EasyChair Preprint

№ 5334

GANN: a Graph Alignment Neural Network for Video Partial Copy Detection

Xiyue Liu, Xin Feng and Pan Pan

April 18, 2021

GANN: A Graph alignment neural network for video partial copy detection

Xiyue Liu
College of Computer Science and Engineering
Chongqing University of Technology
Chongqing, China
e-mail: liuxiyue@2018.cqut.edu.cn

Xin Feng
College of Computer Science and Engineering
Chongqing University of Technology
Chongqing, China
e-mail: xfeng@cqut.edu.cn

Pan Pan
College of Computer Science and Engineering
Chongqing University of Technology
Chongqing, China
e-mail: panpan9314@2020.cqut.edu.cn

***Abstract*—With the advent of we-media era, massive videos have been uploaded by users to the Internet. Such a large volume of video data brings us various information. It, however, contains some fake information created by partial copy videos, which constitute infringement act and are harmful to original authors and common users. In this paper, we propose a graph alignment neural network (GANN) for partial copy videos detection. Through building a graph neural network based on video frame-level feature extracted by a pretrained convolutional neural network and their relationship, GANN automatically integrates the global representation of a video, and learns the intra-similarity between original and copied videos, and the inter-discriminative from other videos by the self-attention and cross-attention mechanism in the graph neural network. We perform experiments on the challenging dataset VCDB, which includes a variety of complex transformations in the real scene. Results demonstrate that our GANN has better detection performance than baseline methods, where the precision of GANN is close to 80%, and the recall rate reaches 65%.**

Keywords—video copy detection, graph neural network, attention embedding, global information

I. INTRODUCTION

With the popularity of social media, a large number of video creators share their video on the Internet. However, a large number of video segments in these uploaded videos have been taken from another full video and altered in a variety of ways. This phenomenon leads to serious copyright problems. Therefore, high precision and robust copyright detection algorithms have become an urgent need of the video big data era.

The current mainstream technology focuses on video similarity detection, near-duplicate video detection. In the past, the algorithm used to describe the whole video with a single feature to determine whether two videos are similar or duplicate, which largely solved the problem of a large number of duplicate videos in the network. However, with the prosperity of short video and we media, a large number of creators have created more and more video works including partial copies through editing and processing the original video. If the proportion of copied video segments in the whole video is negligible, it is quite difficult to detect the copyright video through the above methods. Unfortunately, this is a common phenomenon in social media, where interesting parts of a video are cut, edited, and then randomly pasted onto another video with similar or arbitrary themes. Therefore, video copyright detection technology for time segment location is still a problem to be solved.

As a result, complex content changes or transformation between copied segments and original videos are generated, such as scale and lighting changes, picture-in-picture, filters, stickers, picture stretching and video transcription, etc. As shown in Figure 1.

Figure 1. Different video copy results from the same video clip, including remake (b), illumination Change (c), low quality image (d), Add stickers(e), Image size change(f).

The most effective method of early near duplicate detection is the visual bag-of-words model is recognized as an effective method which is uses clustering and statistical methods to describe the feature of the whole video. However, when the task becomes partial copy video detection, most of the various methods of feature extraction from image matching task is used to get features. The most common method is to use local features (such as SIFT) to match and find similar video frame pairs, and then perform time alignment. It has achieved good results on some simulated data, but in the complex changes of real data, this framework is also difficult to achieve high accuracy, so it still cannot be used to deal with real situations.

Recently, deep learning has been applied in a wide range of fields to near-duplicate video detection, video similarity detection. Although it has made gratifying progress in the feature extraction part, if the selected key frame has undergone motion blur or multiple combined changes, it will not be able to accurately locate copy segments.

In fact, in [1], the robustness of the approaches relies heavily on the stability of visual content. For some cases where all frames of the video are very similar, it is very difficult to locate the time period of the partial copy video. Therefore, the method of identifying the diagonal pattern in the matching matrix by the target detection algorithm in the STRNN [2] method is difficult to achieve satisfactory results.

