Content Based Video Retrieval using SURF Descriptor

Asha S
M.Tech in Computer Science and Information Systems
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT)
Angamaly, India

Sreeraj M
Assistant Professor
Department of Computer Science and Engineering
Federal Institute of Science and Technology (FISAT)
Angamaly, India

***Abstract*— This paper presents a Robust Content Based Video Retrieval (CBVR) system. This system retrieves similar videos based on a local feature descriptor called SURF (Speeded Up Robust Feature). The higher dimensionality of SURF like feature descriptors causes huge storage consumption during indexing of video information. To achieve a dimensionality reduction on the SURF feature descriptor, this system employs a stochastic dimensionality reduction method and thus provides a model data for the videos. On retrieval, the model data of the test clip is classified to its similar videos using a minimum distance classifier. The performance of this system is evaluated using two different minimum distance classifiers during the retrieval stage. The experimental analyses performed on the system shows that the system has a retrieval performance of 78%. This system also analyses the performance efficiency of the low dimensional SURF descriptor.**

***Keywords* - Shot Boundary Detection; Feature Extraction; SURF Descriptor; Fuzzy K Nearest Neighbor.**

I. INTRODUCTION

Now a days, video is one of the most commonly used way of information exchange. A large amount of digital video is now available publicly in an ever increasing speed. Without a proper video retrieval mechanism, it becomes tiresome for the users to retrieve the video content of their interest. This led to the development of an automatic and effective mechanism for video retrieval.

A complete video data management system consists of efficient and effective methods for the storage, indexing and retrieval of videos. Videos must be kept in storage in standard video formats with suitable indexing mechanism. Indexing plays a crucial role in video retrieval. Out of the existing methods for indexing, content based approaches appear to be more efficient.

The content based approach for video retrieval focuses on retrieving similar videos based on the video contents. The content of a video can be represented using either global features or local features. In this paper a CBVR system that uses SURF feature descriptor is presented. Due to the high dimensionality of SURF a stochastic reduction method is applied on the extracted SURF feature descriptor and generates a model data. During video retrieval, the model data of the query clip is compared with that of the stored database clips using a minimum distance classifier. Based on this evaluation, videos in the library are ranked. From this list top four similar videos are retrieved.

This paper is organized as follows. In section II the related works are briefed. Section III introduces the methodology used in the proposed system. Section IV explains the algorithms used in the implementation of different modules. The experimental results and evaluation of the system is given in Section V. Section VI concludes this system by briefly discussing the future work.

II. RELATED WORKS

As reviewed in [1], many content-based video retrieval methods have been proposed. One type of video retrieval methods uses global descriptor. The global descriptors can be color-based, texture-based, or shape-based. In [2], Amir et al. computed color histogram and color moments for video retrieval. Shen et al. [3] introduced a real-time video retrieval system, UQLIPS, which globally summarized each video to a single vector. Another type of methods is based on local descriptors. Among the various local feature descriptor, SURF has proven to be stable, reliable and has high efficiency in information retrieval. S. Huang, C. Cai, F. Zhao in [4], used SURF feature descriptor for extracting the contents of images for wooden image retrieval.

Various methods are proposed for automatic detection of shot boundaries. The simplest approach is to compare the histogram value of consecutive frames. To detect both abrupt and gradual transitions, the twin comparison algorithm proposed by Zhang et al. [5] can be used, which uses a dual threshold approach.

In [1] various key frame extraction methods are presented. From the various proposals made on key frame identification, one of the blind conclusions arrived is to extract the first and last frames in a shot as key frames.

Another major task of the video retrieval deals with finding similar clips. Many approaches like Bag of Features (BoFs) [6] or locality sensitive hashing (LSH) [7], hierarchical indexing structure for efficient video retrieval [8], etc., has been proposed for efficient video search.

In this CBVR system, an auto dual threshold algorithm [5] is used for shot boundary detection. The feature extracted is the SURF descriptor. For feature comparison, the concept of Fuzzy K Nearest neighbor algorithm [10] is adopted that helped in finding K Nearest Video clips.

