ISSN:0975-9646

Survey on Web Scale Based Near Duplicate Video Retrieval

Bhosale Varsharani[#1], D.A.Phalke[*2].

ME-II Student Savitribai Phule Pune University,
D.Y.Patil Collage of Engineering Pune, Maharashtra, India[#1].

Assistant Professor Savitribai Phule Pune University,
*D.Y.Patil Collage of Engineering, Pune, Maharashtra, India[*2].*

***Abstract*:— A large number of videos are uploaded and shared on social websites in each single day. There are a large number of near-duplicate videos (NDVs) on the web are generated in different ways, such as simple reformatting, to different transformations, editions, and mixtures of different effects. The Internet is over flowing with near-duplicate videos, the video duplicates connected with visual and temporal transformations and post productions. Two basic issues, copyright infringement and search result redundancy, are present currently. To overcome these issues, a spatiotemporal pattern-based approach under the hierarchical filter-and-refine structure for efficient and successful near-duplicate video retrieval and localization can be used. As we survey the works in near-duplicate video retrieval, we investigate existing variants of the definition of near-duplicate video retrieval and near duplicate video detection, describe a generic framework and summarize related work.**

Keywords— Near-duplicate video localization, near-duplicate video retrieval, video copy detection, web-scale video analysis

I. INTRODUCTION

The video related services, such as video sharing, monitoring, advertising and recommendation. The video related activities are becoming interesting such as uploading, downloading, Commenting, and searching. A large number of videos are uploaded and shared on social websites in each single day. There are a large number of near-duplicate videos (NDVs) on the Web are generated in different ways, such as simple reformatting, to different acquisitions, transformations, editions, and mixtures of different effects.

User's time spent on video capturing, editing, uploading, searching, and viewing has increased to an large or huge level. The massive publishing and sharing of videos gives rise to the existence of an already large amount of near-duplicate content. This requires urgent demands on near-duplicate video retrieval as a key role in tasks such as video search, video copyright protection, video recommendation, and many more. Therefore near duplicate video retrieval has recently attracted a lot of attention. Duplicates refers to videos that are semantically and visually identical, which means that they have exactly the same story, scenario, etc. Video which are duplicate copies that share exactly the same semantics and scenes with an origin, but differ in visual presentations A segment of video derived from another video, usually by using various transformations such as addition, deletion, modification like color, contrast, encoding. The same scene may differ slightly near the duplicate shares. Near-duplicates videos include minor misalignments. Near-duplicate videos normally present large semantic and visual misalignment in the level of whole videos.

With the surrounding current online video applications and services, copyright video products are easier to access. Thus the copyright video products are protected or exposed to severe risk of being used for unauthorized copying, editing, and redistribution. Content providers are find their video products uploaded in video sharing Web sites without their permission. This may cause financial loss to content providers by decreasing the number of potential customers who initially want to purchase their video products [2].

Recent advances in video technologies have made video editing software easier and less expensive to access and consequently it has granted more possibility of producing video editions with significant variations to their corresponding original copies. Nowadays, user-generated videos may differ from the original ones in many different ways, such as motion, evaluate, glow, subtitle, frame rate, or include additional logo, banner, scene cut, merging, etc. The incapability of predicting which changes could be made to the original copyright textile increases the difficulty and effort to detect these violations, making the assignment more complex than it used to be. For example, due to the open nature of the Internet, a copyright movie trailer published by its content provider on YouTube may result in the spawn of hundreds of near-duplicates editions which may be republished in different ways by Internet users. To eliminate these drawback or violations is a current need of the content provider and video service provider. The issue of copyright is raised by requirement of detecting various forms of near-duplicates including copies.

As bandwidth available to regular users is increasing, video is becoming one of the fastest growing types of data on the Internet. After the modification the user can send the web videos easily and distribute them again. Current web video investigate outcome rely entirely on text keywords or user-supplied tags. A search on typical popular video often returns many duplicate and near-duplicate videos in the top results. Due to the large variety of near-duplicate web

videos ranging from simple formatting to complex editing, near-duplicate discovery remains a difficult issue [15].