Through the analysis of the task and summarize the past methods, we propose a new framework GANN. Inspired by the excellent work of graph matching, we find that in the local

matching problem, one-to-one matching often leads to many conflicts, which leads to some wrong matching. After the global information is fused by graph neural network, the discrimination of single feature can be greatly improved. We use the video global information after selective fusion of graph neural network to do matching. This method further approximates the human's logical judgment (locating the copied video segment through context contrast and multi frame comprehensive information)

To the best of our knowledge, we are the first to apply graph neural networks to solve the video partial copy problem. In this paper, we cast the research on video-pair retrieval to a graph matching problem that refines the expression of embedded information in a single frame by fusing global information. This framework is divided into three parts: feature extraction, feature matching, and time alignment. First, we extract key frames from the video, and treat each frame in the video as a node in the graph, and use the CNN to extract features from these nodes. After that, we learn the local matching features between the two sets of features through the graph matching. Finally, the result obtained by the graph alignment network is subjected to time constraints to obtain the final matching result.

Although cascading multi-layer features is a common method in near duplicate video detection (because it can capture features from multiple scales to improve the matching accuracy). However, the feature dimension generated by the above method is too high, and the amount of computation will increase greatly when processing long video. As a result, the graph neural network cannot be trained. Hence in our method, we only use the vector generated by the last layer of convolution network to describe the features of a single frame. The experiment shows that this method is feasible.

Different from the traditional matching strategy that requires a lot of computing resources to calculate multiple sets of candidate matches, and then perform post-processing to determine the final result (time alignment network), this paper directly uses the graph neural network to capture global information and then learns the most reliable set of matching results. In other words, because we found that the matching result does not need to make the best choice in the huge matching candidate pool, so we discarded the previous method of aligning the network to find the longest path in the topology map, but directly in the matching adding time constraints to the results.

The reminder of this paper is organized as follows. In Section 2, the related works from three different feature levels are introduced. In Section 3, the proposed GANN is described in detail. Experiment setup and result analysis are presented in Section 4, and conclusions are finally drawn in Section 5.

II. RELATED WORKS

The research of partial copy video detection tasks is particularly important in the explosive growth of video data. In fact, similar research has been developed for decades, including video similarity detection, near-duplicate video retrieval, and video retrieval. Generally, these proposed methods can be mainly divided into three genres according to the granularity: video-level, frame-level, hybrid-level.

A. Video-level

This coarse-grained method mainly solves the problem of large-scale near-duplicate video retrieval. The features of all frames are usually merged to embed the video, such as aggregate feature vector [3, 4, 5] or hash code [6, 7, 8], and video matching is based on the calculation of pairwise similarity between corresponding video.

B. Frame-level

However, video-level similarity detection ignores the spatial and temporal structure of visual similarity, because feature aggregation will be affected by low-quality feature extraction and irrelevant content. Some other methods also try to solve these problems by using Dynamic Programming [9, 10,23] which exploits a detect-and-refine strategy, can effectively measure the similarity between videos and localize the similar parts, In Temporal Networks [11, 12] , the time network finds the top-k similar pairing results for each frame to construct the time graph, which transforms the matching problem into the optimal transmission problem, and finds the optimal path from the source node to the sink node , However, these modules are often in the post-processing part of the network, can not produce constraints from the initial feature extraction and matching process. Other methods do histogram statistics by accumulating the results of multiple frames, and search the peak value in a fixed range of time stamp, and do matching around the peak value. However, the features of consecutive frames are often similar, which leads to the explosion of matching, thus causing the deviation of histogram statistical results. In order to alleviate this problem, a reweighting scheme is proposed-Temporal Hough Voting [13, 14]. Another line of research considers spatio-temporal video representation and matching based on the Fourier domain [15, 16, 17] or Recurrent Neural Networks (RNN) [18, 19].

C. hybrid-level

This method combines the advantages of the video-level and frame-level methods. It uses the video-level method to do similarity detection on a large scale, and then followed by a more accurate frame-level similarity measurement to determine the final matching result, such as [9]. Considering that there is a lot of redundancy between consecutive frames, it often leads to the increase of computation and redundancy. In order to speed up the calculation of similarity, some frames with similar visual features are clustered in advance by clustering algorithm and assigned unique symbols. Each video is serialized into a symbol set, and then accurately located in the set.

