Content based Video Retrieval: A Survey

Dipika H Patel
Department of Computer Science and Engineering
L. J. Institute of Technology
Ahmedabad 382210, Gujarat, India

ABSTRACT
Videos are a powerful and communicative media that can capture and present information. The rapidly expanding digital video information has motivated growth of new technologies for effective browsing, annotating and retrieval of video data. Content-based video retrieval has attracted wide research during the last 10 years. Users are more diverted to content based search rather than text based search. These lead to the process of selecting, indexing and ranking the database according to the human visual perception. This paper reviews the recent research in content based video retrieval system. This survey focusing on video structure analysis, like, shot boundary detection and key frame extraction, different feature extraction methods including SIFT, SURF, etc, similarity measure, video indexing, and video browsing. This system retrieves similar videos based on local feature descriptor called SURF (Speeded-Up Robust Feature). For image convolution SURF relies on integral images. In SURF we use Hessian matrix-based measure for the detector and a distribution-based descriptor. SURF can be computed and compared much faster with respect to repeatability, uniqueness and robustness. SURF is better than previous proposed methods as SIFT, PCA-SIFT, GLOH, etc. Finally the future scope in this system is specified.

General Terms
Content based video retrieval, indexing, segmentation, feature extraction

Keywords
Video retrieval, feature extraction, SIFT, SURF, C-SURF, video browsing.

1. INTRODUCTION
Videos have the following characteristics: 1) much richer content than individual images; 2) huge amount of raw data; and 3) very little prior structure [10]. These characteristics make the indexing and retrieval of videos quite complex. In the earlier period, video databases have been relatively small, and indexing and retrieval have been based on keywords. Nowadays volume of these databases has become much larger and content-based indexing and retrieval is required.

1.1 Motivation
The main motivation factor inspiring to do survey of content based video retrieval system are as follow:

1) Without a proper video retrieval mechanism, it becomes annoying for the users to retrieve the video content of their interest.

2) Automated or semi-automated methods can save people's time and money.

3) For effectively and efficiently organizing those video data, video retrieval based on human visual perception and internet multimedia plays an important role in it.

Paper is organized as follows: in section 2, framework of content based video retrieval system is given. Section 3 comprises of methodology used for each step. Survey on content based video retrieval techniques is highlighted in section 4. Finally, future scope and work is concluded in section 5.

2. FRAMEWORK OF CONTENT BASED VIDEO RETRIEVAL
The content based video retrieval system is outline in Fig.1.

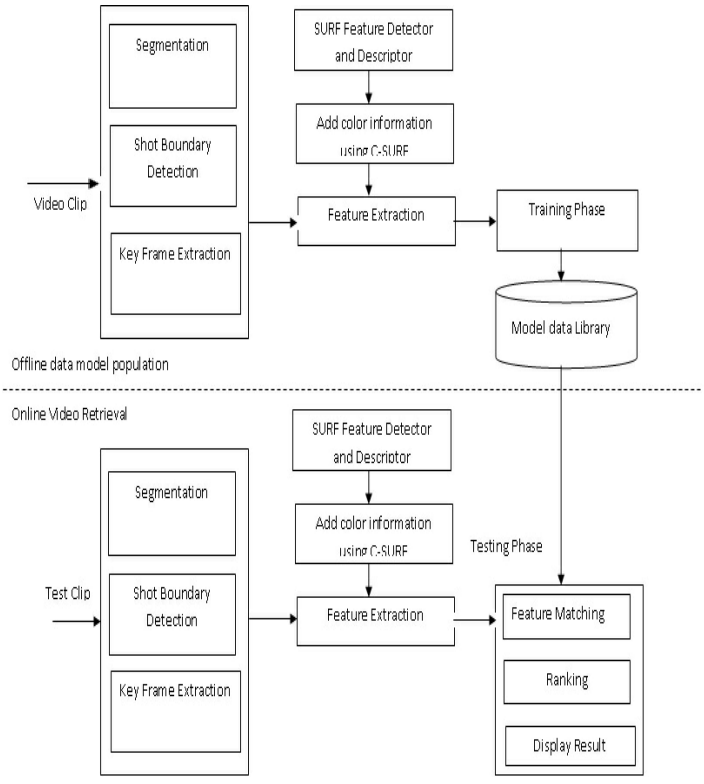

Fig 1: Generic framework for visual content-based video indexing and retrieval

The structure includes following:

1) Structure analysis:

I. Detect shot boundaries: A shot is defined as an image sequence that presents continuous action which is captured from a single operation of single camera. In the editing stage of video production, shots are joined together to form the complete sequence. Shots can be effectively considered as the smallest indexing unit where no changes in scene content can be perceived.

