Invariant Visual Patterns for Video Copy Detection

Xiao Wu[1, 2], Yongdong Zhang[1], Yufeng Wu[1], Junbo Guo[1] and Jintao Li[1]
[1]*Key Laboratory of Intelligent Information Processing,*
Institute of Computing Technology, Chinese Academy of Sciences
[2]*Graduate School of the Chinese Academy of Sciences*
{wuxiao, zhyd , wuyufeng, guojunbo, jtli}@ict.ac.cn

Abstract

Large scale video copy detection task requires compact feature insensitive to various copy changes. Based on local feature trajectory behavior we discover invariant visual patterns for generating robust feature. Bag of Trajectory (BoT) technical is adopted for fast pattern matching. Our algorithm with lower cost is more robust compared to the state-of-art schemes.

1. Introduction

Video copy detection system is designed to judge whether there exits a common segment (copy) in two different videos, which can be viewed as a problem of visual pattern mining and matching. For the existing of various visual transformations on videos (e.g. contrast change or lower frame resolution), video copy segments with identical visual content may not share the same appearance [1]. It is important to research how to uncover the underlying common patterns of visual segments and how to construct stable feature robust to a variety of usual transformations. Meanwhile, a widely noticed fact is that in a video copy detection system, the number of query videos which contain copy segments is much smaller than number of non-copy queries. And the duration of copy segments is properly shorter than the duration of the query video. This is essentially a problem of sparsely pattern matching. Hence efficient filtering scheme and accurate copy segment localization algorithm are also critical in practice. In this paper we attempt at evolving the invariant visual patterns to address above problems for advanced content based video copy detection.

This work focuses on incorporating both spatial and temporal information to model the invariant visual patterns of video content for copy detection. In Section 3, we present the invariant visual patterns based on Harris detector and KLT tracking method which is insensitive to various transformations and efficient to generate. Our algorithm localizes the copy segment at sub-shot level. By adopting visual keywords technical, each sub-shot is represented by the concise pattern keywords and then copy detection task is converted to a pattern matching problem to address (refer to Section 4). In Section 5 the performance is quantified on a benchmark dataset [14] compared to state-of-art works.

2. Related works

It is worth noting that effective features for large scale video copy detection should satisfy three key requirements, i.e. 1) insensitive to frequent spatial and temporal video transformations; 2) low computation complexity and 3) small quantity of disk occupation. A category of previous works are focus on extracting various global and local visual features of video frames [2-7]. A common place of these features is to capture spatial signals for frame matching. Early researches proposed compact signatures for fast matching [2] [3], which are incapable in coping with local frame transformations. Ordinal Measure (OM) [2] is deemed to be the best solution for detection under global transformations [10]. Recent years, the well studied local covariant area detectors [4] and descriptors [5] from computer vision field are introduced for copy detection in [6] [7] to construct features. Though have considerable robustness, local features detectors with high computation cost restricts the practical application.

On another hand, a number of researchers aim to adopt temporal information to construct robust feature for segment alignment [8] [9]. This category of algorithms focuses on utilizing temporal character-istics for sequence localization, which are robust to visual transformations but sensitive to temporal changes such as frame dropping and ratio change.

We aim to extract invariant visual patterns for matching different videos with common segment. Different from previous works, we incorporate both spatial and temporal info to generate the feature. And the localization algorithm is conduct using sub-shots instead of frames or clips.

3. Mining invariant visual patterns

The invariant visual pattern is generated using video trajectory, which has found wide application in previous works of surveillance video analysis. Recent developments on local invariant features [5] boost the usage of trajectory in video content analysis [6] [12]. In this paper we model the invariant visual patterns by extracting stable local feature trajectories and modeling its behavior patterns.

3.1. Extracting stable local feature trajectory

As one of the most robust local covariant area detectors [4], Harris detector is introduced into copy detection work [6] in which its robustness is confirmed. Different from [6] which designed 20 dimension point descriptor to match and construct video trajectory, we attempt at adopting faster and stable point tracking scheme to track Harris points. KLT [11] algorithm is used in previous duplicate detection work [12], but the point number is to be fixed in every frame which is not suitable for describing various visual contents. For example, KLT detects same number of feature points even on the black frame with no visual content. In our work Harris detector is combined with the fast KLT tracking approach to get stable local feature trajectory.

