Video Copy Detection Based on Fusion of Spatio-temporal Features

BAO Wei, JI Lixin, GAO Shilin , LI Xing , Liu Lixiong
National Digital switching System Engineering & Technological R&D Center
Zhengzhou , China
3202004075@163.com

***Abstract*—A video copy detection method based on fusion of spatio-temporal features is proposed in this paper. Firstly, trajectories are built and lens boundaries are detected by SURF features analyzing, then normalized histogram is used to describe spatio-temporal behavior of trajectories, the bag of visual words is constructed by trajectories behavior clustering, word frequency vectors and SURF features with behavior labels are extracted to express spatio-temporal content of lens, finally, duplicates are detected efficiently based on grade-match. The experimental results show the performance of this method is improved greatly compared with other similar methods.**

Keywords-Video copy detection; Speeded-up robust feature; Spatio-temporal behavior of trajectories; Bag of visual words; Grade-match.

I. INTRODUCTION

With the development of multimedia technology and the popularity of broadband services, many websites allow users to upload and share video. Users can create video through mobile phone and digital camera, or directly download website video to edit and then upload it to share. There are a large number of copy videos on the websites, which infringe the copyright of author seriously. In order to protect intellectual property rights, scholars propose content-based video copy detection technology, abbreviated as video copy detection. Video copy detection has been widely used in areas such as copyright protection, business intelligence, ad tracking and content regulation [1].

The validity of video copy detection algorithm is mainly dependent on the robustness and distinguishability of the feature. Most video copy detection algorithms based on global feature extract low-level feature from the video image to represent the video, but these algorithms are sensitive to various copy techniques, so the detection result is not satisfactory. Local feature describes the structure and texture information of neighborhood of the interest point, having a good robustness generally to brightness, viewing angle, geometry and affine transformations. Local feature is widely used in video copy detection in recent years, and has a good detection performance. However, with the improvement of video resolution and the explosive growth of the amount of video, the number of local features extracted from large video database is increasingly large. And a mass of inter-frame redundancies, a huge amount of computation for similarity measure and memory overhead become the main problems that limit its application. In addition, the majority of video copy detection algorithms based on local feature always use the frame-match. These algorithms do not utilize the time domain information of video. Two irrelevant local features may match, which results in false detection.

For the above issues, Law-To [2] proposes a video copy detection algorithm on the trajectory of feature points. First, it extracts feature points from each frame of video and builds the trajectory by matching feature points. Second, cluster spatio-temporal behavior of trajectories and assign labels having semantic information. Finally, create an index for feature points, which not only reduces the probability of mismatch, but also improves the efficiency of match. By the experiment, they proved its superior detection performance [3]. Basing on this, many scholars propose the improved algorithm. These improved algorithms can be summed up into two categories: one is that Shi Chen [4] etc. use U-SURF algorithm to extract feature points, match feature points to build trajectories, and describe spatio-temporal behavior of trajectories by space coordinates and time coordinates of the feature points. Trajectories are divided into four categories, including stationary, horizontal movement, vertical movement and complex movement. Finally they use local sensitive hash (LSH) to create index to accelerate the match process. The other is that Guo Junbo [5], Wu Xiao [6] use Harris combined with KLT algorithm to rapidly extract trajectories. The video is divided into sub-lenses. They quantify relative displacement of adjacent feature points on one trajectory, describe the spatio-temporal behavior of trajectories by normalized histogram or Markov model, and cluster trajectories to build a bag of visual words. A word frequency vector represents a sub-lens for fast video copy detection. These two improved methods both have some flaws, the former drawback is the huge amount of calculation of feature extraction and matching, which cause time-consuming seriously; the shortcomings of the latter is the feature contains only time-domain information but lacks distinguishability and robustness, and detection result is not satisfactory.

Aiming at the drawbacks of the above algorithms, fusing the advantages of the above algorithms, we propose a new video copy detection method. First of all, the SURF features are extracted from each frame, then we build the trajectories and split lens by analyzing SURF features, then use statistical normalized histograms to describe the spatio-temporal behavior of trajectories. Second, cluster trajectories behavior to build a bag of visual words, calculate word frequency vector describing lens, and finally use grade-match strategy of word frequency vector and SURF features with dynamic behavior label for quick match. The experimental result on MUSCLE-VCD-2007 database [7]

Published by Atlantis Press, Paris, France.

shows the method not only maintain good detection effect but also greatly improve the detection speed, more suitable for large and complex databases video copy detection.

