IR Feature Embedded BOF Indexing Method for Near-Duplicate Video Retrieval

Kaiyang Liao, Hao Lei, Yuanlin Zheng, Guangfeng Lin, Congjun Cao, Mingzhu Zhang, and Jie Ding

***Abstract*— Due to the explosive increase in online videos, near-duplicate video retrieval (NDVR) has attracted much researcher attention. NDVR has very wide applications, such as copyright protection, online video monitoring, and automatic video tagging. Local features serve as elementary building blocks in many NDVR algorithms, and most of them exploit the local volume information using a bag of features (BOF) representation. However, such representation ignores potentially valuable information about the global distribution of interest points. Moreover, the discriminative power of the local descriptors is significantly reduced by the quantizer in BOF. Our motivation is that if we use the global features to classify the same or similar keyframes into the same class, it will be very useful in improving the performance of NDVR. In this paper, we present an improved radon transform (IR) feature which captures the detailed global geometrical distribution of interest points. It is calculated by using the 2D discrete Radon transform, and then applying a principal component analysis. Such IR feature is not only invariant to the geometry transformations but also robust to the noises. In addition, we propose a fusion strategy to combine the BOF representation with the global IR feature for further improving the recognition accuracy. Convincing experimental results on several publicly available datasets demonstrate that our proposed approach outperforms the state-of-the-art approaches in NDVR.**

***Index Terms*— Content based retrieval, similarity search, near duplicate video retrieval, video indexing.**

I. Introduction

With the rapid development of the communication techniques, video editing softwares, and video capture devices, the number of online videos is growing exponentially. Meanwhile, the video related applications, such as advertising, video sharing, recommendation and monitoring, inspire online users' interests and participation in the video related activities, including searching, downloading, commenting, and uploading. A large number of videos are shared and downloaded on the internet every day. It showed that there are a great deal of near duplicate videos (NDVs) on the internet, which are produced in different approaches, ranging from transformations, to different editions, acquisitions, simple reformatting, and mixtures of different effects [1]. The appearance of substantial near duplicate videos imposes strong demand for effective NDVR in many new applications, such as video result re-ranking, copyright enforcement, online video usage monitoring, video tagging, cross-modal divergence detection, video database cleansing, and so on. For example, a typical situation could be that a terminal user wants to search out some new videos, but eventually the highest-ranking results returned by a search engine are many NDVs. Another scenario could be that the video producers want their copyrighted videos to avoid sharing on the Internet. Both situations need NDVR technology to help achieve their goals.

In recent years, people have done a lot of research on NDVR. Most of the existing methods usually make use of the following NDVR framework [2]. First of all, the videos are divided into key frame sequences which are extracted by shot boundary detection or time sampling algorithms. Secondly, these key frames are represented by some visual features, such as Local Binary Pattern (LBP), Scale Invariant Feature Transform (SIFT), etc. The sequence of these key frames' features is used as the signature of the original video. Finally, the NDVR systems need to compute the similarity between each dataset video and the query video, and return the name of dataset video that is most similar to the query video. In general, both temporal and spatial information are used to evaluate the similarity between the videos [1], [3]. There are also some existing methods summarized the whole video clip with a single and global feature to achieve real-time retrieval, but they are generally not effective in representing long time videos [4]. In the recent works, the pair-wise frame correlation in two videos has also been used to measure the videos' similarity [5]. A recent survey on near duplicate video retrieval can be found in [6].

Intuitively, the multiple features can complementary to each other, for each of the multiple features of video clips reflects the specific information of video data. At the same time, the use of multiple properties helps to disambiguate. For instance, local features are sensitive to the changes in captions,

Manuscript received May 19, 2016; revised September 10, 2016 and October 5, 2018; accepted November 25, 2018. Date of publication December 5, 2018; date of current version December 6, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61671376 and Grant 61771386 and in part by the Scientific Research Project of Shaanxi Provincial Department of Education under Grant 18JK0556. This paper was recommended by Associate Editor P. Remagnino. *(Corresponding author: Kaiyang Liao.)*

K. Liao and C. Cao are with the Faculty of Printing, Packaging Engineering and Digital Media Technology, Xi'an University of Technology, Xi'an 710048, China, and also with the Printing and Packaging Engineering Technology Research Centre of Shaanxi Province, Xi'an 710048, China (e-mail: liaokaiyang@xaut.edu.cn).

H. Lei, Y. Zheng, G. Lin, and J. Ding are with the Faculty of Printing, Packaging Engineering and Digital Media Technology, Xi'an University of Technology, Xi'an 710048, China (e-mail: 8872080@qq.com).

M. Zhang is with the Department of Public Courses, Xi'an Fanyi University, Xi'an 710005, China (e-mail: 33046404@qq.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2884941

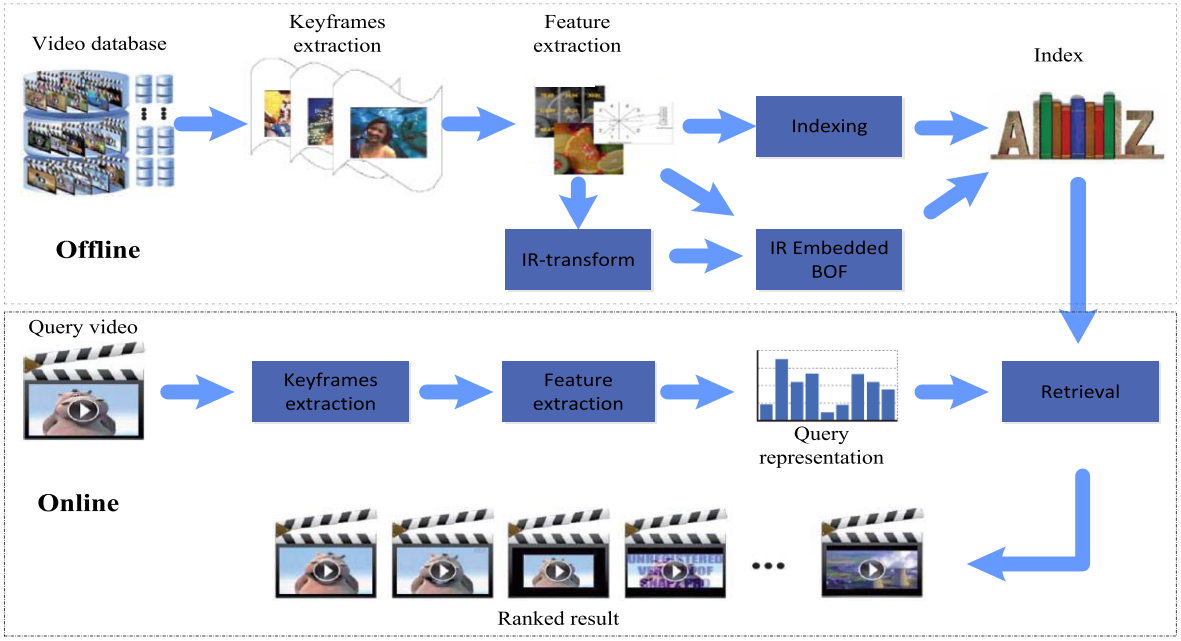


Fig. 1. Overview of the proposed video retrieval methods.

video length, frame rate, and global features are less robust to changes in camera viewpoint, bright-ness, contrast, scale and so on. At present, it is very difficult to find a single visual feature that is robust to all types of distortions and changes. So, it is very important to make use of multiple features for near duplicate video retrieval, in which various distortions and changes could take place between near duplicate videos.