A spatiotemporal pattern-based approach, which is capable of all fully NDVR, partial NDVR, and NDVL, under the hierarchical filter-and-refine framework. Given a reference database and a query video, aim at not only retrieving both fully and partial near-duplicate videos but also localizing precise positions of near-duplicate segments. Firstly low-level features of video frames are symbolized, and sequences of symbols are collected to form index patterns and m-keyframe patterns by sliding windows of different sizes. In the filter stage a spatiotemporal indexing structure utilizing index patterns, termed Pattern-based Index Tree is used to fast filter out non-near-duplicate videos and to retain more partial near-duplicate videos than video-level methods. The PI-tree is more efficient than frame-level methods. In the refine stage, m-Pattern-based dynamic Programming is used to localize near-duplicate segments and to re-rank the results of the filter stage [1].

II. NDVR AND NDVD

After feature extraction is done for a video, the set of features are processed by a particular signature generation algorithm, with the purpose of reducing data to increase search response or improving features representativeness. Normally a type of signature is designed to work with one or many specific NDV types in specific applications, so that the characteristics of the NDV type are used during the signature generation for more accurate representation and better distinguishing power.

It is important to know the relation between Near-Duplicate Video Retrieval (NDVR) and Near-Duplicate Video Detection (NDVD). Retrieval is considered to the situation where a video database is established, and the user inputs a query video into the search interface. The method receives the query video, processes it into features and then signatures, at a final point it matches the query signatures with those in the database.

A ranking list is normally in ascending order by distance or descending order by similarity, is returned to the users. A place list against a true or false matrix that indicates the ND relationship between videos can properly describe the difference between their outcomes. NDVD aims at finding pairs or groups of NDVs within one or more given sets of videos. NDVD and NDVR share many methods and processes, particularly for their feature mining, signature generation, and matching. The difference is that NDVD requires the excessive number of combinations between videos when performing detection, so the time consumed by the detection is relatively long, compared to NDVR[2].

The most important part of the accessible work on NDVs focus on the retrieval task. Generally, there are two techniques used for video copy detection:

1. Digital watermarking
2. Content-based approach.

1. Watermarks

Watermarks are used to introduce an invisible signal into a video to ease the detection of illegal copies. The photographers use this technique widely. Introducing a watermark on a video such that it is simply seen by an audience allows the content creator to detect easily whether the image has been copied. The limitation of watermarks is that if the original image is not watermarked, then it is not possible to know whether other images are copies.

A digital watermark is a category of sign embedded in a noise-tolerant signal such as an audio, video or image data. It is specifically used to recognize ownership of the copyright of such signal. Watermarking is the practice of defeating digital information in a carrier signal. The hidden information should, but does not need to, contain a relation to the carrier signal. Digital watermarks may be used to verify the authenticity or integrity of the carrier signal or to show the identity of its owners.

It is mainly used for tracing copyright infringements and for authentication. Watermarking is about strength against possible attacks, watermark need not be buried [14]. Video is one of the most popular data shared in the Web, and the protection of video copyright is one the interest. In the paper copyright protection on the Web a hybrid digital video watermarking scheme a inclusive approach for protecting and managing video copyrights in the Internet with watermarking techniques is presented. A original mixture digital video watermarking method with twisted watermarks and error correction codes can be used. As a process of intellectual assets protection, digital watermarks have recently moved major interest and become a very active area of research. Video watermarking presents a number of problem not present in image watermarking. Due to the huge quantity of data and natural redundancies between frames, video signals are highly susceptible to attacks such as frame averaging, frame dropping, frame swapping etc.

With a survey of the current watermarking technologies, it is noticed that none of the existing schemes is capable of resisting all attacks. Analyzing the strengths of different watermarking schemes and apply a hybrid approach to form a super watermarking scheme that can resist most of the attacks.

- The one of is a visual-audio hybrid watermarking scheme. As videos consist of both video and audio channels, the robustness of the scheme can be enhanced by including an audio watermark. Also errors correcting codes is embedded a video watermark as a watermark embedding in audio channel and refine the retrieved watermark during the watermark detection.
- The second approach is another hybrid with different watermarking schemes with two alternatives independent watermarking schemes which embeds the watermarks into the frames with special watermarking schemes in different scenes, or reliant watermarking schemes which embeds the watermarks serially in a frame with different watermarking schemes.