III. GANN FOR VIDEO COPY DETECTION

A. The Graph Alignment Nerual Network

In this paper, we proposed to use the method of graph matching to accurately match the partial copy video segments. Graph matching is a basic and important issue in the field of computer vision and pattern recognition. It has a wide range of aspects in many application. From an optimization perspective, the graph matching problem is a discrete combinatorial optimization problem, which makes the problem itself NP (non-deterministic polynomial)-hard. In our graph alignment method, by cascading convolutional neural network and graph neural network, the graph matching problem is turned into an optimal transmission problem, and the optimal solution of this problem can be obtained by Sinkhorn algorithm.

Unlike video similarity detection or near-duplicate video detection can integrate global information to make a comprehensive judgment on video. The partial copy video

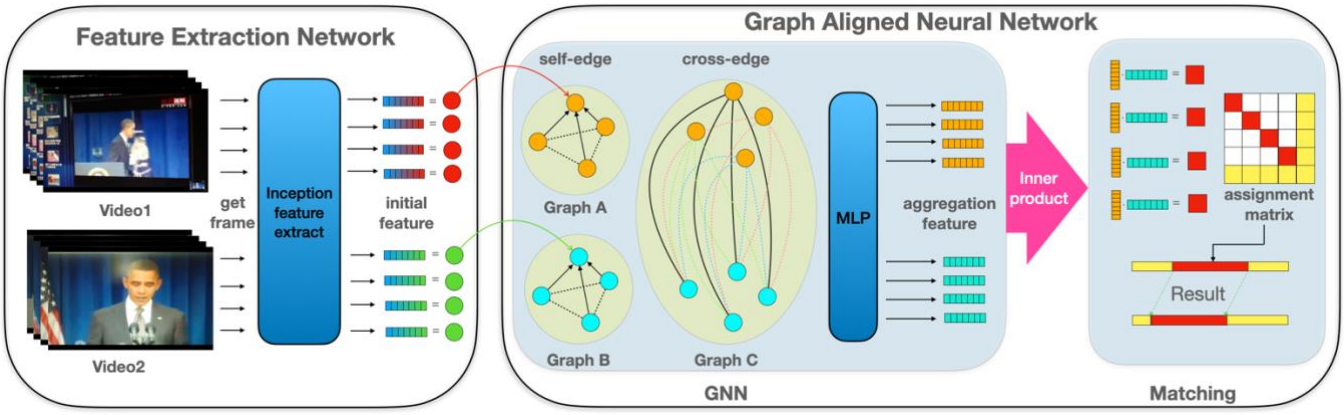

Figure 2. The framework of proposed graph alignment neural network, GANN.

detection method needs to find out which part of the video is copied and combined transformed, we observe that in the traditional partial copy video detection method, most of them consider the pairwise matching between frames, and then use post-processing to constrain the entire match to predict. Although those method can find some local matches between videos through the expression of a single frame, it loses the global information in the video sequence. In this paper, we fully-connected all the frames of the current video and all the frames of the video to be matched through the method of graph neural network (This connection method can be used to suppress the error caused by a single feature expression, because the features of the frame incorporate global information, which can be more accurately described than the single frame) and the method of transformer is embedded in the graph neural network, so that the network will automatically learn global information to generate higher-level semantic expression as show in Figure 2.

B. Feature extract

Since the information between adjacent frames is usually redundant for extracting video features, for example, when all the features of a video with a frame rate of 25 are completely extracted, a large number of feature maps with similar features will be generated every second. In theory, this can indeed improve the accuracy of matching, but the computational cost is also exponentially increased, so some frames can be discarded in the original video data. By observing the data set, we find that the labeling accuracy of the data set is at the second level, so our paper also uses the method of extracting one frame per second.

In traditional methods, SIFT features are usually used as single-frame features for matching. This method relies heavily on the extraction quality of each key point in a single image: such as video transcription, lighting changes, occlusion, etc. Since it is the feature extracted from the key points, SIFT as a single frame feature expression relies heavily on the detection result of the key points, which is likely to cause insufficient feature expression. With the convolutional neural network as a feature extractor, good results have been achieved in various visual tasks.

