Near-Duplicate Video Retrieval: Current Research and Future Trends

JIAJUN LIU, ZI HUANG, HONGYUN CAI, HENG TAO SHEN, The University of Queensland
CHONG WAH NGO, City University of Hong Kong
WEI WANG, The University of New South Wales

The exponential growth of online videos, along with increasing user involvement in video-related activities, has been observed as a constant phenomenon during the last decade. User's time spent on video capturing, editing, uploading, searching, and viewing has boosted to an unprecedented level. The massive publishing and sharing of videos has given rise to the existence of an already large amount of near-duplicate content. This imposes 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. Driven by its significance, near-duplicate video retrieval has recently attracted a lot of attention. As discovered in recent works, latest improvements and progress in near-duplicate video retrieval, as well as related topics including low-level feature extraction, signature generation, and high-dimensional indexing, are employed to assist the process.

As we survey the works in near-duplicate video retrieval, we comparatively investigate existing variants of the definition of near-duplicate video, describe a generic framework, summarize state-of-the-art practices, and explore the emerging trends in this research topic.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*; *indexing methods*

General Terms: Design, Algorithms, Experimentation

Additional Key Words and Phrases: Near-duplicate video, indexing, video retrieval

ACM Reference Format:
Liu, J., Huang, Z., Cai, H., Shen, H. T., Ngo, C. W., and Wang, W. 2013. Near-duplicate video retrieval: Current research and future trends. ACM Comput. Surv. 45, 4, Article 44 (August 2013), 23 pages.
DOI: http://dx.doi.org/10.1145/2501654.2501658

1. INTRODUCTION

Emerging online video-related services such as video sharing, video broadcasting, video recommendation and so on, increasingly bring user interests and participation to video-related activities like editing, publishing, searching, streaming, and viewing. According to a report by comScore.com, a leading company in measuring the digital world, 76.8% of the total U.S. Internet audience viewed online videos and these users viewed 14.8 billion online videos in January 2009 alone, with an average view count of 101 videos and an average view time of 356 minutes per user. It also shows an evident rising demand for online videos, supported by the facts that the view count of January 2009 increased by 4% and average view time up by 15% compared to November 2008.

Authors' addresses: J. Liu, Z. Huang, H. Cai, and H. T. Shen (corresponding author), School of Information Technology and Electrical Engineering, The University of Queensland, Australia; email: shenht@itee.uq.edu.au; C. W. Ngo, Department of Computer Science, City University of Hong Kong, Hong Kong; W. Wang, School of Computer Science and Engineering, The University of New South Wales, Australia.

DOI: http://dx.doi.org/10.1145/2501654.2501658

Evidently, the rapid growth of video-related applications and services subsequently leads to the continuous exponential growth of online video content as noted in Tan et al. [2009], Wu et al. [2007a], and Shen et al. [2005]. To be specific, according to comScore, the number of total Internet videos climbed from 12,677,063,000 of November 2008 to 14,831,607,000 of January 2009, representing 17% growth within 3 months. The average duration per video incremented from 3.2 minutes to 3.5 minutes within the same period. Both figures are expected to keep growing continuously. Clearly, mainstream media is moving to the Web, with their video products provided to consumers by Web applications and services. To name a few, music videos are being published by musician's official channels in YouTube, TV programs are being broadcast by TV broadcasting company's Web sites, movie trailers are being previewed with theme Web sites like the Avatar movie Web site. It is also common for the publisher, user, or third-party business entity to edit and republish modified copies of the original video. On the other hand, similar videos regarding the same event may be acquired and published by multiple publishers or individuals.

All these phenomena contribute to the substantial percentage of near-duplicates in online videos, which are referred as Near-Duplicate Videos (NDVs) in Cherubini et al. [2009], Tan et al. [2009], Wu et al. [2007a], and Shen et al. [2007]. The ratio of NDVs even reaches a surprising 93% for some of the user queries searched from the Web [Wu et al. 2007a]. The presence of massive NDV data imposes heavy demands on Near-Duplicate Video Retrieval (NDVR) as it is crucial to many novel applications such as copyright violation detection, video monitoring, Web video reranking, video recommendation, etc. To list a few, a typical scenario could be that a content provider publishes a copyright video in its own YouTube channel and it wants to enforce its copyright protection by detecting and removing illegal editions of its original copy from YouTube. Another scenario could involve a company, who has invested in a TV commercial, wishing to monitor that the commercial is being broadcast for the right counts during the right time period. Both tasks require NDVR techniques to automatically achieve results. In the following sections we describe a number of scenarios in detail to show how NDVR forms the foundation of some of the applications mentioned before.

Copyright Protection. With the circumstances surrounding current online video applications and services, copyright video products are easier to access than ever. Concurrently, copyright video products are exposed to severe risk of being compromised by unauthorized copying, editing, and redistribution. Content providers often find their video products uploaded in video sharing Web sites without their permission. This may cause further financial loss to content providers by decreasing the number of potential customers who initially want to purchase their video products. 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 resolution, contrast, brightness, subtitle, frame rate, or include additional logo, banner, scene cut, concatenation, etc. The incapability of predicting which changes could be made to the original copyright material increases the difficulty and effort to detect these violations, making the task more complicated 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 violations is a concrete need of the content provider and video service provider. The new requirements of copyright protection bring up the issue of detecting various kinds of near-duplicates including copies [Cherubini et al. 2009; Assent and Kremer 2009; Shen et al. 2007;

Law-To et al. 2007, 2006]. With appropriate NDVR techniques, video service and content providers can potentially eliminate copyright violations by checking uploaded videos and matching them to a list of protected videos.

Video Monitoring. Many online video streaming and live broadcasting services are available now. To find partial videos of interest, such as commercials and inappropriate content, is also important to many applications [Huang et al. 2010b, 2009b; Tan et al. 2009; Yan et al. 2008]. For example, when a company invests in a TV commercial it wants to guarantee the commercial is being broadcast in a correct frequency during an expected time period as contracted with the broadcasting corporations. Manual monitoring is labor intensive and expensive. NDV detection techniques can be used to fulfill the needs of video stream monitoring under such circumstances. The company can input the commercial as a query, and then monitor the TV broadcast. When the commercial appears in the TV program, NDV detection methods will be able to detect the occurrence and record the play count, time spot, and frequency, which is crucial information for the company to protect its investments.

Video Reranking. NDVs take a huge part in video search experience. Very often one wants to see novel/diverse videos from the same keywords but ends up with many duplicates or near-duplicates on the top of the result list. Search service providers want to employ effective techniques to improve ranking results with appropriate handling of NDVs to improve novelty of search returns. NDVR provides an effective way of improving this experience by removing NDVs from top of the list [Wu et al. 2007a] or by clustering NDVs as groups for reranking and rerendering the original results [Huang et al. 2010a]. With the former solution, the novelty of results within the ranking list is greatly increased but at the same time videos which are of users' interests can possibly be repositioned too far from the top, as users often regard NDVs as a complementary source of information where they can obtain additional information like different subtitles, audio, resolutions, and formats [Cherubini et al. 2009]. For the latter, novel results are highlighted while NDVs are grouped into the same cluster to improve novelty of the result list.

Video Recommendation. It has been observed that the probability is high that when a user likes a video, the user will also be interested in other near-duplicate videos [Cherubini et al. 2009]. Inspired by this implicit user need, video recommendation systems can be improved to enhance user experience in video sharing Web sites. However, current personalized recommendation systems tend to match only text similarity derived from titles, comments, descriptions, and other metadata between watched clips and database clips. This introduces considerable noise as a result of totally different videos having similar descriptions. For instance, when a user is watching a video with the title "Cute Dog" which shows a Labrador playing tricks in a pet competition on YouTube, the recommendation system may not be capable of precisely finding other videos containing dog playing scenes during the same competition just because they are not titled with "Cute Dog". The reason is that the recommendation system is text based while a large number of videos are uploaded without informative or precise text metadata. Introducing NDVR techniques to such systems can increase the precision of recommendations by combining textual information with content similarities [Cherubini et al. 2009]. In addition to these possible enhancements, social features [Zhao et al. 2010; Cherubini et al. 2009] can also be included to further improve accuracy.