Due to the rapid growth of the number of videos on Internet and the high complexity of most features in the video content, the scalability of near duplicate video retrieval algorithm has become a more and more important research topic. Because real-time response is very important to NDVR, it is necessary to establish an efficient index structure to facilitate the fast search over large-scale video datasets. In recent years, researchers have proposed a number of indexing algorithms to support image/video search, such as data reduction, tree-structures, hashing and others [7]. The performance of these indexing methods is generally assessed by the tradeoffs between search efficiency and quality, but the efficiency is the major concern. How to better preserve the relationships of NDVs in the index structure is necessary to be further researched.

In order to deal with large video data sets, most of the recent video search systems build upon the BOF (Bag of Features) representation, which is first introduced in "search of videos cast as text retrieval" [8]. Firstly, k-means clustering algorithm is used to quantify the descriptions into visual words. Each descriptor of the video is then assigned to the nearest visual word, so that the video is represented by the frequency histogram of the visual words. Fast access to the feature vectors (frequency histograms) is obtained by an inverted file method. Note that this approach would reduce performance slightly, because it is only an approximate match relative to the direct match of individual descriptors. Compared with other approximate nearest neighbor search methods, the memory occupancy rate of this method is very low, such as the popular Locality Sensitive Hashing (LSH) [9].

Many modern video retrieval methods are based on the combination of the Bag of Features model [10] and the local interest point feature due to its efficiency and simplicity. When using the local information or texture (e.g., HOF [11]) encoded the discriminability of local features in video, however, most BOF based representations ignores the global information or texture of the videos.

Motivated by the above observations, we propose an IR feature embedded BOF indexing method for near-duplicate video retrieval. The framework of our proposed video retrieval method is shown in Fig. 1.

In this paper, we first propose a novel approach to extract global information from the local interest points. We focus on the geometrical distribution of the local interest points, and characterize the local feature points from a geometric point of view. Based on the two-dimensional discrete Radon transform, the properties and form of the IR-transform is deduced, and the IR-transform is applied to the problem of video retrieval for local interest point representation.

Then, a more informative representation is used to make the distance between frequency vectors of visual words more significant. We add IR features to the BOF model, which are usually compared with the Euclidean distance, resulting of an IR feature Embedding (IRE) into the SIFT descriptors. The thought of using additional information has recently been proposed in [12], in which they are used to compress global descriptors. This contribution is to integrate location information into the BOF model for very large datasets. IRE requires storing additional information, and so increasing some memory usage in the process of retrieval. Yet the efficiency of the searching is not significantly increased when the method used.

The IR feature captures the global location information, while the BOF representation encodes the discrimination ability of local features. So, these two features naturally complement to each other. This new IR feature embedded BOF model is more robust to variations than the traditional BOF model, which considers only the pair wise relationships between feature vectors. In summary, the proposed algorithm combines two different characteristics and extracts the underlying related information from video.

The remainder of the paper is organized as following. Section II gives a review of the related works. Section III introduces IR-transform on local interest points. Section IV discusses the IR feature embedded BOF method. Section V reports experimental results on three public datasets. Section VI concludes the paper.

II. Related Work

A. NDVR

NDVR and related applications such as video similarity and copy detection have been actively studied in many real world applications. Recently, various methods have been proposed for NDVR, which use different features or index algorithms to retrieval videos. The existing works on near duplicate video retrieval can be roughly divided into two categories: the first is local feature based approaches and the other is global feature based approaches.

Many of the existing global features based near duplicate video retrieval methods emphasize the faster identification of near duplicate videos. These methods are very effective in handling identical or almost identical videos. In literature [13], for example, the authors use HSV to represent the key frames and further cumulate all the HSV information in the video to create a global video signature. This method obtains higher retrieval accuracy as well as faster retrieval speed in their datasets. Xu *et al.* [14] proposed a discriminative video representation for event detection over a large scale video dataset when only limited hardware resources are available. Zheng *et al.* [15] proposed an accurate image search with multi-scale contextual evidences, they show that CNN feature is complementary to SIFT due to its semantic awareness and compares favorably to several other descriptors such as GIST, HSV, etc. However, the limitation is that the global feature based methods usually become not very effective when it comes to processing near duplicate videos with distortions and variations [16]. Meanwhile, the global feature based methods heavily rely on the types of selected feature.

The local feature based methods, as an alternative of the global feature based methods, have also attracted many researchers' attention. The global features such as color, shape and texture extracted at the image level are further segmented into different regions, which are especially suitable for searching near duplicate videos with complex distortions and variations. There are many popular local feature based methods (e.g., SIFT [17]) used in retrieval near duplicate videos. However, the methods based on local features use pair-wise comparison that is computationally expensive, and they are not suitable for very large scale near duplicate video retrieval. Some of the recent approaches proposed to generate a compact representation from the local descriptors [18]. Although they are very efficient, it will degrade the retrieval quality, for the reason that the compact representations will loss some information.

In literature [16], Wu *et al.* proposed an algorithm constructing a hierarchy structure to take the advantage of both global feature and local feature. According to color histogram, they first filter out some of the videos, and then utilize the pair-wise comparison method to match the interest points between the keyframes. Wengert *et al.* [19] proposed a Bag-of-colors method for improved image search. Although these methods improve the performance, the pair-wise comparison method used to match the interest points between keyframes is still impractical for very large-scale video data sets, because the computation cost is very huge. Moreover, using the global feature (color histogram) only to filter out a large proportion of videos might be inaccurate, for some near duplicate videos may have quite different color histogram feature. Some recent works (e.g., [20]) have also considered to design new similarity measuring methods or new features (such as spatial-temporal features) to improve the performance. Nonetheless, these works are single feature methods.

Some methods have also considered scalability problem (e.g., [21]). They utilize keyframes to query within large scale video datasets. However, their main ideal is to investigate different frame samplings of the reference video dataset to evaluate the possible trade off between scalability and accuracy.

B. Multiple Feature Fusion

As far as it is concerned, a sort of feature is not sufficient to fully bewrite an image or a video. Given that multiple different features are often used to indicate video or image data, multiple feature fusion (properly combining the evidences derived from different features) becomes a significant re-search project. The key problem in multiple feature fusion is how to identify the correlation or similarity between two observations symbolized with multiple features.

Early fusion strategies and late fusion strategies are the conventional multi-source fusion methods. The multiple features are combined at the input stage in early fusion strategies. For example, in literature [7], the authors proposed a multiple feature fusion mothed in which different features are projected in an unified space. literature [22] proposes a coupled Multi-Index (c-MI) framework to perform feature fusion at indexing level. In literature [23], the feature level fusion is used to concatenate all the feature vectors which produced by diverse approaches to form a larger feature vector. Nevertheless, the computation cost is heavy and each individual feature's structural information cannot be well preserved.

Firstly, the late fusion strategies obtain the separation results from distinct features, and then make use of different algorithms to fuse the results together. In reference [24], the kernel-level fusion approach is employed, the multi-kernel classifier is used to combine different features. However, these approaches do not take into account the correlation between multiple features. Besides, the late fusion method is more

computationally expensive for training. Literature [25] proposed a query-adaptive late fusion for image search and person re-identification, which is an effective late fusion method at score level. Some other algorithms (e.g., [26]) use tensor to incorporate multiple features, but these methods focus on transductive learning.