II. OVERVIEW OF THE PROPOSED ALGORITHM PROCESSES

The whole algorithm is divided into offline and online part, as is shown in Fig. 1. The processing steps of Offline part are as follows:

- We firstly extract the SURF features of each frame of video, build trajectory and split lens by analyzing the SURF features;
- Then, quantify and encode the relative displacement of adjacent points along the trajectory, statistically generate normalized histogram to describe spatio-temporal behavior of trajectories;
- Cluster the normalized histograms of spatio-temporal behavior of trajectories for all video, then we regard cluster center as word to build a bag of visual words;
- The spatio-temporal behavior of trajectories along the lens are taking as a word, and the bag of visual words is used for expressing time domain information of each lens as a word frequency vector;
- Word frequency vector and SURF features sets with the label of dynamic behavior are extracted from each lens of the videos in the video library. All these are regarded as the reference video template.

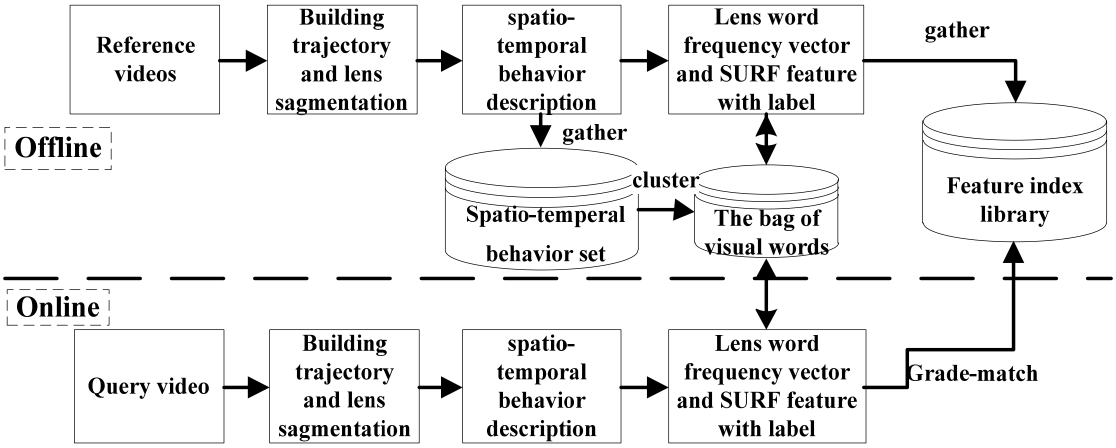
Reference videos
Building trajectory and lens sagmentation
spatio-temporal behavior description
Lens word frequency vector and SURF feature with label
gather
Offline
gather
cluster
Spatio-temperal behavior set
The bag of visual words
Feature index library
Online
Query video
Building trajectory and lens sagmentation
spatio-temporal behavior description
Lens word frequency vector and SURF feature with label
Grade-match

Figure 1. Algorithm flowchart.

Online process consists of feature extraction and grade-match two phases. In the feature extraction phase, the bag of visual words generated in offline module are used to extracted word frequency vector for each lens, and characterize each SURF feature to form a feature set with the label of dynamic behaviors. In the grade-match phase, matching is conducted between lens word frequency vectors of the query video and vectors of each reference video to identify the most similar lens; then SURF features with the label of dynamic behavior are used for exact-match to determine detection results.

III. FEATURE EXTRACTION

Feature extraction includes building trajectory and lens segmentation, spatio-temporal behavior of trajectories description, spatio-temporal feature extraction three steps. The purpose of feature extraction is to extract lens word frequency vector and SURF feature set with label of dynamic behavior to describe each lens' spatio-temporal content.