2. Content-based approach

In video copy detection much more data has to be processed than in image copy detection. Thus ordinal

signature is a popular means of video copy detection. Approximated string matching, which can deal with slight frame-rate changes, is then used for near duplicate detection.

Content-based video indexing and retrieval have a wide range of applications such as quick browsing of video folders, analysis of visual electronic commerce such as analysis of interest trends of users selections and orderings, analysis of correlations between advertisements and their effects, remote instruction, digital museums, news event analysis, intelligent management of web videos which is useful video search and harmful video tracing, and video surveillance.

A video may have an auditory direct as well as a image channel. The accessible information from videos includes the following:

1. Video metadata, which are tagged texts fixed in videos, usually including title, summary, date, actors, producer, broadcast duration, file size, video format, copyright, etc.
2. Audio information from the auditory channel.
3. Transcripts i.e. speech transcripts can be obtained by speech recognition and caption texts can be read using optical character recognition techniques.
4. Visual information contained in the images themselves from the visual channel. If the video is included in a web page, there are usually web page texts associated with the video[8].

Video information indexing and retrieval are required to describe, store, and organize multimedia information and to help people in finding multimedia resources conveniently and quickly. Videos have the characteristics as much richer content than individual images, huge amount of raw data, very little prior structure. These uniqueness make the indexing and recovery of videos quite complex. Before the video databases have been quite minute and indexing and recovery have been based on keywords annotated manually. More recently, these databases have be- come much larger and content-based indexing and retrieval are required, based on the automatic analysis of videos with the minimum of human participation. So generally the near duplicate video retrieval approaches focus on the visual contents of videos [16].

One of the most common forms of copyright infringement is to illegally copy a video or video segment on the video sharing sites. In these cases, any auxiliary textual information associated with the video would have no use when it comes to determine if the video has been illegally copied. A promising approach to tackling this problem is to look directly into the visual content of the video so called content- based copy detection (CBCD).

Unlike watermarking, CBCD extracts a small number of features from the original video content instead of inserting some external information (watermarking) into the video content prior to the distribution. As has been recognized by others, a major challenge of CBCD lies in the truth that a copy is not essentially an alike or a near duplication, but rather a photometric or geometric transformation of the original video [16]

III. GENERAL FRAMEWORK OF NDVR

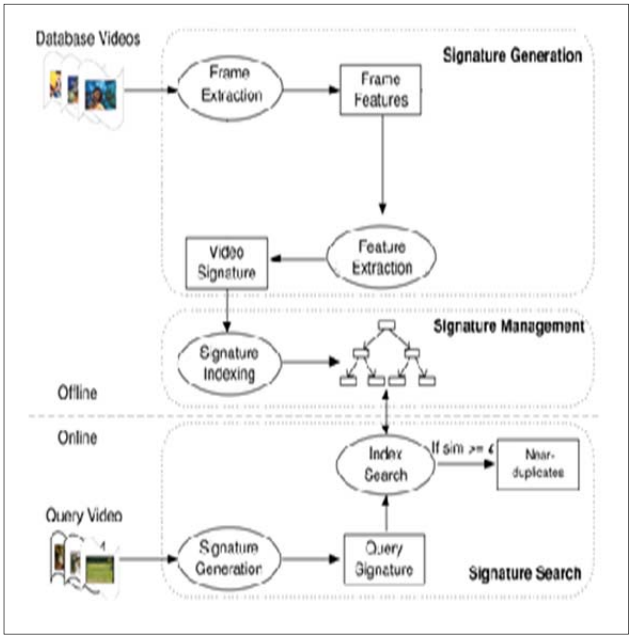

Figure 1: Generic Framework

Figure1 shows a general framework for an NDVR system, including three major components, signature generation, signature management, and signature search.

1. First, a video is divided into a sequence of keyframes extracted by time sampling or shot boundary detection algorithms like histogram based, motion vector based algorithms. These keyframes are represented by their visual features, such as color histogram, local points, local binary pattern.

2. The sequence of the keyframes features is then considered as the signature of the video.

3. To find out NDVR, the system needs to compare the similarity between the signatures of the query video and each dataset video and return the dataset videos which are most similar to the query example. Normally for a given database video, in the signature generation component, low-level features are first extracted. Based on this video summarization can be further performed to generate compact and distinctive video signatures to access or perform the retrieval. Now a-days video databases grow rapidly and the sizes get larger and larger.