The above fusion methods rely only on the pair wise similarities of videos without considering the high order correlations among videos. This may cause sensitivity to outliers and noise of the data.

C. Indexing

To improve retrieval efficiency, hashing and indexing are the two typical solutions. For example, literature [27] uses geometric hashing to generate database indices, while literature [28] perform indexing utilizing tree structure of different features and matching criteria. Sivic and Zisserman [8] described a Video Google system which retrieves videos from a database using bag of features matching. Liu *et al.* [29] reviewed earlier efforts in near duplicate videos retrieval, mostly rely on feature based similar methods or relevance feedback.

Recently, the BOF based retrieval methods are widely applied to the NDVR system. Some recent extensions of the BOF method make the search speed faster by frequency vectors [30] or assigning individual descriptors to visual words [31]. It is also possible to increase the performance by assigning multiple descriptors to the visual words [32] or regularizing the neighborhood structure [30] at the cost of reduced efficiency. Others improve the discriminative ability to recognize the visual words [33], in which the entire database must be known in advance. Finally, post-processing with spatial verification improves the retrieval performance, which is a re-occurring technique in computer vision.

In recent years, several algorithms have been presented to integrate geometrical information into BOF. Using grids [11] or multi-scale pyramids [34] is a common way to produce a coarse description of the feature layout. These methods uniformly divide the feature space into a grid and then calculate the histogram of feature vectors in each sub-volume. This grid structure just captures some simple location information, yet£the richer geometrical distribution information is also discarded.

III. IR Transform on Local Interest Points

A. The Radon Transform

The definition of image's Radon transform is determined by a set of image projections along lines taken from different angles. In this way, each nonzero pixel point on a discrete binary image is projected into a Radon matrix.

According to the definition, the Radon transform is linear. Therefore, the Radon transform can explicitly obtain geometric properties such as curves or straight lines, which can concentrate the energy of the image into several high value coefficients in the transformed domain.

To be useful, an image retrieval framework should remain explicit invariance under the operations of scaling, translation, and rotation. In order to calculate the similarity between the Radon matrices of two images, there is no need to realize the original geometric transformations from one image to the other. However, Radon transform does not have scale, translation, and rotation invariance. In this paper, we propose an adaptation of the Radon transform to overcome the above problem.

B. The Improved Radon Transform

Let the Improved Radon Transform, called IR-transform, be:

$$IR_f(\theta) = \int_{-\infty}^{\infty} T_{R^f}^2(\rho, \theta) d\rho \tag{1}$$

where T_{R^f} is the Radon transform of $f(x,y)$. We can see the following properties:

• Periodicity: $IR_f(\theta \pm \pi) = IR_f(\theta)$, Where the period is therefore set to π.

Poof: From the property of equation (1), we have

$$\begin{aligned}\int_{-\infty}^{\infty} T_{R^f}^2(-\rho, \theta \pm \pi) d\rho &= -\int_{-\infty}^{\infty} T_{R^f}^2(\upsilon, \theta \pm \pi) d\upsilon \\ &= \int_{-\infty}^{\infty} T_{R^f}^2(\upsilon, \theta \pm \pi) d\upsilon \\ &= IR_f(\theta \pm \pi)\end{aligned}$$

using v = −q.

• Rotation: $IR_f(\theta + \theta_0) = \int_{-\infty}^{\infty} T_{R^f}^2(\rho, \theta + \theta_0) d\rho$. A rotation of the image by an angle θ_0 results in a translation of the IR-transform of θ_0.

• Translation: $\int_{-\infty}^{\infty} T_{R^f}^2(\rho - x_0 \cos\theta - y_0 \sin\theta, \theta) d\rho = IR_f(\theta)$. The IR-transform is invariant under a translation of $f(x,y)$ by a vector $\vec{u} = (x_0, y_0)$.

Poof:

$$\begin{aligned}\int_{-\infty}^{\infty} T_{R^f}^2(\rho - x_0 \cos\theta - y_0 \sin\theta, \theta) d\rho &= \int_{-\infty}^{\infty} T_{R^f}^2(\upsilon, \theta) d\upsilon \\ &= IR_f(\theta)\end{aligned}$$

using $\upsilon = \rho - x_0 \cos\theta - y_0 \sin\theta$.

• Scaling: $\frac{1}{\alpha^3} R_f(\theta)(\alpha > 0)$. A scaling of $f(x,y)$ induces a scaling only in the amplitude of the IR-transform.

Poof:

$$\begin{aligned}\frac{1}{\alpha^2} \int_{-\infty}^{\infty} T_{R^f}^2(\alpha\rho, \theta) d\rho &= \frac{1}{\alpha^3} \int_{-\infty}^{\infty} T_{R^f}^2(\upsilon, \theta) d\upsilon \\ &= \frac{1}{\alpha^3} IR_f(\theta)\end{aligned}$$

using $\upsilon = \alpha\rho$.

To summarize, the IR-transform is invariant under scaling and translation if the transform is normalized by a scaling factor (in this paper, we set the area of the IR-transform). A rotation of the image results in a translation of the transform modulo θ. We can see that only the rotation transform modify the function.

Let a single point image with the coordinate (x0, y0) be: $I(x, y) = \delta(x - x_0)\delta(y - y_0)$. According to the definition, the Radon transform is: $TR^I(\rho, \theta) = \delta(\rho - x_0 \cos\theta - y_0 \sin\theta)$. That means, a single point's Radon transform is a set

of non-zero points along sinusoidal curve of equation $\rho = x_0 \cos\theta + y_0 \sin\theta$. Given a set of discrete feature points, the way to compute the transform is to map every non-zero feature point, using the normal parameterization $\rho_i = \mathrm{x}_i \cos\theta_i + y_i \sin\theta_i$, into a Radon matrix. That means, for each feature point (x_i, y_i) of the image, i is fixed and the value of ρ_i is calculated using stepwise increments of θ_i from 0 to π. To avoid aliasing, the increment is defined to follow the Shannon theory. In this paper, we set $\Delta\theta = \Delta\rho = 1$, $\Delta\mathrm{x} = \Delta\mathrm{y} = 1/2$ and the sampled values of ρ_i are calculated by the linear interpolation. The time complexity of the algorithm is O(NM), N feature points and M different angles (here, M = 180). There are two major optimizations can be made to reduce the complexity. For all the possible values of θ_i, the sine and the cosine are computed just once. The values of ρ_i are defined recursively, and the step increment $\Delta\mathrm{x}$ is set to 1/2, so we have: $\rho_{i+1/2} = \rho_i + (\cos\theta_i)/2$. Therefore, when we move in the y-direction ρ is increased by $\sin\theta/2$. Similarly, in the x-direction, ρ is incremented by $\cos\theta/2$.

In order to improve the robustness and reduce the dimension of the IR feature, we apply the $(2D)^2$PCA to the matrix obtained from the IR-transform to obtain the corresponding low-dimensional matrix as the final feature. The $(2D)^2$PCA is introduced in [35], simultaneously calculates PCA in the row and column directions, in this way can obtain higher recognition accuracy than PCA.

If we can use the position information to create an IR feature for each keyframe image, then the IR features can be used to classify the keyframes (the same or similar keyframes can be classified into the same class). We can use this property to reduce descriptor's error matching. According to this principle, we can use the descriptor features to establish a codebook, and then use the IR features to reduce the error matching.