A. building trajectory and lens segmentation

Standard SURF algorithm has been selected to extract local features in this paper. SURF algorithm was put forward by Bay etc. [8] in 2006, with a higher detection speed, characterizing flexibly, more robust. In 2007, Bauer[9], Luo[10], respectively made an experimental comparison between SURF algorithm and other mainstream algorithm. The results show SURF feature's good robustness to the common video copy transformation, and the speed is significantly better than other algorithms.

Video is a sequence of frames. Per second video generally contains 25 to 30 frames. Each frame within one lens contains similar content. If we simply rollup SURF features of each frame in a video, there are a lot of redundancy and huge amount of data in the feature set. It is difficult to measure their similarity. After extracted SURF features from each frame of video, this paper use feature matching method to build trajectory and split lens. Thus the inter-frame redundancy of SURF features will be eliminated and it will also remove singular point. The simple and stable SURF feature set will be gained to represent each lens. Singular point is an isolated feature points that do not match

Published by Atlantis Press, Paris, France.

other SURF features inside the lens. The trajectory building and the lens segmentation processes are as follows:

1) In this step, SURF features of each video frame are extracted and then perform the following steps to the end.

2) Sequentially match the current frame SURF feature sets (number of features as m) with a set of the trajectories and the features do not include trajectories of preceding 15 frames : When matching with one trajectory in the trajectory sets, the feature is added to the trajectory. The trajectory parameters are updated, and the features are removed from the feature sets of current frame. Record the updated number of trajectories as n_1; When a feature matches with a feature in the preceding adjacent 15 frame sets, a new trajectory is built. Then record the number of the new trajectories as n_3 , the total number of the trajectories is n;

3) If $(n_1+n_3)/n<\varphi \&\& (n_1+n_3)/m<\phi$, the lens conversion occurs, then perform step 4), otherwise go to step 2). φ and ϕ should be set according to the experimental result, in this paper we choose 0.2;

4) Calculate the mean of each trajectory SURF feature as a description of main spatial content of the lens. Record the spatial coordinate of each point on the trajectory as the description of the temporal domain's dynamic behavior. Then split lens and repeat step 2) and step 3) in the new lens.

Through the above processes, the video is divided into a group of lenses. Each lens is described with a SURF feature set and a trajectory set. The SURF feature set is the description of the steady content of the video. The spatial coordinate of each point along the trajectory is the description of the target point’s dynamic behavior in subsequent frames. Two together can complete show the video content.

B. The spatio-temporal behavior of trajectories description

Using space coordinates to describe the spatio-temporal behavior of trajectories exists two problems: one is that it is sensitive to some copy transformations, such as picture-in-picture, geometry transformation, local variations, dropping frames and other transformations in the time domain; the other is that the trajectory length is not uniform, it is difficult to measure the similarity. In this paper, we quantify and encode the relative displacement of adjacent points along the trajectory and then, statistically generate a normalized histogram to describe the spatio-temporal behavior of trajectories. The principle is shown in Fig. 2. The spatio-temporal behavior of trajectories are divided into two states, stationary and movement. Stationary means relative displacement is less than a certain threshold. Movement means relative displacement is greater than the threshold. To increase the distinguishability, the relative displacement is divided into three lengths and eight directions. Coupled with stationary, there are a total of 25 states, corresponding with 25 code. The quantification of the relative displacement should be set according to the video resolution and frame rate.

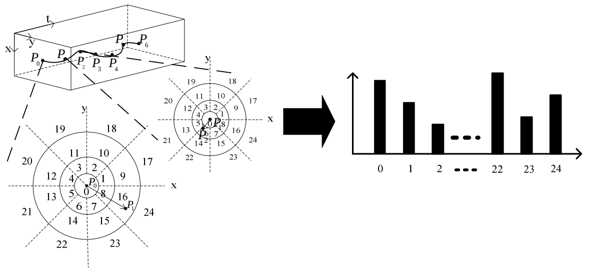

Figure 2. Spatio-temporal behavior of trajectories description diagram.