Video Thread Tracking. The scenario of conveniently browsing NDVs regarding the same event but taken by miscellaneous mass media shows potential significance, as observed in Zhao et al. [2007]. This is especially useful for online news videos about the same event, for the reason that the videos often show very similar visual images and clips

Table I. Various Definitions of NDV

[Wu et al. 2007a]	Identical or approximately identical videos close to the exact duplicate of each other, but different in file formats, encoding parameters, photometric variations (color, lighting changes), editing operations (caption, logo and border insertion), different lengths, and certain modifications (frames add/remove).
[Shen et al. 2007]	Clips that are similar or nearly duplicate of each other, but appear differently due to various changes introduced during capturing time (camera view point and setting, lighting condition, background, foreground, etc.), transformations (video format, frame rate, resize, shift, crop, gamma, contrast, brightness, saturation, blur, age, sharpen, etc.), and editing operations (frame insertion, deletion, swap and content modification).
[Basharat et al. 2008]	Videos of the same scene (e.g., a person riding a bike) varying viewpoints, sizes, appearances, bicycle type, and camera motions. The same semantic concept can occur under different illumination, appearance, and scene settings, just to name a few.
[Cherubini et al. 2009]	NDVs are approximately identical videos that might differ in encoding parameters, photometric variations (color, lighting changes), editing operations (captions, or logo insertion), or audio overlays. Identical videos with relevant complementary information in any of them (changing clip length or scenes) are not considered as NDVs. Two different videos with distinct people, and scenarios were considered to be NDVs if they shared the same semantics and none of the pairs has additional information.

but with quite different interpretation via its audio or subtitles. A good example may be a government press conference. Media at divergent stances would give very different interpretations of the event but their videos' visual content would very likely be close to each other. Discovering and presenting such videos together helps people get a clearer view of the event because of the different perspectives included.

As we have shown earlier, Near-Duplicate Video Retrieval (NDVR), which plays a key role in many new video applications, remains an emerging research problem that has great value. It is important to provide an overview and some insightful discussions with respect to its scope, efficiency, and accuracy. Progress has recently been made in various sectors. In this article we will look into the current state of the problem by exploring various near-duplicate definitions and solutions, along with future trends.

The remainder of this article is organized as follows. First we present existing variants of the NDV definition and compare NDVR with similar tasks to clarify its scope in Section 2, followed by a generic NDVR framework in Section 3. Section 4 will provide a review of state-of-the-art methods with related topics. At the end, a summary and an outlook towards future trends are presented in Section 5.

2. SCOPE

2.1. Various Definitions of NDV

Being a relatively new topic in the research society, NDVR has a variety of understandings and definitions on NDV. Representative definitions include those defined in Wu et al. [2007a], Shen et al. [2007], Basharat et al. [2008], and Cherubini et al. [2009], as shown in Table I.

From Table I it can be observed that Wu et al.'s [2007a] definition of NDV has the strictest scope among the definitions, in which only identical and approximately identical videos for the exactly same story with minor photographic differences and editions are regarded as NDVs. This is technically similar to the definition of video copy in Law-To et al. [2007], in which video content is semantically identical but differs only in photometric or geometric transformations. By contrast the definition in Shen et al. [2007] extends this to videos with the same semantic concept but might differ in various aspects introduced during capturing time, including photometric or geometric settings. Basharat et al. [2008] suggest an even looser definition in which videos with the same

semantic concept under certain circumstances are NDVs. In order to involve human perception into the definition of NDV, Cherubini et al. [2009] have done a large-scale online survey whose results indicate the real users' concept of NDV. While the human perception of NDV matches many of the features presented in its technical definitions with respect to manipulations of visual content in Wu et al. [2007a] and Shen et al. [2007], similar videos differing in overlaid or added visual content with additional information were not perceived as near-duplicates. Conversely, two different videos with distinct people and scenarios were considered to be NDVs because they shared the same semantics, provided that none of near-duplicates has additional information. It is evidenced that users perceive as NDVs those which are both visually similar and semantically identical. This clearly differs itself from conventional Content-Based Video Retrieval (CBVR), because for the latter the content similarity is the only factor it needs to consider, while for the former both content and semantics need to be examined.

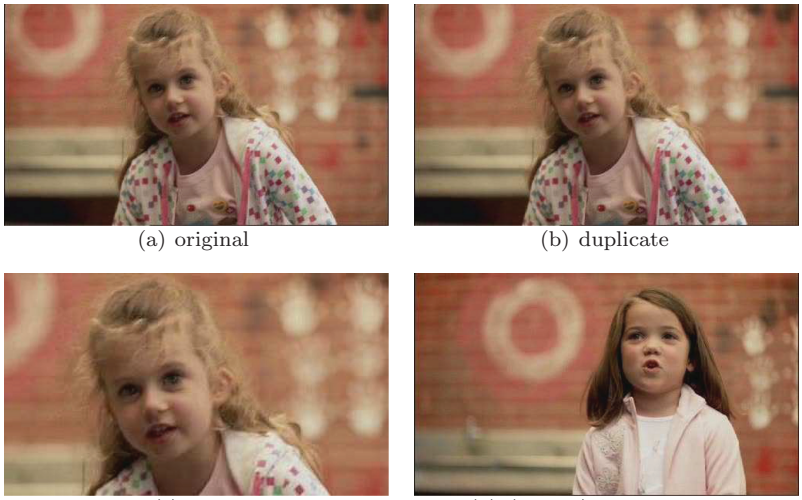
There is not a commonly recognized *unified definition* so far. At this current stage, definitions in Wu et al. [2007a] and Shen et al. [2007] appear in most of the existing works. However, there is a convergence that NDVs should carry highly similar visual information indicating the same semantics. Therefore, NDVR can be regarded as the bridge between traditional content-based similarity retrieval (i.e., videos should have similar visual content regardless of semantics) and semantic-based video retrieval (i.e., videos should have relevant semantics regardless of visual content).

2.2. (Partial) Near-Duplicates, Copies and Duplicates

One may wonder at the differences among the concepts: partial near-duplicates, near-duplicates, copies, and duplicates. In fact they share similar approaches, even though their objectives are not closely alike. We hereby give a brief clarification for each of them, from strict to loose, as follows.

(1) Duplicates refers to videos that are semantically and visually identical, which means they have exactly the same story, scenario, etc.
(2) Copies share exactly the same semantics and scenes with an origin, but differ in visual presentations. They can be the results of a large variety of photometric and geometric transformations and editions. Videos from the same origin but with changes in color, brightness, contrast, frame rate, etc., or with the addition of banners, logos, etc., fall into this category. In summary, they cast exactly the same story and the same scene but with visual differences. For instance, the TRECVID project [Over et al. 2011] defines a copy to be "a segment of video derived from another video, usually by means of various transformations such as addition, deletion, modification (of aspect, color, contrast, encoding, . . .), camcording, etc". Eight transformations, including simulated camcording, picture in picture, insertions of pattern, strong re-encoding, change of gamma, decrease in quality, postproduction, and combination of three randomly selected transformations, were selected in TRECVID 2011 Content-Based Copy Detection task.
(3) Near-duplicates share the same semantics and their scenes may differ slightly. Minor misalignments among near-duplicates can be tolerated. While partial near-duplicates, as a special case of near-duplicates, normally present large semantic and visual misalignment in the level of whole videos, they share one or more near-duplicate segments (or even near-duplicate regions in the image plane) with each other, for example, two long videos showing totally different stories but with a very similar segment (or regions) including scenic pictures for the same place of interests.

In Figure 1 we demonstrate the differences among the concepts of video duplicates, copies, and (partial) near-duplicates by showing and analyzing four variants of a TV



(a) original (b) duplicate

(c) copy (d) (partial) near-duplicate

Fig. 1. The concept of video duplicate, copy, and near-duplicate.

commercial from the dataset used in Huang et al. [2009a]. Clearly video 1(b) is identical in all aspects with its origin video 1(a), hence it is regarded as a duplicate video. For video 1(c), the content has been obviously changed by cropping part of the video, however, the semantics and the scene are exactly the same. Since it is only affected by geometric modification, it is categorized into a copy of video 1(a). However, in video 1(d) the foreground girl has been replaced by another, while it is still advertising the same product. The scene of the video is slightly changed with different foreground. Given such characteristics, video 1(d) is a near-duplicate to video 1(a). Both are two versions of the same commercial.

2.3. Retrieval vs. Detection

It is important to justify the relation between NDVR and Near-Duplicate Video Detection (NDVD). Retrieval refers to the situation where a video database is established, and the user inputs a query video into the search interface. The system receives the query video, processes it into features and then signatures, finally it matches the query signatures with those in the database. A ranking list, normally in ascending order by distance or descending order by similarity, is returned to the users. A ranking list versus a true or false matrix that indicates the ND relationship between videos can properly describe the difference between their outcomes. In fact, the major part of the existing work on NDVs focus on the retrieval task.

For NDVD, two cases are discussed. In the first case, NDVD aims at finding pairs (groups) of NDVs within one or more given sets of videos [Zhao and Ngo 2009; Wu et al. 2009a; Tan et al. 2008]. This task is motivated by the need of database cleansing, topic tracking, copyright infringement detection, recommendation, and so on. Actually in this case, NDVD and NDVR share many techniques and procedures, especially for their feature extraction, signature generation, and matching. The difference is that NDVD suffers from the excessive number of combinations between videos when performing detection, so the time consumed by the detection is relatively long, compared to NDVR.

