The users can capture videos easily even on smart phones and share it on the social media and also on Internet. This results in huge amount of re-posting of same videos creating duplicates on World Wide Web. It has become very easy for the user to make small changes in the video, transform it and upload it which violets the copyright. Such modified videos are called Near Duplicate Videos. Theres a tremendous demand on research in the field of Near Duplicate Video Retrieval for activities like video copy infringement, video search, video recommendation and many more. Many researchers are working in this field to detect such videos which are mere reproduction of original videos. This paper discusses these state-of-the-art approaches used in retrieving Near Duplicate Video Retrieval.

***Key Words*:** Near Duplicate Video Retrieval, Key Frame Extraction, Local Features, Global Features, Performance Measures.

1 Introduction

With the use of Android based smart phones, Internet is reached to a common mans pocket. This has increased the use of Internet Services enormously not only to download popular videos, but also to create personal video contents and upload them. A statistics by [18] says that every minute, approximately 300 hours of videos are being uploaded on YouTube by different users across the globe. All these videos are not original videos, whereas they are Near Duplicate Videos which are created by modifying visual or temporal parameters and/or transformed by either cropping or scaling or re-encoding, or adjusting its contrast and brightness parameters etc. Sometimes, original videos are re-recorded with slow motion or fast forwarding and added with captions or subtitles [1]. As per a statistics YouTube, Yahoo!, Google Videos sites contain nearly 27% of Near Duplicate Videos [2].

Near Duplicate Video Retrieval (NDVR)is the systematic approach to search for such videos, which are created by transforming original videos with or without permission. There are many NDVR methods and algorithms available in the literature which work on retrieving various frames of the query video and matching them with video dataset based on many image parameters. This paper presents the basic framework for Near Duplicate Video Retrieval system in Section 2, followed by discussion on various methods proposed in literature and their performance analysis in Section 2 and 3 respectively. Section 4 concludes the findings.

2 FRAMEWORK FOR NDVR

Retrieval of Near Duplicate Videos can be processes in two main phases as Online Processing and Offline Processing. In Online Processing, features are extracted from key frames which are extracted from query videos. After extracting features from query video, similarity measure is used to retrieve the near duplicate video from video

dataset. Whereas, in offline processing features are extracted from the key frames which are extracted from reference video dataset,and stored for further processing. Framework of generalized NDVR is shown in Fig. 1.

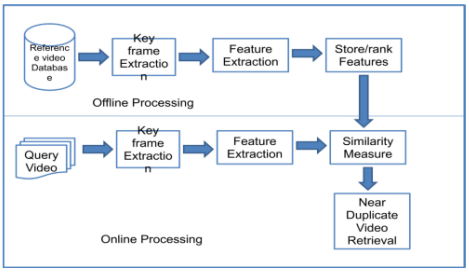

Fig. 1: General Framework for NDVR

NDVR consists of four modules : i) Key Frame Extraction, ii) Feature Extraction, iii) Store or Rank the Extracted Features and iv) Similarity Measure Evaluation. These modules are explained in detail below.

i. KEY FRAME EXTRACTION

Videos have significant structures at frame level, shot level, and scene level. Video is a sequence of frames. Instead of extracting the features from all the frames which increases complexity, only selected frames are extracted which is called as key frame. Key frames are those frames extracted from video where structural information is stored. Table 1 shows methods used by various authors to extract key frames.

Table 1: Approaches for Video Key Frame Extraction

Authors [Reference Paper]	Key frame extraction method
Chien-Li Chou et.al. [1]	Uniform sampling on videos every t frame
Xiushan Nie et al. [3]	Shot based sampling method
Yanbin Hao et al.[4]	Shot based sampling method
Chien Li Chou et al [7]	Every t frame where t = frames per second/3
Chien Li Chou et al [8]	Every t frame where t = 5
Yang Cai e al [12]	1 frame / second

3 ii. FEATURE EXTRACTION

Visual features of extracted key frames are divided in three types: Spatial Features, Temporal Features and Motion Features. Further, Spatial Features are divided as - Global Features and Local Features [9].

