Content-based Video Retrieval

Using image queries

Kinjal Patel[1], Dr. Hiren B. Patel[2]

[1] *Computer Engineering, LDRP_ITR, Gandhinagar, India*

[2] *Computer Engineering, LDRP_ITR, Gandhinagar, India*

ABSTRACT

Given a query image, the objective is to extract, from a database of videos, the relevant videos to the query image. Here relevant means "visual similarity". Majority of the queries on Google search want video answers from YouTube or similar sites. Learners prefer information to be presented to them in audio-visual media than text alone. But there is a huge "semantic gap" between what people want, how they can express, and what is available out there. With lack of satisfaction from textual based video retrieval, the idea of content based video retrieval has been the attention for researchers since long time. Our system has integrated different algorithms to propose a new approach for achieving higher efficiency in the field of content based video retrieval.

Keyword: - ***Video Parsing, Key-frame Selection, Feature Extraction, Similarity Measure etc.***

---

1. INTRODUCTION

In this paper we are presenting Content Based Video Retrieval (CBVR) System where given an image query and the objective is to retrieve relevant videos to that query image. Here, relevance means "visual similarity". There has been a huge amount of multimedia data on the Web. Therefore, for a user it is nearly impossible to find desired videos. Even when the user has found related video data, it is still difficult most of the time for him to judge whether a video is useful by only glancing at the title and other global metadata which are often brief and high level. Therefore, a more effective method for retrieval of video in web is needed. The increasing demand for object tracking on CCTV Surveillance Videos has invoked the research on Content-Based Video Retrieval. Video data possesses a lot of information for those using multimedia systems and applications like digital libraries, publications, education, broadcasting and entertainment. Such applications are useful only when video retrieval systems are efficient enough to retrieve videos.

It includes various steps: Video Segmentation: Segments the video into shots, Key frame Selection: Selects the key frame to represent the shot using HSV color histogram, Feature Extraction: Features are extracted for the key frame and stored into feature vector, Similarity Measure: Asymmetric comparison method used to compare query image and dataset key frames. And finally displays the result to end user.

2. RELATED WORK

This section discussed a literature survey on the use of the feature extraction methods which enable video retrieval by content.

Navdeep Kaur et al. [14] have made the system receive the user query in the form of image. They have used the 2-D Correlation Algorithm to extract the features. This algorithm is compared with the existing technique of surf feature based point matching algorithm.

B V Patel et al. [4] have reviews the interesting features that can be extracted from video data for indexing and retrieval along with similarity measurement methods. They have also identified the present research issues in area of content based video retrieval systems.

Mr. Pradeep Chivadshetti et al. [8] have implemented an approach for content-based video retrieval using combination of different approach in large video archives proposed an end-to-end text detection and recognition system as OCR and also applying ASR. The text detection component uses HOG based on rich shape descriptors such as HOG.

Deepak C R et al. [5] have presented a novel content based video retrieval system using the conventional video features such as color, texture, edge and motion patterns. In this content based video retrieval system query is a video clip.

3. PROPOSED SYSTEM

From the surveillance of previous work done in this field, we have integrated different approaches for video parsing and feature extraction to achieve the higher relevancy in retrieval of videos. The general framework for content based video retrieval is depicted in Fig. 2.

This framework includes the following: 1) video parsing: To detect the shot boundaries and extract key-frames from each shot using sub sampling. 2) Feature Extraction: To extract the features like color feature, texture feature, edge feature, sift feature etc. 3) Similarity measure: compare the features using binarized feature vector.

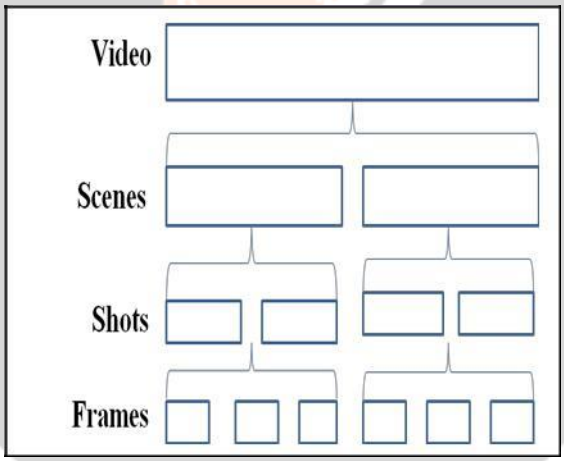

Fig1. General structure of video

Frame: Frame is a smallest unit of a video which appears in the form of an image.

Shot: Frames recorded in a single camera operation.

Scene: Several shots combine together to form a scene.

Video: A Video is composed of different shots, scenes, and sequences arranged according to some logical structure.