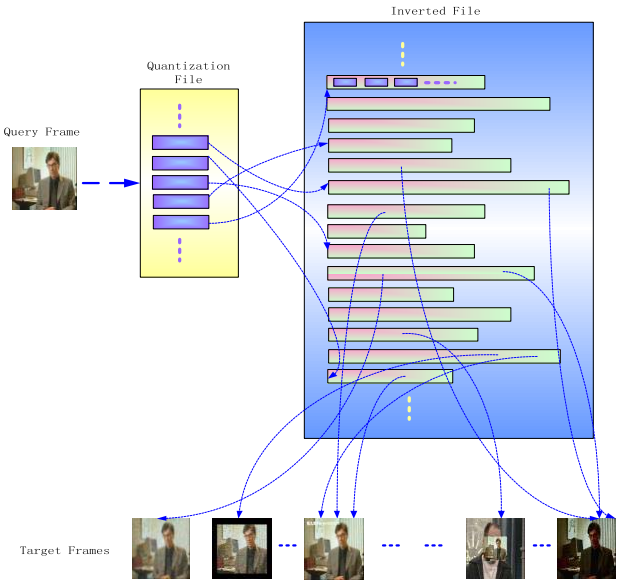


Fig. 2. Illustration of the BOF based voting process.

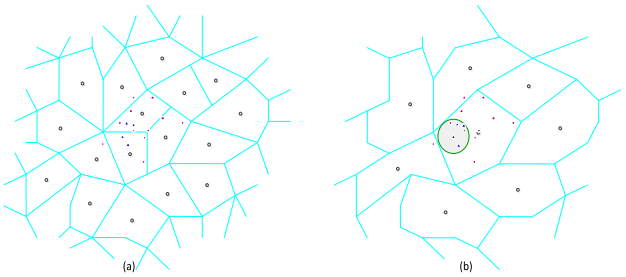


Fig. 3. Illustration of our embedded signature and K means clustering. (a) Fine quantization (with high k). b) Coarse quantization (with low k) and embedded signature. The similarity calculation (within a Voronoi cell) is based on the Euclidean distance. Key: ● = centroid, ✸ = noisy versions of this descriptor, ▲ = descriptor.

IV. IR Feature Embedded BOF for Video Retrieval

A. Weakness of Quantization Based Approaches

The video search engine based on BOF adopts the vector space model of information retrieval. Firstly, segment all the reference videos and select the key frames for those video segments.Then the SIFT [17] features are extracted from those key frames, and the visual vocabularies are generated by clustering the SIFT descriptors. Finally, all the SIFT features are quantized according to the visual vocabularies and stored in an inverted file. The query and each document in the inverted file are symbolized as a sparse vector of term (visual word) occurrences, and the retrieval process uses the Euclidean distance to calculate the similarity between the query vector and each document vector. To speed up the computation, the engine stores visual word occurrences in the inverted file, which maps individual words to the documents where they occur. This can lead to substantial acceleration because only the documents that contain certain vectors presented in the query need to be checked. The scores of each document are accumulated, so they are exactly the same as the explicit calculation of similarity. Fig.2 shows an illustration of the BOF-based voting process.

Video search based on BOF combines the advantages of efficient video frame comparison using inverted files and of local features. However, the quantization method significantly reduces the discriminative power of local descriptors. If two descriptors are assigned the same quantization index, i.e., lie in the same Voronoi cell, they are considered to be matched. Choosing the number of visual words k is a compromise between the descriptor noise and the quantization noise. A coarse quantization obviously leads to many incorrect matches. However, using a finer quantization, many incorrect matches are indeed removed, but at the same time, many correct matches are also removed.

As shows in Fig. 3(a), a high value of k provides better precision for the local descriptor, but the probability that a descriptor with noise is assigned to the same cell is lower. Conversely, a low value of k leads to large Voronoi cells, the probability that a descriptor with noise belongs to the correct cell is high, as Fig. 3(b) illustrated. However, if different descriptors lie in the same cell, this also reduces the discriminative power of the descriptor. Moreover, for a given visual vocabulary, the complexity of assigning the query

descriptors is O($k \times d \times m_l$) (d is dimension of the local descriptors, k is the number of visual words, m_l is the number of descriptors assigned to the visual word), consequently the complexity is significantly higher for larger vocabulary sizes.

B. Improving BOF With IR Features

We use two types of features to represent each video sequence: the BOF representation of the local features and the global IR feature. We first employ the Harris affine detector [36] to detect the local interest point for a given video. Afterward, we use the SIFT descriptor [17] to describe the detected interest points. So, a video V can be denoted as (x_i, α_i), $1 \leq i \leq N$, where N is the total number of interest points detected in the video, α_i is the SIFT descriptor feature of the i^{th} detected interest point, and x_i is the position vector.

Subsequently, two different types of features are extracted to characterize each video. The first is the BOF based representation, which uses the SIFT descriptor feature α_i of each interest point. The second is the global IR feature, which only uses the position feature x_i. We use the new IR-transform to characterize the spatial distribution of the detected interest points and then refine it by $(2D)^2$PCA, as presented in Section 3.2. Evidently, these two features complement to each other. The BOF based representations have very good discriminative power, which benefits from the local descriptors. While the IR features exploit the interest points' global spatial distribution. To make full use of the advantages of these two features, we utilize the global IR feature to make some improvements on the BOF based representation.

In the following, we present an approach that combines the advantages of a fine quantizer (high number of centroids k) with those of a coarse quantizer (low k). The main idea is refining the quantized index $q(x_i)$ with a db-dimensional global spatial feature which encodes the location of the local interest point within the Voronoi cell, see Fig. 3(b).

We propose to apply IR-transform to video frames to exploit the spatial distribution structure of the interest points extracted from a video frame. The minimal window containing all the interest points extracted from a video frame can be regarded as a 2D model. On this 2D model, the binary function $f(X)$ is defined as following:

$$f(X) = \begin{cases} 1 & \text{if interest point } X \subset \mathrm{D} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $X = (x, y)$ is the position of each interest point in the 2D model. The positions of interest points detected in a frame are denoted as $\{(x_j, y_j)\}_{j=1}^{J}$, where J is the total number of interest points in the frame. By Eq. (1) and (2), IR-transform of the frame is given by:

$$\begin{aligned} IR_f(\theta) &= \int_\rho T_f^2(\rho, \theta) d\rho \\ &= \int_\rho [\sum_{j=1}^{J} f(x_j, y_j) \delta(x_j \cos\theta + y_j \sin\theta - \rho)]^2 d\rho \end{aligned} \tag{3}$$

Observed from equation (3), every interest point is projected into all planes with parameters ρ and θ. The IR feature is then obtained by the integral of the square of projections over parameter ρ. Therefore, the IR-transform efficiently describes the geometrical distribution of local interest points. Afterwards, to achieve the robustness to rotation, we normalize the IR transform to get the rotation invariance by:

$$IR'_f(\theta) = \frac{IR_f(\theta)}{\max_\theta \{IR_f(\theta)\}} \tag{4}$$

For convenience, henceforward we use $IR_f(\theta)$ to represent the normalized IR-transform. The IR-transform uses a two dimensional variable $IR_f(\theta)$ to represent the location distribution of the interest points. By sampling the parameter θ, $IR_f(\theta)$ turns out to be a 2D vector. In order to improve the robustness and reduce the dimension of the IR feature, we apply the $(2D)^2$PCA to the vector obtained from the IR-transform. By applying the $(2D)^2$PCA on the obtained vector $IR_f(\theta)$, we obtain the corresponding low-dimensional vector as the final feature.

