COPY DETECTION TOWARDS SEMANTIC MINING FOR VIDEO RETRIEVAL

Shikui Wei[*+†], *Yao Zhao*[*+], *Changsheng Xu*[‡], *Dong Xu*[†]

[*]Institute of Information Science, Beijing Jiaotong University, Beijing100044, China
[†]School of Computer Engineering, Nanyang Technological University, Singapore
[‡]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[+]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing100044, China. Email:shkwei@gmail.com

ABSTRACT

In large-scale video database, lots of different videos frequently share the similar content copied from the same source. Generally, those videos have certain semantic correlations, such as being of similar events and sharing the same topic. Mining these semantic correlations can greatly facilitate video search. However, as a preprocessing step, detecting and localizing the copy pair among videos, i.e. copy detection problem, plays a key role for precise semantic mining. To meet the requirements in semantic mining scenario, we propose a frame fusion based copy detection scheme. In this scheme, the copy detection problem is converted to HMM decoding problem with three relaxed constraints, where Viterbi algorithm is employed to automatically detect the copy pair. The experimental results show that the proposed approach achieves high localization accuracy even when the copied clip undergoes some complex transformations, while achieving comparable performance compared with state-of-the-art copy detection methods.

Index Terms— Semantic Mining, Copy Detection, Viterbi-Like Algorithm, Frame Fusion, HMM

1. INTRODUCTION

Currently, most of video retrieval models implement search procedure by implicitly or explicitly measuring the similarity between the query and database shots in some low-level feature spaces [1]. However, such similarity is not always consistent with human perception due to the limitation of image/video understanding techniques, i.e., semantic gap problem. Therefore, it is necessary to involve more semantic clues for bridging the gap. As an alternative method, mining the semantic correlation among video can greatly enhance the capability of semantic understanding. By analyzing the video archives in large-scale database, some researchers found that certain video content copied from the same source is frequently occurred in lots of different videos due to its popularity or importance [2] , such as popular network video and important news shots. Obviously, the videos containing the same copy have some semantic correlations like sharing similar topic, which can be mined to enhance the semantic understanding of video content. However, before doing this, we must determine which videos share the same copy and where the copy occurs in the videos, i.e., copy detection problem. Our main effect focuses on this key preprocessing step of semantic mining.

Formally, content-based copy detection(CBCD) originally refers to judging whether a query video contains the content originated from copyright protected reference video via some feature extraction and matching techniques [3]. In our context, our purpose is to find the video archives which share the same content from the same source. Hence, each video can serve as both the query video and the reference video. In the previous literature[4, 5, 6, 7], the main effort focuses on the copy detection of a short query video. A common characteristic of those methods is that the query video should be a temporally bounded video archive and all query frames(or key frames) must be parsed beforehand. In other words, at least one of two videos to be detected must be a short video. In semantic mining scenario, however, the videos in database are generally long videos, and it is unpractical to parse them beforehand due to the limitation of computational cost and memory usage. Therefore, existing copy detection methods are not suitable for this task, and we need to develop some new methods for copy detection in long videos.

To this end, we consider a frame fusion based copy detection approach, which combines similar frame search and frame fusion under a temporal consistency assumption. The key idea of this method is to convert copy detection problem to HMM decoding problem in which Viterbi algorithm together with three relaxed constraints can be employed to meet the requirements of copy detection in long videos. In particular, any one video in database is considered as a query video, and the others are the reference videos. To fit the HMM model, the frame(or key frame) sequence of query video is treated as the observation sequence in HMM model, and all frames in database videos are considered as the states. Therefore, the copy detection is equivalent to finding a partial state

sequence(or partial best path) that generates the corresponding observation sequence with high probability. However, if database is large, the set of states (i.e, frames of database video) is extremely huge, which leads to intractable process due to the limitation of computing power and memory. To address this problem, only the frames who are most similar to query frames are retained to form the state set. In addition, the state set is dynamically updated when certain query frames move in or move out, which further reduces the computational cost and memory usage.

2. OVERALL FRAMEWORK

The architecture of proposed method is illustrated in Figure 1, which includes keyframe extraction, feature extraction, similarity matching and frame fusion. As stated above, each video archive in the database is in turn treated as query video and the others are considered reference videos. Since our main work focuses on the frame fusion step, we just use existing methods for the other components. In our scheme, we sample three frames per second, and each frame is described by a Bag-of-Words scheme used in [8]. For similariy measurement, we adopt the Okapi BM25 scoring function proposed in[9]. After the system searches the reference database and

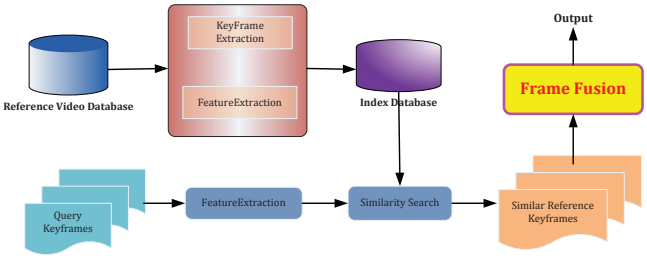

Fig. 1. Framework of proposed video copy detection system

returns a list of similar reference frames for each query frame, the copy pairs can be determined by fusing these returned reference frames. We will detail the fusion process in next section.

