A Text Based Video Retrieval Using Semantic And Visual Approach

Vaidehi K.Bante, PG Student
Department of Information Technology
Sinhgad College of Engineering
Pune, India
vaidehibante@gmail.com

Avinash N Bhute, Associate Professor
Department of Information Technology
Sinhgad College of Engineering
Pune, India
anbhute@gmail.com

***Abstract*- In recent past, the cost for storing multimedia data and the digital multimedia storage grows is cheaper. So there is an abundant number of videos available in the video repositories. Video sharing on the web is growing with a tremendous speed which creates perhaps the most heterogeneous and the largest publicly available video archive. There is a vast amount of video archives, including broadcast news, documentary videos, meeting videos, movies, etc. Therefore video retrieval, searching and retrieving videos relevant to a user defined query is one of the most popular topics in both multimedia research and real life applications.**

Text data present in the videos contains useful information for retrieving and indexing purpose. In current video search, the search results are influenced by the metadata information such as title, captions associated with them. Extracted text can be used to recognize the overlay, scene text, which is used for video retrieval. The main focus of this work is video retrieval using 1) overlay text and 2) semantic feature. The semantic concept: an objective linguistic description of an observable entity. The main contribution of this proposal is utilization of the semantic word similarity measures for the text-based retrieval of videos.

Keywords- CBIR, DWT, key frames, Multimedia Retrieval, Similarity Matching Algorithm, SVM, OCR ,Text frames classification.

I. INTRODUCTION

Comprehensive development and the use of multimedia and networking technology have recently provided a large growth in production and distribution of multimedia files. Standard multimedia databases are not able to now-a-days to serve user's demands for database storage capacity, accessibility and searching functionalities. This problem can be conquered by the use of large multimedia databases with advanced functionalities presented on personal computers or servers all over the Internet. Because of their simple procurement and distribution using the Internet, the images are the most multifarious multimedia files. Today, every user of computers has his own image database hosted on a personal computer or on some of online image databases. In a couple of years, the average user can produce an impressive number of images using a digital camera.

Inquiring images in large image databases can be very hard and finding an efficient searching procedure represents a challenge for researchers. There are two different approaches for image tallying in large image databases. The first approach uses semantic labeling based on a group of words and image file name that describes the content of the image. This approach is very time consuming because every image has to be tested and described. Searching process is based on matching the designate words. The effectiveness of this procedure depends on the language used for a number of used words and image description. The results of the searching task are not usually satisfying for users. The second approach for image labeling is image annotation by content with low level features (color, texture, shape,…), and the systems which use interpret images in this way for searching procedure are named Content Based Image Retrieval Systems (CBIR systems). These systems provide automated annotation of the images and very simple measuring of similarity based on metric distance between feature vectors.

A study of the literature acknowledges that no complete video text extraction system has been developed. Video text means extracting the text data from a collection of images and video. Video text detection and recognition can be classified in the two broad categories graphic text or overlay text and scene text as shown in the Fig 2 and Fig 3. There has been much prior work for video text detection and extraction. Most methods in the research have been developed to extract text from complex color images and have been extended for application to video data.

In this paper the fig. 1 shows the framework for video retrieval system and then gives a block wise system description.The defined framework is used in every video retrieval system.

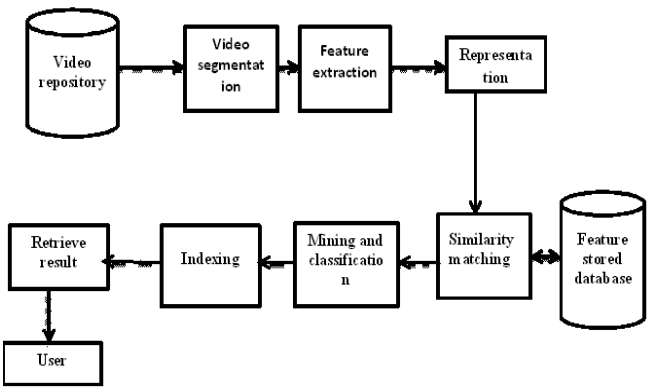


Figure 1. Framework for Video Retrieval System

The framework consists of following steps as,

1) Video detachment or segmentation which includes shot boundary detection,
2) Feature Extraction includes extracting features from segmented video clips,
3) Video mining to leads to create the feature vector,
4) Video interpretation or annotation to build a semantic index,
5) User query and
6) Retrieval of accurate video i.e. the output**.**

Text based video retrieval has a wide range of applications such as, quick browsing of video folders, remote instruction, digital museums, consumer domain applications, news event analysis, video surveillance, and educational applications [10]. These applications motivate the research in text based video retrieval.