III. METHODOLOGY

This system has the generic architecture as shown in Figure 1

978-0-7695-5033-6/13 $26.00 © 2013 IEEE
DOI 10.1109/ICACC.2013.49

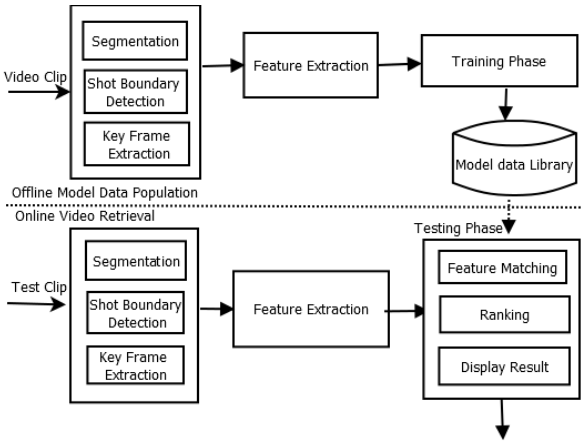

Figure 1. The Proposed System Architecture

Here, during the offline model data vector population stage, the input video clip initially undergoes a preprocessing phase, which includes Segmentation, Shot Boundary Detection and Key Frame extraction modules. During this preprocessing stage, the input video gets converted into a set of key frames. From this identified keyframes, SURF feature descriptor is extracted. This descriptor is then passed into a training phase, where the feature descriptor undergoes a stochastic dimensionality reduction procedure and gets trained into a standard model vector. The model vector is then is then uploaded in the model data library.

On video retrieval, for a given test clip, its model vector is computed and is classified using a minimum distance classifier. During the event of classification, the videos in the model data library are ranked based on its similarity to the test clip. From the list of similar videos the highest ranking videos are retrieved.

IV. IMPLEMENTATION

A generic CBVR system consists of five phases: Segmentation, Shot Boundary Detection, Key Frame Extraction, Feature Extraction and Feature Matching. The algorithm used in each of the five phases is described in the following subsections.

A. Video Segmentation

Video segmentation is the process of segmenting a video into frames. Segmentation of video can be Temporal, Spatial, or Spatio-temporal.

In this section, for a given input video **V** with **n** number of frames, represented as $V = f(Ii, t)$, where $i = 1,2 \ldots n$; temporal video segmentation is applied and get converted into its elementary parts $I_1, I_2 \ldots In$.

B. Shot Boundary Detection

Shot Boundary detection is the process of automatically detecting the boundaries between consecutive shots i.e. identifying where a new transition begins and where it ends. Hence a shot can be defined as a group of consecutive frames that are temporally and visually close to each other.

A shot S consisting of closely related frames can be represented as $S = g(I_k, t), k = i, i+1 \ldots j$, *where* $1 \le i < j \le n$ *and* $I_k \quad V$.

In this system, each shot boundaries *Sj* are automatically identified using an auto dual threshold approach. The algorithm uses a high threshold T_b for finding hard cuts. The starting frame of gradual transition is determined using a lower threshold T_s. From the starting frame, the histogram differences are accumulated. End of the gradual transition is determined if the accumulated difference goes beyond the upper threshold T_b.

C. Key Frame Extraction

Key frames are frames that can characterize the entire frames in a shot. As one or two frames selected from a shot could represent the salient content of the shot, it reduces the amount of information needed to be stored for a video during indexing, storage and retrieval.

This system employs the strategy of extracting the first and last frame in a shot as key frames. Here, from each shot *S* in the given video *V*, we take *Ii* and *Ij*. Hence for a video *V* with *k* shots, there will be $2 * k$ key frames.

D. Feature Extraction

The process of transforming an input frame to a vector that represents the frame's content is referred to as feature extraction. Frame content can be represented efficiently and more distinctively using local feature descriptors [11].

In this system, the SURF [9] feature descriptor is extracted from the key frames. SURF is a robust local feature descriptor. It detects landmark points in an image called interest points, and describe these points by a vector which is robust against rotation, scaling and noise. SURF feature descriptor is a low dimensional (64 dimensional) vector compared to other local feature descriptors, which aids matching algorithms to be performed faster.

The SURF descriptor of an image is a 64 x M vector, where 64 is the number of features; M is the number of interest points identified for that image. So the descriptor for a video will be 64 x M x Q Vector, where Q is the number of key frames.

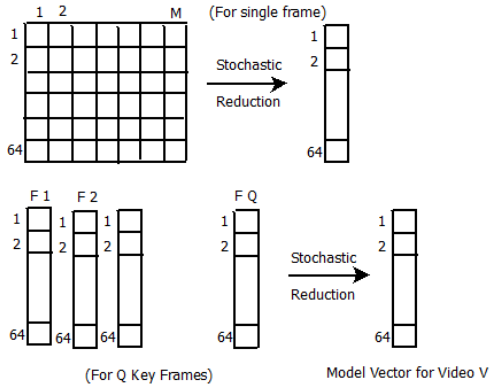

Figure 2. Stochastic Reduction Technique

This high dimensional descriptor is then converted to a more readable, distinct and compact final knowledge vector. This dimensionality reduction is achieved using a stochastic distribution function and then taking its likelihood estimate. Figure 2 depicts the stochastic reduction procedure. After applying the stochastic reduction method, for each key frame a 64 x 1 vector is obtained. Hence, for a video with Q key frames, the feature descriptor will be a 64 x Q vector. On this vector, another stochastic reduction method is applied to get a knowledge row vector. This knowledge vector is then stored as a model data in model data library. The knowledge vector library file contains all the information about the video including video path name, video name, descriptor, preview, size of the video and number of frames.

