Dynamic Segmentation Of Videos Based on Spatio-Temporal Pyramid Matching In Large Scale Video Retrieval System

R.Nalini[1], C.Senthil Kumar[2]

[1]*(Department of Information Technology, SNS College of Technology, India)*

[2]*(Department of Information Technology, SNS College of Technology, India)*

Abstract —**Tremendous growth in the field of multimedia technology has headed to large and detailed multimedia databases. Broadcasting of digital video content on different media brings the search of copies in large video databases to a new critical issue. Content Based Copy Detection (CBCD) presents an alternative to the watermarking approach to identify video sequences and to solve this challenge.Multimedia mining deals with the extraction of implicit knowledge and patterns not explicitly stored in multimedia files. The amount of redundant and duplicate videos present in Youtube, Yahoo, Google and other video search engines is about 27%.Segmentation and graph-based video sequence matching method is used for video copy detection by using SIFT descriptor for video content description. An exhaustive search method will involve high computational cost to detect all possible video copies of various lengths and locations. To resolve this problem, a graph based video sequence matching method is used. The Spatio Temporal Pyramid Matching with optical flow (STPM+OP) method provides the spatio temporal features of the frames that is being collected in the video repository. The Dense SIFT descriptor has been applied at dense grid to perform tasks such as object categorization, image alignment and texture classification more efficiently.**

*Keywords:***Content Based Copy Detection, Dense SIFT, SIFT Descriptor, Video copy detection**.

I. INTRODUCTION

Rapid progress in the field of multimedia technology and storage has led to the fast growing tremendous and large amount of data stored in databases. The essential information may be hidden behind those data and it may be difficult for human beings to extract them without powerful tools. Multimedia mining can automatically extract semantically meaningful information from the multimedia files. There is lot of demand for providing semantic information that are hidden behind those multimedia files, hence a large number of techniques have been proposed ranging from simple measures to more sophisticated systems. The Simple measures ranges from color histogram for image, energy estimates for audio signal to sophisticated systems like speaker emotion recognition in audio, automatic summarization of television programs. The wide development in the multimedia hardware and software technologies, video data collection and creation has become large in numbers. Every day hundreds of video data are generated and published. In these huge volumes of videos, there exist large numbers of copies or nearly-duplicate videos. The statistics of videos show that there are about 27 percent of redundant videos that are duplicate or nearly duplicate present in the search results from YouTube, Google video and Yahoo video search engines. There are two kinds of feature in multimedia mining: description based and content based. The video copy detection aims to determine whether a query video segment is a copy of a video from the video data set. An effective and efficient method for video copy detection has become more important to solve the problem.

Multimedia database systems collects, store and manage a large volumes of multimedia objects, such as video, image, hypertext and audio data. In general, the multimedia files from a database needs to be preprocessed to improve their quality. Later, these multimedia files undergo various transformations and features extraction to generate the important features from the multimedia files.

Using the generated features, mining can be carried out using data mining techniques to discover significant patterns. These resulting patterns are then evaluated and interpreted in order to obtain the final application's knowledge. The videos can be duplicated by performing number of transformations as defined in TRECVID 2008[2].

Some transformations that are made in the videos are picture in picture, brightness (change Increase brightness by 15% -25%), noise addition(Adding 15% random noise), rotation(Rotating up to 90), blurring (Blurring by 20%), horizontal flip (Horizontal mirroring up to 90),vertical flip (Vertical mirroring up to 100), color change(Changing color spectrum), pattern insertion (Pattern is inserted into selective frames), moving caption insertion (Entire video includes moving caption), slow motion (Half the video speed), fast forward (Double the video speed), zooming in, etc. Some transformation such as picture in picture is hard to detect. Hence such video copies that have undergone transformations could be detected by considering the local features by using Scale Invariant Feature Transform (SIFT) and also by using Dense SIFT (DSIFT).In this paper, we focus on detecting transformations and propose a dual-threshold segmentation,extracting feature set matching and also graph-based sequence matching method.

II. RELATED WORK

According to the comparative study of video copy detection there have been huge number of videos uploaded everyday which is a serious challenge to the owners of video web servers [3].The paper deals about Content Based Copy Detection (CBCD) approach to identify video sequence to solve the above challenge. The copy detection is done by using global descriptors and local descriptors. Transformations results are compared based on single transformations and random transformation mixed which results in choosing Ordinal Temporal measure is an efficient for small transformations. J.Jeon has reviewed that image annotation and retrieval can be done automatic using Cross media relevance models (CMRM). The annotation of images can be done by models such Probablistic annotation based cross relevance model (PACMRM), fixed annotation based cross relevance model (FACMRM).For annotation of images the cross relevance model is a good choice.