The tracking approach is to minimize the sum of squared intensity differences between two consecutive frame windows [11]. For a given frame sequence, Harris detector is applied on first frame to generate initiate feature points. Then KLT tracker matches points between adjacent frames to form trajectories (ref. to Figure1). It is observed that trajectories are cut off when visual content dramatically change (e.g. shot boundary). We re-detect the Harris points and extract another group of trajectories. During this process, the frame sequence could be segment into elementary short clips for matching, denoted as the sub-shot (SS).

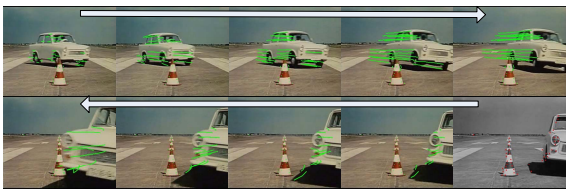


Figure 1. Generating local feature trajectories using Harris detector and KLT tracker

3.2. Modeling trajectory behavior patterns

Visual trajectories represent the behavior patterns of feature points, which are relative invariant despite of various visual transformations. To obtain concise feature of small size, an encoding strategy is designed to convert trajectory behaviors into visual patterns in form of code sequence, as shown in Figure 2. A code number represents the relative position (i.e. behavior trend) of two consecutive feature points on a trajectory.

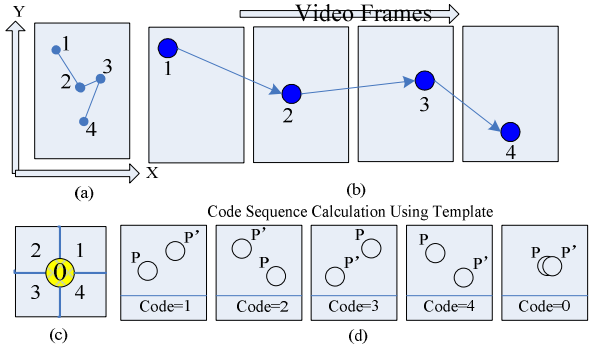


Figure 2. (a) Relative position of a point on continual frames (b) feature point "walks" on a video trajectory (c) quadrant template for point behavior encoding, the center area defines the scale of point repeatability (d) position relations and the corresponding code number

The spatial info is encoded as the individual codes and temporal info is encoded as the order of the sequence. Global transformations are supposed to have little effects on the relative position of two points because of the robustness of Harris [4] [6] and KLT [11]. Though local variations change the amount and position of trajectory, the survived ones are of similar behaviors. However, trajectory under temporal changes (e.g. frame ratio change) will have different code sequences. In addition, code sequences with different length are hard to compare. To address these problems we propose a normalization method to quantify the behavior code sequence into a histogram feature.

$$CH(i) = \frac{\#i}{Length(Code\_Sequence)},\ i = 0,1,...,4 \quad (1)$$

Formula (1) demonstrates the process of normalizing the code sequence into Code Histogram (CH) of 5 dimensions, which preserves the statistical spatio-temporal info and facilitates the comparison between trajectories. For example, the sequence {4, 1, 3} is

converted into a 5-Dimension histogram of {0:0, 1: 0.33, 2: 0, 3:0.33, 4:0.33}. The discrimination of the histogram feature increases when trajectory is longer.

4. Matching invariant visual patterns

After extracting and modeling invariant visual patterns, a video is represented as a group of SSs. Each SS contains amount of CHs. Inspired by the successful visual keywords and Bag of Words (BoW) techniques [13], we view the SS as a document which contains a Bag of Trajectory words (BoT), i.e. CHs. In this Section, based on BoT and CH dictionary we propose an algorithm for invariant visual pattern matching.

4.1. Off-line: generation the CH dictionary

Similar to visual keywords [13], priory to detecting queries, CHs of all the videos in reference dataset are first calculated and then clustered using the Kmeans algorithm. The off-line procedure generates a CH dictionary in which each word is the center of a CH category, denoted as the CH Keyword (CHK). An observation is, there exist large amount of stationary trajectories in BoT, whose CH is zero sequence. Behavior of these trajectories indicates a sub-shot consists of stationary frames. Since the CHs which fall in the stationary category are lack of discrimination power, the most frequent CHs that occur in almost all sub-shots are suppressed using a stop list analogy. Inverted indexing structure is adopted in this work for managing large scale features as it is in BoW works.