In order to enhance the robustness of spatio-temporal behavior feature and the stability of the trajectory, we remove the trajectory whose length is less than a certain threshold and normalize the coding histogram. The ultimate spatio-temporal behavior of trajectories are represented as 25-dimensional feature vector:

$$T=\{t_0,\ldots,t_i,\ldots,t_{24}\}\,. \tag{1}$$

Where $t_i=m_i/M$ indicates the frequency of code *i*, *M* is the length of the trajectory, and m_i is the coding *i*’s occurrences along the trajectory. Description of the Normalized histogram has good robustness, and can effectively reduce the impact of the above transformations and improve detection accuracy. E.g. due to the effect of time domain transform, if the trajectory whose length is M drops *x* frames and adds *y* frames, then the effect of dynamic behavior is $x+y$. But the effect on the normalized histogram is $(x+y+z)/M$. When $z<<M$, It means the effect to be reduced *M*x in general case.

C. The video spatio-temporal features extraction

Copy video and source video may vary considerably on visual content, but the semantic is the same. To extract richer semantic information, we first cluster the spatio-temporal behavior of trajectories and build a bag of visual words based on clustering results. Then each lens is regarded as an article taking the spatio-temporal behavior of trajectories as words. Calculate Frequency vector using (3) to represent lens. In order to eliminate the mismatch, according to the clustering results, we give each SURF feature dynamic behavior label, and then store the sorted SURF features with the dynamic behavior label for quickly matching.

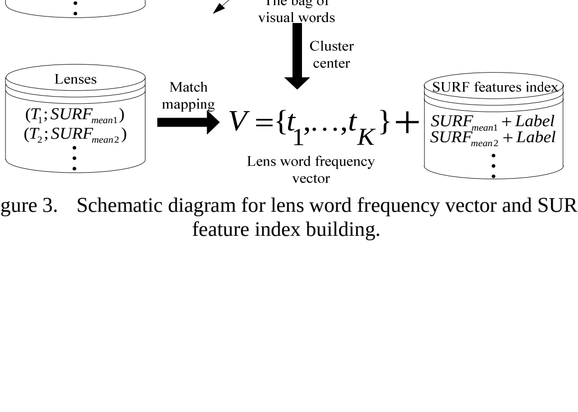

Figure 3. Schematic diagram for lens word frequency vector and SURF feature index building.

Good clustering results can build a more appropriate bag of visual words, and get lens frequency vector and dynamic behavior label easier to distinguish. Compared with K-Means clustering algorithm, AP [11] cluster algorithm is more suitable for processing the large-scale database cluster problem, and can get better clustering results in shorter time [12]. This paper uses AP cluster algorithm improved by Wang kaiJun etc. [12] to cluster spatio-temporal behavior of trajectories. The cluster algorithm can specify the number of clusters. We regard cluster center as visual keyword, each SURF feature in one cluster is marked by it and then build a bag of visual words based on codes from visual keywords. As is shown in Fig. 3, take spatio-temporal behavior of each trajectory along the lens as the words, each lens is represented by a 25dimensional word frequency vector using the bag of visual words:

$$V = \{ t_1, \dots , t_i, \dots , t_K \} . \quad (2)$$

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} . \quad (3)$$

Where n_{id} is the occurrence of spatio-temporal behavior feature i of trajectories along the lens *d*, and n_d is the number of the trajectories along the lens d , n_i represents the number of the lens which contains spatio-temporal behavior feature i of trajectories, while *N* represents the total number of the lens.

IV. GRADE-MATCH

In the former section, the video is divided into a set of lenses. The SURF feature set with dynamic behavior labels and word frequency vector are used to represent each lens. The SURF feature set is a description of the lens spatial information and huge; word frequency vector is an overview representation of the lens time domain information. The lens is one of the continuous shooting of the camera frame sequence and the video’s smallest structural unit. When the query video contains a reference video lens, it is confirmed that the query video is a copy of the reference video. To simplify the match process and improve the match efficiency, this paper uses a grade-match strategy shown in Fig. 4. First, the lens word frequency vector match: If it can determine that a query video lens is a copy of one reference video, then output the result; If unsure, identify the most similar lens, then perform the SURF feature matching, two weighted similarity determine the final inspection results.