II. Extract key frames and segment scenes: Key-frames are still images extracted from original video data that best represent the content of shots in an abstract manner. Key-

frames, if extracted accurately, are a very important visual abstract of video contents and are very useful for fast video browsing. A video outline, such as a movie preview, is a set of selected segments from a long video program that highlight the video content, and it is best suited for sequential browsing of long video programs. Apart from browsing, key-frames can also be used in representing video in retrieval. Video index is constructed based on visual features of key-frames, and queries may be direct.

2) Feature extraction from segmented video units (shots or scenes): To extract features according to video structural analysis results is the base of video indexing and retrieval. These visual features are:

I. Static Features of key frames

i. color-based features

ii. texture-based features

iii. shape-based features

II. Object Features

III. Motion Features

i. Statistics-Based

ii. Trajectory-Based

iii. Objects' Relationship-Based

3) Query: the video database is searched for efficient video using index and similarity measures.

4) Video browsing and feedback: video found in response to the query by matching the feature between train dataset and test dataset and return to the user.

"Content-based" means that the search will be based on the actual content of the video. The term 'Content' here refer to the features such as color, shape, texture of the video. The content based approach focuses on the retrieval of videos by their similarity matching based on its video content. This content can be represented by either: global feature or local feature [4]. Global descriptors detail the overall content of the image but with no information about the spatial distribution of this content. Local descriptors relate to particular image regions and, in combination with geometric properties of these latter, express also the spatial arrangement of the content. Different local descriptor is as follow:

1) Scale Invariant Feature Transform (SIFT)

2) Speed-Up Robust Feature (SURF)

3) Histogram of Oriented Gradient (HOG)

4) Gradient Location Orientation Histogram (GLOH)

5) PCA-SIFT

SURF (Speeded Up Robust Feature) is the local feature descriptor which is scale and rotation-invariant. SURF approximates or even outperforms previously proposed schemes stated above in accordance to repeatability, distinctiveness, and robustness [6].

3. METHODOLOGY

As shown in fig. 1 during the offline stage, first of all input video clip 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 key frames, SURF feature descriptor is extracted. This descriptor is then passed into a training phase. The model vector is then uploaded in the model data library [4]. Next for a given test clip to retrieve the video, model vector of test clip is computed and is classified using a minimum distance classifier. From the list of similar videos the highest ranking videos are retrieved.

1. Shot Boundary Detection [4]

For Shot Boundary Detection the histogram value of consecutive frame can be compared. A shot S consisting of closely related frames can be represented as

$$S = g(Ik, t),\ k = i, i+1....j,\ where\ 1 \le i \le j \le n\ and\ Ik\ V \qquad (1)$$

In this system, each shot boundaries *Sj* are automatically identified using an auto dual threshold approach. The algorithm uses a high threshold *Tb* for finding hard cuts. The starting frame of gradual transition is determined using a lower threshold *Ts.* 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 *Tb.*

2. Key Frame Extraction [4]

There are great redundancies among the frames in the same

Shot; therefore, certain frames that best reflect the shot contents are selected as key frames [12], [13], [14], [15] to succinctly represent the shot.

The simplest and best way to select the key frame is, Key frame can be extracted by selecting the first and last frame as the key frame. i.e. 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.

3. Feature Extraction using SURF [6] [13]

For feature extraction SURF detector and descriptor is used.

First of all 'interest point' are selected at distinctive location in a key frame, such as blobs, corners, and T-junctions.

3.1 Interest Point Detection

Hessian-matrix approximation is used for interest point detection. Integral images are used here.

3.1.1 Integral Images:

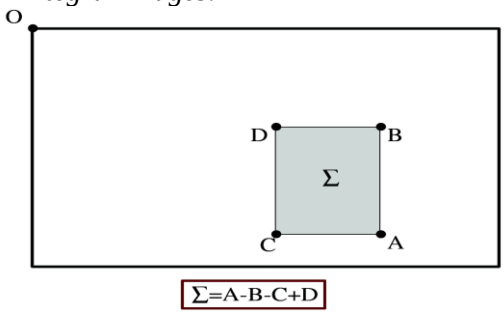

Fig 2: Integral image calculation

Integral Image or summed area tables is an intermediate representation of the image. It contains the sum of intensity values of all pixels in input image *I* within rectangular region formed by origin O = (0, 0) and any point X = (x, y). It provides fast computation of box type convolution filters.

3.1.2 Hessian Matrix Based Interest Points:

Here blob like structure is detect where the determinant is maximum.

