Wireless Pers Commun (2018) 103:401–416
https://doi.org/10.1007/s11277-018-5450-x

Video Copy Detection Based on Deep CNN Features and Graph-Based Sequence Matching

Xin Zhang[1] · Yuxiang Xie[1] · Xidao Luan[2] · Jingmeng He[1] · Lili Zhang[1] · Lingda Wu[3]

Published online: 2 March 2018

Abstract This paper introduces a novel content-based video copy detection method using the deep CNN features. An efficient deep CNN feature is employed to encode the image content while retaining the discrimination capability. Taking advantage of the extremely fast Euclidean distance similarity of deep CNN features, a keyframe-based copy retrieval method that exhaustively searches the copy candidates from the large keyframe database without indexing is proposed. Moreover, a graph-based sequence matching algorithm is employed to obtain the copy clips and accurately locate the video segments. The experimental evaluation has been performed to show the efficacy of the proposed deep CNN features. The promising results demonstrate the effectiveness of our proposed approach.

Keywords Video copy detection · Convolution neural networks · Deep learning · Computer vision

1 Introduction

In recent years, with the rapid development of the video capture devices, the volume of professional and user-generated videos is growing exponentially. Each day, there are tens of thousands videos generated, uploaded and published. Among these humongous video data, there is a large majority of videos belongs to copies or partial copies. This brings the increased concern about copyright issues due to the low cost of copying a video (or a segment of it) and massively distributing it on the Internet. As a consequence, an effective

✉ Yuxiang Xie
yxxie@nudt.edu.cn

[1] College of Information System and Management, National University of Defense Technology, Changsha, Hunan, China

[2] Department of Mathematics and Computer Science, Changsha University, Changsha, Hunan, China

[3] The Key Lab, The Academy of Equipment Command and Technology, Beijing, China

and efficient solution for video copy detection, which aims at automatically identifying copies in the large video dataset, has received significant research attention. This paper addresses the issue of video copy detection and localization. Video copy detection is a challenging issue for the following reasons. First, the complex transformations of the video content makes it difficult to represent the video frame; Second, what makes it more complicated is that the type of copy pattern could be copy or partial copy; Third, the length of copy video clips is ranging from a few seconds to a few hours.

The key of a video copy detection algorithm is to extract robust and discriminative features from video. Lots of researches have employed handcrafted features for video copy detection and localization. Compared with global features, such as color histogram, local binary pattern (LBP) [1] and histogram of gradient (HOG) [2], local features are more robust to the variation of scale, affine and rotation. The local features are usually constructed through extracting a collection of interest points, and projected to a fixed-length description. The scale-invariant feature transform (SIFT) [3] is likely to be the most used local features for video copy detection. In particular, these features are carefully designed for the special tasks. However, it is hard to decide which feature is proper for the special task, and the performance are highly dependent on the experience of feature designer. Lately, deep models have been overwhelmingly successful in a broad range of applications, such as computer vision, machine translation and natural language processing. Deep convolutional neural networks (CNN) [4, 5], in particular, have enjoyed huge success in tackling many computer vision problems in the past few years through the high-level feature learned from the raw data directly. Perkins [6] compared three pre-trained deep networks on near-duplicate video detection task and achieved superior performance.

The other crucial issue is how to detect and locate copy video segments based on frame level matching result. Most current methods employ the time correlation of video data, and treat the copy video detection and location as a graph or a network problem. These approaches consider temporal consistency within video segments, and can eliminate the influence caused by the mismatch of keyframes.

In this paper, we propose a novel video copy detection and localization framework based on CNN feature and graph-based segment matching. The motivation of this framework is jointly considering deep CNN feature based on visual similarity and temporal consistency. The CNN architecture is typical used as a feature extractor at frame level. The graph-based segment matching is used to tackle with the temporal relationship of video clip, and find the retrieved copy/partial copy video segment, as well as the location of the video segments. Figure 1 illustrates the framework of our proposed approach. It consists of two parts. (1) Offline procedure. First, the reference video database is partitioned into a keyframe set by frame sampling technique. Then, the deep features are extracted at each keyframe. All the features are stored for efficient keyframe matching; (2) Online procedure. For each query video, after the keyframe and deep feature extraction, a top-k matching matrix is constructed through the pairwise similarity computation between feature database and query video keyframe set. Last but not the least, the graph-based matching module is used to find and locate the most likely copy/partial copy video segment in reference videos.

The main contributions of the proposed approach are as follows:

1. Deep CNN features, instead of handcrafted feature representation, are used for the image visual content encoding.
2. We present a graph-based fragment matching strategy for video copy detection and localization, which is capable for both copy and partial copy pattern.