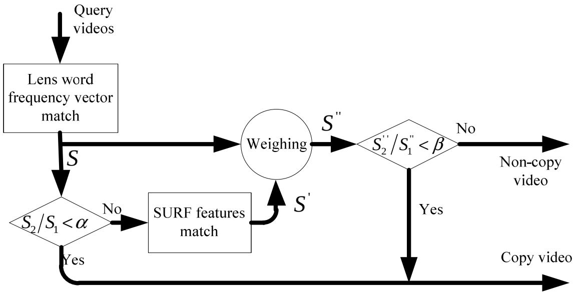

Figure 4. Grade-match flowchart.

The query video takes lens as unit, matches sequentially. In this paper, the cosine distance measures the similarity of the word frequency vector along the lens. Thus obtain the N lenses getting the highest similarity score, the score in descending order are $(D_1 ; S_1),(D_2 ; S_2),\cdots,(D_n ; S_N)$. Where D represents a lens of the reference video library, and S represents the similarity. If S_2/S_1 is less than a certain threshold, it is considered that the query video is a copy of the video which contains lens D_1. Otherwise, in this N lens range, use the following method to measure the similarity of the SURF feature sets: Assume that the two SURF features sets are respectively A and B , and the number of their features is respectively $KeyNum_A$ and $KeyNum_B$. We use the Euclidean distance as the measure, the number of matching features is $KeyMatch_{AB}$. According to Equation (4), calculate the similarity of SURF features sets, the results are respectively $(D_1 ; S_1'),(D_2 ; S_2'),\cdots,(D_n ; S_N')$. We calculate the final similarity with the weighted (5), and the results in descending order are S_1'' , S_2'' ,..., S_N'' . If N is less than a certain threshold value, it is considered that the query video is a copy of the video which contains the most similar lens. Otherwise, the query video is not a reference to a copy of the videos in the video library. According to the experimental results, the *N* is 30.

$$S' = \frac{KeyMatch_{AB}}{\min(KeyNum_A, KeyNum_B)} \times 100\% . \quad (4)$$

$$S'' = \gamma \times S + \eta \times S' . \quad (5)$$

Assuming that the reference video library contains x lenses, and each lens contains y trajectories in average. Query video contains x' lens and y' trajectories per lens. Thus the calculation amount of grade-match strategy is:

$$x' \times (x \times \sigma_K + y' \times y \times N \times \sigma_{64} / K) . \quad (6)$$

If using the feature point matching firstly, and then match trajectory to confirm, the calculation amount is:

$$x^{'} \times y^{'} \times (x \times y \times \sigma_{64} \big/ K + \sigma_{25} \times N) . \tag{7}$$

Where N is the number of the candidate trajectories. K is the number of clusters of the spatio-temporal behavior of trajectories. σ_{64} is the unit computation amount calculating similarity of 64-dimensional SURF features. σ_{25} is the unit calculation amount of similarity of 25-dimensional spatio-temporal behavior of trajectories. σ_{K} is the unit calculation amount of the similarity of K -dimensional word frequency vector. We can find the grade-match can effectively reduce the amount of calculation.

V. Experimental results and analysis

A. Experimental set

Our experiment was conducted on the Intel dual-core 2.9GHz CPU, 2GB RAM computer, Matlab and C language mixed programming. Reference video set is MUSCLE-VCD-2007 database [7].The video set contains 101 reference videos, about 100 hours. Nonreference videos are downloaded from open-video website. The query video is made with TRECVID [1] official program. There are total 300 query videos, about 12 hours. The query videos consist of copy videos of the 10 transformation types. Video format is MPEG-1.

B. Evaluation standard

Two evaluation criteria of TRECVID 2011[1] video copy detection competition are used to measure the performance of the algorithm:

1) Normalized detection cost rate(NDCR):

$$NDCR = P_{miss} + \beta \times R_{FA} . \tag{8}$$

$$\beta = C_{FA} \big/ (C_{miss} \times R_{t\arg et}) . \tag{9}$$

Where P_{miss} is the undetected rate, R_{FA} is the false detection rate, β is a weighting coefficient, $C_{FA} = C_{miss} = 1$ are the cost of false and missed detections, $R_{t\arg et} = 0.005 / h$. The smaller $NDCR$ is, the smaller the cost of video copy detection algorithm, the better the performance.