- Scale Space Representation:

Interest points need to be found at different scales. The scale space is analyzed by up-scaling the filter size rather than down scaling the image size.

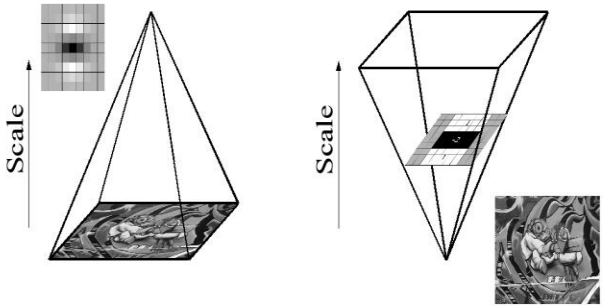

Fig 3: Instead of iteratively reducing the image size (left), the use of integral images allows the up-scaling of the filter at constant cost (right).

3.2 Interest Point Description and Matching

They build on the distribution of first order Haar wavelet responses in x and y direction.

- First step consists of fixing a reproducible orientation based on information from a circular region around the interest point.
- Then, we construct a square region aligned to the selected orientation and extract the SURF descriptor from it.

Finally, features are matched between two images.

3.3 C-SURF: Colored Speeded up Robust Features: [7]

- Color is an important component for objects recognition.
- This will adds the color information into the scale- and rotation-invariant interest point detector and descriptor, coined C-SURF (Colored Speeded-Up Robust Features).
- The first three stages are the same with SURF. In the last stage after calculating the Harr-Wavelet response we also calculate three factors namely $\sum r(x, y)$, $\sum g(x, y)$, $\sum b(x, y)$ for each sub-region.
- The figure below explains how pure gray-based geometric description can cause confusion between two different features.

Fig 4: An example that illustrates the neglecting of color information may confuse the two magnified corners.

4. A SURVEY ON CONTENT BASED VIDEO RETRIEVAL TECHNIQUES

For video retrieval based on content different approaches are used for each video analysis task. For shot boundary detection frame blocking, MI normalization and histogram comparison is used [1] [4]. Key frame can be extracted using mutual information [1]. Feature can be extracted by CSS based shape representation [2], trajectory based motion representation [2], SIFT descriptor [3], SURF detector and descriptor [4], C-SURF descriptor [7]. Comparison of different detector and descriptor is given in [5]. Evaluation of all survey papers for CBVR is given in below table.

TABLE I: paper comparison

Topic Name	Publication/ Year	Algorithm used	Description
Content based video retrieval using information theory (IEEE-2013)[1]	IEEE/2013	Shot boundary detection • Frame Blocking • MI normalization • Block weighting Key frame extraction • Mutual Information Video Indexing • Fuzzy color histogram • K-means algorithm	Video indexing and retrieval is based on hierarchical information theory and speed up the performance.
Combining features for shape and motion trajectory of video objects for efficient content based video retrieval **(IEEE-2009)[2]**	IEEE/2009	Feature Extraction • CSS based shape representation • Trajectory based motion representation	System is tested on synthetic and real world database.

STAR: A Content based video retrieval system for moving camera video shots(IEEE-2013)[3]	IEEE/2013	SIFT descriptor Space-time volume generation EMST-CSS Formation and Feature Matching	STAR uses spatio-temporal features to retrieve similar videos. EMST-CSS is extended for moving camera videos.
Content based video retrieval using SURF descriptor(IEEE-2013)[4]	IEEE/2013	Shot boundary detection • Histogram comparison • Auto dual threshold algorithm Feature Extraction • SURF Descriptor Feature comparison • Fuzzy K Nearest neighbor algorithm	The experimental analysis shows that system provides average precision of 75% with 83% recall. Stochastic reduction and distance classifier is also used to improve efficiency.
Performance comparison of various feature detector-descriptor combinations for content-based image retrieval with JPEG-encoded query images (IEEE-2013)[5]	IEEE/2013	Feature Detector • Hess.-Aff., Harr.-Aff., MSER, MFD, WαSH, DoG, SURF, STAR, BRISK, Kaze Feature Descriptor • SIFT, PCA-SIFT, GLOH, SC, MROGH, LIOP, SURF, Kaze	SIFT and SURF is superior to all other descriptor. Also the Hessian-Affine detector is quite robust.
Speeded-Up robust features (SURF)[6]	Elsevier/2008	Feature Descriptor • SURF	SURF is scale- and rotation-invariant detector and descriptor. SURF outperform in terms of repeatability, distinctiveness and robustness. It is possible to achieve real-time computation without loss in performance is possible.
Real-time traffic sign detection using SURF features(IEEE-2013)[8]	IEEE/2013	Feature Descriptor • SURF with FPGA	In driver assistant system video based traffic sign detection play an important role.