A. Global Features: Global features give invariant information of a frames. It is easy to extract global features, requires low computation cost and can achieve fast retrieval speed however global features failed to differentiate between foreground and background resulting in less accuracy. Following are the global features used in various papers for feature extraction 1) HSV 2) Ordinal etc.

a) HSV Feature: In the HSV representation, the color is determined from hue, saturation is used to determine the intensity of the color and lightness of the image is determined by value. Xiushan Nie et al. [3] and Yanbin Hao et al.[4] used HSV features with 162 dimensions and Chien-Li Chou et.al. [1] used.

b) Ordinal Feature: Here, the image frame is divided as 33 blocks to find the average intensity of each block. This average intensity is then used to compute the comparative ordinal features of every block of the image. If the average intensity of image block P is higher than image block Q, the value of the corresponding feature dimension od(P, Q) is set to 1, and if it is less then it is set to 0. Thus, the number of dimensions of Ordinal Feature are 36. od(P,Q)feature of a frame f can be computed as

$$od^f(i,j) = \begin{cases} 1, & \text{if } I_i > I_j and J > i; \\ 0 & otherwise \end{cases} \quad (1)$$

where i and j are the indexes of the blocks, and Ii and Ij are the average intensity of the ith and jth block. Chien-Li Chou et.al. [1] [8] used the ordinal features with 36 dimensions.

Some other global features are Pyramid Histogram of Oriented Gradients(PHOG), Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), GIST feature, etc [6].

B. Local Features: Color, Background data, Texture, Shape of an object, Spatial layout for image representation etc. are called as Local Features and which are low level features in NDVR. Local features takes more time and storage spaceas compared to that of Global Features. However, being low level features of the im-

age they provide better accuracy than the use of Global Features [11]. Local features are more robust to photometric and geometric changes, in original videos and hence give enhanced performance in NDVR [4].

a) Local Binary Pattern(LBP): LBP works on grayscale images so color frames need to be converted in grayscale. Generally, a set of neighboring pixels is selected in surrounding the current pixel. For each pixel in the neighborhood of the grayscale image, Local Binary Pattern value is calculated by comparing the neighboring pixel values either in clockwise or anti-clockwise direction. If the current pixel value is less than the neighboring pixel value, a bit is set to zero in the binary array. And the pixel value is set to one, if the current pixel value is either greater than or equal to the neighboring pixel value. The corresponding pixel location in the LBP mask is updated after calculating the LBP value of the current pixel. Since there are 8 neighboring pixels for each pixel, the results of the comparisons are stored in a 8-bit binary array[14]. Xiushan Nie et al. [3] and Yanbin Hao et al.[4] used LBP features with 256 dimensions.

b) Speeded Up Robust Features (SURF): SURF is a very strong and powerful Local Feature descriptor [16]. A Hessian Matrix is calculated for detecting the key points in an image. Hessian matrix is built which represents every pixel of the image, according to its position in the image. A determinant of approximated Hessian Matrix is calculated at different scales, to determine the key image points. The SURF key points location and scale, s is obtained with the maxima values. Haar-wavelet response is calculated within its 6 s radius neighborhood to assign the orientation to the obtained key point.

Further, feature descriptors are extracted at the key-points to obtain the SURF feature descriptor. A square of 20 s size, with center at the key-point is constructed, in the direction of the orientation assigned to the key-point. This square is divided into 4x4 sub regions and each sub region is further divided into 5x5 regularly spaced sample points. At each of these sample points, a vertical and horizontal Haar wavelet responses are calculated in the form of 4 dimensional vector:

$$V = (\sum dx, \sum dy, \sum |dx|, \sum |dy|) \tag{2}$$

Such a 64-dimensional vector (4 vectors each for 16 sub regions), is generated at each key-point[7]. SURF is highly robust against geometric transformations and is used by [1] [5] [9] [10] [11] [17].

To get the benefit of global and local features, researchers used both types of vectors. Table 2 shows such papers:

Table 2: Features Extracted

Reference	Features Extracted for use	
	Global	Local
Chien-Li Chou et.al. [1]	Ordinal	HOOF and SURF
Xiushan Nie et al. [3]	HSV	LBP
Yanbin Hao et al.[4]	HSV	LBP
Prinka et al [5]		SURF
Chien-Li Chou et.al. [7]		HOOF

iii. STORE / RANK FEATURES

Extracted features of video are stored in the data structure to compare using similarity measures. Before storing, some of the researchers applied optimization techniques and indexing for fast retrieval of the results.