3. HMM-BASED FRAME FUSION ALGORITHM

Since a copy is usually a small part of the video, we define subsequence for both the query frame sequence and its corresponding list sequence of returned reference frames, which are denoted as follows:

$$Q_{sub}(i,j) = \{(q_i, q_{i+1}, \cdots, q_j) | 1 \le i \le j \le T\} \quad (1)$$

$$L_{sub}(i,j) = \{(L_i, L_{i+1}, \cdots, L_j) | 1 \le i \le j \le T\} \quad (2)$$

where $Q_{sub}(i,j)$ is a temporally successive frame sequence from time instant i to j in query video with length T, $L_{sub}(i,j)$ is the corresponding list subsequence in which L_i is the list of similar reference frames returned for q_i by similarity search.

The purpose of frame fusion is to reconstruct reference frame sequences from $L_{sub}(i,j)$ according to the temporal consistency constraint. If we can construct a sequence $h = \{(h_i, \cdots, h_m, \cdots, h_j) | h_m \in L_m\}$ whose frames are temporally successive in a single video archive, we say that $Q_{sub}(i,j)$ and h are a copy pair. The starting and ending positions of the copy in query and reference videos are determined by $\{p_i, p_j\}$ and $\{h_i, h_j\}$, respectively.

In our scheme, the frame fusion problem is converted to HMM decoding problem in which the Viterbi algorithm can be employed for fast computing. In particular, the query subsequence can be directly treated as the emission sequence E_{seq}, and the reference frames in $L_{sub}(i,j)$ constitute the state set S after $Unique$ operation. Here, the $Unique$ symbol denotes the duplicate-removal operation on $L_{sub}(i,j)$. Then, the fusion problem is how we can find a state sequence h^* in sequence space $H_{sub}(i,j)$ with all possible state sequences, which is most likely to have generated the emission sequence E_{seq}. The conversion model can be formulated as follows:

$$E_{seq} = (q_i, \cdots, q_m, \cdots, q_j) \Leftarrow (e_i, \cdots, e_m, \cdots, e_j) \quad (3)$$

$$\begin{aligned} S &= \text{Unique}\{L_i, \cdots, L_m, \cdots, L_j\} \\ &\Leftarrow \{s_1, \cdots, s_m, \cdots, s_Y\} \end{aligned} \quad (4)$$

$$\begin{aligned} H_{sub}(i,j) = \{(h_i, \cdots, h_m, \cdots, h_j) | 1 \le i \le T, \\ i \le m \le j \le T, h_m \in S\} \end{aligned} \quad (5)$$

$$\begin{aligned} h^* =& \arg\max_{h \in H_{sub}(i,j)} P(E_{seq}, h) \\ =& \arg\max_{h \in H_{sub}(i,j)} P(h) P(E_{seq}|h) \\ =& \arg\max_{h \in H_{sub}(i,j)} \{P((h_i, \cdots, h_m, \cdots, h_j)) * \\ & P((e_i, \cdots, e_m, \cdots, e_j) | (h_i, \cdots, h_m, \cdots, h_j))\} \end{aligned} \quad (6)$$

In our context, $P((h_i, ..., h_m, ..., h_j))$ reflects the transition relationship among returned reference frames, whereas $P((e_i, ..., e_m, ..., e_j)|(h_i, ..., h_m, ..., h_j))$ implies the similarity measurement between the query sequence $(q_i, ..., q_m, ..., q_j)$ and a reference frame sequence $(h_i, ..., h_m, ..., h_j)$. We employ the first-order Markov chain for modeling the transition relationship, which assumes that the present state is only dependent on the previous state. That is, $P(h_m|h_{m-1}, ..., h_i) = P(h_m|h_{m-1}), m = i+1, ..., j$. For similarity measurement, since we perform an independent similarity search for each query frame, $P(e_m|h_m), m = i, ..., j$, are independent of each other. Therefore, we can rewrite objective function (6) as:

$$\begin{aligned} h^* =& \arg\max_{h \in H_{sub}(i,j)} \{P(h_i) P(e_i|h_i) \\ & * \prod_{m=i+1}^{j} P(h_m|h_{m-1}) P(e_m|h_m) \end{aligned} \quad (7)$$

In order to calculate the objective function above, we need to estimate both $P(h_m|h_{m-1})$ and $P(e_m|h_m)$, i.e., the state transition probability $Tr = \{P(s_y|s_x)\}$ and the emission

probability $Em = \{P(e_y|s_x)\}$. To this end, two relaxed constraints are given in the following. Note that we assume $P(s_x)$ follows the uniform distribution. Therefore, $P(s_x)$ is set to $1/Y$, where Y is the total number of states. That is, $P(h_i)$ is set to $1/Y$.

