STRNN: End-to-end deep learning framework for video partial copy detection

Yanzhu Hu, Zhongkai Mu, Xinbo Ai
School of Automation, Beijing University of Posts and Telecommunications, Beijing, 100086, China
axb@bupt.edu.com

Abstract. The task of partial copy detection in videos aims at determine if one or more segments of the query video are already present in the data-set, while giving the information of similar portion time period. At present, most effective algorithms of partial copy detection in videos are designed as three steps: feature extraction, feature matching and time alignment. The separation of feature matching and time alignment module ignores the spatio-temporal information of partial copy to some extent. Therefore, satisfactory performance is not obtained. In order to reduce this loss, this article does not decompose it into two separate tasks, but using a single convolution neural network to solve these two aspects. First, we sample video frames and extract CNN features, calculate the spatio-temporal relationship matrix of the source video and the query video, and then graphically map the matrix and train the convolution neural network based on the object detection task of the RefineDet model. Finally, in the query phase, the time period of the partial copy is deduced based on the detection result. In this paper, we evaluate the performance of the algorithm on the real complex video copy detection data-set VCDB which is significantly improved compared with the state-of-the-art partial copy detection framework.

1. Introduction

With the popularity of video capture devices and network sharing activities, a large number of videos have been spread on online media. However, many videos are copied from another video and are similarly converted as shown in Fig.1, which also brings more copyright issues. Therefore, for video copy identification in big data sets, network video copy detection has received extensive attention. The application scenario of the technology is to use the existing video source to find the same or similar copy segment in the massive video, in addition to solving the above copyright protection problem, it can also deal with other problems such as video monitoring counting.
Fig. 1 Examples of copy mode

Early technology of partial copy detection in videos mainly used various traditional methods to extract features. The most popular method is to use local features such as SIFT to match and find similar video frame pairs, and then time alignment. In this type of method, visual bag-of-word model is widely recognized as an effective method. In fact, it also shows good performance on some simulated data sets, but when in face of more complex visual transformations in real copy data sets the performance of this method is far from satisfactory. Recently, with the success of deep learning methods, many algorithms of partial copy detection in videos based on deep neural network models have emerged, although they have achieved better results in feature representation than traditional methods, but the performance of time alignment is still disappointing. Therefore, as Jiang [1] evaluated, the precise time segmentation of partial copy detection in videos is still an open question in the research field.

Based on the above observations, this paper proposes a novel space-time feature fusion framework. We extend the research on image retrieval to generate video frame sequence feature maps to solve the problem of missing timing information in the matching process. The entire framework is divided into an offline phase and an online phase. In the offline preparation phase, video frames are first extracted from the video database, and then we encode key frames to CNN-based features, thereby convert videos into compact two-dimensional array description. Secondly, the paired two-dimensional array is used to calculate the spatio-temporal relationship matrix, and then the grey-scale map is used to generate the spatio-temporal relationship grey-scale map. The similarity values in these feature maps are measured by the cosine distance, and the brightness of the local area after grey-scale mapping is proportional to the similarity. Through experiments, we have found that this is the local highlighting feature that the copied segments is easy to present. Therefore, we manually label these highlighted block of the copied video pairs to train the convolutional neural network for the object detection task to make the model remember this pattern; In the online retrieval phase, we generate the grey-scale map of the space-time relationship between the search video and the data sets videos in the same way, and use the detection model to detect the feature block with the copied behaviours. Finally, we use post-processing to infer the copied segments of the original video.

2. Related works

The research of video copy detection technology has been more than ten years. At present, some important work proposed at home and abroad can be divided into two different detection levels according to different tasks: global copy detection and partial copy detection. If researchers only consider whether there is copied behaviour in query video, this detection is regarded as global copy detection. Relatively speaking, partial copy detection is for tasks that also need to know the time period of a specific copy. Although partial copy detection has higher precision, the retrieval process is usually more complicated and the retrieval efficiency is lower. Most existing studies divide partial
copy detection into three separate modules: feature extraction, feature matching, and time alignment. The three modules of feature extraction, feature matching and time alignment will be explained as follows.