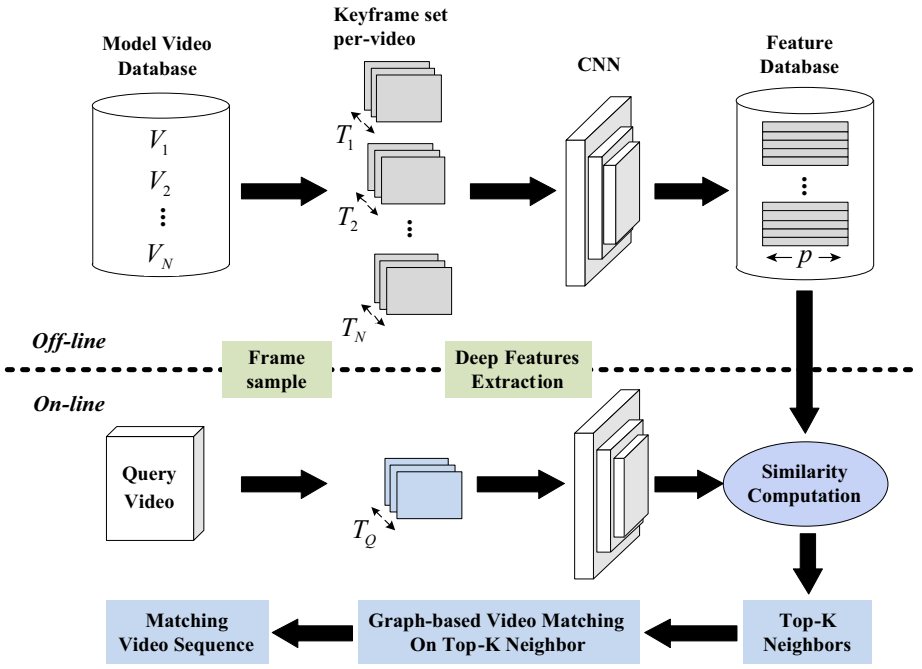

Fig. 1 The framework of our proposed approach

3. We construct a full stage framework for video copy detection and localization. It performs better than other methods which are based on handcrafted features.

The rest of the paper is organized as follows. Section 2 reviews related work on video copy detection and deep learning. Section 3 presents the proposed algorithm framework and describes each stage in detail. The proposed approach is evaluated under our copy detection dataset (Sect. 4). Finally, we summarize our findings in Sect. 5.

2 Related Works

A few representative approaches on video copy detection are first reviewed in this section. And then we briefly discuss some popular deep learning methods.

2.1 Video Copy Detection Approaches

The definition of copy video indeed varies depending on the target applications. Generally, a video copy is a segment of video sequence that is transformed from reference video by means of inserting patterns, compression, change of gamma, decrease in quality, camcording, and so on [7].

Most video copy detection approaches contain the following procedure: feature extraction from frames, frame-level matching, and final copied segments matching through temporal alignment. The most important issue in video copy detection research is to find the robust and distinctive features that can reflect the essential content of videos. Features

for video copy detection followed two main broad categories. One type is global descriptors. In general, global descriptors build frame level representation using the whole area of video frames. Global features such as shape, color histogram [8] and texture features like local binary pattern (LBP) [1, 9, 10] are greatly applied to video copy detection. In [11], each video frame was divided into several blocks and improved block histogram were then used to represent the frame content. In [12], color correlation on summarized videos was used for efficient computation. Meanwhile, this method was capable of handling videos with different frame rate without processing video on the whole. Huang et al. [13] used global features such as color histograms and the reference point to represent video frame, which is robust for the resize, shift and flip transformations. Shen et al. [14] encoded both video spatial (RGB and HSV color histograms) and temporal information into a single vector, and introduced a real-time near-duplicate video detection system named UQLIPS.

Local invariant descriptors are another category of widely used features in video copy detection. The local descriptors on corners, shapes, points and lines play a significant role in image and video copy detection. Among them, descriptors based on points such as SIFT are widely used [15, 16]. Bag-of-words representation of local descriptors was invited for video frame content description [17]. In [16], a SVD-based method was proposed to match two video frames with SIFT point set descriptors. Similarly, Liu [18] proposed a computationally efficient algorithm based on SURF descriptors and local harsh indexing.

Methods based on global feature were computationally efficient and can be computed in real-time, but they lost the capacity on detecting copies with complicated transformations. The local features can handle with harsh transformations, such as affine distortion, change of viewpoints, and additive noise, but they were expensive in space and time and weak to noise and frame insertion. Researchers tried to combine global and local features together, and the experiments showed its potential on video copy detection. Through applying feature fusion strategies in the early stage to multiple features, bag of words (BoW) based method will improve the capacity and discriminability of the feature representation, which was widely applied and validated in image retrieval [19] and image copy detection [20], as well as instance-Level object retrieval and recognition [21]. Wu et al. [22] presented a hierarchical approach for near-duplicate web video search. Especially, color histograms was first used for fast sampling, and then local feature based near-duplicate detection was employed for further accurate duplicate analysis.

Video segments matching is another major task of video copy detection using the temporal consistence [8, 11, 16, 17]. Two new sequence-matching techniques were proposed for copy detection respectively based on motion and histogram by Hampapur [8]. In [11], a dynamic video sequences matching was employed to accelerate the matching procedure. Based on the pair-wise constraints generated from keypoint matching, Tan [17] converted partial alignment into a network flow problem through constructing a temporal network. Similarly, the video sequence matching problem was converted into finding the longest path in the frame matching result graph by Liu et al. [16].