Transition Constraint: For any two reference frames (states) s_x and s_y, s_x can transfer to s_y only if s_x and s_y are in the same shot or in two adjacent shots. If the transition exists, its transition probability $P(s_y|s_x)$ is set to 1(in the same shot) or 0.8(in two adjacent shots), otherwise 0.

Emission Constraint: For emission probability $P(e_y|s_x)$, we directly calculate it using scoring function for similarity search. All similarity scores are normalized beforehand.

As indicated above, a key problem is how to distinguish copies from non-copy video clips, i.e., determining the positions of q_i and q_j. To this end, we introduce an additional gap constraint to localize the boundaries of copies.

Gap Constraint: Given the query subsequence $E_{seq} = (q_{i-\Delta t}, ..., q_i, ..., q_m, ..., q_j, ..., q_{j+\Delta t})$ and a reference frame sequence $h = (h_{i-\Delta t}, ..., h_i, ..., h_m, ..., h_j, ..., h_{j+\Delta t})$, if h_i doesn't have any transition to Δt reference frames in the past, the time instant i is a possible starting point of a copy. The constraint means that we can determine the starting instant of a copy based on the transition relationship among similar reference frames at different time instants. Likewise, we can determine the ending instant in the same way.

As mentioned in section 1, the state set is dynamically updated. In our scheme, only the reference frames in $\{L_{i-\Delta t}, \cdots, L_i, L_{i+1}, \cdots, L_j, \cdots, L_{j+\Delta t}\}$ are used to form the state set, which is updated when query frames comes into or out of the set. In this way, the size of state set always keeps small, which greatly reduces computational cost and memory requirement.

4. EXPERIMENTS

4.1. Experimental Setup

In the proposed scheme, one of remarkable features lies in that it can detect video copy in unbounded query videos rather than short query as existing methods do. However, to facilitate comparison with previous work, we adopt the Sound & Vision dataset used in TRECVID 2008 CBCD task [10] for testing, where the query videos are relatively short. In our experiment, five original query video streams are generated for testing, in which four of them indeed have copies. To test the robustness to various video transformations, these queries are further transformed separately by applying ten complex transformations [10]. Then total 50 query video streams are generated for testing.

For evaluation criteria, we employ the miss rate and the false alarm rate to evaluate the detection precision. Moreover, we introduce additional two criteria for evaluating localization precision in query stream and reference video.

Copy Overlap Degree: This criterion measures the overlap degree in time duration between the detected copy and its ground truth.

Reference Overlap Degree: This criterion measures the overlap degree in time duration between the asserted reference clip of a copy and its ground truth.

In our experiment, three leading copy detection systems [5, 7], which achieve the best detection performance in TRECVID 2008 CBCD task, are used for comparison. These three systems are named as Lear-Strict, Lear-Soft, and IMedia-Fusion, respectively. For the proposed scheme, the length M of L_m is fixed to 200, and the gap Δt is set to 3.

4.2. Localization Precision

In this section, we compare these systems on the localization performance for different transformations. The histograms of both query and reference overlap degrees are plotted in Figure 2 and 3, respectively. For the copy localization, the proposed method obtains the best performance for four transformations, i.e., T1, T3, T5, and T10. The performance for other transformations except T8 is also good. For the copies undergoing T8 transformation, we find that almost no true relevant reference frames are returned for the copy frames near the boundaries. This means that only a part of copy is detected, which leads to low query overlap degree. The main reason lies in that the SIFT descriptor used for feature extraction is not robust to flip transformation which is commonly occurred in T8. In our future work, we will take the flip transformation into account when designing feature extraction scheme.

For the reference localization, the proposed method achieves the best performance for six of ten transformations. The performance for other transformations is also comparable with the best ones. This means that our proposed copy detection method is indeed robust to various video distortions. Encouragingly, the proposed method achieves the best localization performance for the most complex transformation T10 in both the query and reference videos. That is, our method can tolerate severe signal distortions.