5. FUTURE DEVELOPMENTS

Although a large amount of work has been done in visual content-based video indexing and retrieval, many issues are still open and deserve further research, especially in the following areas [10].

1) Motion Feature Analysis: The effective use of motion information is essential for content-based video retrieval. To distinguish between background motion and foreground motion, detect moving objects and events, combine static features and motion features, and construct motion-based indices are all important research areas.

2) Hierarchical Analysis of Video Contents: One video may contain different meanings at different semantic levels. Hierarchical organization of video concepts is required for semantic based video indexing and retrieval.

3) Combination of Perception with Video Retrieval: It is interesting to simulate human perception to exploit new video retrieval approaches. The research in visual perception shows

6. CONCLUSION

It is concluded form the survey that for shot boundary detection dual threshold based histogram comparison algorithm gives good performance. Also for key frame extraction Fuzzy K-Nearest neighbor algorithm shows promising results. For feature extraction SURF (Speeded-Up Robust Feature) outperforms the other. SURF provides scale- and rotation-invariant detector and descriptor. Repeatability, distinctiveness and robustness are unique features of SURF. The main drawback of SURF is that both detector and descriptor not use color information. C-SURF adds color information to the existing SURF method.

7. REFERENCES

[1] Yarmohammadi, H.; Rahmati, M.; Khadivi, S., "Content based video retrieval using information theory," Machine Vision and Image Processing (MVIP), 2013 8th Iranian Conference on , vol., no., pp.214,218, 10-12 Sept. 2013

[2] Dyana, A.; Subramanian, M.P.; Das, S., "Combining Features for Shape and Motion Trajectory of Video

Objects for Efficient Content Based Video Retrieval," Advances in Pattern Recognition, 2009. ICAPR '09. Seventh International Conference on , vol., no., pp.113,116, 4-6 Feb. 2009

[3] Chattopadhyay, C.; Das, S., "STAR: A Content Based Video Retrieval system for oving camera video shots," Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on , vol., no., pp.1,4, 18-21 Dec. 2013

[4] Asha, S.; Sreeraj, M., "Content Based Video Retrieval Using SURF Descriptor," Advances in Computing and Communications (ICACC), 2013 Third International Conference on , vol., no., pp.212,215, 29-31 Aug. 2013

[5] Jianshu Chao; Al-Nuaimi, A.; Schroth, G.; Steinbach, E., "Performance comparison of various feature detector-descriptor combinations for content-based image retrieval with JPEG-encoded query images,"Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on , vol., no., pp.029,034, Sept. 30 2013-Oct. 2 2013

[6] Herbert Bay; Andress Ess, Tinne Tuytelaars,Luc Van Gool, "Speeded-Up robust features (SURF)" Vol. 110, No. 3, pp. 346--359, June 2008.

[7] Jing Fu, Xiaojun Jing, Songlin Sun, Yueming Lu, Ying Wang," C-SURF: Colored Speeded Up Robust Features", International Conference, I SCTCS 2012, Beijing, China, Volume 320, pp 203-210. May 28 – June 2, 2012

[8] Jin Zhao; Sichao Zhu; Xinming Huang, "Real-time traffic sign detection using SURF features on FPGA," High Performance Extreme Computing Conference (HPEC), 2013 IEEE , vol., no., pp.1,6, 10-12 Sept. 2013

[9] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features", Computer Vision–ECCV 2006.

[10] Weiming Hu; Nianhua Xie; Li Li; Xianglin Zeng; Maybank, S., "A Survey on Visual Content-Based Video Indexing and Retrieval," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on , vol.41, no.6, pp.797-819, Nov. 2011

[11] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Trans. Multimedia, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[12] K. W. Sze, K. M. Lam, and G. P. Qiu, "A new key frame representation for video segment retrieval," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 9, pp. 1148–1155, Sep. 2005.

[13] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," ACM Trans. Multimedia Comput., Commun. Appl., vol. 3, no. 1, art. 3, pp. 1–37, Feb. 2007.

[14] D. Besiris, F. Fotopoulou, N. Laskaris, and G. Economou, "Key frame extraction in video sequences: A vantage points approach," in Proc. IEEE Workshop Multimedia Signal Process., Athens, Greece, Oct. 2007, pp. 434–437.

[15] D. P. Mukherjee, S. K. Das, and S. Saha, "Key frame estimation in video using randomness measure of feature point pattern," IEEE Trans. Circuits Syst. Video Technol., vol. 7, no. 5, pp. 612–620, May 2007.