Meanwhile, many current researches adopted the spatio-temporal feature-based approaches which were widely used in video copy detection. A novel method was proposed in [23] to address the news web video event mining issues, and a compact spatio-temporal feature was introduced to represent each video segment. Specifically, the spatio-temporal feature was detected by Harris3D detector and represented by a set of feature vectors with HOG/HOF descriptors. Zhu et al. [24] described a temporal-concentration SIFT (TCSIFT) for large-scale video copy retrieval, which encoded with temporal information by tracking the SIFT.

2.2 Deep Learning

Deep convolutional neural networks (CNN), in particular, have tackling various computer vision tasks over the past few years. The powerful capacity of deep learning lied in the robust, distinctive and scalable features that were learned directly from raw training data through neural network architecture instead of handcrafted. In 2012, Krizhevsky et al. [5] trained a 7 layers CNN on 1.2 million labeled images. The proposed AlexNet won the first prize (Top-5 error 15.3%) on ILSVRC image classification competition, which surpassed the second (Top-5 error 26.1%) by a large margin. The tremendous success rekindled interest in CNN, and CNN was widely applied to a broad range of computer vision applications, such as object detection, object segmentation and so on. In the past 5 years, CNN architectures have seen tremendous development, AlexNet [5], Clarify [25], VGGNet [26], NIN [27], GoogLeNet [28], ResNet [29]. The bloom of the deep learning gives some new hints to the video copy detection. Wang et al. [30] proposed a efficient video copy detection method based on the compact CNN features. In [6], features separated extracted from AlexNet, R-CNN, GoogLeNet were tested for near-duplicated video detection. In [31], a large-scale video copy database (VCDB) on partial video copy detection was introduced. Based on VCDB, Jiang et al. evaluated two neural networks; the experiment showed that the CNN features performed well on partial video copy detection.

3 Methodology

Given the reference video database, and a set of query videos that are generated by applying some transformations on the corresponding reference videos, our task is to detect the correct copy, partial-copy or claim no copy is found for each query video from the reference video database using the video content information. In this paper, we propose a novel approach which employs deep CNN features to represent frame content, and temporal consistency is considered to detect and locate video copy segment. In the following section, we will introduce each step in detail.

3.1 Video Frame Sampling

Video data always consists of a great number of frames, and these frames usually restore lots of redundancy information. For example, a 10 min video contains approximately 15,000 frames. Extracting all video frames is time consuming and contributing little to the final results. To efficiently capture the video characters, keyframe-based sampling is widely taken to reduce the frame number to be processed. There are two commonly seen sampling methods, one is keyframe extraction based on shot boundary detection, and the other is sampling frames at a fixed sampling rate. Since the shot boundary detection based techniques is time expensive, we use the sampling method to extract keyframe from video data. In this paper, we adopt a certain sampling ratio of 1 frame/s.

3.2 Deep Feature Extraction

As mentioned in Sect. 2, global feature-based methods are weak to detect copies with complicated transformations such as picture in picture. It has been validated by many studies that local feature-based image retrieval is quite efficient in both space and time

when combined with vector quantization with a large visual vocabulary (e.g., of 1 M visual words) and an inverted index. Unlike global features and the activations of a fully connected layer, local features are "local" and so usually robust as regards partial occlusion (e.g., caused by picture in picture), viewpoint changes, etc. In order to get rid of visual vocabulary learning, as well as simplify the detection architecture, we introduce a deep CNN feature-based approach which is robust to diversified transformation. As shown in Fig. 2, for each sampled keyframe, a 4096 dimension feature vector is extracted using the Caffe [32] implementation of the AlexNet [5]. The AlexNet architecture is independent trained on ImageNet [33] dataset. The video keyframes are directly rescaled into 227 × 227, and a mean value is subtracted. We directly use the output of the sixth fully connected layer of AlexNet as the keyframe level representation. The deep CNN features contain both global information and local description hierarchically, which builds a comprehensive description of the keyframe. The detail of AlexNet architecture is beyond our scope, please refer to [5] for more detail information. After feature extraction, Euclidean distance is used to measure the similarity between two video keyframes.

3.3 Temporal Consistency in Video Segment Matching

Video data is not just a collection of continuous frames, the inherent temporal consistency between adjacent frames play a key role for video copy detection. Since there are errors in keyframe level matching results, the inherent temporal consistency of the video data is employed to eliminate the keyframe level error. In this paper, the graph-based video sequence matching method proposed by Liu [16] is employed to cope with the copy segment detection and localization. In this section, we will briefly introduce the graph-based video sequence matching method. The method is presented as follows:

Stage 1: *Frame level matching matrix generation*. Supposing that $Q = \{F_1^Q, F_2^Q, F_3^Q, \ldots, F_n^Q\}$and $M_c = \{F_1^M, F_2^M, F_3^M, \ldots, F_m^M\}$ are the keyframe set of query video and reference videos, respectively. For each keyframe F_i^Q in the query video, compute the similarity $sim(F_i^Q, F_j^M)$ with every keyframe in reference video database, and return k largest matching results. For each keyframe in query video, top-k unique matching