The second case is real-time online detection, where the query is in the form of a continuous video stream. Such scenario suggests that, unlike retrieval, the system will not obtain the full video nor the complete information about the video because buffering

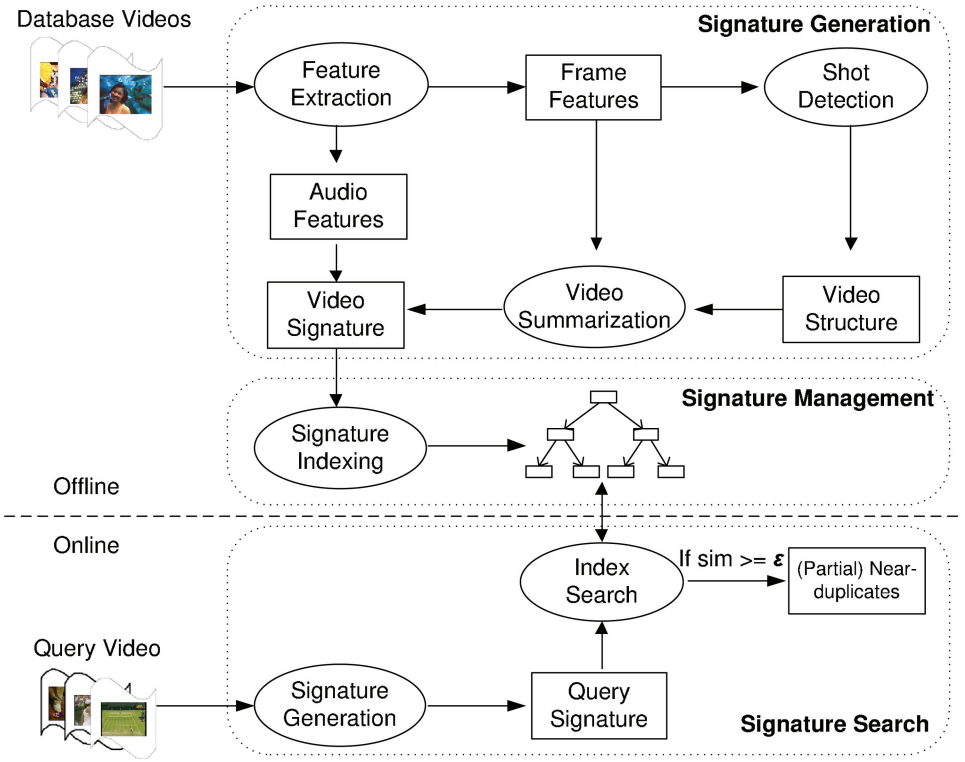


Fig. 2. A general framework of NDVR.

the whole video is not feasible since the stream itself may never stop, for example, in the case of monitoring a TV channel or a surveillance camera. From the detection system perspective, at each particular spot of time, the data and information will be fragmented and incomplete. Another challenge is that due to the continuous arrival of new stream data, previously received data must be processed in a timely manner. Consequently, with these constraints, in detection it may not be able to employ techniques which are typically used in the retrieval case because of the lack of information. Compared to NDVR and the first case of NDVD, few studies are conducted on real-time online NDVD. To name a few, there are Xie et al. [2010], Shen et al. [2007], Huang et al. [2009b], and Yan et al. [2008] that propose solutions for the continuous NDVD over video streams.

Moreover, in many cases the exact location (starting/ending timestamp) of the near-duplicate video needs to be found in some applications for near-duplicate video detection.

As NDVR and NDVD share similar principles and a lot of techniques in NDVR or NDVD can be applied to the other, in this survey we will look into the existing work for both tasks.

3. FRAMEWORK

A video is typically regarded as a sequence of frames (sometimes other information like audio and surrounding text is also included). Due to the inefficiency manipulating low-level features, compact signatures (or summaries) are normally necessary for practical NDVR. Figure 2 depicts a general framework for an NDVR system, including

three major components, signature generation, signature management, and signature search, for a typical NDVR task, and an optional component graph/network-based mining for the more specified partial detection/reranking task.

Normally, given a database video, in the signature generation component, low-level features are first extracted, based on which video summarization can be further performed to generate compact and distinctive video signatures to facilitate retrieval. As videos usually have meaningful structures at levels of frame, shot, and scene, video signatures can also take structural information into consideration by performing shot detection to find shots/scenes. Nowadays video databases grow rapidly and the sizes get larger and larger. To achieve fast retrieval, in the signature management component, the generated video signatures are organized by effective indexing structures to avoid extensive data accesses. The signature generation and indexing of database videos are typically offline processes. Given a query video, in the signature search component, its signature is first generated like that in the signature generation component. The obtained query signature is then searched in the database index to find NDVs, based on signature matching. If the similarity between a database video's signature and the query signature is greater than or equal to a predefined similarity threshold- ε, the database video is regarded as an NDV of the query. The signature search component is an online process. The preceding procedure describes the principles of an NDVR task. In real life, the existence of partial NDVs imposes additional requirements to the procedure. With the assistance of some specific techniques, such as fusing frame-level or segment-level video relevance to enable subsequence retrieval, the framework can be capable of detecting partial NDVs as well.

In the following section, most of the components in the NDVR framework will be discussed as well as state-of-the-art techniques.

4. STATE-OF-THE-ART

We review the state-of-the-art research in NDVR in this section, from the aspects of low-level feature extraction, video signature, signature indexing, to evaluation criteria. Here by *low-level feature* we are referring to those forms of data that carry the raw visual information of video frames, such as color moment, color histogram, local features, etc., while *video signature* denotes the forms or structures that represent videos (including shots and segments) on which we actually perform search. A video signature is normally a higher-level video summarization derived from low-level frame features.

4.1. Low-Level Feature Extraction

We can conclude that almost every work conducted on NDVR involves various low-level features, based on which higher-level signatures are generated. Indeed these methods of low-level feature extraction are derived in the studies of image processing and retrieval. Although there seems to be tens of feature types available from the image field, for example, color moment, color correlogram, color histogram, edge direction histogram, wavelet texture, various local features, in NDVR only a small number of them appear in existing works. This is mainly because of the constraints brought by the enormous data involved in NDVR. Unlike in image retrieval, low-level feature extraction is normally not considered enough for NDVR with reasonable efficiency and effectiveness because of the large number of frames involved in videos. Video signature, a higher-level video representation, is often derived after low-level feature extraction to provide more compactness and distinctiveness. This has been the major research focus in NDVR.

Two most commonly used low-level features in NDVR are color histogram and local feature. Color histogram is a representation of the color distributions in an image and can be extracted by concatenating the counts of the number of pixels whose color value

falls into the corresponding range [Huang et al. 2010a, 2009a; Wu et al. 2009a, 2007a; Liu et al. 2007; Zobel and Hoad 2006; Law-To et al. 2006]. It is compact and computationally efficient. However, it contains no geometric, shape, or texture information and is sensitive to color changes. Similar histograms may be generated from very different objects or scenes if they have similar color patterns. Local feature is generated by detecting and describing identifiable and characteristic interest points or regions from an image. A number of different detectors and descriptors are available in various works [Zhao and Ngo 2009; Wu et al. 2008; Zhao et al. 2007; Wu et al. 2007c, 2007a; Zhu et al. 2008; Zhou et al. 2009; Jiang and Ngo 2009; Poullot et al. 2008; Douze et al. 2010]. It captures object-level information and is invariant to scale and affine transformations. However, it is computationally expensive to match a large number of local features for image similarity computation.

4.2. Video Signature

Existing video signature generation techniques are usually derived and extended from image summarization techniques and additional spatio-temporal models. After low-level 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 boost search response or improving low-level feature's representativeness. In signature generation, different works focus on different aspects of the video.

One can find that numerous types of video signatures have been developed by the researchers. The reason is that since the establishment of this research topic, a major challenge of NDVR has been that the presence of an NDV could be of a very arbitrary type. According to the definitions we previously discussed in Section 2, the form of an NDV could be a transformed video, in which case tens of different transformations need to be considered. It could also be a slightly different scene of the same story. Additionally, partial NDVs only show some subsequences of closely related visual content. The unprecise nature of NDV's definition determines that there does not exist a single type of signature effective for all types of NDV. On the other hand, applications often involve different types of NDVs with divergent constraints. For instance, video monitoring requires very high time efficiency to make real-time processing possible, while in video event tracking the objects in the videos need to be closely examined for object matching. The diversity of NDV's forms and its applications imply that an NDVR approach may be able to serve a specific application for the best. Here we summarize existing video signatures into four categories.