To achieve fast recovery, in the signature management factor, the generated video signatures are organized by effective indexing structures which avoid extensive data accesses. The signature generation and indexing of database videos are the offline processes.

Given a query video, in the signature search component, its signature is first generated like that in the signature generation component. The gained query signature is then searched in the database index to find NDVs, based on signature matching. If the similarity between a database videos signature and the query signature is greater than or equal to a predefined similarity threshold the database video is called as an NDV of the query. The signature search component is an online process.

RELATED WORK

Existing works in the fields of NDVR can be classified into four categories: frame-level, spatiotemporal, video level, and hybrid hierarchical methods. Normally a type of signature is used to work with one or many specific NDV types in specific applications. So that the characteristics of the NDV type are considered during the signature generation for more accurate representation and better distinguishing power.

Frame-level methods require significantly high computation cost. As spatiotemporal methods are more effective and efficient than frame-level methods, they are still not well suited for web-scale applications. Video-level methods are faster but cannot achieve NDVL [2].

A. Frame-level signatures

Using frame-level local signatures, changes in the aspects of the videos like the angle of viewport when recording the video can be properly recognized where low matching speed is expected. It suits to applications that are not focus to speed but aims to find videos that are captured from different viewports about the same scene.

A frame-level local signature represents the local information on individual frames with local features like local keypoint descriptors. Such usage of local information in NDVR is very closely related to the near-duplicate image retrieval task. Local keypoints contain so much information that makes it not easy practical to directly perform a search.

After the detection of keypoints, these keypoints are organized by their corresponding frames, the same as it does in local keypoint-based image retrieval. The distances between frames are measured simply by a composition of the Euclidean distance between keypoint descriptors. A frame-level global signature represents a whole frame with a single signature. Due to the time complexity in matching a large number of local keypoints, a more practical way is to represent a frame with a single global signature which is more efficient to be matched. Standard representations like color histograms carry any local information. To consider local information, local keypoints are often used to generate global signatures.

Frame-level global signatures are a compromise between speed and accuracy. It stills keeps some ability of detecting aspect changes while the approaches using them run as fast as video-level signature-based approaches.

A key challenge to the successful detection of a copied video lies in the design of effective video content descriptor. Frame-level descriptors are the most popular and many such schemes have been proposed in the existing work. These descriptors can be classified into global and local descriptors.

- The global descriptors are generally more efficient to compute, more compact to storage, but less accurate in terms of their retrieval quality. On the other hand, local descriptors are relatively more robust to image transformations, such as occlusion, cropping, etc.
- Local descriptors based copy detection systems, a large number of descriptors each has to be individually searched. These approaches are therefore memory and computation demanding.

One novel frame level global descriptor for CBCD in which new descriptor is derived from the covariance of visually salient features. Compared to other approaches in the literature, this new descriptor is much more compact and is also very robust as well. It introduces and verifies that data independent fast standard transforms such as Hadamard Transform can be successfully used to replace conventional data dependent and computationally expensive transforms such as independent component analysis in the calculation of information theoretic based visual saliency map. It then introduces the covariance of visually salient uses a covariance matrix distance for image and video copy detection.

To compute the visual saliency map the saliency of a location is quantified by the self information of an m x n local image patch centered on that location. From the pixel vector of a local mxn patch and p(X) the probability density function of X, the saliency of the patch is calculated. Hadamard transform contains only binary coefficients (+1 and -1), the projections therefore only involve addition operation which is much simpler than doing real value multiplication. Hence using hadamard transform is computationally much more efficient than independent component analysis measured the true saliency (TS) as the percentage of saliency scores that are greater than a threshold T and coincide with the eye fixation points. Using the saliency score at location L(x, y), and L(x, y) is a small local window around (x, y), TS is calculated.

Visual saliency has recently attracted a lot of interests in the computer vision community and various methods have been developed to exploit visual saliency for various tasks such as object recognition. One of the key issues is how to use the visual saliency score to derive feature vectors to effectively describe the image contents.