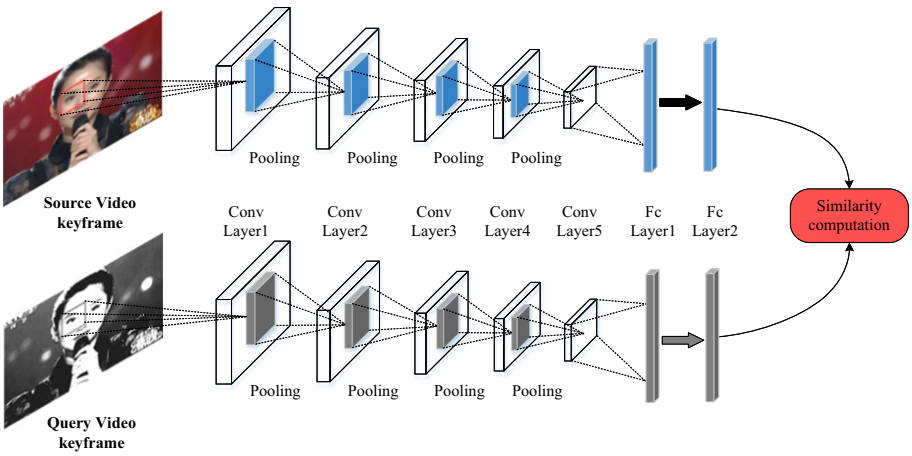

Fig. 2 An illustration of our CNN feature based similarity computation

frames are selected from reference video keyframe set. Then we will get a $n \times k$ matching matrix, where k is set to 5 based on our empirical study.

Stage 2: *Convert matching matrix into hierarchical directed acyclic graph and find the longest path* The $n \times k$ matching matrix can be converted into a hierarchical directed acyclic graph. In Fig. 3, the node $F_{i,j}^M$ means a matching between keyframe F_i from query video and keyframe F_j from reference video. There exists an edge between two nodes if the following two criteria are satisfied at the same time.

Time orientation consistency For $F_{i,j}^M$ and $F_{k,l}^M$, if $(i-k) \times (j-l) > 0$, then the two nodes satisfy the time orientation consistency.

Time span degree The time span degree between $F_{i,j}^M$ and $F_{k,l}^M$ is defined as:

$$\Delta t_{i,j}^{k,l} = \max\{|i-k|, |j-l|\} \tag{1}$$

Video data follows inherent time direction. If a query video is defined as a copy video, then the time direction of query video and reference video must satisfy the time consistency, which is reasonable for real applications. If the time span meets $\Delta t > \tau$ (τ is a preset threshold based on the experiment), then we consider there is no link between these two matching results.

As illustrated in Fig. 3, the solid lines are those satisfy both the two criteria. The blue line satisfies the first condition but the red line does not. After the hierarchical directed acyclic graph is built, there exists more than one path or only one path. The copy videos can be detected through finding the longest path in graph, which can be well settled by dynamic programming methods, such as Floyd [34]. The longest path can be determined by both the location and time length of the copy video. For example, there are two available longest paths as follows:

$$F_{1,224}^M \rightarrow F_{2,228}^M \rightarrow F_{3,229}^M \rightarrow F_{4,230}^M \rightarrow F_{5,231}^M \rightarrow F_{6,232}^M \rightarrow F_{7,233}^M$$
$$F_{1,227}^M \rightarrow F_{2,228}^M \rightarrow F_{3,229}^M \rightarrow F_{4,230}^M \rightarrow F_{5,231}^M \rightarrow F_{6,232}^M \rightarrow F_{7,233}^M$$

Stage 3: *Output final detection result* It may have more than one longest path in directed acyclic graph. For each path, we compute the similarity of video sequence and select the highest as the final result. According the start and end frame of video segment, the location of video copy can be obtained at the same time.

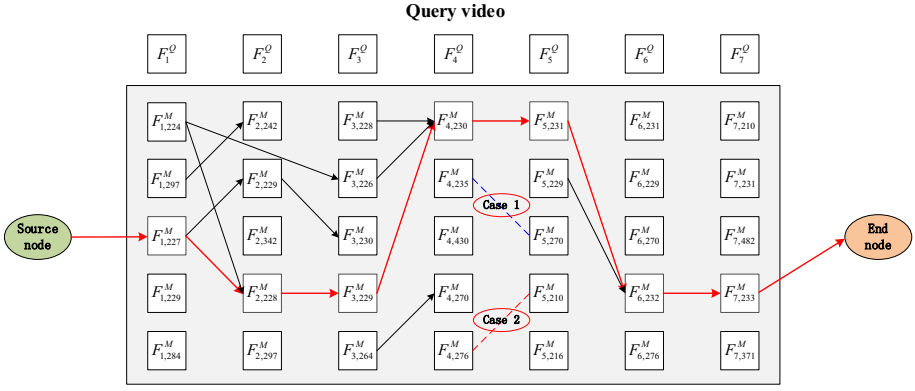

Fig. 3 Matching result graph based matching matrix between query video and reference videos

4 Results and Discussion

In this section, we will validate the effective and efficient of the proposed algorithm on the video copy detection dataset. Two key techniques will be evaluated in this section. The first experiment is to examine the effectiveness of the deep CNN feature. Second, the effectiveness of the graph-based video sequence matching method is validated. Furthermore, we study optimal parameters for graph-based matching method.