4.2. On-line: matching based on sub-shots

In online procedure, query video is segment into many SSs as described in Section 3.1. Invariant visual patterns of SSs are first encoded into CHs and then classified into CHKs using CH dictionary for matching.

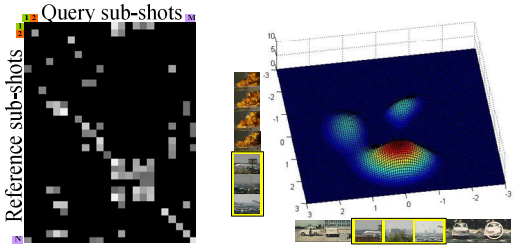
4.2.1. Sub-shot matching using cosine distance. We adopt the visual analogy of document retrieval to sub-shot retrieval. Suppose the size of CH dictionary is K, each sub-shot is represented as a K-vector, i.e. $V_s = \{t_1, t_2, \ldots, t_i, \ldots, t_k\}$. t_i quantifies the weighted word frequency which is computed using the weighting scheme (i.e. tf-idf) described as in Formula (2), where n_{id} is the number of occurrences of CHK i in SS d, n_d is the total number of CHKs in the SS d, n_i denotes the number of occurrences of CHK i in the whole reference dataset and N is the number of CHKs in the

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \qquad (2)$$

$$Sim(SS_a, SS_b) = Co\sin e\_Sim(V_a, V_b) \qquad (3)$$

whole dataset. Using cosine distance between V_s (ref. to Formula (3)), on-line sub-shot retrieval is conduct at very low computation cost. For each query sub-shot, similar sub-shot candidates from reference dataset are ranked according to the cosine similarity.

4.2.2. Video matching based on sub-shot similarity matrix. To locate the copy segment coexists in query and reference videos, we search "ridge" and "plateau" in the similarity matrix using watershed algorithm as in Figure 3. "Ridge" denotes the matching of continual SSs and "plateau" denotes matching of a group of SSs. Video segment with largest total similarity is located.



(a) (b)

Figure 3. (a) Example similarity matrix of query and referent sub-shots. (b) The copy clip (airplane) is located by discovering the "ridge" using watershed algorithm. Solo "peaks" of smaller similarities are discarded

5. Dataset and comparison experiments

Video collection [14] for CIVR 2007 video copy detection evaluation and TRECVID 2008 CD task [15] is used as the reference corpus. Following the official instruction of query building [15], 20 segments (duration range from 3 Secs to 2 Mins) are randomly extracted from this corpus and embedded into 20 TRECV 2007 high-level videos to generate the queries. 10 TRECV videos are selected as non-reference noises. Each query suffers one kind of transformations such as contrast, gamma and re-encoding as in [15].

Table 1. Information of videos processing

Info / videos	#video	Total duration	#Sub-shot	time cost	Disk space
Reference videos	101	58hrs	652,500	25 hrs	227.69 mb
Query videos	30	2hrs 20mins	25,500	56 mins	7.22 mb

Table 1 demonstrates info of processing reference and query videos in our detection system. As discussed

in Section 1, low computation and storage cost of our algorithm facilitate practical large scale copy detection.

First, we evaluate robustness of our algorithm using the evaluation measure of TRECVID 2008 CD task [15], i.e. detection precision, location accuracy which are defined as follow and the curves are drawn in Figure 4 with various K. $\#Overlap(Det, Copy)$ denotes the number of positive copy frames we detected. Our algorithm achieves both good detection precision and location accuracy when reference data is well clustered.

$$Det\_Prec = \frac{\#Correct\_Det}{\#All\_Det} \quad loc\_Accu = \frac{\#Overlap(Det, Copy)}{\#Copy\_Frames}$$

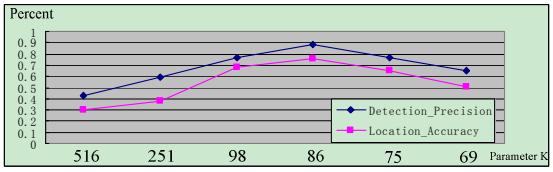


Figure 4. Curves of average detection precision and location accuracy of our algorithm using different K

Second, we compare results of proposed approach against Chiu's algorithm [8], which claimed to be superior to [2] [3]. The results are plotted in Figure 5. Different from schemes using ordinal measure [2] [8], trajectory method uses Harris detector which is robust to severe local changes (e.g. occlusion, crop, shift and picture in picture). We also exhibit sub-shot examples in Figure 6 to illustrate the invariance of trajectory behavior under several common changes.