In Table II we tie the existing signature types and the corresponding approaches in which they are used. We also describe a general relationship between different types of signatures and their corresponding NDV types here. 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 exploited during the signature generation for more precise representation and better distinguishing power. For instance, video-level global 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 omission of frame orders leads to great robustness to any NDV that is created with temporal changes.

Using frame-level local signatures, changes in the aspects of the videos (the angle of viewport when recording the video) can be properly recognized, however, low matching speed is expected. It suits applications that are insensitive to speed but aims to find videos that are captured from different viewports about the same scene. Frame-level global signatures are a compromise between speed and accuracy. It stills keeps some

Table II. Categories of Video Signature

Signature Types	Existing Works
Video-level Global Signature: represents a whole video with a single signature	—Bounded Coordinate System [Huang et al. 2009a; Shen et al. 2007] —Accumulative HSV Histogram [Wu et al. 2007a] —Reference Video-Based Histogram [Liu et al. 2007] —Cluster Representative [Shen et al. 2005]
Frame-level Local Signature: represents local information at frame level	—Local Keypoint Descriptor [Wu et al. 2007a, 2009a; Zhao et al. 2007; Zhu et al. 2008; Zhou et al. 2009]
Frame-level Global Signature: represents a whole frame with a single signature	—Bag-of-Words [Jiang and Ngo 2009] —Glocal Descriptor [Poullot et al. 2008]
Spatio-temporal Signature: represents a video using spatial and temporal information	—CE and LBP-based Spatio-temporal Signature [Shang et al. 2010] —Spatio-temporal Post Filtering [Douze et al. 2010] —Video Distance Trajectory [Huang et al. 2010b] —Shot-length, Color-shift and Centroid methods [Zobel and Hoad 2006] —Video Sketch [Yan et al. 2008] —Local Descriptor-based Trajectory [Law-To et al. 2006]

ability of detecting aspect changes while the approaches using them run as fast as video-level global signature-based approaches.

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.

In the following, we discuss each type of signature in detail. Note that in image retrieval, global signature often refers to a type of signature that contains the information of the whole image. Here in NDVR, we use the term *frame-level global signature* to represent the same concept, and we use *video-level global signature* to indicate a type of signature that summarizes a whole video with a single signature, regardless of any frame-level information.

4.2.1. Video-Level Global Signature. A video-level global signature represents a whole video with a single signature. Video-level global signature has been considered as an efficient form of representing videos and has been widely used in many works [Huang et al. 2009a; Liu et al. 2007; Wu et al. 2007a; Shen et al. 2005]. 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. The general idea of generating global signature is to summarize the global content distribution information into different forms, such as principal components, various histograms, and cluster representatives.

Bounded Coordinate System. A statistical model called Bounded Coordinate System(BCS) is proposed in Huang et al. [2009a] and Shen et al. [2007] to summarize short video clips for real-time near-duplicate video clip retrieval. The basic assumption is that a short video clip should have a single theme represented by its content. Derived and extended from Principal Component Analysis (PCA), the concept of this scheme is to capture the dominating distribution and changing trends of video frames by generating a coordinate system from principal components. To construct the BCS signature, the low-level feature matrix is processed by PCA, and the principal components with top significance are regarded as the video signature. To compute the similarity between two BCS signatures, translation and rotation are both considered for matching two

principal components. This method is stated to be robust to various transformations which yield little changes in distribution, including reformatting, frame rate changes, scaling, resizing, mirroring, inserting, deleting, swapping, etc.

Accumulative Histogram. Color histogram is widely used as a signature that accumulates and combines all raw color features into bins. In Wu et al. [2007a], the authors propose a hierarchical framework which employs a filter-and-refine scheme with the combination of global signatures and pairwise comparison for near-duplicate video elimination. The idea is to seek better search efficiency while at the same time obtain good accuracy to certain transformations which yield major difference in color features. In the framework, global video signatures are first used to perform fast approximate search, by the end of which a set of candidates is identified. At this point, the ones with remarkable confidence will be directly marked as near-duplicates while the ones with very low confidence will be identified as novel videos. Except for those already classified, the ones left will be applied with expensive pairwise comparison to the query video as a complementary solution to the framework.

For global signature generation, an accumulative HSV histogram, denoted as VS, is defined as

$$VS = (vs_1, vs_2 \ldots vs_m),\ vs_i = \frac{1}{n}\sum_{j=1}^{n} h_{ij}, \tag{1}$$

where h_{ij} represents the i^{th} bin of the HSV color histogram at the j^{th} keyframe of all n keyframes in a video and m is the number of bins. Given the definition of this video signature, the distance between two video signatures can then be simply computed based on the Euclidean distance.

Reference Video-Based Histogram. A reference-based histogram is proposed as a video signature in Liu et al. [2007]. In this work, each video is compared with a set of reference videos (seed videos) using 2-dimensional PCA signature, and then the percentage of video frames which are closest to the corresponding reference video is recorded and combined as a histogram. Let $SV = (sv_1, sv_2, \ldots, sv_m)$ be a set of feature vectors to represent seed videos, the video histogram is defined as a m-dimensional vector VS

$$\begin{aligned} VS &= (vs_1, vs_2, \ldots, vs_m), \\ vs_i &= \frac{\sum_{j=1}^{n} I(sv_i = arg\min_{1<=k<=m} d(f_j, sv_k))}{n}, \end{aligned} \tag{2}$$

where $I(x)$ is a binary function which equals to 1 if x is true and zero otherwise, f_j is the j^{th} frame, and n is the number of frames in the video. $d(f_j, sv_k)$ is the Euclidean distance. Each dimension's value of the signature means the percent of the videos frames which are closest to the corresponding seed vector with the same order in the SV. Clearly VS is normalized and it satisfies $\sum_{i=1}^{m} vs_i = 1$. Then given signatures VS_1 and VS_2 of two videos, their similarity can now be defined as

$$\begin{aligned} Sim(VS_1, VS_2) &= 1 - \frac{1}{\sqrt{2}} d(VS_1, VS_2), \\ d(VS_1, VS_2) &= \sqrt{\sum_{i=1}^{m} (vs_{1i} - vs_{2i})^2}. \end{aligned} \tag{3}$$

As can be observed, critical dependency to reference videos exists in this method as an essential issue which can heavily affect the quality of signatures. To solve this issue,

the work then presents an intuitive clustering method by maximizing the distances between reference videos.

Cluster Representative. In Shen et al. [2005], a clustering-based approach is presented to summarize a video sequence. The idea is inspired by the observations that keyframe-based representation loses some information and similarity measurement is expensive and heavily dependent on video length. Intuitively, the video similarity can be estimated by the percentage of similar frames in two sequences and thus it is also robust to temporal order of frames. This measurement is obviously expensive. In this work a new clustering-based approach is proposed. Each video sequence is summarized into a small number of clusters, each of which contains similar frames and is represented by a compact model called Video Triplet (ViTri). ViTri models a cluster as a tightly bounded hypersphere described by its position, radius, and density. The ViTri similarity is measured by the volume of intersection between two hyperspheres multiplying the minimal density, that is, the estimated number of similar frames shared by two clusters. The total number of similar frames is then estimated to derive the overall video similarity. Hence the time complexity of the video similarity measure can be reduced greatly.

Global signatures capture the overall video information and are efficient for indexing and searching because they are very compact. The reduction in content redundancy brings various benefits in storage, management, computation, and retrieval. It puts most of the effort into offline processing and accelerates online matching with relatively low complexity. The main drawback of video-level global signatures is that the local information on objects/regions in videos is often ignored. Due to the outstanding need for compact signatures, video-level global signature is rarely generated from local features. More often, simpler features like color histograms are used. On the contrary, frame-level signatures have an advantage on this point. Local features are frequently used to generate higher-level signatures. Local information is better preserved in such cases. Some approaches use low-level features directly as their video-level global signature, like in Wu et al. [2007a]. However, this is normally used only for the filtering stage in a filter-and-refine framework.

In terms of NDVR, the loss of local information in video-level global signature makes it difficult to distinguish two totally irrelevant videos with similar color distributions. Hence very different videos may have similar global signatures. For instance, a video that gives close-ups on a black carpet with a white object in the middle may have very similar signature to another video which shows a man wearing light-colored clothes in a dark night, just because they have similar color distributions, that is, black colors as background and light colors as foreground. This issue requires the assistance from more localized, but more expensive descriptors, as follows.

4.2.2. Frame-Level Local Signature. 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 as in Chum and Matas [2010] and Chum et al. [2008]. The frame-level local signature generated from frames is the most recent and popular form of signature in the field of NDVR. The foundation of this technique is the development of local keypoints (or local interest points) detection and description [Laptev and Lindeberg 2003; Ke and Sukthankar 2004; Bay et al. 2006] from computer vision and pattern recognition area. Its performance is well-recognized because of its good capability in detecting variations and major changes (especially geometric changes) introduced by complex editions, despite the fact that they are much more computationally expensive than global signatures due to the large number of local keypoints and their high dimensionality.