It is designed that the distance between two features x and y lying in the same cell reflects the Euclidean distance d(x, y) is small. So that the distance of IR feature's between a descriptor and its NNs in the Euclidean space is also small. At this point, a descriptor is represented by q(x) and b(x), q is a quantizer and b is the IR feature. We can now define the IRE matching function as

$$f_{IRE}(x, y) = \begin{cases} (tf - idf(q(x)))^2 & if q(x) = q(y) \\ & and\ d(b(x), b(y)) \leq h_t \\ 0 & otherwise \end{cases} \tag{5}$$

where d is the Euclidean distance and h_t is a fixed threshold. Quantizer can be set sufficiently high to ensure that x and its NNs match, and h_t can be set sufficiently low to filter many improper points that lie in a same Voronoi cell.

Note that this approach is different from the E2LSH (Euclidean version of LSH) [37], which generates several hash keys for each descriptor. E2LSH assumes that two descriptor vectors are similar if at least one of their hash functions has the same hash values. That is to say, if the descriptors lie in the same cell as a space partition, they are similar. In contrast, IRE defines a single partition of the descriptor space and measures their similarity by utilizing the Euclidean metric between R features in the embedded space.

For the descriptor vectors $y_l, l = 1, ..., m_1$ of the query video frame, and for the descriptor vectors $x_{i,j}, i = 1, ..., m_2$ of the database, compute the score s_j of the corresponding frame image by

$$s_j = \sum_{l=1}^{m_1} \sum_{i=1}^{m_2} f_{IRE}(x_{i,j}, y_l) \tag{6}$$

The final image matching score s_f used for ranking can be obtained from s_j by applying a post processing function as:

$$s_f = s_j / \sqrt{m_1 m_2} \tag{7}$$

which takes into account the influence of image descriptor numbers.

V. EXPERIMENTS

The proposed method is evaluated through comprehensive experiments on three datasets, including CC_WEB_VIDEO [13], MUSCLE VCD 2007 [38] and TRECVID-CBCD 2011 [39]. Our experiments are conducted on a Dell desktop computer under Windows 10 64 bits OS with I7 3.4 GHz CPU and 16 GB RAM.

To evaluate NDVR accuracy and efficiency, we present an extensive comparison of the proposed method with a set of existing NDVR methods, which are briefly described as follows. A text retrieval approach to object matching in videos proposed by Sivic and Zisserman [40] , referred to as 'BOF'. An mage-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering proposed by Douze *et al.* [3], referred to as 'BOF + HE'. Multiple feature hashing proposed by Song *et al*. [41], referred to as 'MFH'. Multiple feature hashing with frame group information proposed by Song *et al*. [42], referred to as 'G-MFH'. Frame Fusion for Video Copy Detection proposed by Wei *et al.* [43], referred to as 'FF'. video copy detection using inclined video Tomography and bag-of-visual-words proposed by Min *et al.*. [44], referred to as 'TBOW'. And our proposed method referred to as 'G-BOF'. The baselines and the proposed scheme are implemented in C++ and run on single core. We rewrite the Matlab scripts of MFH [41] into C++ codes. Other methods are implemented by following the algorithms in the papers.

A. Datasets

CC_WEB_VIDEO is a well-known NDVR benchmark, which consists of 12,790 reference videos and 24 queries videos downloaded from Yahoo!, Google, and You Tube Video. Many state of the art methods show their great performances on the CC_WEB_VIDEO dataset.

MUSCLE VCD 2007 contains 10 GB videos of about100 hours. There are two query sets in the dataset. ST1 dataset is for for NDVR, which contains 15 query videos of about 2.5 hours, and ST2 dataset is for for NDVL, which contains 3 query videos of about 45 minutes. The queries are reproduced with the following transformations: blurring, color adjustment, re-encoding, horizontal flipping, frontal and non-frontal camcording, cropping, border, resizing, and addition of subtitles. The execution time and performance of NDVR are calculated from the retrieval results of dataset ST1.

TRECVID-CBCD 2011dataset is one of the most complex and largest benchmarks for video copy detection, which consists of about 420 hours 11,503 reference videos. Query videos include three types: a non-reference video only, a reference video only, and a reference video embedded into a non-reference video. Only the last two types of query videos are NDVs. One of the eight visual trans-formations listed in Table 1. The query videos are made by the tool downloaded from NIST.

TABLE I
LIST OF VIDEO TRANSFORMATIONS IN TRECVID-CBCD 2011

	Description
T1	Simulated camcording: by perspective transform, automatic gain control, and blurring effects
T2	Picture-in-picture (PiP): The original video is embedding in front of the background video.
T3	Insertions of pattern
T4	Strong re-encoding
T5	Change of gamma
T6	Decrease in quality: 3 random selected combination of blur, gamma, frame dropping, contrast, compression, ratio, white noise
T8	Post production: 3 random selected combination of crop, shift, contrast, text insertion, vertical mirroring, insertion of pattern, picture-in-picture (the original video is in the background)
T10	Three randomly selected transformations chosen from T2-T6 and T8

B. Performance Evaluation Metrics

Mean average precision (MAP) is utilized for evaluating NDVR on CC_WEB_VIDEO and MUSCLE VCD 2007 datasets. Precision-recall (PR) curves are adopted to show the trend of tradeoff between recall and precision on these datasets.

In order to measure the performance of approaches on TRECVID -CBCD 2011, we use the evaluation metrics utilized in TRECVID-CBCD 2011 contest [39]. The Normalized Detection Cost Rate (NDCR) is used to measure the accuracy of NDVR for every video transformation, which is defined as follows

$$NDCR = \frac{FN}{N_{t\arg et}} + \frac{C_{FA}}{C_{Miss} \times R_{t\arg et}} \times \frac{FP}{T_{refdata} \times T_{query}} \tag{8}$$

where FP and FN indicate the numbers of false positives and false negatives, N_{target} is the number of target videos, C_{Miss} and C_{FA} represent the respective costs of miss detections and false alarms. A prior target rate R_{target} is set to 0.5/hr^2. T_{query} and T_{refdata} indicate the queries for a transformation and the total lengths (in hours) of entire reference dataset, respectively. The smaller NDCR indicates the better retrieval performance.

The mean processing time (MPT) is use for the computational efficiency comparison, which is the mean time (in second) to execute a query for the retrieval and localization. The execution times of shot detection, sampling, and feature extraction are excluded from MPT calculation.

C. Parameter Evaluation of the IR Feature

There is one parameter θ in the IR-transform during the computation of the proposed IR feature. Therefore, we evaluate the parameter in our method on the MUSCLE VCD 2007 dataset. Moreover, we test whether the performance is enhanced by using $(2D)^2$PCA to process the feature obtained from the IR-transform.