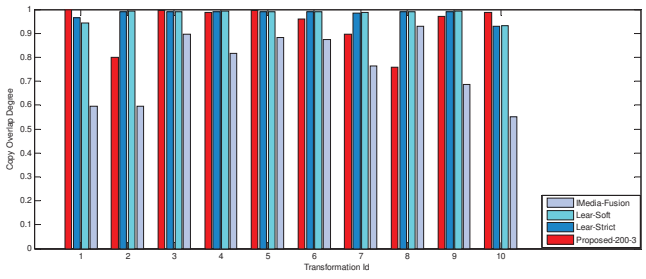

Fig. 2. Comparison with state-of-the-art copy detection systems on copy overlap degree with varied transformations

4.3. Detection Precision

Here, we compare all systems on the overall detection performance, i.e., the miss rate and the false rate. Table 1 lists all

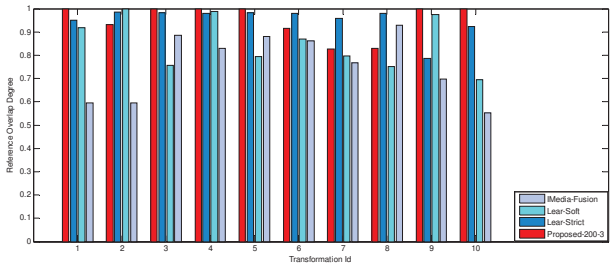

Fig. 3. Comparison with state-of-the-art copy detection systems on reference overlap degree with varied transformations

the evaluation results. Although the performance of the proposed algorithm is not as good as other systems in the miss rate, it still achieves a comparable performance with the system IMedia-Fusion, especially in the false alarm rate. As discussed in [6], a complete copy detection system comprises a few key components including the sampling rate of key frames, feature extraction, similarity search as well as frame fusion results. The overall performance of such a system depends on the aggregated result of all the constituents. In our scheme, we focus mainly on the frame fusion stage. Although the proposed frame fusion method achieves high localization precision, it indeed slightly reduces the detection precision. In fact, this issue can be solved by enhancing the other components. For example, the detection performance can be notably improved by simply changing the size of visual vocabulary since there is a tradeoff between the robustness and discriminability of bag-of-features [8]. A larger size of visual vocabulary means better discriminability capability. Again, we

Table 1. Comparison on Miss rate and False alarm rate

System	R_{Miss}	R_{FA}
Lear-Strict	0.000000	0.166667
Lear-Soft	0.075000	0.663636
IMedia-Fusion	0.200000	0.360000
Proposed-200-3	0.225000	0.288900

want to emphasize that the main goal of this work is to prepare for the semantic mining among database videos based on the copy occurrence. In this scenario, copy detection system must have the capability for detecting copy pair among long videos. Our main effort in this work just focuses on this problem.

5. CONCLUSION

In this paper, we propose a frame fusion based copy detection approach, which involves similar frame search and frame fusion. The key idea is to convert copy detection problem to HMM decoding problem so that Viterbi algorithm can be employed for fast computation. A remarkable feature of this algorithm is that we can dynamically determine the starting and ending points of copies in videos. In this way, the detection of copy pair among unbounded videos can be carried out easily, paving the way to mining semantic correlation among videos. The experimental results show that the proposed approach achieves high localization accuracy in both the query stream and the reference videos and provides good tolerance to some difficult video transformations.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (61025013), Sino-Singapore JRP (2010DFA11010), 973 Program (2011CB302204), the Open Foundation of National Laboratory of Pattern Recognition (2009JBZ006-3).

6. REFERENCES

[1] M.S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM TOMCCAP*, vol. 2, no. 1, pp. 1–19, 2006.

[2] Yan Ke and R. Sukthankar, "Pca-sift: a more distinctive representation for local image descriptors," in *IEEE conference on CVPV*, 2004, vol. 2, pp. 506–513.

[3] X. Yang, Q. Sun, and Q. Tian, "Content-based video identification: a survey," in *International Conference on Information Technology: Research and Education*, 2003, pp. 50–54.

[4] C.Y. Chiu, C.S. Chen, and L.F. Chien, "A framework for handling spatiotemporal variations in video copy detection," *IEEE TCSVT*, vol. 18, no. 3, pp. 412–417, 2008.

[5] M. Douze, A. Gaidon, H. Jegou, M. Marszalek, and C. Schmid, "INRIA-LEARs video copy detection system," *Proceedings of TRECVID*, 2008.

[6] N. Gengembre and S.A. Berrani, "A probabilistic framework for fusing frame-based searches within a video copy detection system," in *ACM conference on CIVR*. ACM, 2008, pp. 211–220.

[7] A. Joly, J. Law-to, and N. Boujemaa, "Inria-imedia trecvid 2008: Video copy detection," *Proceedings of TRECVID*, 2008.

[8] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE conference on CVPR*, 2006.

[9] S.E. Robertson, S. Walker, S. Jones, MM Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *the Fourth Text Retrieval Conference*, 1995, pp. 109–126.

[10] "Trec video retrieval evaluation.http://www-nlpir.nist.gov/projects/trecvid," 2008.