After the detection of keypoints, these keypoints are organized by their corresponding frames, the same as it does in local keypoint-based image retrieval [Lowe 2004]. The distances between frames are measured simply by a composition of the Euclidean distance between keypoint descriptors. However, precise as they might be, local keypoints contain so much information that makes it barely practical to directly perform a search. We now give a quick example to obtain a more concrete understanding. Most local interest point detection techniques tend to locate remarkable number of interest points (e.g., around 1,000 points for a 800×640 image), each of which is represented by a 128D SIFT [Lowe 2004] descriptor or 36D PCA-SIFT [Ke and Sukthankar 2004] descriptor. To match two frames, pairwise interest point comparisons are needed, resulting in quadratic time complexity. Assume every video consists of 5 keyframes, and there are 100,000 videos in the video database, there exist 500,000,000 keypoints, each of which is recorded as a high-dimensional vector. Clearly, the data size and the time complexity are already beyond the capability of any ordinary system of performing feasible similarity search. This makes it impractical to perform exhaustive matching of keypoints in NDVR.

Existing works in which local signatures are used can be classified into two classes. The first one is to use pairwise frame matching with local keypoint descriptors [Wu et al. 2007a]. The scheme contains highly intensive computation. Acceleration is inevitable in order to achieve reasonable time performance. In Wu et al. [2007a], local keypoints are used only as complementary information. Exhaustive matching of keypoints is only performed after a filtering step in which the majority of irrelevant videos are excluded according to the global color histograms of the frames. The second one is to generate a compact frame-level global signature from local keypoints.

4.2.3. Frame-Level Global Signature. A frame-level global signature represents a whole frame with a single signature. Due to the prohibitive 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 merely carry any local information. To consider local information, local keypoints are often used to generate global signatures.

Bag-of-Words. Proposed in Sivic and Zisserman [2003], BoW (also referred to as Bag-of-Visual-Words) has become popular in recent years in both image and video retrieval. With low-level features of a set of videos, to generate BoW signatures, all the keypoint descriptors are put together, regardless of which frame they belong to. These points are then clustered, with the objective of getting highly coherent, lowly coupled partitions of them. After that, a unique "visual word" is assigned to each cluster to represent all the local keypoints in that cluster. All the visual words together form a visual vocabulary. Finally, a frame can be represented as a histogram of the occurrences of the visual words in that frame. BoW has quickly emerged in the NDVR community with its capability of providing object-level recognition while the compactness of its signature facilitates search. Given the characteristics of compactness, it is considered competitive in searching large-scale databases.

Recent studies [Grauman 2010; Jegou et al. 2010] have shown excellent scalability and satisfactory precision of BoW in near-duplicate media detection. Basically, BoW aims to achieve a trade-off between efficiency and quality.

Glocal Descriptor. An interesting frame-level global descriptor generated from local keypoints with spatio-temporal information is presented in Poullot et al. [2008]. The idea can be understood as hierarchically scattering high-dimensional points into hyper-rectangles with given depth to form a histogram-like signature. To be specific, up to 20 interest points are detected by improved Harris detector for each video frame,

where for each interest point a 20-dimensional signature is generated by composing the normalized 5-dimensional vector of first- and second-order partial derivatives of the grey-level brightness for this point and 3 other neighboring points in the frames. Such a signature can be seen as a high-dimensional point in the $[0, 255]^{20}$ space. And the space is then partitioned with a limited depth h, resulting 2^h cells assigned with cell numbers given that a new interval is partitioned into two at every level. After this hierarchical partition, the glocal signature of a frame can be generated as a binary vector, where the i^{th} bit is set to 1 if one or more local signature falls into the i^{th} cell. For example, given depth 4, the 20-dimensional space is hierarchically partitioned for 4 iterations, resulting 2^4 cells. Consequently local signatures are distributed by their leading 4 dimensions into cells. Dice coefficient is used to measure the similarity between two glocal signatures.

4.2.4. Spatio-Temporal Signature. 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 [Wu et al. 2008; Law-To et al. 2006; Satoh et al. 2007]. 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.

CE- and LBP-Based Spatio-Temporal Signature. The proposed spatio-temporal signature in Shang et al. [2010] is inspired by ordinal measure, a robust feature in image correspondence. Ordinal measure describes the pairwise ordinal relations between blocks in terms of average gray-level values, which can be understood as the ranks of blocks according to their gray-level values. The authors proposed a feature selection method using Conditional Entropy (CE) in the information theory. The method performs feature selection on all ordinal relations, with an objective of maximizing the CE of the selected subset of relations. After the selection, a video is modeled as a set of *w-shingling* (a contiguous sequence of *tokens*) in text retrieval, with the frame-level ordinal features hashed into binary numbers and function as *token*s. Temporal information is encoded in the *w-shingling*s. These *w-shingling*s form a visual vocabulary and can be further processed in to BoW signatures. With the same framework, Local Binary Pattern (LBP) can also be selected and transformed into the BoW signatures, only that the selection objective is changed.

Spatio-Temporal Postfiltering. The framework in Douze et al. [2010] matches individual frames of videos by checking the spatio-temporal consistency between them. It uses BoW presentation but enhances it with the employment of the Hamming embedding and weak geometry consistency. The process begins with the extraction of local keypoints, and then they are quantized into the BoW representation. These keypoints are embedded into binary signatures by the Hamming embedding method. In the matching phase, signatures are searched by an inverted file-like structure which utilizes the weak geometry consistency to assist verification of spatial consistency between frames. After that, matched frames are grouped into sequences and estimated using the approach presented in Law-To et al. [2007].

Video Distance Trajectory. Huang et al. [2010b] propose to transform a video stream into a one-dimensional Video Distance Trajectory (VDT) monitoring the continuous changes of consecutive frames with respect to a reference point, which is further segmented and represented by a sequence of compact signatures called Linear Smoothing

Functions (LSFs). LSFs of each subsequence of the incoming video stream are continuously generated and temporally stored in a buffer for comparison with query LSFs. LSF adopts compound probability to combine three independent video factors for effective segment similarity measure, which is then utilized to compute sequence similarity for quick near-duplicate detection.

Video Sketch. Yan et al. [2008] apply the min-hash method on discrete cosine features of video frames to construct video sketches for copy detection over video streams. A bit-vector signature of the sketch is further proposed to achieve two optimization objectives: CPU cost and memory requirement. In order to handle multiple continuous queries simultaneously, an indexing structure is also presented for the query sequence. Similar to VDT, this method is also designed to detect a near-duplicate subsequence from continuous video streams.

Shot-Length, Color-Shift and Centroid Methods. Zobel and Hoad [2006] introduce several compact video signatures which are sequence of numbers, thus a string matching technique could be used to rank the results. The shot-length method summarizes a video clip into a sequence of numbers, each representing a shot length. This method is efficient but loses the content information completely. Blurry shot boundaries or very limited number of shots may lead the accuracy to deteriorate greatly. A color-shift method uses color distributions to produce a signature that represents the inter-frame change in color in the video over time. Centroid-based signature represents the inter-frame change in spatial movement of the lightest and darkest pixels in each frame over time. Color-shift and centroid-based signatures can also be combined together. All the changes are formalized into numbers so that each video is summarized into a sequence of numbers. However, all the methods only indicate the consecutive inter-frame difference within each individual sequence without carrying content information. In Video Distance Trajectory (VDT) [Huang et al. 2010b], some content information can be preserved due to the usage of the reference point in generating the trajectory.

Local Descriptor-Based Trajectory. A local descriptor-based trajectory summarization and matching scheme is introduced in Law-To et al. [2006]. In this work, a Harris detector is employed to locate all keypoints. The reason why Harris detector is chosen is because compared to other detectors (e.g., SIFT [Lowe 2004]), it generates a relatively smaller feature set, making it feasible to process with the proposed method within reasonable time and with reasonable space.

A 20D signature representing a description of points is generated as

$$\vec{S} = \left(\frac{\vec{s_1}}{||\vec{s_1}||}, \frac{\vec{s_2}}{||\vec{s_2}||}, \frac{\vec{s_3}}{||\vec{s_3}||}, \frac{\vec{s_4}}{||\vec{s_4}||} \right), \tag{4}$$

where $\vec{S_i}$ are the 5D subsignatures computed from different spatial positions around the interest point and each $\vec{S_i}$ represents a differential decomposition of gray signal $I(x, y)$.

$$\vec{s_i} = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I^2}{\partial x \partial y}, \frac{\partial I^2}{\partial x^2}, \frac{\partial I^2}{\partial y^2} \right) \tag{5}$$

In the tracking part, a simple and basic tracking algorithm is chosen as they focus on low-cost techniques. The space information is further processed into a mean signature $\vec{s}_{mean}$ as signal description and the time information is extracted into trajectory description. At this point a video can be described by its spatial description and time description. In addition, a label of trends is proposed to be associated with the descriptions defined before. This method considers both spatial and temporal information by a fast voting algorithm. However, critical information in the keyframe might be lost.