4.1 Experiment Settings

Image Retrieval Dataset The image retrieval dataset is Columbia's TRECVID2003 dataset [35], which consists of 600 keyframes with 150 near-duplicate image pairs and 300 non-duplicate images extracted from the TRECVID2003 corpus.

Video Copy Detection Dataset To evaluate our proposals, we use two video datasets. One is (CC_WEB_VIDEO), which is created by the Video Retrieval Group (VIREO) of City University of Hong Kong and Infomedia Group of Carnegie Mellon University. The other one is the dataset from Video Copy Detection in 2014 Specific Audio and Video Retrieval Challenge. CC_WEB_VIDEO contains 12,790 videos of 85G in total. The dataset of 2014 Audio and Video Challenge is about 100–200 h, 100G in total. We randomly choose 200 videos of 10–30 min from these two data sets. Then cut out a 2-min clip from each video to build the reference video. The test copy video clip is constructed manually through adding 6 transformations (T1, Change of gamma; T2, Change of contrast; T3, Tensile and add black border; T4, Occlusion; T5, Crop and Tensile; T6, Combination of random five transformations among all the transformations described above.) to reference videos. The test video dataset contains 627 transformed videos of copy subsequence and non-copy sequence combined, and the length of query videos varies from 30 s to 1 min, the location of copy clip is uncertain. Among them, 407 query video are partial copy videos of reference videos, 220 videos are non-copy videos. All the reference videos are unique. Each partial copy video only contains 1 copy clip. The videos are stored in mp4 format, and include four contents such as movies, news, documentary. Figure 4 illustrate one samples for each topic category on various transformation patterns.

Fig. 4 Example copy frames from the videos, ordered from left to right by transformations (T0–T6) and top to bottom by topic categories: movies, news, documentary, tv series

All codes are written in C++ based on Caffe and conducted on an Intel Core i5-6200U (4 Core 2.3 GHz CPU and 8 GB) in a laptop.

Evaluation criteria To evaluate the performance of our proposed approach, we use the precision and recall measures to compare the effectiveness of our method with traditional local feature-based approaches.

Segment precision and recall is defined as below:

$$\mathrm{VP} = \frac{|\textit{correctly retrieved segments}|}{|\textit{all retrieved segments}|} \tag{2}$$

$$\mathrm{VR} = \frac{|\textit{correctly retrieved segments}|}{|\textit{groundtruth copy}|} \tag{3}$$

While the frame precision (FP) and recall (FR) are defined as:

$$\mathrm{FP} = \frac{|\textit{correctly retrieved frames}|}{|\textit{all retrieved frames}|} \tag{4}$$

$$\mathrm{FR} = \frac{|\textit{correctly retrieved frames}|}{|\textit{groundtruth copy frames}|} \tag{5}$$

The frame-level measures are introduced as auxiliary criteria to show the accuracy of copy video detection and location method. The final recall and precision of the method is:

$$\mathrm{Recall} = \mathrm{RV} \times 0.9 + \mathrm{FV} \times 0.1 \tag{6}$$

$$\mathrm{Precision} = \mathrm{RP} \times 0.9 + \mathrm{FP} \times 0.1 \tag{7}$$

And the score is defined as follows, and here $\beta = 0.3$:

$$\mathrm{Score} = \mathrm{Recall} \times (1 - \beta) + \mathrm{Precision} \times \beta \tag{8}$$

4.2 Feature Comparison

In this subsection, we compare the performance of deep CNN features with two existing local feature based copy video detection methods, which are briefly described as follows.

SIFT based method We extract SIFT feature detectors from each keyframe, and bag of words (BoW) method is employed to encode SIFT detectors. There are two important parameters in SIFT based method, the number of visual words N^v and the number of keypoints N^k extracted from each image. All the feature vocabularies are learned offline. The time cost of feature representation based on vocabulary and image matching is also evaluated.

Table 1 shows the Top-5 accuracy and image matching time at different parameter settings. If the number of visual words N^v becomes larger, we can observe an obvious decrease on the Top-5 accuracy as well as an increase on time cost. Moreover, the selection of N^k is crucial for the retrieval accuracy. The accuracy first increase then drop down along with the increase of the number of keypoints N^k. From Table 1, it can be seen that when $N^v = 500$, $N^k = 200$, SIFT feature-based approach achieved the best retrieval accuracy.

SURF based method We extract SURF feature detectors from each keyframe, and bag of words (BoW) method is employed to encode SURF detectors. There are two important parameters in SURF based method, the number of visual words N^v and the Hessian

Table 1 Comparison on SIFT + BOW method on Columbia's TRECVID2003 datasets, best result highlighted in bold

Number of visual words N^v	Number of keypoints N^k	Top-5 accuracy (%)	Time (s)
500	100	82.6	10.33
500	200	**84.8**	10.67
500	Unlimited	82.6	11.1
1000	200	78.3	13

Number of visual words: The number of visual words used in Bow (bag of words) method; Number of keypoints: the number of keypoints extracted from each image

minimal threshold T_H. All the feature vocabularies are learned offline. The time cost of feature representation based on vocabulary and image matching is also evaluated.