Parameters θ is sampled in the range of [0, 180]. Fig. 4 shows the experimental results of seven different numbers of samples for θ, namely $(\theta) = [1 : 3 : 180]$, [1 :5 : 180], [1 : 10 : 180], [1 : 15 : 180], [1 : 20 : 180], [1 :25 : 180] and [1 : 30 : 180]. The green curve is recognition accuracy using the $(2D)^2$PCA to refine the IR-transform feature, and the red one is the obtained recognition accuracy using the

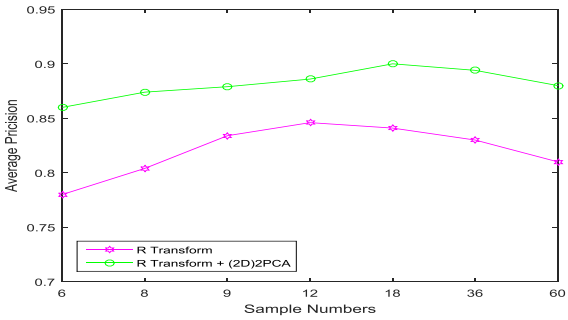


Fig. 4. Retrieval average precision obtained by the (2D)2PCA IR feature and the IR-transform feature with respect to seven different samplings of the parameter θ on MUSCLE VCD 2007 dataset.

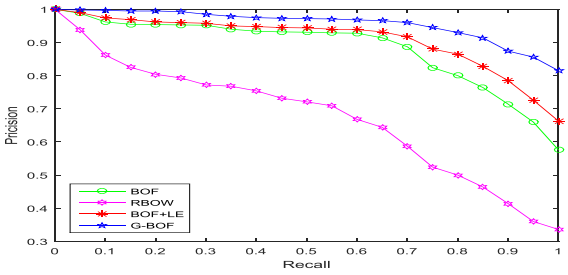


Fig. 5. Evaluate the fusion method on the MUSCLE VCD 2007 dataset.

IR-transform feature without $(2D)^2$PCA. From Fig.4, the following points can be observed: 1) the sampling frequency has a little influence on the final experimental result, and the best average precision of 90.87% is obtained under $(\theta) = [1 : 10 : 180]$; 2) the features refined by $(2D)^2$PCA gain a higher recognition precision in most cases than the IR-transform features on their own. The $(2D)^2$PCA refined features achieve 88.19% average recognition accuracy, and the features without refined achieve 82.07%. It demonstrates that IR-transform feature has a good description ability and the $(2D)^2$PCA further improves the discriminating ability of the IR-transform feature. We set $(\theta) = [1 : 10 : 180]$ in all other experiments on both datasets, and utilize $(2D)^2$PCA on IR-transform as the final IR feature. The final IR feature is an 8×8 matrix.

D. Parameter Evaluation of the Feature Fusion

We evaluate the fusion method on the MUSCLE VCD 2007 dataset. In these experiments the size of codebooks are set to 10000. The IR feature can be used for building codebook and the local feature for embedding to BOF, referred to as 'RBOF + LE'. Alternatively, the local feature can be used for building codebook and the IR feature for embedding to BOF, which is the proposed method referred to as 'G-BOF'. Besides, we test single feature based methods. Namely, we employ the single feature (local feature or IR feature) for building codebook, referred to as 'BOF' or 'RBOF' respectively. The above four experiments are shown in Fig.5. Moreover, Fig.5 shows that the IR feature boosts the retrieval performance by 5.79% with respect to the 'BOF' averagely.

E. Experimental Results on Different Datasets

MUSCLE VCD 2007: The experimental results comparison in terms of MPT and MAP among all the methods listed above is presented in Table 2. As illustrated in Table 2, BOF + HE, which achieves 100% MAP, meaning that all the near duplicate videos are retrieved. By involving BOF, MAP of the proposed G-BOF is still 100% and MPT is only raised from 0.61 seconds (BOF) to 0.73 seconds. That is, the retrieval process is almost kept the same efficiency without any precision loss. The MPT of BOF + HE takes 170% of that of G-BOF. It demonstrates the effectiveness and efficiency of the proposed G-BOF. As illustrated in Fig. 6(a), the precision-recall curves of TBOF, BOF + HE, and the proposed G-BOF dominate the curves of BOF, FF, MFH, and G-MFH. As elaborated in Table 2, the proposed G-BOF is efficient than FF, G-MFH and TBOF with the same outstanding MAP. MFH and G-MFH can also retrieve all the near-duplicate videos efficiently. They are very efficient but not so effective as the proposed G-BOF, since they encode all information of a long video into one hash code. In the evaluation of efficiency, our proposed G-BOF is more efficient than other approaches compared and just next to MFH and BOF.

TABLE II

PERFORMANCE COMPARISON IN TERMS OF MAP AND MPT AMONG ALL THE METHODS COMPARED ON MUSCLE VCD 2007 AND CC_WEB_VIDEO

Method	MUSCLE VCD 2007		CC_WEB_VIDEO	
	MAP	MPT	MAP	MPT
BOF[8]	0.891	0.61	0.867	0.89
BOF+HE[3]	1.000	1.24	0.955	1.67
MFH[41]	0.948	0.64	0.933	0.95
G-MFH[42]	0.956	0.82	0.942	1.18
FF [43]	0.978	1.75	0.953	1.97
TBOW [44]	0.986	1.86	0.954	2.15
G-BOF	1.000	0.73	0.968	1.04

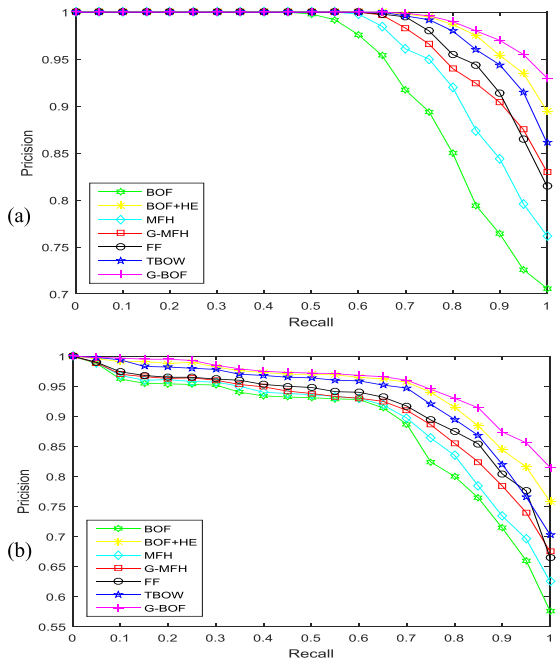


Fig. 6. Precision-recall curves among the proposed G-BOF and the other methods compared on (a) MUSCLE VCD 2007, (b) CC_WEB_VIDEO.