Feature extraction is an important part of copy detection. Before the emergence of deep learning, partial feature-based methods were adopted by most works. For example, Douze [2] proposed the feature extraction scheme based on the STF-LBP algorithm. The algorithm clusters partial features into visual words and then uses Bow Model [3] maps to global features. However, a single video frame can produce more than a thousand partial features, which is a heavy burden for storage and retrieval. After the popularity of deep learning, CNN demonstrated its absolute advantages in image representation and high-speed processing. Therefore, CNN features are increasingly recognized, such as literature [3, 4], which have achieved better results than previous.

There are two types of methods for feature matching. One of the method is to globally fuse the video frame features first and then compare them. This feature fusion methods include feature clustering, histogram statistics, etc. The advantage of this type of method is that the feature description is relatively streamlined and computationally intensive, but the local information such as video or region changes appearing in the video is ignored. The second method is based on video frame comparison, which is to compare one by one in key frames. At present, most copy detection scenarios are more suitable for using the second method.

Time alignment is an operation performed in video partial copy detection to determine which of the two videos are copied to each other. For any two videos, there are one-to-one, one-to-many, many-to-many, and cross-corresponding alignments. The above four cases are briefly described in Fig.2. In order to solve the above-mentioned various alignment problems, previous workers have proposed a variety of algorithms, such as sliding window based algorithms [5], tree structure based algorithms [6] and graph model based algorithms [7]. The sliding window algorithm considers the time information first between two of the video frames, and then is the similarity information of them. The obvious disadvantage of this method is that it is not stable enough due to the difference in video frame rate and threshold. The latter two algorithms consider the similarity information between video frames before the time information. Although their accuracy is better than the sliding window algorithm, the disadvantage is that more computation is generated.

![Fig.2 Four alignments mode of video partial copy clips](image)

Fig.2 Four alignments mode of video partial copy clips

In summary, most of the previous work failed to unify feature matching and time alignment into a module that needs to be considered comprehensively. In summary, most of the previous work failed to unify feature matching and time alignment into a module that needs to be considered comprehensively, but split them into two separate steps, which led to a reduction in the overall performance of the model. In order to overcome this shortcoming, our primary goal is to find a video frames feature representation method that can aggregate the similarity information and timing information, and then select the appropriate model to learn this feature to ensure the final copy detection performance. In order to verify the effectiveness of our proposed method, we conducted a comprehensive experiment on the most authoritative video copy detection data set VCDB [10]. The results show that our method is better than state-of-the-art partial copy detection algorithm.
3. Our approach
In this section, we will introduce the framework of copy detection in videos based on STRNN (spatio-temporal relationship neural network) in detail for videos. The outline of the framework is shown in Fig.1, which is mainly divided into two steps: feature extraction and encoding of video frames, match alignment of video frame features. Match alignment also includes a mapping process that detects the result to the copy time period. In this paper, we are more concerned with the steps in the red box in Fig.3. Feature extraction and coding steps will briefly introduce the feature extraction techniques and the dimension reduction techniques we use.

Fig.3 The framework of partial copy detection in videos based on STRNN

3.1 Feature extraction
The removal of redundant frames is required before feature extraction, because the frame rate of most videos is 20-30 frames/s. If all video frames are used for feature extraction, the efficiency will be affected. To balance accuracy and efficiency, we evenly sample video frames at a fixed time interval (1 frame/0.5s) for each video.

The purpose of feature extraction of video frames is to quantize the rich content information in the image into feature vectors that can calculate similarity. Choosing the right extraction algorithm is conducive to improve the accuracy of retrieval and matching. This paper uses the depth separable convolution Xception structure proposed in the literature [8] to extract single frame features, and Xception is an improvement of Google's Inception v3[9]. In general, convolution on a set of feature maps requires a three-dimensional convolution kernel. That is, the convolution kernel needs to learn spatial correlation and channel correlation at the same time. Explicitly separating these two correlations is the main idea of the Inception module. Fig.4 shows the basic structure of the Inception module. However, the channel correlation and spatial correlation of the Inception module are still not completely separated, that is, the 3*3 or 5*5 convolution kernel is still multi-channel input. Xception's infrastructure is designed as an extremely dense Inception module as shown below, allowing the front-level input to learn the inter-channel correlation by M (channel number) 1*1 convolution kernels. Increase the number of convolution branches by 3*3 based on the Inception structure, making it equal to the number of output channels N convolved by 1*1 to learn the spatial correlation of each channel separately, and achieve a complete separation of channel correlation and spatial correlation, refining the learning objectives of each convolution kernel.
During the experiment, we extracted the video frame features using the Global Average Pooling of the Xception model. Since the dimension of 2048-D is too high, we add a fully connected layer containing M neurons between this layer and the output layer and call it a hash layer. The activation function of this layer is set to tanh [10], and the binary mapping process is added to the extracted hash layer output. The formula is as follows:

$$s_i^m = \begin{cases} 0, & f_i^m < 0, \ \forall m \in M, i \in 1,2,...,n \\ 1, & f_i^m > 0, \ \forall m \in M, i \in 1,2,...,n \end{cases}$$ (1)

Where f is the output feature of the ith M-d sample frame after passing through the hash layer. n is the total number of video frames sampled at a fixed frame rate. S is the mapped hash vector.

During the training phase, we freeze some parameters of the Xception model pre-trained by the ImageNet data-set, and only fine-tune the hash layer parameters. The test phase includes two working states, offline and online. When offline, the video library features are extracted and stored in the database. When the online query is performed, only the query video is extracted, and whether the database is selected according to actual needs. In order to reduce storage pressure, we only store the binary features of video frames in the database.

3.2 Index and Hamming Embedding

We use the inverted file structure to index the quantized hash feature vectors for efficient suspected copy frame filtering. In addition, our framework introduces Hamming embedding to reduce the computational complexity of the framework. The key idea of Hamming embedding is to approximate the similarity between feature vectors using Hamming distance. Since the feature extraction model we designed can directly extract the binary hash feature, we calculated whether the Hamming distance between each frame and the video library feature was lower than the threshold in the query phase. In order to limit the Hamming distance to a controllable range, we use the following formula to measure the scores of two feature similarities:

$$\text{Score} = e^{-\text{dis}^2}$$ (2)

Where dis is the Hamming distance between the two features. Finally, the suspected copy video is roughly filtered according to the number of frames below the threshold of each video in the video library.

3.3 Feature matching and time alignment

In order to obtain a prior Information of temporal alignment, the traditional partial copy detection framework usually use the algorithm step of “match first and then align”. This is the idea abandoned in this paper. Our starting point is to hope that the time series information and the matching information affect each other, so that the comprehensive information is transformed into the representation of the video partial copy relationship. The conversion formula is as follows:

$$F = 255 \times (R \times Q^T)$$ (3)

Given the query video Q and the reference video R containing q frames and r frames, respectively, we multiply their feature arrays q*M and M*r then divide by the modulus length to calculate the
cosine correlation matrix $F$. The elements in the matrix are the cosine values of the unit vectors computed by $Q$ and $R$, so the magnitude of each value represents the correlation of the corresponding frame, the larger the value, the more similar (ranging from 0 to 1). Then, we make a graphical mapping of the cosine correlation matrix to obtain a spatio-temporal correlation feature map. Fig. 5 is a few examples of the spatio-temporal correlation feature maps found in our experiments. In order to obtain a model that can recognize this pattern, we manually labelled about 6,000 spatial-temporal correlation maps between 528 pairs of partial copy detection videos according to the actual situation, and trained the object detection network model based on RefineDet to learn this law. In the online retrieval phase, first, the suspected copy video set obtained by using the index is paired with the query video to generate a spatio-temporal correlation feature map. Then, we use the trained model to detect partial copy regions on these related feature maps. Finally, the corresponding time period is predicted according to the coordinates of the detection frame.

Fig. 5 Spatio-temporal correlation feature map and partial copy regions detected by the model

4. Experiments and analysis

4.1 VCDB data-set

Fig. 6 Some examples of copy frames in VCDB data-set
In this experiment, we used the VCDB data-set to comprehensively evaluate our proposed method. VCDB is the most recognized copy detection data-set, which includes the core data-set and background data-set collected by YouTube and MetaCafe. The core data set contains carefully selected 528 videos with 9236 pairs of partial copies and manual annotations, and the background data-set has more than 100,000 interference videos. Fig.6 is an example of copying video frames in the VCDB data set. This paper mainly evaluates the performance of the algorithm in the core data set. Like the benchmark method in [8], each video is used as the query video, and 9236 pairs of copied segments are used as the ground truth object. Performance is measured by standard recall and precision, which is a widely used indicator to reflect the capability of copy detection system. If the detected copy segment and the ground truth segment contain overlapping time windows, then the detected pair of copy segment is considered correct. We do not set a threshold of the minimum overlap area (e.g. 0.5 or 0.75) as usual object detection tasks, because in practical application (e.g. copyright protection), it is enough to use a single frame hit ground truth bonding box. In this paper, the accuracy and recall rate are more accurately defined as follows:

\[
\text{precision} = \frac{|\text{correctly retrieved segments}|}{|\text{all retrieved segments}|} \quad (4)
\]

\[
\text{recall} = \frac{|\text{correctly retrieved segments}|}{|\text{ground-truth copy segments}|} \quad (5)
\]

4.2 Experimental results and analysis

We give the results of this method and compare some representative methods. Fig.7 shows the benchmark system proposed in literature [2], the standard CNN with VGG-16/Xception and SCNN [3], and the comparison of the effects of our algorithm. It can be seen that our method has achieved better results.

The blue curve represents the benchmark method proposed in literature [2]. This method uses the common SIFT descriptor and the word bag model for video frame feature extraction. The temporal alignment algorithm selects the common temporal network model. This system has been widely used in previous work and has shown good performance.

The yellow and green curves are the results of the SCNN model and the standard VGG-16 model implemented in [3], respectively, which have some improvement over the benchmark algorithm. VGG-16 achieves good results because of its high number of layers and strong descriptive ability, while SCNN performs generally. This may be because it has no pre-trained parameters as a prior support for training compared with the classical CNN structure. These two methods used temporal network as time alignment algorithm.

Red and purple curves are the algorithms used in our experiment, and their feature extraction models are Xception. Compared with the red curve, the purple curve replaces the traditional feature matching and time alignment steps with the space-temporal relationship network proposed in this paper.
5. Conclusion
This paper focuses on the information loss caused by the traditional method of separating the feature matching module from the time alignment module in the video partial copy detection task, and we propose a new method of end-to-end spatio-temporal feature fusion representation to solve this problem. The experimental results on the VCDB show that compared with the recent CNN-based partial copy detection system, the method achieves a significant improvement in accuracy and recall rate, thus verifying the effectiveness of the system and making up for the lack of matching information in existing methods. It has practical and promotional value in the field of video copy detection.

Acknowledgments
This work was mainly funded by the development of digital copyright protection technology scheme based on digital image tamper identification technology, Key Project of Beijing, China (No.Z18110000618006), and also supported by data-driven security production supervision image construction technology and security management situation research (No.Z181100009018003).

References
[1] Jiang Y G, Jiang Y, Wang J. VCDB: A large-scale database for partial copy detection in videos. [J]. 2014.357-371
[2] Jiang Y G, Wang J. Partial Copy Detection in Videos: A Benchmark and an Evaluation of Popular Methods [J]. IEEE Transactions on Big Data, 2016, 2(1):32-42
[3] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. 2015.
[4] Jing Z, Zhu W, Bing L, et al. Image Copy Detection Based on Convolutional Neural Networks [M]// Chinese Conference on Pattern Recognition. 2016.
[5] Roopalakshmi R, Reddy G R M. A novel spatio-temporal registration framework for video copy localization based on multimodal features [J]. Signal Processing, 2013, 93(8):2339-2351.
[6] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in Proc. 19th ACM Int. Conf. Multimedia, Nov. 2011, pp. 423-432.
[7] Song J, Yang Y, Huang Z, et al. Effective Multiple Feature Hashing for Large-Scale Near-Duplicate Video Retrieval [J]. IEEE Transactions on Multimedia, 2013, 15(8):1997-2008.
[8] Esmaeili M, Fatourechi M, Ward R K. A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting [J]. IEEE Transactions on Information Forensics & Security,
2011, 6(1):213-226.

[9] Chiu C Y, Chen C S, Chien L F. A framework for handling spatiotemporal variations in video copy detection [J]. IEEE Transactions on Circuits & Systems for Video Technology, 2008, 18(3):412-417.

[10] Gordo A, Almazán J, Revaud J, et al. Deep Image Retrieval: Learning Global Representations for Image Search [M]// Computer Vision – ECCV 2016. 2016.