Kristen Grauman [4] has proposed kernel based classification method for learning complex decision boundaries. The proposed method addresses all issues by Kernel based learning algorithms. Pyramid matching kernel is the proposed method that converts input sets into multi-resolution histograms. The idea of the proposed method is to map the sets of features into multi-resolution histogram and then compared with weighted histogram intersection measure in order to approximate the similarity of the best partial matching between feature sets [5].Beyond bags of features paper presents the method to recognize the natural scene categories based on the global geometric correspondence. And the scene is represented as 'spatial pyramid' which is a computationally efficient extension of an order less bag of features image representation. The image features are considered by Torralba's "gist" and Lowe's SIFT descriptors. Lazebnik presents a "holistic" approach for categorization of image which repeatedly sub divides an image and computing histograms of image features.[6]

III. EXISTING SYSTEM

In existing system the copy detection could be analysed by using local features of SIFT. Visual information of the video frames are temporally redundant. So, video sequence matching is not necessary to be carried out using all the video frames. The way to reduce unnecessary matching is to extract certain keyframes to represent the video content and the matching of two video sequences can be first performed by matching the keyframes.

The clustering of video frames by considering the similarity between neighboring frames and can choose a keyframe from each cluster to represent it .However, the extracted keyframes cannot represent the temporal information among frames. Some methods was proposed to detect video shots and extract keyframes from each shot to represent the video content . Matching two video sequences based on extracted keyframes from the segments can meet the requirement of two videos being in different frame rates.

A. Auto Dual-Threshold Method

In the Existing system, an auto dual-threshold method was used to eliminate redundant video frames. The method continuously cuts down the video frames into video segments by eliminating temporal redundancy of the visual information of continuous video frames. The method has two characteristics:

First, two thresholds are used. One threshold is used for detecting abrupt changes of visual information of frames and another for gradual changes.

Second, the values of two thresholds are determined adaptively according to video content. Auto dual threshold method eliminates the near-duplicate frames along the video time direction and does not take into account the concept of the shot, also does not require post processing to obtain the actual shots .

B.Single Value Decomposition

The singular value decomposition (SVD) is a powerful linear algebraic technique.The SVD posses geometric structure of a matrix which is an important aspect of matrix calculations. The SVD is used in many applications based on its key properties. It reflects the relation to the rank of a matrix and its ability to approximate matrices of a given rank.

The SVD method has been used widely in data compression, signal processing and pattern recognition.SVD method is used to match SIFT features Point sets features, by measuring the similarity between two SIFT feature point sets, and emphasize the similarity of "frame-to-frame."

The SVD theorem matrix can be described as follows:
If A € $R^{m \times n}$ (based m > n), ran (A) = r, then there exists two orthogonal matrices U, V and a diagonal matrix makes the establishment of the following equation:

$$A = U\text{⍰}T$$

Where
$U = [u_1, u_2, u_3, . . . ,u_m]$ € $R^{m\times m}$, $UU^T = I$;
$V = [v_1, v_2, v_3,. . . , v_n]$ € $R^{n\times n}$, $VV^T = I$;
⍰= $[\lambda_1, \lambda_2, ..., \lambda r,0,.....,0]$ € $R^{m\times n}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

The proposed single value decomposition method consists of three steps.

Step 1: Matrix A $^{N\times m}$ = (A_1, A_2,…..,$A_{m)}$ represents the feature point set of image A and matrix B $^{N\times n}$= (B_1,B_2,...,B_n) represents the feature points set of image B, respectively. In another word, A_i (i=1,….m) and B_j(i=1,…,n) represent SIFT feature points in image A and B, respectively. The dimension of A_i and B_j is N (N = 128).

Step 2: A d-dimensional linear subspace of A and B is represented by an orthonormal basis matrix P_A € $A^{N\times d}$and P_B € $B^{N\times d}$, respectively, s.t. $AA^T \cong P_A \text{⍰}_A P^T{}_A$ and $BB^T \cong P_B\ \text{⍰}_B P^T{}_B$, where ⍰$_A$ and ⍰$_B$ are the eigen value dialog matrices of the *d* largest eigen values, P_A and P_B the values of eigen vector matrices of the *d* largest eigen values.

Step 3: Make the singular value decomposition for $P^T{}_A P_B$ € $R^{d\times d}$, i.e., $P^T{}_A P_B = USV^T$, so the similarity between A and B is *sim*(A, B) = *trace*(S).

The proposed SVD-based SIFT feature point set matching method can obtain a better tradeoff between the detection effectiveness and detection time cost.