2) The average detection time T_{mean} :

$$T_{mean} = T_{full} \big/ N_{quers} . \tag{10}$$

Where T_{full} is the time of the whole process for all video from decoding to outputting a detection result, N_{quers} represents the number of the query videos. The smaller the T_{mean} , the faster the algorithm detection is.

C. Performance Analysis

We extracted 31,473,425 trajectories from 58 hours reference videos, which contain 33,804 lenses. In order to improve the stability of the trajectories, the short trajectories whose lengths are less than 3 have been removed. The remaining 29,461,003 trajectories are used for AP cluster to build a bag of visual words. 50 query videos have been selected randomly. Compare the detection results of different number of clusters, and determine the optimal number of clusters. Fig. 5 shows when K is 82, the detection results are best.

In order to verify the effectiveness of our algorithm, under the above experimental environment, we made an experimental comparison between algorithm [2], [4], [6] and our algorithm in this paper. The evaluation criteria are $NDCR$ and T_{mean} . The results are shown in Fig. 6 and Table I. We can see that the detection of our algorithm is better than the algorithm proposed in [6] and [2]. The result is near with the algorithm of [4]. However, the average detection time is much shorter than [2] and [4]. As is shown in Fig. 6, our algorithm is optimal to the following transformations: affine transformation (T1), re-encoding (T4) and a slight drop quality (T6). The reason is we used the standard SURF algorithm to extract the local features which have good robustness to affine transformation. Describing the spatio-temporal behavior of trajectories by normalized histogram can effectively reduce the influence of time-domain variation. Fig. 6 also shows our algorithm is not as good as the algorithm of [4] for some complex transformations (T9), (T10). The reason is that word frequency vector cannot completely eliminate the impact of complex transformations, which cause miss detection. The results can be improved by increasing the number *N* of candidate lens, but it will reduce the efficiency of detection. To set the value of *N*, it should be based on the complexity of copy transformations and the size of video libraries.

TABLE I. DETECTION RESULT OF THE ALGORITHMS

Algorithm	Average NDCR	The average detection time (s)	Average time for feature extraction (s)	Average time for feature matching (s)
Algorithm of Ref [2]	0.3615	1094.31	15.42	1078.89
Algorithm of Ref [4]	0.2403	2517.46	1110.37	1407.09
Algorithm of Ref [6]	0.5368	114.57	49.36	65.21
Our algorithm	0.2451	651.94	584.51	67.43

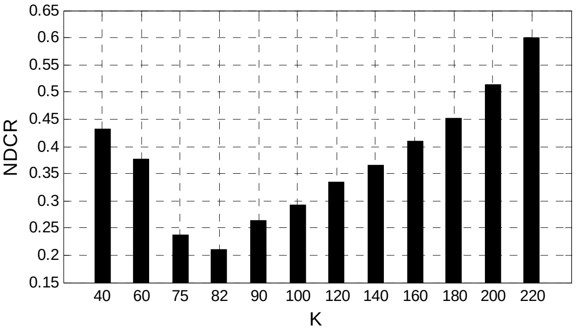

Figure 5. Detection performance under different number of clusters.

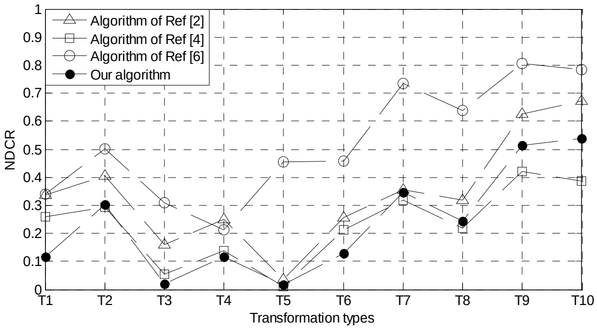

Figure 6. Comparison of NDCR Among the 10 kinds of the various algorithms for copies transformation.

