VISUAL WORDS BASED SPATIOTEMPORAL SEQUENCE MATCHING IN VIDEO COPY DETECTION

Huamin Ren[1,2,3], Shouxun Lin[1], Dongming Zhang [1], Sheng Tang[1], Ke Gao[1,2]

[1]Laboratory of Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Graduate University of Chinese Academy of Sciences, Beijing, China
[3]Information Center, Beijing University of Chinese Medicine

ABSTRACT

This paper proposes a novel content-based copy retrieval scheme for video copy identification. Its goal is to detect matches between a doubtful video and the ones stored in the database of the legal holders of the videos. Due to various transformations the copy may has, we use visual words vector as a representation of a frame which is based on SIFT descriptor. Unlike traditional Bag-of-Words (BoW) based approach applied in semantic retrieval, in which the temporal variation during the video is always neglected, our matching algorithm takes into account spatial and temporal distances between a query clip and the one in database. Experiments show robustness and effectiveness of our approach according to various single and compound transformations.

1. INTRODUCTION

Growing broadcasting of digital video content on different media brings the search of copies in large video databases to a new critical issue. Videos collected through different facilities with common content may differ in color, resolution, contrast and noise; meanwhile, source videos may be transformed purposely, such as cam-cording, flip transformation, crop, shift and picture in picture (PIP). Figure 1 illustrates frames from source videos and that from their corresponding copies with single and compound transformations.

By definition from Trecvid, a copy is a segment of video derived from another video, usually by means of various transformations such as addition, deletion, modification (of aspect, color, contrast, encoding...), cam-cording, etc [1]. Therefore, the intrinsic task of Content Based Copy Detection (CBCD) is to determine whether a video clip is a copy or not. Detecting such complicated and diverse transformed copies in mass web data is significant though difficult. It can provide powerful solutions for copyrights protection. Besides, the techniques used in CBCD can also help in improving the performance of Content Based Video Retrieval (CBVR). Nowadays video retrieval systems always provide results with large number of duplications. If duplications and copies can be declined, we could get further step in improving precision and meeting user's needs.

Recently, methods for CBCD can be categorized into two: global and local. Global methods extract global features based on color histogram or spatial-temporal distribution of intensities, which can detect the integrated variations and perform well in searching copies with transformation in color, contrast, etc. Such methods contain [2-4]. Local methods, such as [5-6], focus on interest points detector and a description of the local region around each interest point. Searching for copies from database is a process of comparison between feature descriptors or trajectory of interest points. When considering post-production transformation, like insert of caption, cam-cording and PIP, local methods exceeded global ones in precision and scalability.

(a) (b) (c) (d)

Fig.1. Frames from source video and its copy. (b) is a copy of (a) with single transformation: PIP; (d) is a copy of (c) with compound transformation: cam-cording & highlight

Among local feature based approach, BoW has been widely used in the task of object recognition, content-based image retrieval (CBIR) and near duplicated video retrieval. It models an image as an unordered bag of visual words, which are formed by vector quantization of local region descriptors, such as SIFT [7]. Extracting local features could grasp spatial variations while utilizing bag of words could avoid high computation that large number of interest points brings. However, detecting copies is a difficult task, due to different types of transformations in the database and high rate of mismatch. The discrimination of BoW is not enough to distinguish copies from reference videos. It is extraordinary important to utilize temporal information in video retrieval and copy detection.

In this paper, we propose a spatiotemporal sequence matching method of visual words (STW) to handle a wide range of modification that the video signal might undergo. Our approach contains two processes: offline and online (See Figure 2). With offline processing, we extract features from each frame of video from the reference video database to create feature descriptors (for simplicity and computation, we extract every ten frames in our experiments). K means clustering is then used for generating vocabulary with K words. Each feature is then quantized into its nearest visual word after comparing with each word in vocabulary, a vector with words frequency is formed to represent a frame and vectors of frames constitute representation of a video. During online processing, similar descriptors and visual words representation at the frame level is then built. Recognition of a copy can then be conducted by spatial and temporal matching of the descriptor vectors (visual words representation) between query video and reference videos.

The rest of this paper is organized as follows. Section 2 provides the details of generating visual word representations. Section 3 describes a sequence matching method of visual words. Section 4 outlines several experiments, followed by discussion. And conclusions and plans for future work are drawn in section 5.