In this paper, we use the feature vector of the last layer of fully -connected layer of the inceptionv4 network like [8]. The feature vector is used as the description of the frame to obtain a 1536-dimensional feature vector. Using the above-mentioned feature extractor can better capture the features of the entire image to deal with the problem that it is arduous to extract key points from high-quality images and to extract enough number of key points from low-quality images.

C. The Video Gragh

For a pair of videos, we need to build two kinds of graphs, which are self-connected graph and cross-connected graph. Two self-connected graphs A and B are used to aggregate the global features of the video, so that the context information can be considered in the learning process. A cross-connected graph is used to aggregate the features of pair video, which improves the accuracy of frame to frame matching. Compare two videos A and B, after extracting frames, we get two different sets of frames, which are represented as $S_A$and S_B. Then the trained convolutional neural network is used to represent the features of each frame to get the node description of the original graph, We use n_A^i and n_B^j to represent i-th frame of video A and j-th of video B , $i \in [1, |S_A|]$、 $j \in [1, |S_B|]$.

$$G_A^{self} = \left(\sum_{i=1}^{|S_A|} n_A^i , \sum_{i=1}^{|S_A|} e_{ii} \right) \tag{1}$$

$$G_B^{self} = \left(\sum_{j=1}^{|S_B|} n_B^j , \sum_{j=1}^{|S_B|} e_{jj} \right) \tag{2}$$

$$G_C^{cross} = \left(\sum_{i=1}^{|S_A|} n_A^i , \sum_{j=1}^{|S_B|} e_{ij} \right) \tag{3}$$

where e is the edge connecting two nodes and "$|S_A|$,$|S_B|$" is the number of nodes in two different sets. Equations (1) and (2) represent the video self-connected graph respectively, and Equation (3) represents the video cross-connected graph. By combining these graphs, we can realize the feature aggregation of graph neural network.

D. Graph alignment:

When human determine whether there are copy video segment in two videos, they often refer to the multiple frames before and after the query video segment to determine which part is the target segment, and also need to look back-and-forth at both video, combining the content of the context to finally determine which video is the final goal we are looking for. After careful analysis of the matching process, we found that

the graph neural network is a perfect tool to solve the above problems. Use the previously obtained features as the input of GNN, these feature vectors are used as nodes in the graph to construct a graph describing the video. All nodes in the graph and across the graph are fully connected, and these edges are divided into self-connection and cross-connection. The self-connection edge is used to aggregate the feature of the frame in the query video (original video without processing), and the cross-connection edge aggregate the feature of the frames in the answer video（transformed partial copy video）. These two different connection methods imitate the process of referring to the preceding and following frames in the video and comparing all the frames of the candidate video when humans judge the matching time period. Attention embed graph neural network, however, only relying on graph neural networks is not enough, and not all the extracted key frames have the same effect on the description of a single frame, which implies an iterative process that can focus its attention on a specific location. Therefore, an attention mechanism needs to be added to deal with this problem. As shown in the figure2, when the basic graph neural network aggregates the information of all nodes in the graph to obtain the initial features, it is also necessary to use the attention mechanism to filter out which node has a decisive influence on the final matching result to make the network more robust. Here, we refer to the method in [21] to add self-attention and cross-attention to the graph neural network. An attention mechanism performs the aggregation and computes the message $m_{\varepsilon \to i}$, self-edges are based on self-attention [20] and cross-edges are based on cross-attention. Akin to database retrieval, i and j represent the nodes in the two graphs respectively, the query q_i, retrieves the values v_j of some elements based on their attributes keys k_j. It computes the message as weighted average of the values:

$$m_{\varepsilon \to i} = \sum_{j:(i,j)\in\varepsilon} \alpha_{ij} v_i \tag{4}$$

where the attention weight α_{ij} is the softmax over the key-query similarities. This value is used as a weight of the current feature to control the importance of the feature. In this way, the attention mechanism is embedded in the graph neural network:

$$\alpha_{ij} = Softmax_j(q_i^{\mathrm{T}} k_j) \tag{5}$$

Partial assignment: This part includes two step: produces a partial assignment matrix and time alignment. Generally speaking, graph matching can be defined as the assignment matrix P obtained by computing a score matrix $S \in \mathbb{R}^{M\times N}$ for all possible matches and maximizing the total score $\sum_{i,j} S_{i,j} P_{i,j}$. The first step is to calculate an inner product for all potential matching nodes to form a score matrix:

$$S_{i,j} = < f_i^A, f_j^B >, \forall (i,j) \in A \times B \tag{6}$$

where $<\cdot,\cdot>$ is the inner product. As shown in the figure 2, we need to find the maximum value of the column and the maximum value of the row in the obtained score matrix. The red filled box represents the maximum value of the row, which means that each frame in the A video has found the most similar frame in the B video. And the green box indicates that each frame in B video is found in the most similar frame in A video. When some boxes are filled with two colors, it means that the current two frames are the most similar frames in the two candidate videos. Use this method to continue to find all the matching pairs that are the most similar frames to each other. Then we get a set of frame-pairs.

The final step is to make threshold and time constraints for these matching pairs. The threshold constraint requires that even if the two frames are the most similar to each other, the matching score must be greater than a certain fixed value, so as to ensure that the matching pair of the found frame is not an accidental coincidence. After satisfying the threshold constraint, we take the longest ascending sequence from the matched frame sequence to ensure that the two candidate video frames are continuously matched in time.

As shown in the figure 3, the intuitive result of the time constraint is: when the first matching frame is found, the next frame of the sequence can only continue to be searched in the right or downward box, so the matching sequence appears is an approximate diagonal shape.

IV. Experiments and Result Analysis

A. Dataset

In this paper, we use the VCDB dataset to evaluate the proposed method. VCDB is the most recognized copy detection dataset, it is different from the previous simulation dataset through fixed combination of transformation to generate data to evaluate the performance of the algorithm. But collecting data from real scenes, not only greatly increases the difficulty of algorithm identification, but also makes the subsequent related algorithms more close to the real scene. VCDB contains video data with a very large time span, ranging from the shortest 2 seconds video to the longest 44 minutes. The video types include business, film, music, public speaking, sports, etc. The statistical discovery of transformation types for multiple duplicated videos includes around 36% of them that contain “insertion of patterns”. An the 18% are from “camcording”, 27% have scale changes, and 2% contain “picture in picture” patterns. These percentages are quite different from that in the simulated datasets. Many “insertion of patterns” copies exist in the practical scenario because of the logos of different TV channels, and the “picture in picture” patterns frequently seen in the simulated copies do not seem to be popular in real cases.

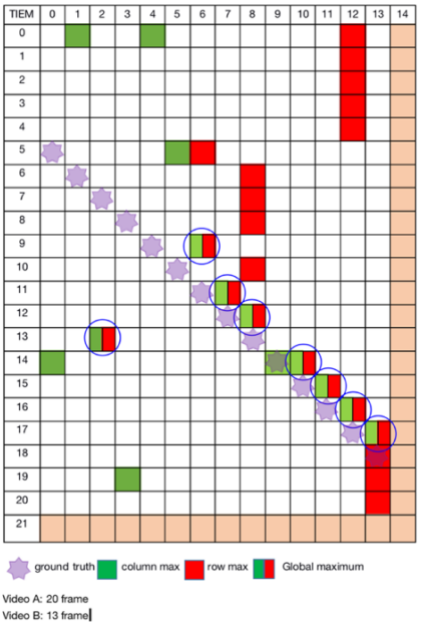

Figure 3. Assignment matrix

The core dataset contains 528 carefully selected videos, 9236 partial copies and manual annotations, and the background dataset has more than 100000 interfering videos. This paper mainly evaluates the performance of the algorithm in the core data set.