It can be seen from Table 2 that if the number of visual words N^v becomes larger, we can observe an obvious decrease on Top-5 accuracy and an increase on time cost. Moreover, the Hessian minimal threshold T_H also have a deep impact on the representation. From Table 2, it can be seen that when $N^v = 500$, $T_H = 800$, SURF feature-based approach achieved the best retrieval accuracy.

Finally, we employ the parameter settings which achieve the best performance based on SIFT and SURF features separately, and compared with AlexNet and VGGNet. The Top-5 accuracy of all the compared methods is showed in Table 3 on the Columbia's TRECVID2003 dataset. Specifically, Euclidean distance is employed for similarity computation. First of all, we can observe that deep feature based on CNN (AlexNet) achieves the best result (97.8%), which performs better than VGGNet, the original SIFT (BOW) and SURF (BOW) features. Moreover, the performance of SURF (BOW) feature is worse than the SIFT (BOW) feature. Thus, we can conclude that the deep CNN features preserve the discriminative capability of the original features by taking advantage of Euclidean distance.

4.3 Parameter Sensitivity Study

In this subsection we will study the performance variation at different parameter settings. As described in Sect. 3.3, there are two important parameters in the graph-based sequence matching method, the time span threshold τ and the minimal length degree k in the experiment.

Table 4 shows the performance variation at different parameter setting of graph-based sequence matching approach. If time span threshold τ becomes smaller, there will be a

Table 2 Comparison on SURF + BOW on Columbia's TRECVID2003 datasets, best result highlighted in bold

Number of visual words N^v	Number of keypoints N^k	Top-5 accuracy (%)	Time (s)
500	800	**73.9**	14.4
500	1600	54.3	13.6
1000	1600	30.4	15.3

Number of visual words: The number of visual words used in Bow (bag of words) method; Number of keypoints: the number of keypoints extracted from each image

Table 3 Comparison on different features on Columbia's TRECVID2003 datasets

Method	Top-5 accuracy (%)	Time (s)
AlexNet	97.8	14.8
VGGNet	93.5	37.8
SIFT + BOW	84.8	10.67
SURF + BOW	73.9	14.4

Table 4 The performance at different τ and k combination, best result highlighted in bold

τ	k	VP (%)	VR (%)	FP (%)	FR (%)	Score (%)
3	10	77.1	83.4	77.3	83.7	81.5
4	10	84.6	84.5	80.7	80.3	84.1
5	10	**84.0**	**85.7**	**97.7**	**76.1**	**84.9**
5	5	81.0	84.8	73.6	76.7	82.9

τ : time span threshold; k: minimal length of accepted copy video clips; VP: video precision; VR: video recall; VR: video recall; FP: frame precision; FR: frame recall; Score: the final evaluation score

smaller tolerance for the adjacent keyframe matching result, leading to a more precisely keyframe matching result. On the contrary, the matching stage will put more weight to the temporal consistency than visual information for copy video detection; Parameter k indicates the minimal length of accepted copy video clips. Ideally, the length of copy video varies from 1 s to hours. From Table 4, it can be seen that when $\tau = 5$ and $k = 10$, the graph-based matching approach achieved best detection performance.

4.4 Video Copy Detection Comparison

In this subsection, we will introduce the comparison of different features with graph-based matching method. We have submitted two local features using graph-based sequence matching approach described in Sect. 3. One employed the SIFT feature and the other used the SURF features, both the two features are encoded with BOW.

The best performance parameters setting of SIFT and SURF features are employed from Sect. 4.2. Compared with SIFT/SURF based methods, the proposed deep feature is more representative and distinctive for video copy detection. Meanwhile, The proposed method is based on the off-the-shelf VGGNet, which is used to generate the 4096-d feature representation without fine-finetuning. Table 5 shows the performance of three different features with graph-based sequence matching. The experimental results demonstrate the advantage of CNN feature over SIFT based method [14] and SURF based method [16]. From Table 5, it can be seen that our deep CNN (AlexNet) features based method obtains

Table 5 VP, VR, FP, FR, score (higher is better) and time (lower is better) results of three different features with graph-based sequence matching method

Method	VP (%)	VR (%)	FP (%)	FR (%)	Score (%)
Proposed method	84.0	85.7	97.7	76.1	84.9
SIFT(BOF) + graph [14]	79.7	77.8	74.7	58.9	76.9
SURF(BOF) + graph [16]	70.0	71.0	63.2	56.4	69.5

Table 6 VP, VR, FP, FR and score (higher is better) results of the proposed approach on different transformation patterns

Transformation pattern	VP (%)	VR (%)	FP (%)	FR (%)	Recall (%)	Precision (%)	Score (%)
T0	100	100	100	100	100	100	100
T1	100	100	100	78.6	97.9	100	99.2
T2	100	100	100	58.5	95.9	100	97.1
T3	33.3	33.3	100	25	32.5	40.0	34.8
T4	100	100	100	80.6	98.1	100	98.7
T5	100	100	100	97.7	99.8	100	99.8
T6	66.7	66.7	81.0	58.6	65.9	68.1	66.7
Overall	84.0	85.7	97.7	76.1	84.8	85.4	84.9

the promising copy detection performance for all the criteria. The deep CNN features can preserve most of the essential data information for the majority transformation patterns. Meanwhile, it is worth mentioning that we only use the basic AlexNet architecture trained on ImageNet without specific fine-tuning. Our proposed deep features and graph-based sequence matching method contribute to a desirable performance. Moreover, the proposed method never fails to a certain transformation, which shows the great generalization capability of the deep CNN features with the presented pre-processing techniques. Respectively, the proposed approach does not perform well for the picture in picture pattern because the picture in picture patterns may affect the deep feature representation.