A. Video Parsing:

Shot Boundary Detection:

A shot is an unbroken sequence of frames and also a basic meaningful unit in a video. It usually represents a continuous action or a single camera operation. A video usually contains more than one shot, so there must be shot boundaries among those shots.

We use the HSV histogram instead of the traditional RGB color histogram [8].HSV color histogram is more intuitive than RGB color histogram. Shot boundary detection is performed by comparing HSV histograms using L1 distances.

Key frame selection:

Video is rich in content and it results in a tremendous amount of data to process. This can be made easier by only processing some frames, such as the key frames of video. Sivic et al. [12] takes the middle frame of a shot as a key frame. This fails since the middle frame may not be representative of the shot.

Frames are selected by subsampling the 1 fps stream in regular intervals within the shot. In case there are fewer frames in a shot than the selected number of frames, we simply use all frames from the given shot.

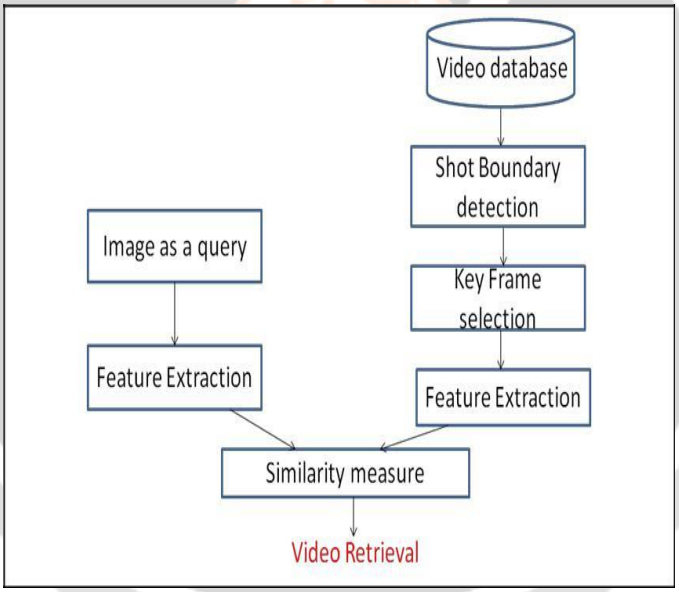

Fig2. CBVR system Architecture

B. Feature Extraction

We focus on the visual features suitable for video retrieval.

Color features:

The basic features of any frame include color and texture. Navdeep kaur et al.[14] have used RGB color histogram to extract color features. We have used HSV color histogram to generate color features.

The transformation from the RGB color space to the HSV color space is as follows:

$$V = \max(R, G, B)$$

$$S = \begin{cases} \dfrac{(V - \min(R, G, B))}{V} * 255, & if\ V \neq 0 \\ 0, & else \end{cases}$$

Here, the range of R, G, B is [0, 255]. If H < 0, we set

H = H + 180. Hence the ranges of H, S and V are [0, 180], [0, 255] and [0, 255].

Texture features:

Texture properties are extracted using gabor filter. Gabor function is the only function who can reach the time-frequency uncertainty bounds, so it can work as a filter to realize texture segmentation and determine the edge of texture under optimal meaning time-frequency.

A set of Gabor filters with different frequencies and orientations may be helpful for extracting useful features from an image. In the discrete domain, two-dimensional Gabor filters are given by,

$$G_c[i,j] = Be^{-\frac{(i^2+j^2)}{2\sigma^2}} \cos(2\pi f(i\cos\theta + j\sin\theta))$$

$$G_s[i,j] = Ce^{-\frac{(i^2+j^2)}{2\sigma^2}} \sin(2\pi f(i\cos\theta + j\sin\theta))$$

where B and C are normalizing factors to be determined.

f defines the frequency being looked for in the texture. By varying , we can look for texture oriented in a particular direction. By varying , we change the support of the basis or the size of the image region being analyzed.

Edge features:

Edge information gives the information about the orientation of the objects present in the key frame. many edge detection methods has been proposed in the last half century, such as Laplacian operator, Roberts operator, Sobel operator, Prewitt operator, Kirsch operator, Marr operator, Canny operator and so on. But all these operators have no automatic zoom function and cannot show the edges of an image in different scales. We have used Gabor filter to get the edge information. When using Gabor filter, it will not be too sensitive to the effects of local lightening due to the removed dc component, as a result, we can select different directions and scales to get corresponding results, most of the edge points in different directions can be captured by the results of the filter.