B. Result Analysis

Follow the benchmark method in [22,24,25], we use standard recall ratio and precision ratio to measure the performance of copy detection system. If the detected copy segment and ground truth segment contain overlapping time windows, the detected copy segment pair is considered to be correct. We don't set the threshold of minimum overlap area (such as 0.5 or 0.75) as usual in target detection tasks, because in practical applications (such as copyright protection), it is sufficient to use a single frame hitting the ground truth binding box. In this paper, the definitions of accuracy and recall are as follows:

$$\text{precision} = \frac{|correctly\ retrieved\ segments|}{|all\ retrieved\ segments|} \quad (7)$$

$$\text{recall} = \frac{|correctly\ retrieved\ segments|}{|ground-truth\ copy\ segments|} \quad (8)$$

And in the process of our training, we found some data labeling errors, and the length of the two video clips matched with each other has a multiple difference. In this paper, in order to verify the effectiveness of the graph neural network, we eliminate this part of data, and the remaining annotation data contains 6006 video pairs. We validated our results in these selected data. Here is a comparison of our results with Other methods：

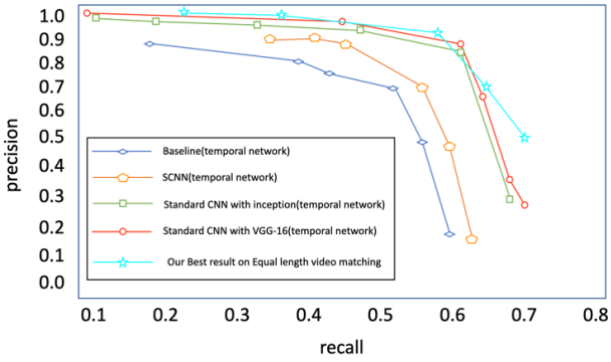

Figure 4: Comparison results between baseline methods and our method.

The result of blue line is the result obtained by using SIFT feature and time alignment network directly. The orange, green and red lines indicate that the convolution network is used to extract features, and the time aligned network is cascaded to get the matching results. It can be seen that the recall rate and accuracy rate are not enough to reach the level of practical application. Cyan lines are our experimental results on selected equal length matching data. After replacing the two-step operation of "convolution time alignment" in the traditional method with graph aligned network, we propose a new method of graph aligned network. From the results of precision and recall, we can see the effectiveness of the network.

In order to better assess the proposed method, we further show the results of some sample video pairs of video partial copy detection in terms of confusion matrix in Figure 5-7. In these result figures, the obvious diagonal lines in the graph indicating the good matching result, and in other words, the partial copied video segment is detected. From the resulted confusion matrix, we can see that our network is not affected by the length of video.

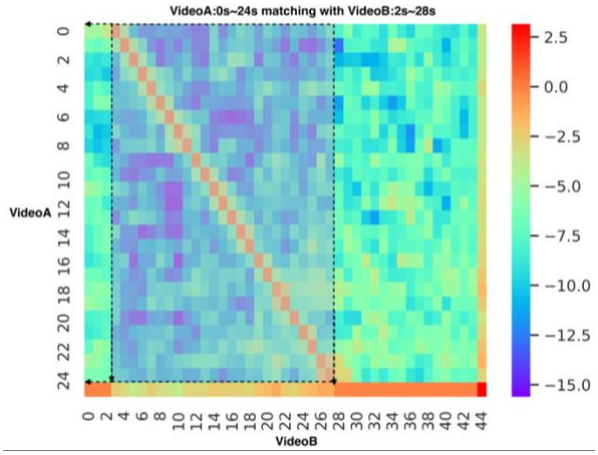

Figure 5. **Confusion matrix of video A and video B.** In the case of partially copied video between video A and video B. Taking the upper left corner of the matrix as the origin, the number of blocks from top to bottom indicates the total number of frames of video A, and the number of blocks from left to right indicates the total number of frames of video B. The value of each grid is obtained by the inner product of the features of the frame corresponding to the values of abscissa and ordinate. The larger the value is, the more similar the two frames are. The right most bar chart in the figure shows the degree of similarity. The closer the red end is, the more similar it is. Otherwise, the more dissimilar it is.