Table 6 shows the results on each transformation pattern. We can see that our method performs well in most transformation pattern except type T3 (Tensile and add black border) and T6 (combination of disturb). It is because that the transformation T3 add black border to the original video, thus make the visual content a small partition of the feature representation, and which is difficult to recognize.

We show video copy detection results on a single video in Fig. 5. The length of the reference video is 130 s, and the copy video is transformed from reference video (10–40 s) by means of inserting. The dash lines indicate the ground truth location between reference

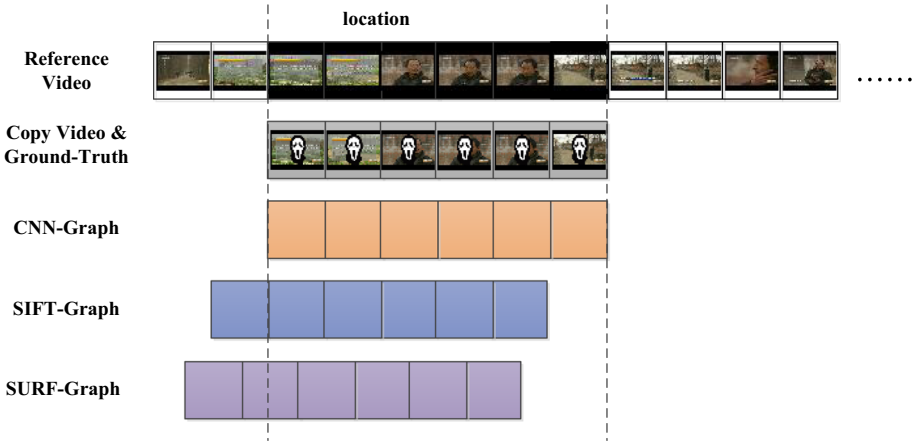

Fig. 5 Comparison of video copy detection results are represented in color bar form. The results of CNN-Graph, SIFT-Graph and SURF-Graph

video and copy video. We can see that the video copy detection by the CNN-Graph method is relatively accurate. It detected all keyframes in reference video. The SIFT-Graph method is inferior to CNN-Graph method for the video fragments are misclassified. The copy detection result of SURF-Graph is bad.

Obviously, the CNN feature has a more compact and informative representation, and the deep CNN features with graph-based sequence matching algorithm shows the leading performance according to the results. Meanwhile, the CNN feature can still describe the copy video keyframe with sufficient accuracy. On the contrary, the SIFT and SURF features cannot be well modeled with graph-based sequence matching algorithm, which leads to a poor performance. The representation metrics of keyframes is the key to the success of the proposed method. From above all, we can conclude that the desirable results are mainly due to the capability and the generalization capability of the deep CNN features.

5 Conclusion

In this paper, we presented a novel deep CNN feature approach to the content-based video copy detection using visual information. We used deep CNN features to describe video visual content. Accordingly, we developed a deep CNN feature based keyframe retrieval algorithm to exhaustively search the video copy candidates. What's more, a graph-based sequence matching method is employed to cope with copy video detection and localization. The extensive experiments demonstrated the effectiveness of our proposed video copy detection framework. Moreover, deep learning features perform better than SIFT, SURF handcrafted features, which shows a promising research direction for copy video detection. However, the method proposed in this paper doesn't work well for all types of transformation. For example, the proposed method may not perform such well on the transform type T3 and T6. In the future, significant efforts will be devoted into training better networks specifically for the copy detection problem, as well as improving the robustness of features on complex transformations.

Acknowledgements I would like to thank Jun Lei for helpful discussions and encouragement. This work has been supported by the National Natural Science Foundation of China under Contract Nos. 61571453, 61202336 and by the Natural Science Foundation of Hunan province under Contract No. 14JJ3010.

References

1. Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition, 29*(1), 51–59.
2. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE conference on computer vision and pattern recognition*, pp. 886–893.
3. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.
4. LeCun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
5. Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.
6. Perkins, L. N. (2015). Convolutional neural networks as feature generators for near-duplicate video detection.
7. Wu, C., Zhu, J., & Zhang, J. (2012). A content-based video copy detection method with randomly projected binary features. In *IEEE conference on computer vision and pattern recognition workshops*, pp. 21–26.