Past methods have employed techniques such as sequential saccade based feature-vector representation which takes samples from randomly-selected locations in the saliency map. This kind of method is usually complex, not compact and random. A covariance matrix based descriptor is thus more effective. The salient features are those F(x, y) with a corresponding saliency score S(x, y) greater than a threshold and using this covariance matrix of the salient features is determined. To use the covariance matrices as the feature vectors, there is need a metric to measure the similarity of the features. As the covariance matrix forms a Riemannian manifold, the similarity of two covariance matrices can be computed by a distance [21].

B. Video-level signatures

Video-level signatures are generally working well in applications where the NDVs include changes of the video colors like color, contrast, encoding, etc. as well as the temporal orders of frames, and response speed is a key requirement. The abstraction of video colors makes it tolerant to minor color changes, while the droping of frame orders leads to robustness to any NDV that is created with temporal changes. A video-level signature represents a whole video with a single signature. Video-level signature

is considered as an efficient form of representing videos and widely used in many works. A great advantage of global signature lies in the small data size, thus it can be efficiently stored, managed, and retrieved, in the sense of disk space and computational time.

We can regard a video as a sequence of keyframes. Local signature is less robust to the changes in frame rate, video length, captions. While global signature is sensitive to changes in contrast, brightness, scale, rotation, camera viewpoint, Different forms which can only make use of single feature to learn the hash code of keyframes, while the individual structural information of each feature type is preserved in using multifeature hashing. Once the hash codes are obtained, only Hamming distance calculation is performed to compute the similarity between videos. Such an operation is very fast, making it possible to perform NDVR over large scale video datasets in real time. Hashing methods implement fast nearest neighbor search in sub-linear time by mapping highly similar data points together.

- Recent compact binary code approache as self-taught hashing are possible to perform real-time search due to the quick similarity computation by using bit XOR operation in the Hamming space. The primary challenge on binary code methods is how to generate the compact binary codes for the data points. A good code requires that similar data in the original space should be mapped into similar binary codes.
- Spectral hashing utilizes spectral graph partitioning in the learning phase to get the hash codes of training data, which is similar to self-taught hashing. But to calculate the binary codes for a new data point, spectral hashing assumes that the data are uniformly distributed in a hype-rectangle, which is very restrictive, and self-taught hashing has to learn new classification models based on the result of learning phase.
- Locality sensitive hashing (LSH) uses a family of locality sensitive hash functions composed of linear projection over random directions in the feature space. The intuition behind is that for at lease one of the hash functions, nearby data points have high probability of being hashed into the same state.
- MFH learn a series of s hash functions each of which generates one bit hash code for a keyframe according to the given multiple features. Each keyframe has s bits. Using the derived hash functions, each keyframe for a dataset video can be represented by the generated s-sized hash codes in linear time. In the second phase which is online, the query videos keyframes are also represented by s-sized hash codes mapped from the s hash functions. NDVR can be efficiently achieved where only efficient XOR operation on the hash codes is performed to compute the similarity between two videos [6].

A fingerprint is a content-based signature derived from a video or other form of a multimedia asset so that it specifically represents the video or asset. To find a copy of a query video in a video database, one can search for a close match of its fingerprint in the corresponding fingerprint database. Closeness of two fingerprints represents a similarity between the corresponding videos.

- Color-space-based fingerprints are used for video fingerprinting. They are mostly derived from the histograms of the colors in specific regions in time or space within the video. Since color features change with different video formats, these features have not been very popular. Another drawback of color features is that they are not applicable to black and white videos. For this reason, most of the video fingerprinting systems are designed so that they can be applied to the luminance (the gray level) value of the frames.
- Temporal fingerprints are extracted from the characteristics of a video sequence over time. These features usually work well with long video sequences, but do not perform well for short video clips since they do not contain sufficient discriminant temporal information. Because short video clips occupy a large share of online video databases, temporal fingerprints alone do not suit online applications.
- Spatial fingerprints are features derived from each frame or from a key frame. They are widely used for both video and image fingerprinting. There is a large body of research in the area of image fingerprinting and many researchers have extended the concepts developed for image fingerprinting to the video fingerprinting field. Spatial fingerprints can be further subdivided into global and local fingerprints. Global fingerprints depict the global properties of a frame or a subsection of it (e.g., image histograms), while local fingerprints usually represent local information around some interest points within a frame (e.g., edges, corners, etc.) [17].