As is shown in Table I, in the case of detection results are similar, our method is most efficient, especially on the matching time. Point trajectory asynchronous match strategy is adopted by [2]. It only need extract local feature from keyframes(every 29 frames to extract one). The feature extraction speed of [2] is very fast, but it require to match a large number of feature points with large-scale (tens of millions) video database's trajectories. Although the index structure can speed up the match process, time-consuming is still serious. Reference [4] uses the current frame feature sets to match adjacent 15 frames feature sets to build trajectories. It firstly matches the feature points, then confirm the result by trajectory coordinates. Thus feature extraction and matching both consume a lot of time. Reference [6] uses Harris and KLT algorithm to quickly build the trajectory. It divided lens into sub-lens basing on the difference between current frame and the first frame of the lens. But the number of sub-lens is large, and it needs integrate the matching result of sub-lens to determine the final detection results, and the matching is time-consuming. But [6] improves the match efficiency by index, and the detection speed can satisfy real-time requirement. In this paper, the SURF feature set and the trajectory set of current frame have been used to match with adjacent 15 frames feature set that unjoin in trajectories to quickly build the trajectory. Divide lens basing on the difference of the current frame and the contents of the entire lens. When matching, hierarchical policy has been used to avoid massive SURF features matching, and this effectively reduces the match complexity, and improves the detection efficiency. Whether the query video is reference video copy or not can be determined according to the test results of a single lens. This is why our algorithm is less time-consuming when matching.

VI. Using the Template

In this paper, we propose the video copy detection method integrated spatio-temporal features on previous work. The method utilizes the fusion of spatio-temporal features for analysis, and extracts word frequency vector and SURF feature set with the label of dynamic behavior to describe each lens. It also simplifies the complexity of matching by classifying match strategy and improves the detection efficiency. The experimental results show that the method dramatically improves the detection while ensuring detection effect. The next step of the research focuses on how to improve the speed of building trajectories and copy video locating accuracy.

Acknowledgment

This paper is originated from a project numbered as 2011AA010603, which is supported by national 863 foundations.

References

[1] Guidelines for the TRECVID 2011CD task Evaluation[OL]. http://www-nlpir.nist.gov/projects/tv2011/tv2011.html.

[2] Law-To J, Buisson O, Gouet-Brunet V, et al. Robust Voting Algorithm based on Labels of Behavior for Video Copy Detection[C]. In ACM Multimedia, 2006.

[3] Law-To J, Li Chen, Alexis Joly, et al. Video Copy Detection: A Comparative Study[C]. In Proc of ACM International Conference on Image and Video Retrieval, 2007.371-378.

[4] Shi Chen, Jinqiao Wang, Yi Ouyang, et al. Multi-Level Trajectory Modeling for Video Copy Detection[C]. In Proc of IEEE International Conference on Acoustics Speech and Signal Processing, 2010, 2378-2381.

[5] Guo Junbo, Li Jintao, Zhang Yongdong, et al. Video Copy Detection Based on Trajectory Behavior Pattern [J]. Journal of computer-aided design & computer graphics, 2010,22, 943-948.

[6] Wu Xiao, Li Jintao, Tang Sheng, Guo Junbo. Video Copy Detection Based on Spatio-Temporal Trajectory Behavior Feature [J]. Journal of computer research and development, 2010, 47, 1871-1877.

[7] The origin of the video database is MUSCLE-VCD-2007[EB/OL]. http://wwwrocq.inria.fr/imedia/civr-bench/index.html.

[8] Bay H, Ess A, Gool L.,etc. SURF: Speeded Up Robust Features[J]. Computer Vision and Image Understanding, 110, 346-359.

[9] Bauer .J, Sünderhauf .N, Protzel .P. Comparing Several Implementations of Two Recently Published Feature Detectors[C]. In Proc. of the International Conference on Intelligent and Autonomous Systems, Toulouse, France, 2007.

[10] Luo Juan, Oubong Gwun. A Comparison of SIFT, PCA-SIFT and SURF[J]. International Journal of Image Processing,3,143-152.

[11] Frey .J, Dueck .D. Clustering by Passing Messages between Data Points[J]. Science, 2007, 315,972-976.

[12] Wang Kaijun, Zheng Jie. Fast Algorithm of Affinity Propagation Clustering under Given Number of Clusters [J]. applications of the computer systems, 2010, 19, 207-209