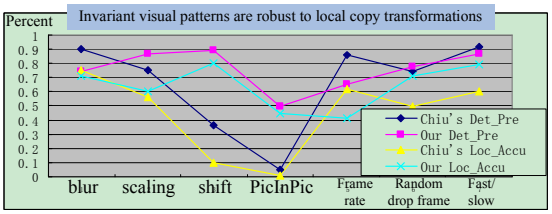


Figure 5. Curves of our algorithm and Chiu's [8] under various copy changes (average results). As in previous BoW works, our K is tuned for best performance

6. Conclusion and future works

How to extract and match invariant visual features is the key problem of video copy detection. In this paper robust visual feature is generated using behavior of local feature trajectory. Based on Bag of Words technical and sub-shot matching scheme, compact features are matched for locating copy segments at low computational cost. Our future works will focus on developing finer trajectory behavior feature to improve the discrimination power of invariant visual patterns.

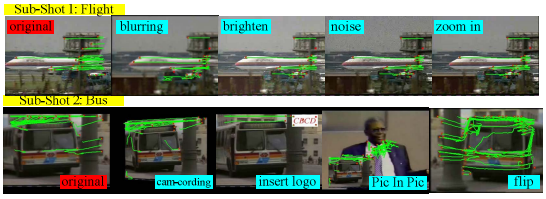


Figure 6. Green curves in an image represent trajectories in a sub-shot. Trajectory behaviors are relatively similar despite of various usual copy transformations

7. Acknowledgement

This work was supported by the National Nature Science Foundation of China (60773056), the National Basic Research Program of China (973 Program, 2007CB311100) and the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416).

Reference

[1] A Joly, O Buisson, C Frelicot. Content-based Copy Retrieval using Distortion-based Probabilistic Similarity Search. IEEE Trans. on Multimedia, 2007.

[2] X. S. Hua, X. Chen, and H. J. Zhang, Robust video signature based on ordinal measure, ICIP, 2004.

[3] C. Kim and B. Vasudev, Spatiotemporal sequence matching for efficient video copy detection, IEEE Trans. on Circuits and Systems for Video Technology, 2005.

[4] K Mikolajczyk, T Tuytelaars, C Schmid, A Zisserman. A Comparison of Affine Region Detectors. IJCV, 2005.

[5] K Mikolajczyk, C Schmid. A Performance Evaluation of Local Descriptors. IEEE Trans. on PAMI, 2005.

[6] J Law-To, V Gouet-Brunet, O Buisson, N Boujemaa. Local Behaviours Labelling for Content Based Video Copy Detection. ICPR, 2006.

[7] X Wu, AG Hauptmann, CW Ngo. Practical Elimination of Near-Duplicates from Web Video Search. MM, 2007.

[8] CY Chiu, CH Li, HA Wang, etc. A Time Warping Based Approach for Video Copy Detection. ICPR, 2006.

[9] A. Hampapur, K.-H. Hyun, and R. M. Bolle, Comparison of sequence matching techniques for video copy detection, The SPIE Conference on Storage and Retrieval for Media Databases, 2002.

[10] J Law-To, L Chen, A Joly, I Laptev, O Buisson. Video copy detection: a comparative study. CIVR, 2007.

[11] J Shi, C Tomasi. Good Features to Track. CVPR, 1994.

[12] M Takimoto, J Adachi. Scene duplicate detection from videos based on trajectories of feature points. MIR, 2007.

[13] J Sivic, A Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. ICCV, 2003.

[14] The origin of the video corpus is MUSCLE-VCD-2007 http://www-rocq.inria.fr/imedia/civr-bench/index.html

[15] Guidelines for the TRECVID 2008 CD task Evaluation. http://www-nlpir.nist.gov/projects/tv2008/tv2008.html