C. Video Sequence Matching Method Using Graph

The steps involved in the sequence matching method are as follows
Step 1: Segmentation of the video frames and extract features of the keyframes. By dual-threshold method the videos sequences are segmented, in addition to this extraction of SIFT features of the keyframes is done.

Step 2: Match the query video and target video. The contents and the bag of features are compared such that redundant frames could be found.

Step 3: Generate the matching result graph according to the matching results. To determine whether there exists an edge between two vertexes, two measures are evaluated.

- Time direction consistency
- Time jump degree

Step 4: Search for the longest path in the matching result graph. The method searches the longest path between two arbitrary vertexes in the matching result graph.These longest paths can determine not only the location of the video copies but also the time length of the video copies.

Step 5: Output the result of detection. Each vertex of the matching resulting graph, has more than one path or no path.

The graph based video sequence matching method results based on visual features of the video frames. The goal of the graph-based video sequence matching method is to refine and order the segment matching results by incorporating the temporal information. The graph based video sequence matching method can automatically find optimal sequence matching result and also automatically remove the noise caused by visual feature matching.

The advantage of the graph based method is that it can detect multiple copies existed in the detected video and it is adaptive to changing video frame rate. In real application, the query video is normally short. But when the query is large video clip query then the it is slightly perform the task for large those thousands of frames.

Drawbacks

- An exhaustive search method will involve high computational cost to detect all possible video copies with various lengths and locations.
- Image alignment is more complicated in case where the dynamic video sequenced in an orderly manner. Incompleteness of object sequence is also a problem that many correlations will be missing
- Semantic analysis techniques are based on this similarity, it sometimes fails.

IV. PROPOSED SYSTEM

In the proposed system the drawbacks of the exhaustive search method will reduce the computational cost of detecting the copied videos of various length.

A. Spatio Temporal Pyramid Matching (SIFT+OP)

The video retrieval can be performed efficiently from a large video database by matching the video contents by using Spatial temporal pyramid matching (STPM) [2]. STPM method recursively divides a video clip into a 3D spatial-temporal pyramidal space and compares the features in different resolutions. With the intention of improving the retrieval performance, both static and dynamic features of objects are considered.

By providing a sufficient condition in which the matching can get the additional benefit from temporal information. It has been proven experimentally that STPM performs better than the other video matching.

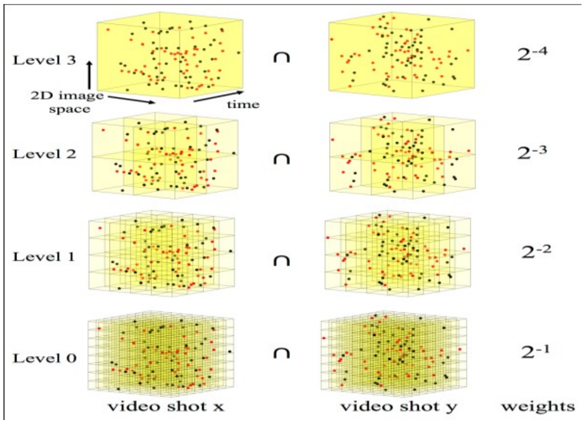
STPM kernel is an extension of Spatial Pyramid Matching (SPM), which is personalized to the 2D image matching. For STPM, we hold the time dimension in a sequence of video frames in a video clip to build a 3D pyramid kernel, which is obviously appropriate for video matching. First, an STPM kernel is built by constructing a (L + 1)-level pyramid so that the input video clip sits on level L. The Fig 4.1 is an example for L =3.

From the coarsest(top) to the finest (bottom) of the pyramid, the 2D space and 1D time dimensions are recursively divided into a half. Each separated volume element or voxel corresponds to a bag of features and significantly the top-level voxel corresponds to a bag of features of a given video clip.

To compare the similarity of two video clips in the same region, STPM represents the set of features in each video as a histogram, in which each bin corresponds to a specific feature, and the height of the bin represents the occurrence of the feature.

We refer $H^l r$ (X) to the histogram of video X in region r at level l.

ISSN: 2278 – 7798



. Fig.4.1 Hierarchical structure of spatio-temporal pyramid matching.

One of the main operations in our video matching is the histogram intersection, which is :

$$H^l_r(X) \cap H^l_r(Y) = \sum_{f€F} \min(H^l_r(X)\,[f],H^l_r(Y)[f]$$

where F is the set of features (i.e., the set of all bins), and $H^l_r(\cdot)[f]$ is the height of the histogram $H^l_r(\cdot)$ at feature *f*. Fig. 4.2illustrates three histograms extracted in different levels of the pyramid.Each histogram has counts for features, e.g., diamond and circle in the example. Note that, in each level, a histogram will be calculated for each cubic voxel. Then, the matching score of video X and Y at level l, Cl(X, Y) is defined as: where Rl is the set of regions at level l.