Figure 2. Scene Text

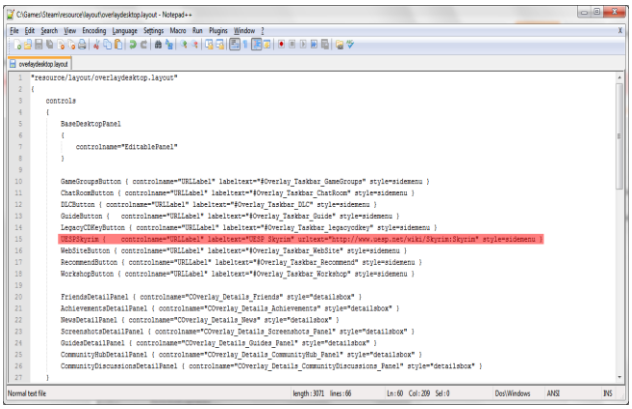


Figure 3. Overlay Text

The rest of this paper is organized as follows: In the next section, Section II addresses the related work. In Section III, Figure 1 shows system architecture an overview is obtained. Each module in the figure is a complex module having own ways of implementation and understanding, the various methods to be considered for text based video retrieval. Section IV describes the feature extraction algorithms. Finally, section V performance evaluation of the system and section VI describes conclusion i.e. summarizes the paper.

II. RELATED WORK

A. A Semiautomatic Method To Generate Annotation For Cricket Videos

According to Dr. Sunitha Abburu [12] presents a paper which is based on indexing and, semantic video analysis and retrieval are necessary for the adequate usage of video repositories. For the proper source to extract semantics of the video, the superimposed text which will increase the efficiency of retrieval system. This paper proposed a semi-automatic method to generate annotation for cricket videos and an automated tool DLER, to extract the semantics of cricket video.

This paper focused on the primary aim is to propose novel techniques for video text detection, localization, extraction, and recognition. The first step, i.e. the text extraction in video frames is difficult because of complex background, unknown text character color, and various stroke widths. Although many methods have been proposed for preprocessing, here the author proposed a fully automatic method, which is a simple approach for preprocessing which integrates all the steps involved in detection, localization, extraction, and recognition as a simple and single tool .

B. Semantic Multimedia Retrieval Using Lexical Query Expansion

According to Milind R. Naphade, Alexander Haubold, Apostol (Paul) Natsev, [13] presents methods to enhance text based search retrieval of visual multimedia content by developing a set of visual models of semantic concepts from a lexicon of concepts allow uniform for the collection via queries of words or fully qualified sentences text search is performed, and results are returned in the form of ranked video clips.

The proposed approach presents methods which involves a query expansion stage, in which query terms are compared to the visual concepts for which authors independently build classifier models. During expansion, this advantages a synonym dictionary and wordnet similarities. In particular, authors spotlight on lexical query analysis and expansion mapping query words and phrases to concepts and build a ranked list of matching shots based purely on automatic concept detection scores and automatically computed query-to-concept relevance scores.

C. Latent Semantic Indexing

According to Roshan Fernandes, Rashmi M, [16] concentrated on the concept of the classifying web based videos are an important yet challenging task. Therefore, this paper focused on the accuracy of retrieval system which depends on the method used for detecting shots and scenes, kind of key frames etc. video features used for

retrieval. For this kind Semantic Video Indexing is a step towards automatic video indexing and retrieval, therefore a Latent Semantic Indexing (LSI) technique is proposed. In this method, it is stated that LSI is a method that exploits the idea of vector space model and singular value decomposition (SVD). LSI uses SVD to reduce noise and dimensionality in the initial term document representation and to capture latent relationships between the terms and the document. Here the proposed system works by analyzing the key frames in video shots and extracting the different visual features from these frames. Then the feature matrix is formed by combining the different types of features from all shots of a video. Then the latent semantic indexing is performed in the feature matrix.