Visual-Temporal Network. The authors of Tan et al. [2009, 2010] proposed a framework to detect and locate ND segments from two videos. Given a query video Q and a reference video R, the framework translates the problem of searching the alignment between two videos in increasing temporal order with maximum similarity into a network flow maximization problem. Specifically, to initialize the network G, top-k neighbors of keyframes in Q are retrieved from R. These keyframes, plus two artificial nodes, *source* and *sink*, represent all the nodes in network G. Then, keyframes among the top-k lists are linked by directed edges which follow strict temporal order, that is, a keyframe can only link to another keyframe when the other keyframe has a timestamp equal to or greater than its own. Each directed edge is then assigned a weight, namely the similarity of the destination node to its corresponding query keyframe in Q. Given that, the flow of a path is the accumulated weights of the edges traversed from *source* to *sink*. By finding the path with maximum flow, it identifies the best match of ND segments between two videos with joint visual-temporal consistency.

One major concern about spatio-temporal signatures is the high complexity of their similarity measures since temporal order is considered.

4.3. Signature Indexing

Signature management aims at formulating distinctive structures to store signatures and their distances or similarities to enable efficient retrieval. Index structures usually vary with signature types and distance metrics. In the field of NDVR, high-dimensional index structures are commonly employed to cope with the fast retrieval of various types of signatures. Most indexing techniques heavily rely on the form of signature. Often, there are approaches that are developed specially for one corresponding type of signatures. Generally, indexing for multimedia is highly related to range and nearest-neighbor search approaches. Various categories of high-dimensional indexing methods have been proposed including tree-like structures, transformation methods, and hashing methods.

4.3.1. Tree Structures. Many tree structures have been proposed to index high-dimensional data. Recently in Böhm et al. [2007], each shot in a video is represented by a Gaussian distribution function over the employed feature space, which is indexed by the Gauss-tree. The Gauss-tree is an index structure for managing Gaussian distributions to efficiently answer video queries. However, generally tree structures are considered less efficient for the high dimensions we normally handle in NDVR, as when the dimensionality increases, the space splitting becomes very ineffective with excessive overlaps.

4.3.2. One-Dimensional Transformation. iDistance [Jagadish et al. 2005] is one of the recently proposed approaches for indexing high-dimensional vector-alike features. In multimedia these features include histograms or local point descriptors. The idea is to partition objects into clusters and index all the objects by their distance to the reference point inside a cluster using a B^+-tree. This enables range search and KNN search as well, in an efficient manner. In Shen et al. [2005], a method to choose an optimal reference point for one-dimensional transformation is proposed to maximally preserve the original distance of two high-dimensional points, which leads to the optimal performance of B^+-tree. As a further improvement of Shen et al. [2005], a two-dimensional transformation method called Bi-Distance Transformation (BDT) is introduced in Huang et al. [2009b] to utilize the power of two far-away optimal reference points. It transforms a high-dimensional point into two distance values with respect to two optimal reference points.

4.3.3. Locality-Sensitive Hashing (LSH). LSH has been widely used and recognized as efficient in removing the curse of dimensionality. It is an approximate similarity search technique which performs well with even very high-dimensional data. The basic idea is to use a family of locality-sensitive hash functions composed of linear projections over random directions in the feature space. And the rationale behind it is that for at least one of the hash functions, nearby objects have a high probability of being hashed into the same bucket. Improvements to LSH have been made in the past decade, regarding its accuracy, time, and space efficiency by improving the hashing distribution [Grauman 2010], enforcing its projection method [Tao et al. 2009], and amending the probe of buckets in the search stage [Joly and Buisson 2008].

4.3.4. Inverted Files. Recently, some text retrieval techniques have been applied to index large-scale image/keyframe databases. Typically, the bag-of-features (or bag-of-words) approach is first used, where an image is represented as a bag-of-visual-words [Jegou et al. 2010; Wu et al. 2007c]. Since each frame in a video is transformed into a set of visual words, it is very similar to text data, and hence inverted files can be adopted to facilitate search. A typical inverted file is constructed by recording the frequency and occurrence position of each visual word, establishing a structure of inverse association between these visual words and occurred videos. A recent proposal [Shang et al. 2010] combines fast histogram intersection kernel and inverted file to achieve accurate and efficient calculations for histogram intersection measurement between videos.

Undoubtedly, distance metrics help the measure of how similar two signature are, and they are the foundation of any index methods. For NDVR, when the signatures are considered as and only as high-dimensional vectors, the use of L_p-norms as distance metric is a natural choice. However, some specific types of signatures are encoded with additional information and hence require specific distance metrics to handle, for example, when cross-bin similarity is considered, Earth Mover's Distance (EMD) [Jiang and Ngo 2009] offers better accuracy for assessing histogram similarity; when the signature is treated as string-like data, Edit distance [Huang et al. 2010b] is then used. Other than the well-known metrics, we find that in some NDVR works there are more signature-specific metrics/matching schemes, like one-to-one symmetric matching in [Wu et al. 2009a, 2007a; Zhao et al. 2007; Zhu et al. 2008; Zhou et al. 2009] and other metrics in [Law-To et al. 2006].

So far we have discussed the state-of-the-art video signatures and the indexing methods for efficient near-duplicate search. Different near-duplicate video applications may have different requirements on video signatures and indexing methods. For example, video monitoring typically needs compact spatial-temporal signatures and near-duplicate search has to be performed in an online and continuous fashion.

4.4. Evaluation Criteria

NDVR is a research topic where performance matters. There is a common practice that any proposed approach should be evaluated under certain criteria, which are usually presented in the experiment section. We briefly introduce a general procedure and describe the elements involved.

A general evaluation framework can be understood as given a video dataset with ND ground truth, one needs to perform the proposed approach on the dataset with certain queries, and then the performance aspects like precision, recall, response time, and scalability are examined. Note that the ground truth is usually identified by human. Some works that focus on detection precision sometimes ignore the efficiency factor of their approaches.

4.4.1. Dataset. The general framework begins with the action that one or more datasets are first prepared. We notice that there exist a couple of standard (public)

datasets, like TRECVID[1] [Over et al. 2011; Smeaton et al. 2006], MUSCLE-VCD-2007[2], CityU CC_WEB_VIDEO[3] [Wu et al. 2007a], Bing Videos [Shang et al. 2010], and Tiny Videos [Karpenko and Aarabi 2011]. The first three are more widely used.

TRECVID is a U.S. government-supported project for video retrieval. It publishes a new dataset on an annual basis, each of which contains thousands of videos, or hundreds of hours of videos. The dataset is organized in such a way that all videos are divided into original videos and "transformations", where ground truth of the ND relationship is also given. It appears in the works Douze et al. [2010], Wu et al. [2009a], and Zhou et al. [2009]].

MUSCLE-VCD-2007 is a dataset for video copy detection. It consists of 100 hours of videos which include Web video clips, TV archives, and movies with different bitrates, different resolutions, and different video format. A set of original videos and their corresponding transformed videos are given for the evaluation of copy detection algorithms. Tan et al. [2009] and Yeh and Cheng [2009] conduct their experiments on this dataset.

CC_WEB_VIDEO, on the other hand, contains 12,790 videos (398,015 keyframes), which are collected from video Web sites including YouTube, Yahoo! Video, and Google Video. They are collected based on a sample of 24 popular queries. An interesting observation of the dataset is that it contains a very high rate of near-duplicates. On average there are 27% redundant videos that are near-duplicates to the most popular version of a video in the search results. For certain queries, the redundancy can be as high as 93%. The dataset is used in a series of works [Shang et al. 2010; Tan et al. 2009, 2008; Wu et al. 2007a, 2007b, 2007c].

In TRECVID and MUSCLE-VCD-2007 datasets, the near-duplicates are produced artificially by using video edition tools, while 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. One larger dataset collected from Bing Videos has also been used in Shang et al. [2010]. Some methods have also been tested on video databases [Böhm et al. 2007; Huang et al. 2009a]. For other studies, most of the authors collect their own datasets, mainly from video Web sites as well. Tens of thousands of videos are often used in the experiments.