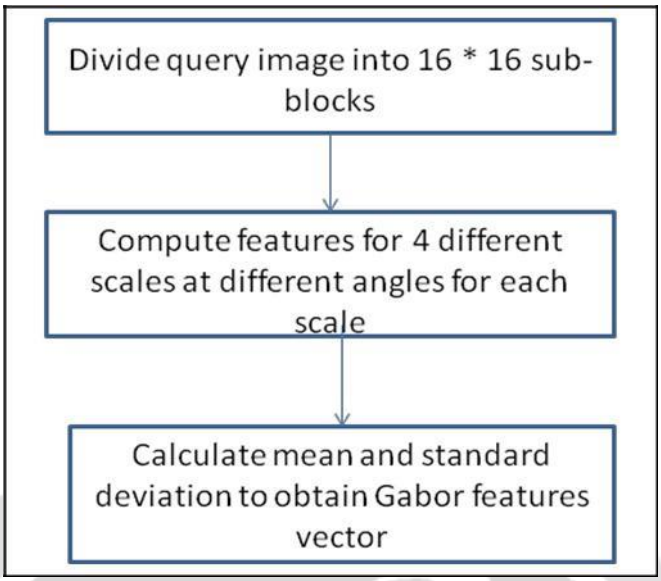

Fig3. Gabor filter algorithm

Sift features:

Video retrieval can be done using SIFT feature[10]. Video is divided into frames, and frames are divided into images. The object is separated from the image by the segmentation of the image. The segmented object is a part of image. Feature is extracted from the segmented image. This method the features are extracted by using the Scale Invariant Feature Transform (SIFT) and are used to find the key points from the images because they are invariant to image.

In this method first the video is converted into images. Features are retrieved from the object image using the SIFT algorithm. Feature matching is then performed on database features by comparing FVs to retrieve the video from database.

SIFT (Scale Invariant Feature Transform) features are used in object recognition are of very high dimension. They are invariant to changes in scale, translation and rotation transformations.

For a given image I(x,y), its linear scale-space representation:

$L(x, y, \square) \square G(x, y, \square) * I(x, y)$

Where $*$ is the convolution operation in x and y, and
$G(x, y, \square) \square \frac{1}{2\square\square^2} e^{\square(x2 \square y2)/2\square 2}$

For each image sample, L(x,y), at this scale, the gradient magnitude, m(x,y), and orientation, θ(x,y), is precomputed using pixel differences:

$$m = \sqrt{(L_{x+1,y} - L_{x-1,y})^2 + (L_{x,y+1} - L_{x,y-1})^2}$$

$$\theta = \tan^{-1}((L_{x,y+1} - L_{x,y-1})/(L_{x+1,y} - L_{x-1,y}))$$

By assigning a consistent orientation to each Keypoint based on local image properties, the Keypoint descriptor can be represented relative to this orientation and therefore achieve in-variance to image rotation.

Where $*$ is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma_2} e^{-(x_2 + y_2)/2\sigma_2}$$

For each image sample, L(x,y), at this scale, the gradient magnitude, m(x,y), and orientation, θ(x,y), is precomputed using pixel differences:

$$m = \sqrt{(L_{x+1,y} - L_{x-1,y})^2 + (L_{x,y+1} - L_{x,y-1})^2}$$

$$\theta = \tan^{-1}((L_{x,y+1} - L_{x,y-1})/(L_{x+1,y} - L_{x-1,y}))$$

By assigning a consistent orientation to each Keypoint based on local image properties, the Keypoint descriptor can be represented relative to this orientation and therefore achieve in-variance to image rotation.

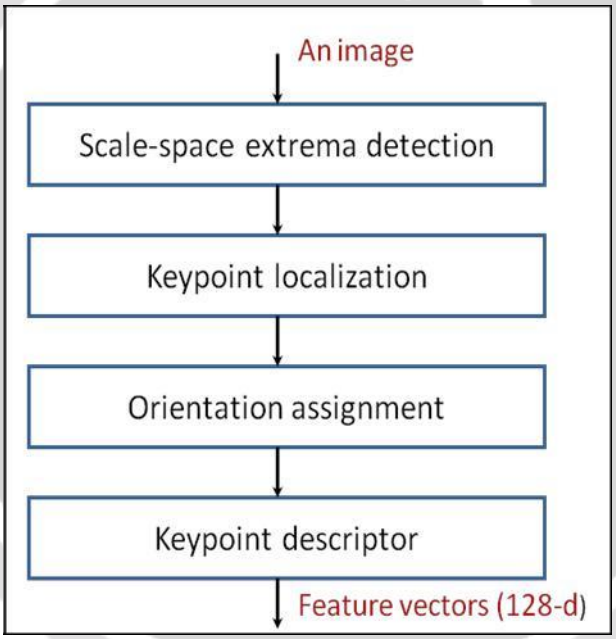

Fig4. SIFT algorithmc steps

C. Similarity measure

Video similarity measures play an important role in content-based video retrieval. In content based video retrieval, a query and its matching dataset entry often have a „containment relationship": for example, a large part of a query image may be contained in a relevant dataset shot, while the reverse is not true. This asymmetry can be exploited to boost retrieval performance. So, an asymmetric comparison method for binarized Fisher Vectors is used to compare the query image and dataset entry.