Feature Optimization or Aggregation: The extracted global and local features are optimized according to objective function for fast video retrieval. Genetic algorithm is used for feature optimization [3]. Giorgos Kordopatis-Zilos et al [6] extracted frame level histogram and aggregated by the summation to derive video level histogram using CNN architecture.

Indexing: Inverted index is nothing but a hash table, wherein each index key is an indexed term. The value corresponding to each index key is a document list containing the indexed term. Inverted index is used to speed up the process of retrieval [12] [14].Xiushan Nie et al. [3] uses a Natural Binary Code (NBC) approach to learn the frame hash and then creates a video-level hash matrix. Hash codes are used by [4] [13] Chien-Li Chou et.al. [1] [7] [8] applied k-means clustering to generate symbol sequence. These symbols are called as I-pattern and in a pre-fix tree structure for fast retrieval.

iv. SIMILARITY MEASURES

In online processing of figure 1, query video is given as input to extract the features in the same way as offline processing. Once the

features are extracted, the similarity measures are used to check the similarity based on features between query video and videos from dataset. Euclidean distance[1][5][7], Manhattan distance [3] and hamming distance [4] are widely used methods for distance calculation of the extracted features.

Euclidean distance is calculated to rank the retrieved videos. The video from the database corresponding to the frame similar to the query frame is higher in rank if the Euclidean distance is smaller. The distance between two frames with the coordinates (x1,y1) and (x2,y2) is given by

$$Dist((x1, y1), (x2, y2)) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2} \quad (3)$$

Manhattan distance is used to measure the similarities of different videos. If there are two vectors x = [x1; x2; _ _ _ ; xn] and y = [y1; y2; _ _ _ ; yn], the Manhattan distanced(x; y) is calculated as -

$$d(x, y) = \sum_{i=1}^{n} n|xiyi| \quad (4)$$

4 PERFORMANCE EVALUATION

The performance of Near Duplicate Video Retrieval approaches can be evaluated using Mean Average Precision [4][5][6][12][14], Precision and recall [5] [6][7][8] [12][14] and F-measure [7] metrics.

The formula to evaluate performance is :

Precision can be defined as fraction of the retrieved relevant videos.

Recall =

Recall can be defined as the fraction of the relevant videos that are retrieved.

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (5)$$

F Measure is a combined measure which assesses precision and recall tradeoff.

$$MAP = \frac{\sum_{m}^{max}}{Q} \quad (6)$$

MAP is the calculated as average of mean of precision.

DATASET

Most of the authors have used CC_WEB_VIDEO dataset for their experimentation. CC_WEB_VIDEO is a most popularly known NDVR benchmark. CC_WEB_VIDEO consists of 24 query videos and 12,790 videos downloaded from Google, YouTube, and Yahoo. There are 27% redundant videos that are duplicate or near duplicate in CC_WEB_VIDEO [1] [3] [4] [6] [7] [12] [14].

Jingkuan Song et al. created a larger video dataset UQ_VIDEO by adding CC WEB VIDEO to their video dataset downloaded from YouTube [4] [7] .

About 100 hours of video materials coming from different sources like TV archives, web video clips, movies with different bitrates, different video format, and different resolutions are contained in MUSCLE-VCD-2007. This dataset contains, a set of original videos and their corresponding transformed videos for the evaluation of copy detection algorithms. [7] [8] [9][19].

In CC WEB VIDEO the near-duplicate transformations are all done by the real Web users, which reflects the real user behavior on generating near-duplicates whereas in MUSCLE-VCD-2007 dataset, the near duplicates are produced artificially by using video edition tools[19].

5 CONCLUSION

Near Duplicate Video Retrieval is playing an important role in every video application domain. The need of the same is growing further stronger, with the increasing drift of online video applications. As per the existing literature studied in this paper, global features are popular for fast retrieval whereas less accurate for results. Whereas, local features are more accurate for results. Combination of both local and global features are proved more effective for NDVR with scene duplicates, in which two videos are presenting the same scene with different angles or viewpoints.

References

[1] Chien-Li Chou, Hua-Tsung Chen and Suh-Yin Lee," Pattern-Based Near-Duplicate Video Retrieval and Localization on Web-Scale Videos", IEEE Transactions on Multimedia, Vol. 17, No. 3, March 2015, pp.382-395.