4.4.2. Performance Metric. In NDVR standard precision, recall, and Mean Average Precision (MAP) in information retrieval are the most used measures for effectiveness evaluation. Efficiency is measured by the total response time [Huang et al. 2010b, 2009a; Zhao and Ngo 2009; Wu et al. 2007c], which means the elapsed time from the time the query is issued to the time when the results are returned. Scalability [Shang et al. 2010; Tan et al. 2009; Poullot et al. 2008; Law-To et al. 2009] concerns the tendency of changes in total response time when the search space grows.

We note that the results from different works are not always comparable, because the objectives of the experiments sometimes misalign for different applications. Some of them aim at finding all ND pairs in a video set [Zhao et al. 2007]. Some retrieve ND keyframes for given query keyframes [Wu et al. 2007b], and some detect all the NDVs of a given query video in a dataset [Wu et al. 2007a]. The TRECVID Content-Based Copy Detection task provides a good opportunity for different research groups to conduct experiments and compare performance. In 2011, there were 22 participants with different focuses. Several observations were made, including that the detection results were generally better than 2009 and 2010, and audio transformations are harder than

[1]trecvid.nist.gov.

[2]www-rocq.inria.fr/imedia/civr-bench/data.html.

[3]vireo.cs.cityu.edu.hk.

video transformations. More participants were attracted and the community got stable in size. The TRECVID platform has been a good starting point for the near-duplicate retrival/detection task [Over et al. 2011].

5. TRENDS AND FUTURE DIRECTIONS

As an emerging topic lately, NDVR is playing a crucial role in almost all fields of video-related applications, and this position is growing stronger along with the dramatically growing trends of online video applications. NDVR is highly related to conventional topics of computer science, including image indexing and retrieval, pattern recognition, video copy detection, indexing, and nearest-neighbor search in high-dimensional spaces, in which great research efforts and advances have been made over the years. According to existing literature, global signatures are popular for fast retrieval as demonstrated in Huang et al. [2009a], Liu et al. [2007], and Shen et al. [2005], as efforts are being constantly made to design more compact and discriminative signatures. Local signatures are competitive in scenarios where intensive transformation has been made to NDVs, while both local signatures and spatio-temporal signatures are proved effective with scene duplicates, in which two videos are presenting the same story scene but from different angles or viewpoints. Heuristic solutions which tend to utilize global signatures along with local signatures as well as spatio-temporal signatures are emerging in recent works like Tan et al. [2009], Basharat et al. [2008], and Wu et al. [2007a], to achieve a reasonable balance between effectiveness and efficiency. At the end of this survey, we address some possible topics as outlook for the near future.

5.1. User Perspective

User-centric aspects for NDVR have recently drawn attention in various works [Cherubini et al. 2009]. These works provide more information and awareness of user perception in the research of NDVR. User behaviors on video Web sites and their experience of browsing, searching, and watching videos is exploited. For instance, in Cherubini et al. [2009] it is observed from large-scale online survey that the major way of accessing online videos is by using video search services. Also, unlike the practice observed in the influential TRECVID project, where audio is often included to assist NDV identification, the authors of Cherubini et al. [2009] state that audio information is not used at all for users to identify NDVs. In addition, in contrast to what has been believed for years, NDVs are sometimes valuable to users as they provide complementary information to satisfy various user needs. But once the most satisfying copy is identified, users prefer not to be provided with other near-duplicates for the same query. Some views may seem uncertain at the current stage of research, and need further survey, research and examination. For example, the effect of audio remains unclear, and the treatment of NDVs after identification is quite different in different applications and research works. These perceptions toward the definition and position of NDVs introduce more directions of research and will be substantially useful to help researchers better understand actual users' needs.

5.2. Scalability Issue

Motivated by the need to manage an unprecedented number of online videos, scalability is becoming an even more significant topic as video applications in the real world nowadays are normally dealing with millions of videos or even more. Under such circumstances, scalability becomes a key aspect to consider for any NDVR technique. As can be observed in various works developed for large video databases [Huang et al. 2009a; Law-To et al. 2009; Tan et al. 2009; Poullot et al. 2008; Shen et al. 2005], several principles can be followed to achieve scalability of NDVR techniques: using compact and generic signatures, building indexing structures to enable fast search, and

applying approximation for complex similarity search to boost performance. Literature shows that a heuristic coarse-to-fine search scheme can often accelerate the search greatly with a simple but coarse technique, while maintaining good effectiveness by applying additional complicated but accurate techniques [Wu et al. 2007a]. Usually this is achieved by building two or more indexes for different levels of signatures. Global signatures are often used for the first-pass filtering, in which candidates with high probability are identified while those remaining unidentified candidates are passed into second-pass search, where more complicated signatures like local signatures or spatio-temporal signatures are measured to determine final results. Good balance can be achieved by the appropriate selection of combinations of approaches with experiments. Needless to say, scalability consideration will be another dominant aspect to judge NDVR approaches in the future given the continuous growth of videos.

5.3. Multimodality and Multifeature NDVR

Most NDVR approaches today are still focusing only on visual information, which is often captured by a specific type of feature. However, as observed by some researchers, it is possible to make NDVR more capable with multimodal and multifeature approaches. Other than the visual content itself, audio and textual information can be included to assess videos from more aspects. For example, more semantic meaning can be obtained by investigating online videos with its contextual information. As ordinary practice in video Web sites, videos embedded in Web pages are often accompanied with their corresponding textual information, which may include title, annotation, descriptive text, sometimes view count for some video Web sites like YouTube, or sometimes audio transcript for some news videos. In Wu et al. [2007c], the authors proposed two components for assessing visual similarity and textual similarity, respectively. In Wu et al. [2009a], the authors also utilize contextual information including view counts, time duration, and thumbnail image for fast identification of NDVs. On the other hand, audio offers the ability to detect NDVs where vision-based approaches fail to do so due to heavy visual changes or misalignments, as shown in the TRECVID project. It is reported that by examining these aspects the search accuracy is improved to a great extent. Besides, interestingly as shown in Wu et al. [2009a], multiple types of features are prepared and then used against different type of transformations. According to the transformation, the best feature is always chosen to perform NDV determination. This also shows promising results.

5.4. NDVR in Social Networks

Social networks as today's hot-spot in the information technology industry have been providing more and more video-related services. At the same time, conventional video Web sites are transiting more and more into social networks. For instance, Facebook provides utilities to share videos with friends, after which those videos can be directly accessed on its Web pages. On the other hand, social features are also available where a user can add friends, share videos with friends, follow another user's video channel, comment on any published video, and so on. With such background, video content, supplemented by its social features, have great potential to provide an outline of hot events and attention inside the social networks. By extracting these events and attention a lot of works can be done based on it, like monitoring user abuses, detecting copyright materials, and user-centric recommendations in social networks. This could be a promising future direction, with little effort made so far.

5.5. Correlation-Based Detection

All existing approaches identify near-duplicates mainly based on feature or signature similarity. It has been noticed that more and more online near-duplicate videos

are generated by amateur users. Many user-generated videos are edited from several sources with heavy changes. Very often, some near-duplicate videos exhibit great content changes, while the user perceives little information change, for example, color features change significantly when transforming a color video into a black-and-white video. These feature changes contribute to content video similarity computations, making conventional similarity-based near-duplicate video retrieval techniques incapable of accurately capturing the implicit relationship between two near-duplicate videos with fairly large content modifications. Different from existing near-duplicate video retrieval approaches which are based on video content similarity, some new dimensions for near-duplicate video retrieval can be further exploited. For example, the strong correlation between two video sequences might be useful in near-duplicate identification. The intuition is that near-duplicate videos should preserve strong information correlation in despite of strong content changes. Replacing or integrating video content similarity measures with information correlation analysis is another potential direction worthy of study.

REFERENCES

ASSENT, I. AND KREMER, H. 2009. Robust adaptable video copy detection. In *Proceedings of the 11th International Symposium on Advances in Spatial and Temporal Databases (SSTD'09)*. 380–385.

BASHARAT, A., ZHAI, Y., AND SHAH, M. 2008. Content based video matching using spatiotemporal volumes. *Comput. Vis. Image Understand. 110*, 3, 360–377.

BAY, H., TUYTELAARS, T., AND GOOL, L. V. 2006. Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision (ECCV'06)*. 404–417.

BOHM, C., GRUBER, M., KUNATH, P., PRYAKHIN, A., AND SCHUBERT, M. 2007. Prover: Probabilistic video retrieval using the gauss-tree. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE'07)*. 1521–1522.

CHERUBINI, M., DE OLIVEIRA, R., AND OLIVER, N. 2009. Understanding near-duplicate videos: A user-centric approach. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*. 35–44.

CHUM, O. AND MATAS, J. 2010. Large-scale discovery of spatially related images. *IEEE Trans. Pattern Anal. Mach. Intell. 32*, 2, 371–377.