TABLE III
COMPAREISON IN TERMS OF MINIMAL NDCR AMONG THE PROPOSED METHOD AND OTHER APPROACHES ON TRECVID-CBCD 2011

	T1	T2	T3	T4	T5	T6	T8	T10	Average
BOF	0.743	0.417	0.214	0.143	0.158	0.341	0.552	0.451	0.377
BOF+HE	0.252	0.253	0.03	0.054	0.065	0.187	0.136	0.291	0.159
MFH	0.741	0.52	0.058	0.085	0.066	0.256	0.468	0.683	0.360
G-MFH	0.712	0.514	0.062	0.064	0.071	0.243	0.384	0.649	0.337
FF	0.623	0.316	0.126	0.052	0.124	0.201	0.355	0.404	0.275
TBOW	0.554	0.213	0.031	0.083	0.102	0.182	0.233	0.255	0.207
CNN-F	0.359	0.385	0.024	0.045	0.064	0.143	0.132	0.215	0.171
CNN-A	0.383	0.456	0.029	0.041	0.052	0.153	0.144	0.236	0.187
G-BOF	0.211	0.26	0.032	0.042	0.063	0.119	0.138	0.212	0.135

TABLE IV
COMPAREISON OF MPT AMONG THE PROPOSED METHOD AND OTHER APPROACHES ON TRECVID-CBCD 2011

	T1	T2	T3	T4	T5	T6	T8	T10	Averge
BOF	1.04	0.94	0.76	0.88	0.74	0.78	0.85	0.97	0.87
BOF+HE	1.96	1.44	1.48	1.35	1.37	1.75	1.84	1.98	1.65
MFH	0.91	0.95	0.87	0.86	0.94	0.92	0.96	1.16	0.95
G-MFH	1.512	1.41	0.82	0.94	0.79	1.24	1.38	1.45	1.19
FF	2.04	2.16	1.83	1.86	1.86	1.97	1.99	2.12	1.98
TBOW	2.65	2.51	1.89	1.88	1.76	1.99	2.38	2.44	2.19
CNN-F	1.85	1.82	1.67	1.62	1.59	1.64	1.78	1.85	1.73
CNN-A	1.81	1.79	1.66	1.58	1.60	1.61	1.71	1.78	1.69
G-BOF	1.18	1.09	0.94	0.88	0.96	0.99	1.08	1.09	1.03

CC_WEB_VIDEO: As shown in Table 2, the MPTs of BOF, MFH and the proposed G-MFH are very closing on CC_WEB_VIDEO, showing little performance difference. Fig. 6(b) shows the precision-recall curves of the seven methods on CC_WEB_VIDEO, wherein our proposed G-BOF is still the best one. Comparing with the state-of-the-art approaches BOF, FF, BOF + HE, TBOW, MFH, and G-MFH, the proposed G-BOF still achieves the highest MAP, as illustrated in Table 2. G-MFH slightly outperforms MFH in terms of MAP. As shown in Table 2, the MPT of G-BOF takes 62% of that of BOF + HE. Although MFH and G-MFH can retrieve in a shorter time of 0.93 seconds, they are incapable of partial NDVR.

TRECVID-CBCD 2011: As shown in Table 3, the high average NDCR value indicates the low NDVR precision. From Table 3 and 4, we can see that, in terms of average NDCR, the proposed G-BOF outperforms the state of the art methods. The NDCRs of the proposed G-BOF on T3 and T8 are higher than those of BOF + HE since BOF + HE uses the SIFT geometric registration technique, which is effective but time consuming in matching the interest points. The performance illustrated in Table 3 show that our proposed G-BOF achieves promising results on most of the transformations. Table 4 shows the MPTs of the methods compared. The MPT of BOF takes 83% of that of G-BOF. The proposed G-BOF is 1.6 times more efficient than BOF + HE. Though G-MFH can accomplish the retrieval in one second, it can not achieve better NDVR performance in the TRECVID-CBCD 2011 dataset. Pooling on CNN features is a very good video representation method. Here, we compare with two CNN methods: CNN feature with average pooling (CNN-A) and CNN feature with fisher vector (CNN-F). The CNN methods are achieved good results. However, in NDVR, the length of reference video and query video may be very different, and this global video feature will contain unnecessary information, so the accuracy of these global video representations is limited. Moreover, CNN is also a little time consuming.

TABLE V
THE MEMORY COSTS (MB) OF THE REFERENCE CODEBOOK ON THREE DATASETS

	BOF	BOF+HE	TBOW	G-BOF
MUSCLE VCD 2007	432	432	568	476
CC_WEB_VIDEO	1288	1288	1695	1420
TRECVID-CBCD 2011	1475	1475	1941	1627

BOF, BOF + HE, TBOW and G-BOF adopt the inverted index structure. The memory costs of the reference codebook on three datasets are list on Table 5. BOF, BOF + HE have the lowest memory cost, and the proposed method need extra memory to deal with R feature. MFH, G-MFH, FF, CNN-F and CNN-A are not using the inverted index structure, so, Table 5 has not list their memory costs.

Overall Performance Discussion: Although the proposed G-BOF is not always the best in terms of NDCR and MPT, the experiment results show its competitiveness to the state of the art methods compared. We can see that most of the state of the art methods can deal with some but not all datasets. For instance, MFH and G-MFH show their competitive performance on MUSCLE VCD 2007 and CC_WEB_VIDEO but cannot perform well on the TRECVID-CBCD 2011 dataset, since it cannot deal with videos with partial geometric transformations.

F. Influence on the Number of Centres

For evaluating the influence on the number of centres, we conduct another experiment on CC_WEB_VIDEO.

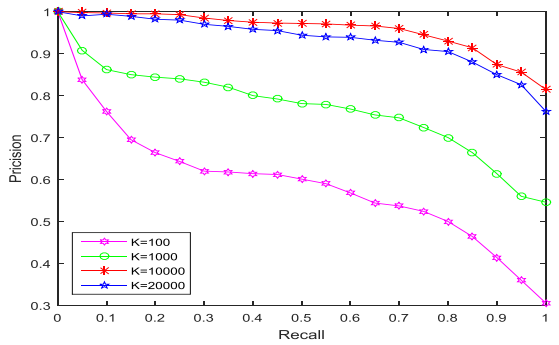


Fig. 7. PR curves of G-BOF by changing the number of centres K on CC_WEB_VIDEO.

As shown in Fig. 7, the proposed approach can achieve the best result when the number of centre K = 10000. When K = 1000 or 100, the dissimilar video frames might be assigned to the same centre, resulting in a poor performance. On the other hand, the similar video frames may be assigned to different centres because of the over fitting problem when K = 20000. The numbers of centres of TRECVID-CBCD 2011and MUSCLE VCD 2007 are 50000 and 2500 respectively.

VI. Conclusion and Future Work

In this paper we proposed a new NDVR framework based on IR features embedded BOF. In order to capture the global geometrical distribution information, we first presented a new holistic video representation, the 2D IR-transform on local interest points. We then proposed a new fusion strategy (IR features embedded BOF) to combine the global IR feature and the local feature for NDVR. Convincing experimental results on several publicly available datasets demonstrate that our proposed approach outperforms the state-of-the-art approaches in near duplicate video retrieval.

There still exist some limitations in our approach. For example, the near duplicate videos with high speed fast forwarding, super-slow motion or picture-in-picture transformation cannot be accurately retrieved and localized. This opens a door for new exploration that a multi-feature mechanism and more robust features may be involved to further make better the effectiveness of the NDVR system. In the future we intend to incorporate the temporal correlation of the keyframes within the same video to further improve the performance.

References

[1] C.-L. Chou, H.-T. Chen, and S.-Y. Lee, "Pattern-based near-duplicate video retrieval and localization on Web-scale videos," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 382–395, Mar. 2015.

[2] R. Fernandez-Beltran and F. Pla, "Latent topics-based relevance feedback for video retrieval," *Pattern Recognit.*, vol. 51, pp. 72–84, Mar. 2016.

[3] M. Douze, H. Jegou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 257–266, Jun. 2010.

[4] X. Zhou, L. Chen, and X. Zhou, "Structure tensor series-based large scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1220–1233, Aug. 2012.