E. Feature Comparison

Here, Fuzzy K Nearest Neighbor (FKNN) Algorithm [10] is used as the minimum distance classifier. In FKNN, in addition to classifying the query clip to a particular class; it assigns membership value, which indicates the extent to which the query can be classified to each video clip in the database. The Fuzzy K-Nearest Neighbor algorithm is as follows

- Initialize the value of K.
- Compute the distance between the model vectors to the test vector.
- Find the minimum distance K model vectors.
- Calculate the membership value of the test clip to each of the selected K model vector.

Based on this distance evaluation, the model data in the knowledge vector library is ranked. From this ranked list, top four videos are retrieved.

V. RESULT AND DISCUSSIONS

The experimental evaluation of the system is presented in this section. The system is evaluated under three scenarios. First, the performance of this system is evaluated using statistical measurements like precision, recall and F-measure. Second, the retrieval efficiency of the system is compared using different combinations of distance algorithms during feature classification. Third, the effectiveness of the proposed low dimensional SURF descriptor is evaluated.

For computational accuracy TRECVID 2011 dataset is chosen as the bench mark. For that purpose, the database is populated with one hundred of TRECVID dataset videos which forms the reference data set. The size of the videos ranges from 200 KB to 1000 KB. The duration of these videos is about 5 seconds to 10 seconds. The reference dataset is populated with clips belonging to different categories like sports, cartoon, nature, news and movie clips.

From the training set consisting of 100 videos a number of test clips are constructed. The test set consists of 40 random clips belonging to all groups. The query clips contains relevant and irrelevant videos. Thus the system is tested using clips belonging to both training set and test set. Table 1 shows the performance of the system considering six sample test clips belonging to different categories. The Precision and Recall curve in Figure 4 illustrate the performance of this CBVR system.

TABLE I. PERFORMANCE EVALUATION OF THE PROPOSED SYSTEM

Video	Metric	Recall	Precision	F-Measure
Video1	Euclidean	0.83	0.5	0.624
	Manhattan	0.83	0.625	0.713
Video2	Euclidean	0.8	0.8	0.8
	Manhattan	0.6	1	0.75
Video3	Euclidean	1	1	1
	Manhattan	0.4	1	0.57
Video4	Euclidean	1	0.85	0.918
	Manhattan	0.83	1	0.91
Video5	Euclidean	0.5	0.67	0.572
	Manhattan	0.25	1	0.4
Video6	Euclidean	0.75	1	0.857
	Manhattan	0.75	1	0.857

Next, the retrieval efficiency of the system is evaluated. Under this scenario, for each test clip, the performance is evaluated using two different minimum distance metrics during similarity measurement. The two main minimum distance metrics used are Euclidean Distance and Manhattan Distance. Figure 3, plots the average performance of using different distance metric in similarity measurement. It shows that both the distance metrics provides a reasonable retrieval rate. However, the system provides better statistical measurements when using Euclidean distance during the testing phase.

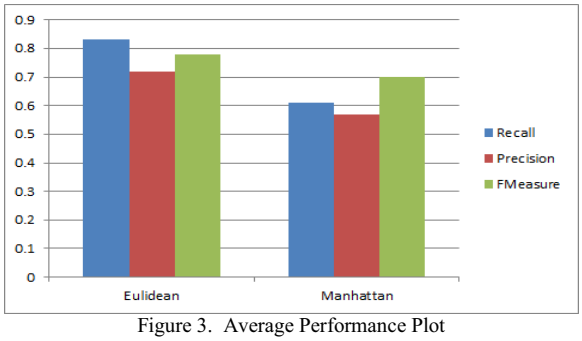

Figure 3. Average Performance Plot

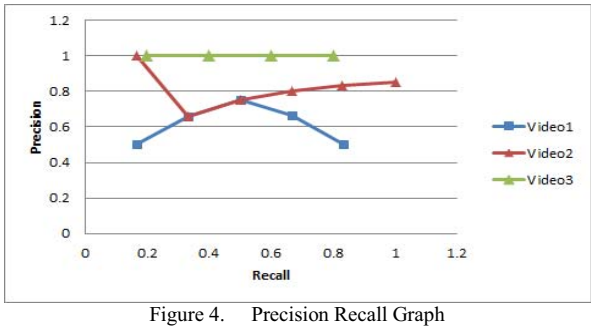

Figure 4. Precision Recall Graph

The system is also compared for its effectiveness of using distance evaluation classifier for similarity measurement. The Table II shows this comparison.