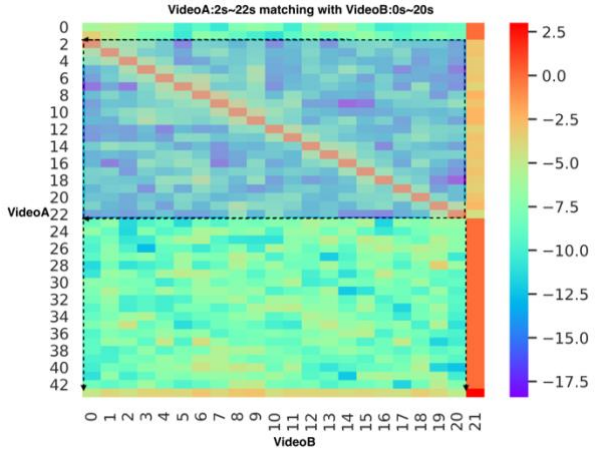

Figure 6. **Confusion matrix of video A and video B.** The difference from above is that the length of the query video (24 seconds) in the above figure is shorter than that of the answer video (44 seconds), while there are 42 seconds for the query video and 21 seconds for the answer video.

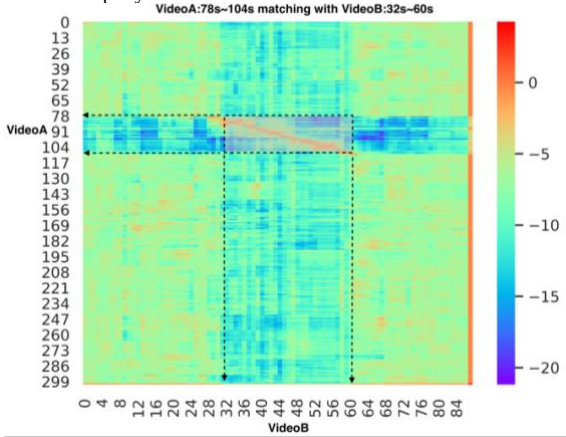

Figure 7： **Confusion matrix of video A and video B.** The video length is 299 seconds long.

If we want to calculate which segments of the two videos have repetition. We only need to find an obvious diagonal in

Figure 6: Visualization of video copy detection results of same sample videos. (a) Video “Titanic”.The first row is the original video and the second row is copied video with the watermark is removed. (b) Video “Troy”. The copied video in the second row is compressed horizontally from the original video in the first row. (c) Video “ Zidane Headbutt”. The copied video in the second row performs border clipping. Green boxes show the original video clip from query video, the red box represent the partial copied clip detected by our proposed GANN.

the matrix, and record the beginning and end coordinates of the diagonal, which corresponds to the beginning and end time of the copied segment. It can be seen from the above similarity matrix examples that the graph aligned network can still capture the copied video without the assistance of time alignment method. An example of a correct match is shown in Figure 6.

V. CONCLUSION

This paper focuses on how to integrate the global information of video to improve the discriminant ability of feature expression. Different from the limitation of single frame feature representation of traditional methods, we use the method of combining graph neural network and attention mechanism to globally capture the useful information for the current frame matching. The experimental results also show that the graph alignment network achieves satisfactory results in our selected data.

VI. REFERENCES

[1] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua, “Scalable detection of partial near-duplicate videos by visual-temporal consistency,” in *ACM MM*, 2009.

[2] Hu, Yanzhu , Z. Mu , and X. Ai . "STRNN: End-to-end deep learning framework for video partial copy detection." *Journal of Physics Conference Series* 1237(2019):022112.

[3] K. Elissa, X.Wu,A.G.Hauptmann,and C.-W.Ngo."Practicalelimination of near-duplicates from web video search". In *Proceedings of the 15th ACM international conference on Multimedia*, pages 218–227. ACM, 2007.

[4] L.Liu,W.Lai,X.-S.Hua,and S.-Q.Yang.Videohistogram:"A novel video signature for efficient web video duplicate detection". In *International Conference on Multimedia Modeling*, pages 94–103. Springer, 2007.

[5] Z.Huang,H.T.Shen,J.Shao,X.Zhou,andB. "Cui.Bounded coordinate system indexing for real-time video clip search". *ACM Transactions on Information Systems (TOIS)*, 27(3):17, 2009.