8. Hampapur, A., Hyun, K., & Bolle, R. M. (2001). Comparison of sequence matching techniques for video copy detection. In *International society for optics and photonics electronic imaging*, pp. 194–201.
9. Ojala, T., Pietikäinen, M., & Mäenpää, T. (2001). A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. In *Conference on advances in pattern recognition*, pp. 399–408.
10. Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. In *IEEE transactions on pattern analysis and machine intelligence*, pp. 971–987.
11. Jun, W., Lee, Y., & Jun, B. M. (2016). Duplicate video detection for large-scale multimedia. *Multimedia Tools and Applications, 75*(23), 15665–15678.
12. Thomas, R. M., & Sumesh, M. S. (2015). A simple and robust colour based video copy detection on summarized videos. *Procedia Computer Science, 46,* 1668–1675.
13. Huang, Z., Shen, H. T., Shao, J., et al. (2010). Practical online near-duplicate subsequence detection for continuous video streams. *IEEE Transactions on Multimedia, 12*(5), 386–398.
14. Shen, H. T., Zhou, X., Huang, Z., et al. (2007). UQLIPS: A real-time near-duplicate video clip detection system. In *VLDB endowment international conference on very large data bases*, pp. 1374–1377.
15. Dong, W., Wang, Z., Charikar, M., et al. (2012). High-confidence near-duplicate image detection. In *ACM international conference on multimedia retrieval*.
16. Liu, H., Lu, H., & Xue, X. (2013). A segmentation and graph-based video sequence matching method for video copy detection. *IEEE Transactions on Knowledge and Data Engineering, 25*(8), 1706–1718.
17. Tan, H. K., Ngo, C. W., Hong, R., et al. (2009). Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *ACM international conference on multimedia*, pp. 145–154.
18. Liu, D., & Yu, Z. (2015). A computationally efficient algorithm for large scale near-duplicate video detection. In *International conference on multimedia modeling*, pp. 81–490.
19. Jiang, F., Hu, H. M., Zheng, J., et al. (2016). A hierarchal BoW for image retrieval by enhancing feature salience. *Neurocomputing, 175,* 146–154.
20. Zhou, Z., Wang, Y., Wu, Q. M. J., et al. (2017). Effective and efficient global context verification for image copy detection. *IEEE Transactions on Information Forensics and Security, 12*(1), 48–63.
21. Wang, S., & Jiang, S. (2015). INSTRE: A new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 11*(3), 37.
22. Wu, X., Hauptmann, A. G., & Ngo, C. W. (2007). Practical elimination of near-duplicates from web video search. In *ACM international conference on multimedia*, pp. 218–227.
23. Zhang, C., Liu, D., Wu, X., et al. (2016). Near-duplicate segments based news web video event mining. *Signal Processing, 120,* 26–35.
24. Zhu, Y., Huang, X., Huang, Q., et al. (2016). Large-scale video copy retrieval with temporal-concentration sift. *Neurocomputing, 187,* 83–91.
25. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833.
26. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.
27. Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.
28. Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*, pp. 1–9.
29. He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pp. 770–778.
30. Wang L, Bao Y, Li H, et al. (2017). Compact CNN based video representation for efficient video copy detection. In *International conference on multimedia modeling*. Springer, Cham, pp. 576-587.
31. Jiang, Y. G., & Wang, J. (2016). Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data, 2*(1), 32–42.
32. Jia, Y., Shelhamer, E., Donahue, J., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM international conference on multimedia*, pp. 675–678.
33. Russakovsky, O., Deng, J., Su, H., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.
34. Hougardy, S. (2010). The Floyd–Warshall algorithm on graphs with negative cycles. *Information Processing Letters, 110*(8–9), 279–281.
35. Zhang, D. Q., & Chang, S. F. (2004). Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proceedings of the 12th annual ACM international conference on multimedia*, pp. 877–884.

Xin Zhang received his B.S. and M.S. degrees from National University of Defense Technology, Changsha, China, in 2011 and 2013 respectively. He is currently pursuing his Ph.D. degree in Control Science and Engineering from National University of Defense Technology. His research interests include image/video processing and deep learning.

Yuxiang Xie received her B.S., M.S. and Ph.D. degrees from National University of Defense Technology in 1998, 2001 and 2004 respectively, all in Systems Engineering. She is currently an associate professor in School of Information System and Management, National University of Defense Technology. Her research interests include image and video analysis, classification and retrieval.

Xidao Luan received his B.S. degree in Applied Mathematics in 1998, M.S. and Ph.D. degrees in System Engineering in 2005, 2009 respectively, all from National University of Defense Technology. He is currently an associate professor in Changsha University. His research interest is multimedia information processing and retrieval.

Jingmeng He received his B.S. degree from National University of Defense Technology in 2015. He is currently pursuing his M.S. degree in System Engineering from National University of Defense Technology. His research interests include image and video analysis and classification.

Lili Zhang received her B.S. degree from National University of Defense Technology in 2016. She is currently pursuing the M.S. degree in System Engineering from National University of Defense Technology. Her research interests include image and video analysis.

Lingda Wu received her Ph.D. degree in management science and engineering from the National University of Defense Technology, Changsha, China. She is currently a professor at the Key Laboratory, the Academy of Equipment Command and Technology, Beijing, China. Her research interests include multimedia information systems and virtual reality technology.