C. Spatio Temporal Technique

By introducing temporal information into the video signatures, spatio-temporal signatures are often more robust to heavy changes in the colors or in the video geometry e.g. video color is shifted and video is cropped. However, it is more sensitive to temporal changes in the video, because such changes disrupt the very essential information based on which it is generated.

Spatio-temporal signatures represent videos with spatial and temporal information. In recent years, spatio-temporal signature has drawn attention in NDVR for its better invariance to viewpoint changes compared to global signature-based techniques and for its relatively better efficiency compared to local signature techniques.

This type of extraction methods focuses on the changes of frames, motion of pixels, or trajectory of interest points. By tracking changes of video content along the time axis, it is considered specially suitable for scene near-duplicates, in which the same scene is played but the viewpoint may vary due to differences introduced during capturing time. Many existing methods are also tested for partial near-duplicate detection [2].

A spatiotemporal pattern-based approach, which is capable of all fully NDVR, partial NDVR, and NDVL, under the hierarchical filter-and-refine framework. Given a

reference database and a query video, aim at not only retrieving both fully and partial near-duplicate videos but also localizing precise positions of near-duplicate segments.

Firstly low-level features of video frames are symbolized, and sequences of symbols are collected to form index patterns (I-pattern) and m-keyframe patterns (m-pattern) by sliding windows of different sizes. In the filter stage, a spatiotemporal indexing structure utilizing index patterns, termed Pattern-based Index Tree (PI-tree), is used to fast filter out non-near-duplicate videos and to retain more partial near-duplicate videos than video-level methods. The PI-tree is more efficient than frame-level methods. In the refine stage, m-Pattern-based Dynamic Programming is used to localize near-duplicate segments and to re-rank the results of the filter stage [1].

There are great redundancies among the frames in the same shot. Therefore, certain frames that best reflect the shot contents are selected as key frames to succinctly represent the shot. The extracted key frames contain as much salient content of the shot as possible and avoid as much redundancy as possible. The features used for key frame extraction generally include colors (particularly the color histogram), edges, shapes, optical flow, MPEG-7 motion descriptors such as temporal motion intensity and spatial distribution of motion activity, MPEG discrete cosine coefficient and motion vectors, camera activity, and features derived from image variations caused by camera motion.

Histogram of Oriented Optical Flow (HOOF) features used to represent human activities. These novel features are independent of the scale of the moving person as well as the direction of motion. Extraction of HOOF features does not require any prior human segmentation or background subtraction. First, optical flow is computed at every frame of the window. Each flow vector is binned according to its primary angle from the horizontal axis and weighted according to its magnitude. Finally, the histogram is normalized to sum up to 1[10]. To accelerate the frame similarity computation, keyframes containing similar visual features are clustered by K-means clustering, and each cluster is assigned a unique symbol. Keyframes within a cluster share the same symbol. Therefore, each reference video can be transformed into a sequence of symbols.

An index pattern consisting of two symbols p1p2 is generated by extracting two symbols by a window sliding over a symbol sequence with the window size two. Longer patterns may provide better retrieval accuracy but require more memory space and higher computational cost. For examples of I-pattern generation. For the video, we can generate 5 I-patterns EA, AC, CB, BA, and AF from the symbol sequence of EACBAF. The m-pattern is generated by applying a non-overlapping sliding window with a window size of on the symbol sequence of a video. A Pattern- based Index Tree, abbreviated as PI-tree used to fast filter out non-near-duplicate videos. A PI-tree is constructed by 2-level prefixed queues, and a symbol is regarded as a prefix of a prefixed queue.

CONCLUSION

There are different aspects of the near-duplicate video retrieval. The different techniques that has been developed and used in past years such as watermarking, content based retrieval. The massive publishing and sharing of videos has given rise to the existence of an already large amount of near-duplicate content. This gives urgent demands on near-duplicate video retrieval as a key role in novel tasks such as video search, video copyright protection, video recommendation, and many more Frame-level methods require significantly high computation cost. As spatiotemporal methods are more effective and efficient than frame-level methods, they are still not well suited for web-scale applications. Video-level methods are faster but cannot achieve NDVL. A spatiotemporal pattern-based approach under the hierarchical filter-and-refine framework is useful for near-duplicate video retrieval. Pattern-based Indexing Tree in the filter stage is capable of to efficiently filter out non-near-duplicate videos and substantially reduce the search space.