[6] J.Song,Y.Yang,Z.Huang,H.T.Shen,andR.Hong."Multiple feature hashing for real-time large scale near-duplicate video retrieval". In *Proceedings of the 19th ACM international conference on Multimedia*, pages 423–432. ACM, 2011.

[7] Y. Hao, T. Mu, R. Hong, M. Wang, N. An, and J. Y. "Goulermas. Stochastic multiview hashing for large-scale near-duplicate video retrieval. IEEE Transactions on Multimedia", 19(1):1–14, 2017

[8] J.Song,Y.Yang,Z.Huang,H.T.Shen,andJ.Luo. "Effective multiple feature hashing for large-scale near-duplicate video retrieval". *IEEE Transactions on Multimedia*, 15(8):1997– 2008, 2013.

[9] Chien-LiChou,Hua-TsungChen,and Suh-YinLee. "Pattern-based near-duplicate video retrieval and localization on web-scale videos". *IEEE Transactions on Multimedia*, 17(3):382– 395, 2015

[10] Hao Liu, Qingjie Zhao, Hao Wang, Peng Lv, and Yanming Chen. "An image-based near-duplicate video retrieval and localization using improved edit distance". *Multimedia Tools and Applications*, 76(22):24435–24456, 2017.

[11] Hung-Khoon Tan, Chong-Wah Ngo, Richard Hong, and Tat-Seng Chua. "Scalable detection of partial near-duplicate videos by visual-temporal consistency". In *Proceedings of the 17th ACM international conference on Multimedia*, pages 145–154. ACM, 2009.

[12] Yu-Gang Jiang and Jiajun Wang. Partial copy detection in videos: "A benchmark and an evaluation of popular methods". *IEEE Transactions on Big Data*, 2(1):32–42, 2016.

[13] Matthijs Douze, Herve´ Je´gou, and Cordelia Schmid. "An image-based approach to video copy detection with spatio-temporal post-filtering". *IEEE Transactions on Multimedia*, 12(4):257–266, 2010.

[14] Yu-Gang Jiang, Yudong Jiang, and Jiajun Wang. "VCDB: a large-scale database for partial copy detection in videos. In *Proceedings of the European Conference on Computer Vision*, pages 357–371. Springer, 2014".

[15] Je´roˆme Revaud, Matthijs Douze, Cordelia Schmid, and Herve´ Je´gou. "Event retrieval in large video collections with circulant temporal encoding". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2459–2466, 2013.

[16] Se´bastien Poullot, Shunsuke Tsukatani, Anh Phuong Nguyen, Herve´ Je´gou, and Shin’Ichi Satoh. "Temporal matching kernel with explicit feature maps". In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 381–390. ACM, 2015.

[17] LorenzoBaraldi,MatthijsDouze,RitaCucchiara,andHerve´ Je´gou. "LAMV: Learning to align and match videos with kernelized temporal layers". In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7804–7813, 2018.

[18] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. Video re-localization. In *Proceedings of the European Conference on Computer Vision*, pages 51–66, 2018.

[19] Yaocong Hu and Xiaobo Lu. Learning spatial-temporal features for video copy detection by the combination of cnn and rnn. *Journal of Visual Communication and Image Representation*, 55:21–29, 2018.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko- reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[21] Sarlin, Paul-Edouard, Daniel DeTone, Tomasz Malisiewicz and Andrew Rabinovich. “SuperGlue: Learning Feature Matching With Graph Neural Networks.” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 4937-4946.

[22] 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):213-22

[23] Gai K , Qiu M . Optimal resource allocation using reinforcement learning for IoT content-centric services ☆ [J]. Applied Soft Computing, 2018, 70:12-21.

[24] Chen M , Zhang Y , Qiu M , et al. SPHA: Smart Personal Health Advisor Based on Deep Analytics[J]. IEEE Communications Magazine, 2018, 56(3):164-169.

[25] Gai K , Qiu M , Zhao H , et al. Resource Management in Sustainable Cyber-Physical Systems Using Heterogeneous Cloud Computing[J]. IEEE Transactions on Sustainable Computing, 2018:1-1.