[5] J. Liu, Z. Huang, H. T. Shen, and B. Cui, "Correlation-based retrieval for heavily changed near-duplicate videos," *Trans. Inf. Syst.*, vol. 29, no. 4, Dec. 2011, Art. no. 21.

[6] J. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo, and W. Wang, "Near-duplicate video retrieval: Current research and future trends," *Comput. Surv.*, vol. 45, no. 4, Aug. 2013, Art. no. 44.

[7] B. Cui, A. K. H. Tung, C. Zhang, and Z. Zhao, "Multiple feature fusion for social media applications," in *Proc. Int. Conf. Manag. Data*, Jun. 2010, pp. 435–446.

[8] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 591–606, Apr. 2009.

[9] M. Dater, N. Immorlica, P. Indyk, and V. S. Mirrpkni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Symp. Comput. Geometry*, Jun. 2004, pp. 253–262.

[10] H. Wang, M. Muneeb Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2009, pp. 124.1–124.11.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[12] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[13] X. Wu, A. G. Hauptmann, and C. W. Ngo, "Practical elimination of near-duplicates from Web video search," in *Proc. 15th Int. Conf. Multimedia*, Sep. 2007, pp. 218–227.

[14] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1798–1807.

[15] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 1–13, Oct. 2016.

[16] X. Wu, C.-W. Ngo, A. G. 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. 196–207, Feb. 2009.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[18] M. Douze, H. Jégou, C. Schmid, and P. Pérez, "Compact video description for copy detection with precise temporal alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 522–535.

[19] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *Proc. 9th Int. Conf. Multimedia*, Dec. 2011, pp. 1437–1440.

[20] Z. Huang, H. T. Shen, J. Shao, B. Cui, and X. Zhou, "Practical online near-duplicate subsequence detection for continuous video streams," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 386–398, Aug. 2010.

[21] S. Poullot and S. Satoh, "Detecting screen shot images within large-scale video archive," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3203–3207.

[22] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1939–1946.

[23] P. Jain, B. Kulis, and K. Grauman, "Fast image search for learned metrics," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[24] P. Li, M. Wang, J. Cheng, C. Xu, and H. Lu, "Spectral hashing with semantically consistent graph for image indexing," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 141–152, Jan. 2013.

[25] L. Zheng, S. Wang, L. Tian, H. Fei, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.

[26] Z. Wu, S. Jiang, and Q. Huang, "Near-duplicate video matching with transformation recognition," in *Proc. 17th Int. Conf. Multimedia*, Oct. 2009, pp. 549–552.

[27] C. G. M. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 638–647, Aug. 2005.

[28] G.-H. Cha, "Capturing contextual relationship for effective media search," *Multimedia Tools Appl.*, vol. 56, no. 2, pp. 351–364, Jan. 2012.

[29] J. Liu, Z. Huang, H. Cai, H. T. Shen, C. W. Ngo, and W. Wang, "Near-duplicate video retrieval: Current research and future trends," *Comput. Surv.*, vol. 45, no. 4, Aug. 2013, Art. no. 44.

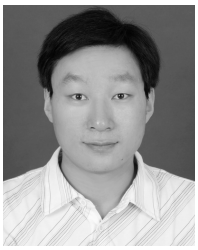
[30] H. Jegou, H. Harzallah, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[32] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[33] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.
[34] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2169–2178.
[35] Z. Zhang and Z. H. Zhou, "Letters: (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, Dec. 2005.
[36] K. Mikolajczyk and C. Schmid, "Scale affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, Oct. 2004.
[37] G. Shakhnarovich, T. Darrell, and P. Indyk, "Nearest-neighbor methods in learning and vision," *IEEE Trans. Neural Netw.*, vol. 19, no. 2, p. 377, 2008.
[38] L. To, A. Joly, and N Boujemaa. (2007). *Muscle-VCD-2007: A Live Benchmark for Video Copy Detection*. [Online]. Available: http://www.rocq.inria.fr/imedia/civr-bench/.
[39] W. Kraaij and G. Awad. (Nov. 2011). *TRECVID 2011 Content-Based Copy Detection: Task Overview*. [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.ccd.slides.pdf
[40] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2003, pp. 1470–1477.
[41] 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. 9th ACM Int. Conf. Multimedia*, Dec. 2011, pp. 423–432.
[42] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.
[43] S. Wei, Y. Zhao, C. Zhu, C. Xu, and Z. Zhu, "Frame fusion for video copy detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 1, pp. 15–28, Jun. 2011.
[44] H. s. Min, S. M. Kim, W. De Neve, and Y. M. Ro, "Video copy detection using inclined video tomography and bag-of-visual-words," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 562–567.

Kaiyang Liao received the B.S. degree in computer science from Xidian University, Xi'an, China, in 2004, the M.S. degree in computer science from the University of Science and Technology at Liaoning, Anshan, China, and the Ph.D. degree in information and communication engineering from Xi'an Jiaotong University, Xi'an, in 2013. He is currently a Full Lecturer with the School of Printing and Packaging Engineering, Xi'an University of Technology, Xi'an. His research interests include data mining, pattern recognition, video analysis, and retrieval.

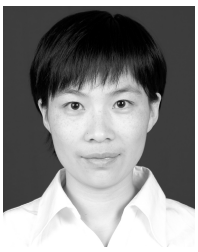
Hao Lei received the B.E. degree in printing engineering from the Xi'an University of Technology, Xi'an, China, in 2018, where he is currently pursuing the master's degree. His research interests include image retrieval and digital image processing.



Yuanlin Zheng received the B.S. degree in printing engineering from the Zhuzhou Institute of Technology, Zhuzhou, China, in 1999, the M.S. degree in pulp and papermaking engineering from the Xi'an University of Technology, Xi'an, China, in 2002, and the Ph.D. degree in pulp and papermaking engineering from the Tianjin University of Science and Technology, Tianjin, China, in 2007. He is currently a Full Associate Professor in printing engineering with the Xi'an University of Technology. His research interests include color management, evaluation of quality of color image and color science, and pattern recognition.

Guangfeng Lin received the Ph.D. degree in control theory and control engineering from the Xi'an University of Technology. He is currently a Lectorate with the Department of Information Science, Xi'an University of Technology. His research interests include digital image processing and pattern recognition. He is a CCF Professional Member and an ACM member.

Congjun Cao received the B.S. degree in printing machinery and technology and the M.S. degree in printing engineering from the Xi'an University of technology, Xi'an, China, in 1992 and 1998, respectively, and the Ph.D. degree in computer software and theory from Northwest University, Xi'an, in 2008. She is currently a Full Professor with the School of Printing, Packaging Engineering and Digital Media Technology, Xi'an University of Technology. Her research focuses on color management technology, quality control technology of printing image reproduction and functional printed materials, video analysis, and retrieval.



Mingzhu Zhang received the B.S. degree in computer science from Xidian University, Xi'an, China, in 2004, and the M.S. degree in management science and engineering from Xi'an Technological University, Xi'an, in 2011. She is currently a Full Lecturer with the Department of Public Courses, Xi'an Fanyi University, Xi'an. Her research interests include data mining, pattern recognition, video analysis, and retrieval.

Jie Ding is currently pursuing the bachelor's degree with the Xi'an University of Technology, Xi'an, China. Her research interests include image retrieval and digital image processing.