TABLE II. PERFROMANCE EVALUATION OF DISTANCE CLASSIFIER

Distance classifier	Metric	Recall	Precision
Yes	Euclidean	0.83	0.7
	Manhattan	0.61	0.6
No	Euclidean	0.75	0.75
	Manhattan	0.4	1

Third, the system is compared with other system for evaluating the efficiency of using stochastic reduction method in Feature vector processing. Table III depicts the performance of using stochastic reduction method. The table III shows that the recall rate of the system using low dimensional SURF descriptor founds to be better compared to a 64 dimensional SURF feature descriptor. However, the low dimensionality reduces the distinct nature of feature descriptors; the precision rate degrades when using the stochastic reduction approach.

TABLE III. PERFROMANCE EVALUATION OF STOCHASTIC REDUCTION METHOD

Stochastic Reduction Method	Metric	Recall	Precision
Yes	Euclidean	0.83	0.7
	Manhattan	0.61	0.6
No	Euclidean	0.86	1
	Manhattan	0.74	1

Another metric used for performance evaluation is the Receiver Operating Characteristic (ROC) curve.

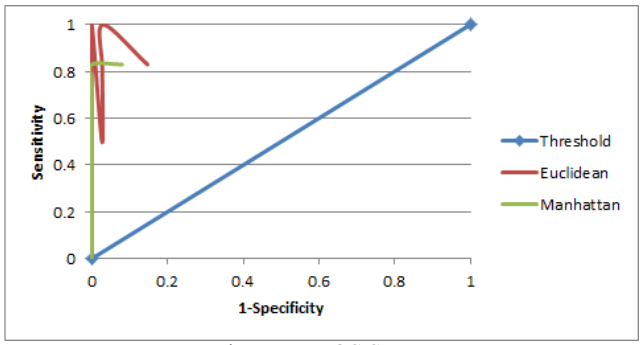

Figure 5. ROC Curve

From the various analysis performed in this section concludes that this system using a low dimensional SURF feature descriptor gives satisfactory performance in the retrieval of clips belonging to all groups.

VI. CONCLUSION AND FUTURE SCOPE

A CBVR system that uses SURF descriptor which efficiently retrieves video clips similar to a given query clip is presented. The experimental analysis shows that the system provides an average precision of 75 % with 83 % recall. The proposed system is also analysed for evaluating its efficiency in using stochastic reduction method and distance classifier method. The comparison shows that the proposed system achieves precision of 70% with a recall of 83 % compared to other system. These experimental analyses conclude that the system is effective and accurately retrieves clips belonging to different categories.

In future, first, the presented system can be extended using any flip invariant feature descriptor. Second, shot boundary can be identified by considering various camera motions, which improves the accuracy of identified shots. Third, the proposed key frame extraction phase can be extended by implementing a suitable algorithm. Finally, the accuracy of the system can be enhanced by using relevance feedback.

REFERENCES

[1] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A Survey on Visual Content-Based Video Indexing and Retrieval", IEEE Trans. On Systems, Man, and Cybernetics, Parts C: Applications and Reviews, vol.41, no.6, pp.797 – 819, NOV. 2011

[2] A. Amir, W. Hsu, G. Iyengar, C. Y. Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang, "IBM research TRECVID-2003 video retrieval system," in Proc. TREC Video Retrieval Eval., Gaithersburg, MD, 2003. Available: http://wwwnlpir.nist.gov/projects/tvpubs/tvpapers03/ibm.smith.paper.final2.pdf.

[3] H. T. Shen, X. Zhou, Z. Huang, J. Shao, and X. Zhou, "Uqlips:A real-time near-duplicate video clip detection ystem," in VLDB, 2007, pp. 1374–1377.

[4] S. Huang, C. Cai, F. Zhao, "An Efficient Wood Image Retrieval using SURF Descriptor", Proc.International Conference on Test and Measurement, 2009, pp.55- 58, doi:978-1-4244-4700-8/09.

[5] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," Multimedia Syst., vol. 1, no. 1, pp. 10–28, Jun.

[6] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition.*. Berlin, germany:Springer, 2006, pp. 127–144.

[7] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in Proc. ACM Symp. Theory of Computing, 1998, pp. 604–613.

[8] H. Lu, B. C. Ooi, H. T. Shen, and X. Xue, "Hierarchical indexing structure for efficient similarity search in video retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 18,no. 11, pp. 1544–1559, 2006.

[9] Bay, H., Tuytelaars, T., and Van Gool, L. : SURF: Speeded Up Robust Features . In: 9th European Conference on Computer Vision, (2006).

[10] J. M Kellerr, M. R Gray, James A Givens, "A Fuzzy K-Nearest Neighbor Algorithm," IEEE Transactions on SYSTEMS,MAN and CYBERNETICS, VOL.SMC-15,No.4,JULY/AUGUST 1985

[11] J. Li and N. M. Allinson "A comprehensive review of current local features for computer vision", Neurocomputing, vol. 71, pp.1771 - 1787,2008