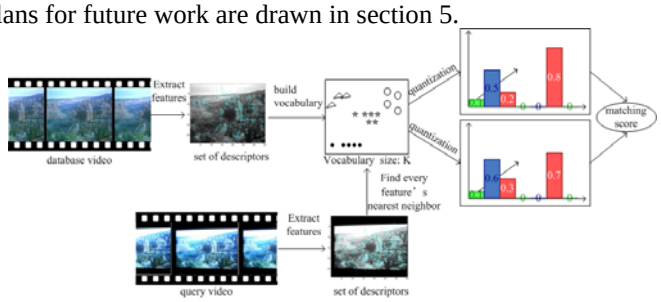

Fig.2. Framework of STM

2. FRAME LEVEL REPRESENTATION

This section describes the procedure of building visual words vocabulary and generating representation for each frame, and then introduces various factors that can affect the performance of utilizing visual words in CBCD.

2.1. Building vocabulary and generating representation

Visual words representation for each frame is generated as follows:

STEP 1: Extract Feature and Descriptor. Typically each frame is represented by a set of overlapping regions each represented by a vector computed from the region's appearance. There already exits many algorithms to detect interest points and generate descriptors, such as Shape Adapted (SA), Maximally Stable (MS), GLOH, SURF and SIFT. As compared in [8], SIFT is superior to SA and MS due to its invariant property to a shift of a few pixels in the region position. Besides, SIFT descriptors are generally recognized as a superior feature in affine invariant transformation, though a little bit more time consuming than GLOH and SURF. Taking our goal of copy detection into account, we choose SIFT for its higher accuracy over speed.

STEP 2: Building visual words vocabulary. A vocabulary is generated by clustering the detected key points in their feature space and treating each cluster as a unique visual word.

STEP 3: Assigning the descriptor to particular visual word. For each feature in feature space, find its nearest neighbor in vocabulary and assign it to that word; then count the number of visual words in each frame and normalize to a vector with *K* dimensions as a representation, where *K* is the vocabulary size.

2.2. Some considerations utilizing visual words in CBCD

First is vocabulary size [9]. Size is a big problem when BoW is applied in object categorization and semantic video retrieval. A small vocabulary may lack the discriminative power since two features may be assigned into the same cluster even if they are not similar to each other, while a large one is less generalized. When applied to copy detection, which contains large number of features and more classifications, the problem goes even tougher since small vocabulary lack information to indentify copies and large one need numerous and heavy calculation in step 2 and 3 above. After observation, we choose K=60 for high efficiency in our experiments.

Second, the role of visual words in CBCD is quite different from that in semantic retrieval, thus we have to consider feature selection and matching cautiously. For semantic retrieval, the key is to find out saliency features, which can represent notable characteristics of video streams, during the course of which features from background are always ignored; On the contrary, copies share the same semantic content with reference videos, only changing in some aspects like color and etc. Visual words from background provide useful contextual information and could assist identifying copies, therefore, cannot be ignored. Meanwhile, words with less representative meanings should be rid off.

Third, comparison of visual words at the frame level is not discriminative enough to indentify copies, which calls for other matching scheme making use of its temporal characteristics.

3. SPATIOTEMPORAL SEQUENCE MATCHING OF VISUAL WORDS

Two questions are put up according to the problems above. First, how to select features and represent a frame; second, how to utilize temporal information that the video has. As a solution to question 1, we separately use normal words scheme and weighting words scheme. As a solution to question 2, we use spatiotemporal sequence matching.

3.1. Problem formulation

We consider video as finite sequences of frame, each of them is represented by visual words.

$$V = \left\langle F_i \right\rangle, i = 1,2,\ldots, n.$$

where n is the number of frames in a video and F_i is a vector with K dimensions.

$$F_i = \left\langle w_{i,h} \right\rangle, i = 1,2,\ldots, n, h = 1,2,\ldots, K.$$

where $w_{i,h}$ denotes the h^{th} visual word in frame i and K denotes the size of vocabulary.

Formally, copy detection is defined as: Given a query clip Q, $Q=<F_1,\ldots\ldots F_N>$ and a series of reference videos in the database, each of them is represented as $R_j=<F_1,\ldots\ldots F_M>$, $N<<M$. Frame i of video Q is represented by $<qw_{i1},\ldots.....qw_{ik}>$ and of R by $<rw_{i1},\ldots.....rw_{ik}>$. The problem is: how to calculate the similarity between Q and R_j and how to judge whether Q is copy of R_j or not.