Fisher vector is a State-of-the-art technique for large-scale retrieval. It represents a set of local descriptors by a compact fixed-length vector. Two images can be compared by comparing their Fisher vectors.

Construction: describe an image with aggregated Fisher scores of its local descriptors. Local descriptor distribution: Gaussian Mixture Model (GMM). Usually only Gaussian means are taken into account. It is an Extension of Bag-of-Words technique.

4. EXPERIMENTAL SETUP

Our experiments make use of the VCDB core dataset. The VCDB core dataset was collected using 28 carefully selected queries from YouTube and MetaCafe.

In order to collect representative partial copies, they started from 28 carefully selected queries, covering a wide range of topics such as commercials, movies, music videos, public speeches, sports, etc. We downloaded the top returned search results of the queries from the two websites, and manually picked on average around 20 videos per query. These videos are all relevant to the query and many of them share partial copies. In total we have 528 videos (approximately 27 hours) in the core dataset.

We use SIFT detector and descriptors [10] in all experiments. The 128D SIFT descriptors are reduced to 32D with a PCA step.

5. PERFORMANCE EVALUATION

Performance is evaluated using mean Average Precision (mAP) over a ranked list of the top 70 retrieved scenes.

Image Queries	**mAP(mean Average precision)**
Query 1	1.0
Query 2	0.50
Query 3	0.17
Query 4	1.0
Query 5	1.0

Table 1: Experimental result

Mean average precision percentage calculated from 17 queries and result is found to be 73.67

6. CONCLUSION

In proposed system, We have detected the shot using HSV histogram and obtained the key frames from shot using sub sampling. And we have extracted the color, texture, edge and SIFT features using different algorithms to propose a new approach for achieving higher efficiency in the field of content based video retrieval.

7. REFERENCES

[1] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, Maybank S., "A Survey on Visual Content-Based Video Indexing and Retrieval", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41-6,797-819, 11/2011

[2] M. Petkovic, W. Jonker, "Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events", Proceedings of IEEE Workshop on Detection and Recognition of Events in Video, pp. 75-82, 2001.

[3] Hong Jiang Zhang, Jianhua Wu, Di Zhong, Stephen W. Smoliar, "An integrated system for content-based video retrieval and browsing", Pattern Recognition, Pattern Recognition Society, Published by Elsevier Science Ltd., Vol. 30, No. 4, pp. 643~658, 1997

[4] B V Patel, B B Meshram, "Content Based Video Retrieval Systems", International Journal of UbiComp, vol 3, No. 2, pg 13-30, 2012

[5] Deepak C R , Sreehari S , Sambhu S Mohan "A Novel Approach for Query by Video Clip" IOSR Journal of Computer Engineering (IOSR-JCE) IOSR Journal e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 5 (Mar. - Apr. 2013), PP 32-35

[6] Dr. H.B. Kekre, Dr. Dhirendra Mishra, Ms. P. R. Rege "Survey on Recent Techniques in Content Based Video Retrieval" International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume-3, Issue-5, May 2015.

[7] Aasif Ansari and Muzammil H Mohammed "Content based Video Retrieval Systems - Methods, Techniques, Trends and Challenges " International Journal of ComputerApplications (0975 – 8887) Volume 112 – No. 7, February 2015.

[8] Mr. Pradeep Chivadshetti , Mr. Kishor Sadafale and Mrs. Kalpana Thakare "Content Based Video Retrieval Using Integrated Feature Extraction and Personalization of Results" 2015 International Conference on Information Processing (ICIP).

[9] Fang Liu and Yi Wan "Improving the Video Shot Boundary Detection Using the HSV Color Space and Image Subsampling" 7th International Conference on Advanced Computational Intelligence Mount Wuyi, Fujian, China; March 27-29, 2015

[10] David G. Lowe "Distinctive Image Features from Scale-Invariant Keypoints" January 5, 2004

[11] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: Video shot retrieval for face sets," in Proc. Int. Conf. Image Video Retrieval, Jul. 2005, pp. 226–236

[12] R. Visser, N. Sebe, and E. M. Bakker, "Object recognition for video retrieval," in Proc. Int. Conf. Image Video Retrieval, London, U.K., Jul. 2002, pp. 262–270.

[13] D.-D. Le, S. Satoh, and M. E. Houle, "Face retrieval in broadcasting news video by fusing temporal and intensity information," in Proc. Int. Conf. Image Video Retrieval,(Lect. Notes Comput.Sci.),4071,Jul.2006, pp. 391–400.

[14] Navdeep Kaur and Mandeep Singh"Content Based Video Retrieval with Frequency Domain Analysis Using 2-D Correlation Algorithm" International Journal of Advanced Research in Computer Science and Software Engineering 4(9), September - 2014, pp. 388-393