B .Weight Assignment On Features

For comparing the video clips which contain persistently moving objects, we choose two features, Shift Invariant Feature Transform (SIFT) and optical flow. SIFT has been widely used to match two still imagesmovement of objects in the frames. Thus, we also use the optical flow to capture the movement of objects in the sequence of frames in a video clip.To calculate the similarity of two video clips, we first calculate two STPMs, $K_{SIFT}$for SIFTs and $K_{OP}$for optical flows.

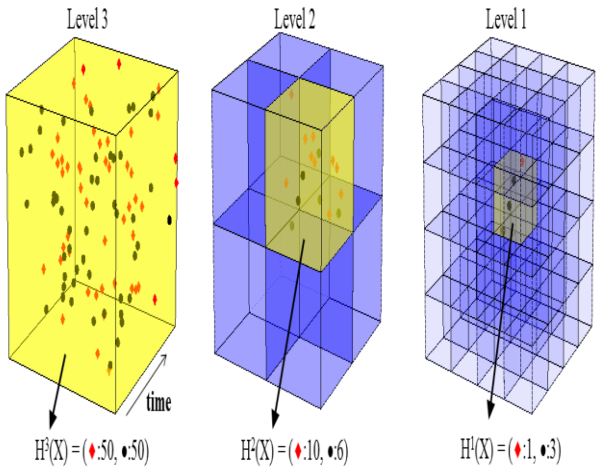


Fig.4.2 Three examples of feature histograms in different levels of spatio-temporal pyramid matching

Then, we use a linear sum of two kernels as follows:

$K_{STPM} = W_{SIFT} * K_{SIFT}(X,Y)+W_{OP} *K_{OP}(X,Y)$ where$W_{SIFT}$and $W_{OP}$are tunable parameters and $W_{SIFT}+ W_{OP}= 1$.

C.Shot Similarity Matching

When we compare two video clips, typically they have different lengths in time,so selecting the same number ofrepresentativevideo frames from the two videos unlikely happens for video matching. In order to construct a uniform pyramidal structure in temporal domain from a video clip, we extract *m* frames out of the video clips, where m equals 2^l. Thus, a video clip is represented as cubes in a $2^l \times 2^l$ spatial grids and 2^l temporal intervals. Each volume element in the cube includes a set of features, which is represented by a histogram of feature bins.

D. Dense SIFT (DSIFT)

The additional feature added to match the frames is Dense SIFT (DSIFT).Scale invariant feature transform often finds stable scales in only a few image pixels.

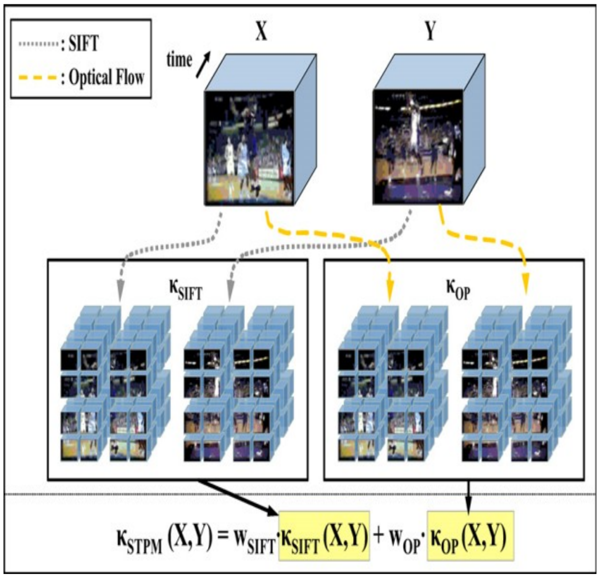

: SIFT
: Optical Flow
time
X
Y
κ_{SIFT}
κ_{OP}
$\kappa_{STPM}(X,Y) = w_{SIFT} \cdot \kappa_{SIFT}(X,Y) + w_{OP} \cdot \kappa_{OP}(X,Y)$

Fig.4.3 Video matching scheme for spatio-temporal pyramid matching with 3 levels (L=2).

As a result, methods for feature matching it characteristically choose one of two extreme option: By matching a sparse set of scale invariant features, or dense matching using arbitrary scales.Applying the SIFT descriptor to tasks such as object category classification or scene classification, experimental evaluations show that betterclassification results are often obtained by computing the SIFT descriptor over dense grids in the image domain as contrasting to sparse interest points as obtained by an significant operator. A basic justification for this is that a larger set of local image descriptors computed over a dense grid usually provide more information than subsequent descriptors evaluated at a much sparser set of image points.