3.2. Visual words with and without weighting

The *term frequency* (TF) weighting is often used in information retrieval and text mining. Here we use *TF* to evaluate how important a visual word is in a collection of videos. The *TF* of visual word in the given database is the frequency of a given word appears in the database.

$$tf(word_i) = \frac{\#word_i}{\sum_{j=1}^{K} word_j} \quad (1)$$

Here $\#word_i$ is the number of occurrences of the considered visual word in the database, and the denominator is the number of occurrences of all visual words in the database.

By using *TF* weighing scheme, we expect to find how weighting scheme performs in detecting copies. As could see in section 4, weighted visual words don't surpass visual words without weighting scheme, which is obviously different from semantic retrieval.

3.3. Spatiotemporal sequence matching

Traditional BoW matching techniques have relied on frame correspondence. The distance between two video sequences is computed by combining dissimilarities of corresponding frames. However, we note the failure to notice the temporal variation of videos may lead to higher false detection of copies. We use spatiotemporal matching approach proposed in [2] to combine spatial distance of visual words in each frame and temporal distance of signatures from the temporal trails of the visual words.

Matching procedure is: Each time compare *Q[1:N]* with *SR*. Here *SR[1:N]=R[p: p+N-1]*, let p start with 1 and *p++* after one comparison. We say video Q is a copy of R if the dissimilarity between Q and SR is less than a noise threshold. We define the dissimilarity measure D as follows:

$$D(Q,SR) = \alpha_1 * D_s(Q,SR) + \alpha_2 * D_t(Q,SR) \quad (2)$$

Ds(Q,SR) and *Dt(Q,SR)* separately defines spatial and temporal dissimilarity between Q and SR, α_1 and α_2 are influence factors for balancing between spatial and temporal distances, here $\alpha_2 = 1 - \alpha_1$.

$$D_s(Q,SR) = \frac{\sum_{i=1}^{N} d(Q_i, SR_i)}{N} \quad (3)$$

d(Qi, SRi) is the dissimilarity of visual words between the i^{th} frame of Q and SR, defined as follows.

$$d(Q_i, SR_i) = \frac{1}{C}\sum_{j=1}^{K}(qw_{ik} - rw_{ik}) \quad (4)$$

where C is a constant number. If visual words have been normalized, let $C=1$. *Dt* focus on the variation of visual words frequency, and defined as:

$$D_t(Q,SR) = \frac{1}{K*(N-1)}\sum_{j=1}^{K}\sum_{i=2}^{N}(\delta_{ij}^{q} - \delta_{ij}^{sr}) \quad (5)$$

where K is vocabulary size and δij is defined as:

$$\delta_{ij} = \begin{cases} 0.5, & if \quad w_{i,j} - w_{i-1,j} > threshold \\ 0, & if \quad \left| w_{i,j} - w_{i-1,j} \right| < threshold \\ -0.5, & if \quad w_{i-1,j} - w_{i,j} > threshold \end{cases}$$

4. EXPERIMENTS

4.1. Database

Table 1. Various transformations to ten video clips from CVPR07. Each line includes source video number, copy video number and transformation type to make that copy.

group	video #	copy #	type of transformation
single transformation (group I)	1	1c	blur
	2	2c	resolution
	3	3c	gamma
	4	4c	noise
	5	5c	picture in picture
compound transformation (group II)	6	6c	blur & highlight
	7	7c	cam-coding & highlight
	8	8c	cut & blur
	9	9c	highlight & subtitle & stretch
	10	10c	blur & noise

Ten video clips are selected from CVPR07 dataset, each with 10 seconds. These ten videos are classified into two groups and different transformations are produced manually on them using Premiere. See table1. Single transformation is put on video 1-5 (group I) and compound transformations are put on video 6-10 (group II).

4.2. Performance and discussions

First, we calculate precision and recall with different α to find best balancing value between spatial and temporal distances. We modify α of STW using *TF* weighting according to group I. Precision and recall are commonly used index to evaluate the performance, they are defined as follows:

$$\Pr ecision = \frac{\text{number of matched copy clips}}{\text{number of matched clips}} \quad (6)$$

$$\text{Re}call = \frac{\text{number of matched copy clips}}{\text{number of total copy clips}} \quad (7)$$

Results can be seen in figure 3. It turns out α =0.5 is the best. Thus spatial and temporal distances are equally important.