[2] X. Wu, C. W. Ngo, A. Hauptmann, and H. K. Tan, Real-time near duplicate elimination for web video search with content and context, IEEE Trans. Multimedia, vol. 11, no. 2, pp. 196207, Feb. 2009.

[3] Xiushan Nie, Xiaoyu Li, Jiande Sun, Yilong Yin, " UFvH: Unified Feature Video Hashing for Near-Duplicate Video Retrieval", VSCC17, ACM,October 23, 2017,pp. 17-23.

[4] Yanbin Hao, Tingting Mu, Richang Hong, Meng Wang," Stochastic Multiview Hashing for Large-Scale Near- Duplicate Video Retrieval", IEEE Transactions On Multimedia, vol. 19, no. 1, January 2017, pp. 1-14.

[5] Prinka, Vikas Wasson, "An efficient Content Based Image Retrieval Based on Speeded up Robust Features (SURF) with Optimization Technique", 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, pp. 730-735.

[6] Kordopatis-Zilos, G. , Papadopoulos, S. , Patras, I., and Kompatsiaris, Y. , 2017, Near- Duplicate Video Retrieval by Aggregating Intermediate CNN Layers, Springer International Publishing, Cham, Switzerland, pp. 251263.

[7] Chien-Li Chou, Hua-Tsung Chen, Chun-Chieh Hsu, Chien-Peng Ho, and Suh-Yin Lee," Near-Duplicate Video Retrieval By Using Pattern-Based Prefix Tree And Temporal Relation Forest"

[8] Chien-Li Chou, Hua-Tsung Chen, Chun-Chieh Hsu, Chien-Peng Ho, and Suh-Yin Lee, "Near-Duplicate Video Retrieval By Using Pattern-set Based Dynamic Programming"

[9] S. R. Shinde, G. G. Chiddarwar, "Recent Advances in Content Based Video Copy Detection", IEEE International Conference on Pervasive Computing, 2015.

[10] A. H. Vardhan, N. K. Verma, R. K. Sevakula, A. Salour, "Unsupervised approach for object matching using Speeded Up Robust Features", 2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1-8, 2015.

[11] Tang-You Chang, Shen-Chuan Tai, Guo-Shiang Lin, "Manipulation classification for near-duplicate videos", IEEE International Conference on Consumer Electronics-Taiwan 2016, (ICCE-TW), Nantou, 2016, pp. 1-2.

[12] Y. Cai and L. Yang, "Large-Scale Near-Duplicate Web Video Retrieval: Challenges and Approaches," in IEEE MultiMedia, vol. 20, no. 2, pp. 42-51, April-June 2013.

[13] Yan Ke, Rahul Sukthankar, Larry Huston, Yan Ke, Rahul Sukthankar," Efficient near-duplicate detection and sub-image retrieval", ACM Multimedia, MM04,October 10-16, 2004.

[14] Lifeng Shang, Linjun Yang, Fei Wang, Kwok-Ping Chan, Xian-Sheng Hua, "Real-time Large Scale Near- duplicate Web Video Retrieval", Proceedings of the 18th ACM international conference on Multimedia, October 25 - 29, 2010, pp. 531-540.

[15] J. Li, "Parallel two-class 3D-CNN classifiers for video classification," 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen, 2017, pp. 7-11.

[16] David G Lowe, Distinctive image features from scale-invariant key points, in International Journal of Computer Vision, vol. 60, no. 2, 2004, pp. 91110.

[17] Hana Gharbi, Sahbi Bahroun, Mohamed Massaoudi and Ezzeddine Zagrouba, " Key Frames Extraction Using Graph Modularity Clustering For Efficient Video Summarization", ICASSP 2017, IEEE, 2017, pp. 1502-1506.

[18] https://fortunelords.com/youtube-statistics/

[19] Jiajun Liu, Zi Huang, Hongyun Cai, Heng Tao Shen, Chong Wah Ngo, Wei Wang, "Near-Duplicate Video Retrieval: Current Research and Future Trends", ACM Computing Surveys, Vol. 45, No. 4, Article 44, Publication date: August 2013, pp - 44:1-44:23.