CHUM, O., PHILBIN, J., AND ZISSERMAN, A. 2008. Near duplicate image detection: Min-hash and tf-idf weighting. In *Proceedings of the British Machine Vision Conference*.

DOUZE, M., JEGOU, H., AND SCHMID, C. 2010. An image-based approach to video copy detection with spatio-temporal post-filtering. *IEEE Trans. Multimedia 12*, 4, 257–266.

GRAUMAN, K. 2010. Efficiently searching for similar images. *Comm. ACM 53*, 6, 84–94.

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.

HUANG, Z., SHEN, H. T., SHAO, J., CUI, B., AND ZHOU, X. 2010b. Practical online near-duplicate subsequence detection for continuous video streams. *IEEE Trans. Multimedia 12*, 5, 386–398.

HUANG, Z., SHEN, H. T., SHAO, J., ZHOU, X., AND CUI, B. 2009a. Bounded coordinate system indexing for real-time video clip search. *ACM Trans. Inf. Syst. 27*, 3, 17–33.

HUANG, Z., WANG, L., SHEN, H. T., SHAO, J., AND ZHOU, X. 2009b. Online near-duplicate video clip detection and retrieval: An accurate and fast system. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE'09)*. 1511–1514.

JAGADISH, H. V., OOI, B. C., TAN, K.-L., YU, C., AND ZHANG, R. 2005. Idistance: An adaptive b^+-tree based indexing method for nearest neighbor search. *ACM Trans. Data. Syst. 30*, 2, 364–397.

JEGOU, H., DOUZE, M., AND SCHMID, C. 2010. Improving bag-of-features for large scale image search. *Int. J. Comput. Vis. 87*, 3, 316–336.

JIANG, Y.-G. AND NGO, C.-W. 2009. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Comput. Vis. Image Understand. 113*, 3, 405–414.

JOLY, A. AND BUISSON, O. 2008. A posteriori multi-probe locality sensitive hashing. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*. 209–218.

KARPENKO, A. AND AARABI, P. 2011. Tiny videos: A large dataset for non-parametric video retrieval and frame classification. *IEEE Trans. Pattern Anal. Mach. Intell. 33*, 3, 618–630.

KE, Y. AND SUKTHANKAR, R. 2004. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04)*. 506–513.

LAPTEV, I. AND LINDEBERG, T. 2003. Space-time interest points. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03)*. 432–439.

LAW-TO, J., BUISSON, O., GOUET-BRUNET, V., AND BOUJEMAA, N. 2006. Robust voting algorithm based on labels of behavior for video copy detection. In *Proceedings of the 14th Annual ACM International Conference on Multimedia (MULTIMEDIA'06)*. 835–844.

LAW-TO, J., BUISSON, O., GOUET-BRUNET, V., AND BOUJEMAA, N. 2009. Vicopt: A robust system for content-based video copy detection in large databases. *Multimedia. Syst. 15*, 6, 337–353.

LAW-TO, J., CHEN, L., JOLY, A., LAPTEV, I., BUISSON, O., GOUET-BRUNET, V., BOUJEMAA, N., AND STENTIFORD, F. 2007. Video copy detection: A comparative study. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR'07)*. 371–378.

LIU, L., LAI, W., HUA, X.-S., AND YANG, S.-Q. 2007. Video histogram: A novel video signature for efficient web video duplicate detection. In *Proceedings of the 52nd Annual Conference on Magnetism and Magnetic Materials (MMM'07)*. 94–103.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis. 60*, 2, 91–110.

OVER, P., AWAD, G., MICHEL, M., FISCUS, J., KRAAIJ, W., SMEATON, A. F., AND QUENOT, G. 2011. Trecvid—An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of the TRECVID Workshop*.

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 (MM'08)*. 61–70.

SATOH, S., TAKIMOTO, M., AND ADACHI, J. 2007. Scene duplicate detection from videos based on trajectories of feature points. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR'07)*. 237–244.

SHANG, L., YANG, L., WANG, F., CHAN, K.-P., AND HUA, X.-S. 2010. Real-time large scale near-duplicate web video retrieval. In *Proceedings of the ACM Conference on Multimedia (MM'10)*. 531–540.

SHEN, H. T., OOI, B. C., AND ZHOU, X. 2005. Towards effective indexing for very large video sequence database. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (*SIGMOD'05)*. 730–741.

SHEN, H. T., ZHOU, X., HUANG, Z., SHAO, J., AND ZHOU, X. 2007. Uqlips: A real-time near-duplicate video clip detection system. In *Proceedings of the 33rd International Conference on Very Large Databases (VLDB'07)*. 1374–1377.

SIVIC, J. AND ZISSERMAN, A. 2003. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th International Conference on Computer Vision (ICCV'03)*. 1470–1477.

SMEATON, A. F., OVER, P., AND KRAAIJ, W. 2006. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*. 321–330.

TAN, H.-K., NGO, C.-W., AND CHUA, T.-S. 2010. Efficient mining of multiple partial near-duplicate alignments by temporal network. *IEEE Trans. Circ. Syst. Video Technol. 20*, 11, 1486–1498.

TAN, H.-K., NGO, C.-W., HONG, R., AND CHUA, T.-S. 2009. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *Proceedings of the ACM Conference on Multimedia (MM'09)*. 145–154.

TAN, H.-K., WU, X., NGO, C.-W., AND ZHAO, W. 2008. Accelerating near-duplicate video matching by combining visual similarity and alignment distortion. In *Proceedings of the ACM Conference on Multimedia (MM'08)*. 861–864.

TAO, Y., YI, K., SHENG, C., AND KALNIS, P. 2009. Quality and efficiency in high dimensional nearest neighbor search. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (*SIGMOD'09)*. 563–576.

WU, X., HAUPTMANN, A. G., AND NGO, C.-W. 2007a. Practical elimination of near-duplicates from web video search. In *Proceedings of the ACM Conference on Multimedia (MM'07)*. 218–227.

WU, X., NGO, C.-W., HAUPTMANN, A., AND TAN, H.-K. 2009a. Real-time near-duplicate elimination for web video search with content and context. *IEEE Trans. Multimedia 11*, 2, 196–207.

WU, X., TAKIMOTO, M., SATOH, S., AND ADACHI, J. 2008. Scene duplicate detection based on the pattern of discontinuities in feature point trajectories. In *Proceedings of the ACM Conference on Multimedia (MM'08)*. 51–60.

WU, X., ZHAO, W., AND NGO, C.-W. 2007b. Efficient near-duplicate keyframe retrieval with visual language models. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'07)*. 500–503.

WU, X., ZHAO, W., AND NGO, C.-W. 2007c. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR'07)*. 162–169.

WU, Z., JIANG, S., AND HUANG, Q. 2009b. Near-duplicate video matching with transformation recognition. In *Proceedings of the ACM Conference on Multimedia (MM'09)*. 549–552.

XIE, Q., HUANG, Z., SHEN, H. T., ZHOU, X., AND PANG, C. 2010. Efficient and continuous near-duplicate video detection. In *Proceedings of the 12th International Asia-Pacific Web Conference (APWeb'10)*. 260–266.

YAN, Y., OOI, B. C., AND ZHOU, A. 2008. Continuous content-based copy detection over streaming videos. In *Proceedings of the IEEE 24th International Conference on Data Engineering* (*ICDE'08)*. 853–862.

YEH, M.-C. AND CHENG, K.-T. 2009. Video copy detection by fast sequence matching. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'09)*.

ZHAO, W. AND NGO, C.-W. 2009. Scale-rotation invariant pattern entropy for keypoint-based near duplicate detection. *IEEE Trans. Image Process. 18*, 2, 412–423.

ZHAO, W., NGO, C.-W., TAN, H.-K., AND WU, X. 2007. Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Trans. Multimedia 9*, 5, 1037–1048.

ZHAO, W. L., WU, X., AND NGO, C. W. 2010. On the annotation of web videos by efficient near-duplicate search. *IEEE Trans. Multimedia 12,* 5, 448-461.

ZHOU, X., ZHOU, X., CHEN, L., BOUGUETTAYA, A., XIAO, N., AND TAYLOR, J. A. 2009. An efficient near duplicate video shot detection method using shot-based interest points. *IEEE Trans. Multimedia 11*, 5, 879–891.

ZHU, J., HOI, S. C. H., LYU, M. R., AND YAN, S. 2008. Near-duplicate keyframe retrieval by nonrigid image matching. In *Proceedings of the ACM Conference on Multimedia (MM'08)*. 41–50.

ZOBEL, J. AND HOAD, T. C. 2006. Detection of video sequences using compact signatures. *ACM Trans. Inf. Syst. 24*, 1, 1–50.

Received December 2011; revised February 2012, May 2012; accepted July 2012