ACKNOWLEDGMENT

The authors would like to thanks the publishers and researchers for making their resources available. We also thanks the college authority for providing the required information the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] Chou, Chien-Li, Hua-Tsung Chen, and Suh-Yin Lee. "Pattern-Based Near- Duplicate Video Retrieval and Localization on Web-Scale Videos." Multimedia, IEEE Transactions on 17.3 (2015): 382-395.

[2] J. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo, and W. Wang, Near duplicate video retrieval: Current research and future trends, ACM Comput. Surveys, vol. 45, no. 4, pp. 123, Aug. 2013.

[3] 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.

[4] Kim, Changick, and Bhaskaran Vasudev. "Spatiotemporal sequence matching for efficient video copy detection." Circuits and Systems for Video Technology, IEEE Transactions on 15.1 (2005): 127-132.

[5] Hampapur, Arun, and Ruud M. Bolle. "Comparison of distance measures for video copy detection." null. IEEE, 2001.

[6] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, Multiple feature hashing for real-time large scale near-duplicate video retrieval, in Proc. 19th ACM Int. Conf. Multimedia, Nov. 2011, pp. 423432.

[7] Douze, Matthijs, et al. "INRIA-LEARs video copy detection system." TREC Video Retrieval Evaluation (TRECVID Workshop). 2008.

[8] BASHARAT, A., ZHAI, Y., AND SHAH, M. 2008. Content based video matching using spatiotemporal volumes. Comput. Vis. Image Understand. 110, 3, 360377.

[9] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, Video Copy Detection: A Comparative Study, in Proc. CIVR, 2007, pp. 371378.

[10] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recog., Jun. 2009, pp. 19321939.

[11] Cai, Yang, et al. "Million-scale near-duplicate video retrieval system." Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.

[12] CHUM, O. AND MATAS, J. 2010. Large-scale discovery of spatially related images. IEEE Trans. Pattern Anal. Mach. Intell. 32, 2, 371377.

[13] HUANG, Z., HU, B., CHENG, H., SHEN, H. T., LIU, H., AND ZHOU, X. 2010a. Mining near-duplicate graph for cluster-based reranking of web video search results. ACM Trans. Inf. Syst. 28, 4, 22.

[14] P. P. W. Chan, M. R. Lyu, and R. T. Chin, Copyright protection on the web: A hybrid digital video watermarking scheme, in Proc. 13th Int. World Wide Web Conf. Alternate Track Papers Posters, May 2004, pp. 354355.

[15] Tan, Hung-Khoon, et al. "Scalable detection of partial near-duplicate videos by visual-temporal consistency." Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009.

[16] BASHARAT, A., ZHAI, Y., AND SHAH, M. 2008. Content based video matching using spatiotemporal volumes. Comput. Vis. Image Understand. 110, 3,360377.

[17] M. M. Esmaeili, M. Fatourechi, and R. K. Ward, A robust and fast video copy detection system using content-based fingerprinting, IEEE Trans. Inf. Forensics Security, vol. 6, no. 1, pp. 213226, Mar. 2011.

[18] Zhang, Dong-Qing, and Shih-Fu Chang. "Detecting image near-duplicate by stochastic attributed relational graph matching with learning." Proceedings of the 12th annual ACM international conference on Multimedia. ACM, 2004.

[19] Cherubini, Mauro, Rodrigo De Oliveira, and Nuria Oliver. "Understanding near-duplicate videos: a user-centric approach." Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009.

[20] POULLOT, S.,CRUCIANU,M., AND BUISSON,O. 2008. Scalable mining of large video databases using copy detection. In Proceedings of the 16th ACM International Conference on Multimedia (MM08). 6170.

[21] L. Zheng, G. Qiu, J. Huang, and H. Fu, Salient covariance for near duplicate image and video detection, in Proc. 2011 IEEE Int. Conf. Image Process., Sep. 2011, pp. 25372540.

[22] Hampapur, Arun, Kiho Hyun, and Ruud M. Bolle. "Comparison of sequence matching techniques for video copy detection." Electronic Imaging 2002. Inter- national Society for Optics and Photonics, 2001.