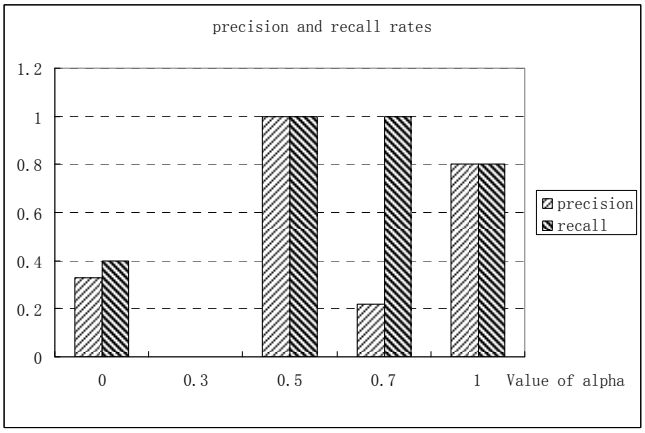

Fig.3. Precision and recall of STW using weighting towards different α

We then compare detection performance of two schemes, STW with and without weighting. Weighting scheme could detect 1c, 3c, 4c, 5c in group I and 8c, 9c, 10c in group II while STW alone could detect 2c, 3c, 4c in group I and 7c, 8c, 9c in group II. The precision and recall of weighting scheme are 1 and 0.8 in group I, while 1 and 0.6 in group II, compared to non-weighting scheme which is 0.8 and 0.8 in group I, while 1 and 0.6 in group II. The detection performance of two schemes is approximate, but the discrimination in type of transformation is quite different. Considering single transformation, the number of features could be detected in small window of PIP is also small since the size of the window is less than 1/4 of the original one, but utilizing weighting scheme could emphasis these features who appear along the time. However, serve blurring loses lots of features, so identification of copies failed utilizing weighting scheme. These characteristics could also be seen in group II. Weighted features with highlight mislead the detection procedure while non-weighted features fail to find copies with blurring.

At last, we compare our approach with ordinary measure used in [2] and verify the effectiveness of our approach. Ordinary measure performs well in global transformations, such as noise, gamma, but performs poorly in any local attack. Its precision and recall is 0.625 and 0.5, both smaller than the results in section 4.2.

5. CONCLUSION AND FUTURE WORD

STW worked effectively due to its several benefits. Firstly, visual words representation, which could grasp the local variance of features and still not so time consuming, show robustness in CBCD. Secondly, spatiotemporal matching fully uses intrinsic characteristics of videos and improves precision in CBCD. In future work, we will focus on modifying clustering strategy since K means is time consuming, enlarging K value to strengthen discrimination of visual words and drawing feature selection to get features more saliency.

6. ACKNOWLEDGEMENT

This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), the National Nature Science Foundation of China (60873165，60802028), Beijing New Star Project on Science & Technology (2007B071) , Co-building Program of Beijing Municipal Education Commission.

7. REFERENCES

[1] http://www-nlpir.nist.gov/projects/tv2008/tv2008.html#4.5

[2] C. Kim and B. Vasudev. Spatiotemporal sequence matching techniques for video copy detection. IEEE Transactions on Circuits and Systems for Video Technology, 1(15):127–132, Jan. 2005.

[3] L. Chen and F.W.M. Stentiford. Video sequence matching based on temporal ordinal measurement. Technical report no. 1, UCL Adastral, 2006.

[4] A. Hampapur and R. Bolle. Comparison of sequence matching techniques for video copy detection. In Conference on Storage and Retrieval for Media Databases, pages 194–201, 2002.

[5] J. Law-To, O. Buisson, V. Gouet-Brunet, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. In ACM Multimedia, MM'06, 2006.

[6] J. Law-To, V. Gouet-Brunet, O.Buisson, N. Boujemaa. Local Behaviours Labelling for Content Based Video Copy Detection. 18th International Conference on Pattern Recognition, 2006. Volume 3, 0-0 0 Page(s):232 - 235

[7] D. G. Lowe. Distinctive image features from scale-invariant key points. International Journal of Computer Vision, 2004.

[8] Sivic. J and Zisserman. A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of ICCV '03, pp. 1470-1477 vol.2, 2003.

[9] Jiang. YG, Ngo. CW. Towards optimal bag-of-features for object categorization and semantic video retrieval. In ACM CIVR, 2007.