Dense SIFT when applied to object categorization tasks in practice, the computation of dense SIFT descriptorsis usually accompanied with a clustering stage,where the individual SIFT descriptors are reduced to a smaller vocabulary of visual words, which can then be combined with abag-of-wordsmodel or related methods. For the task of establishing image correspondences between initially unrelated different images of a 3D object or a 3D scene, the detection of sparse interest points is, however, still important an important pre-processing step to keep down the complexitywhen establishing image correspondences.

E.Motion matching

Content based copy detection schemes extract signatures from the original media. The same signatures are extracted from the test media stream and compared to the original media signature to determine if the test stream contains a copy of the original media. The significant advantage of content-based copy detection over watermarking is the fact that the signature extraction can be done after the media has been distribute. Themotion-based signatureexploits only the changed facts in the video, the ordinal signatureis a function of both the color (intensity) and spatial properties of the video, however thecolor signatureuses only the color properties without using the spatial information.

The following is the experimental procedure used for testing the signature matching.
1. Extract signature from the reference video (R).
2. Extract signature from the test video (V).
3. Set test clip length = *L*.
4. Select a random point (*P*) in test video V.
5. Select a clip (*C*) of length *L* around *P*.
6. Find the best match location *Ml* of *C* against *R* and match score Ms
7. Repeat Steps 4-6, 100 times.
8. Repeat Steps 3-7, for different clip lengths *L*.

F. Ranking Scheme

Multimodal and multilevel ranking approach is used for retrieving best matching video from the video database. First of all, we show how to represent video configurations by graphs, to which a diversity of graph based learning techniques can be applied to solve the video ranking problem. The proposed high-level ranking architecture, implements our multimodal and multilevel rankingframework.

Multilevel framework consists of four ranking stages:
1) Text-based Ranking.
2) Nearest Neighbour Re ranking.
3) Large Margin Supervised Reranking.
4) Multimodal Semi-Supervised Reranking.

The proposed framework not only achieves considerably better retrieval performance than traditional approaches, but also is practically efficient for large-scale applications.

V. CONCLUSION

In this paper, we address the problem of detecting the copied video clips for content-based video query. In future the clip boundaries can be found using a strong classifier from a boosting algorithm on peak of weak classifiers. Then, the similarity of video clips is calculated by our spatio-temporal pyramid matching kernel which includes temporal dimension into the matching schema. By using a personal computer for computation, our proposed approach can take only tens of milliseconds to find the relevant videos present in the data base.

REFERENCES

[1] Hong Liu, Hong Lu, Xiangyang Xue, 'A Segmentation And Graph-Based Video Sequence Matching Method For Video Copy Detection', IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8,Pp.1706-1718,August 2013

[2] Jaesik Choi , Ziyu Wang , Sang- Chul Lee , Won J. Jeon (2013) 'A Spatio- Temporal Pyramid Matching For Video Retrieval', Journal on Computer Vision and Image Understanding,Vol.117, pp.660-669

[3] J. Law-To, C. Li, and A. Joly, "Video Copy Detection: A Comparative Study," Proc. ACM Int'l Conf. Image and Video Retrieval, pp. 371-378, July 2007.

[4] J.Jeon, V. Lavrenko and R. Manmatha(2003) 'Automatic image annotation and retrieval using cross media relevance models', IEEE trans. on pattern analysis and machine intelligence.

[5] Grauman, K.Darrell (2005),' The Pyramid Match Kernel: Discriminative Classification With Sets Of Image Features',IEEE International Conference onComputer Vision,Vol.2

[6] Lazebnik, S.Schmid, C. Ponce (2006),'Beyond Bags Of Features: Spatial Pyramid Matching For Recognizing Natural Scene Categories', IEEE Computer Society Conference on Computer Vision and Pattern Recognition,Vol.2

[7]Steven C. H. Hoi, and Michael R. Lyu (2008),'A Multimodal and Multilevel Ranking Scheme for Large Scale Video Retrieval',IEEE transactions on multimedia, VOL. 10, NO. 4

[8]MatthijsDouze, HerveJegou and CordeliaSchmid(2010),'An Image-based Approach To Video Copy Detection With Spatio - Temporal Post-filtering',IEEE Transactions on Multimedia,VOL.12,NO.

[9] ArunHampapur,Ki-Ho Hyun and RundBolle,'Comparison of sequence matching techniques for video copy detection',IEEE Conf. on multimedia computing and systems,1997