The proposed method performs well when there is complex background and it becomes more reliable as the scene contains more edges in the background. The proposed method is robust to different in respective feature characteristics like character size, position, contrast and color.

III. PROPOSED SYSTEM

The proposed system is based on the following functionalities and features:

A more general overview of the overall process of a video indexing and retrieval framework which is outlined in Fig. 4. The framework includes the following:

1. Structure analysis: to detect shot boundaries, extract key frames, and segment scenes;
2. Feature extraction from segmented video units (shots or scenes). These features include static features in key frames, object features, motion features, etc.;
3. Video data mining using the extracted features;
4. Video annotation: using extracted features and mined knowledge to build a semantic video index. The semantic index together with the high-dimensional index of video feature vectors constitutes the total index for video sequences that are stored in the database;
5. Query: the video database searches for the desired videos using the index and the video similarity measures;
6. Video browsing and indexing. The videos found in response to a query are returned to the user to browse in the form of a video summary.

The detailed description of the proposed system :-

- The important part of the system is subjected to feature extraction process where the attributes of the image and if the input query is video then different scene and caption text as well as both visual such as shape, color, texture and semantic such as intentional, clicks, labels, etc. are extracted from the feature database using convenient methods and by extracting the text the system tries to find concept and synonyms from the dataset present.
- The next module, i.e. in similarity measurement process, the query's feature is compared with the features vectors of digital image are used.
- The query image can be any of the popular formats. To feature extraction process and query stored in feature database the query image is subjected. The distance between the two features is calculated and weights are determined.
- The retrieved videos are then sorted and indexed, so that most similar images and videos can be displayed to the user.

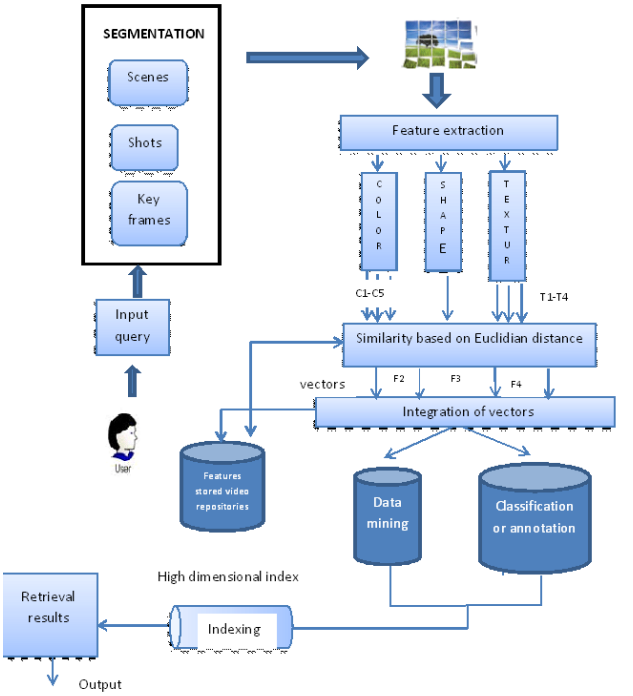


Figure 4. System Architecture

A. *Algorithms Used In Proposed System:*

1. Key-Frame Selection
   Selects the key frame among the extracted frames of the video, to represent the shot using Euclidian Distance Algorithm.

2. Euclidian Distance
   Euclidean distance is used as a similarity measure between two feature vectors and minimum Euclidean distance yields the best similarity.

3. Feature Extraction
   Features are extracted for the key frame and stored into feature vector. Features are of two types that are spatial and temporal.Spatial features includes color, shape, texture and the respected algorithms are discussed in next section.Temporal features are also further classified as motion and audio.

4. Classification
   Classification of video contents is done based on Support Vector Machines (SVM). The support

vector machines associate each set of data points in the multidimensional feature space to one of the classes during training.

5. Clustering
For Clustering we are using K- Mean's clustering algorithm. This is a widely used clustering algorithm. It assumes that we know the number of clusters k. This is an iterative algorithm which keeps track of the cluster centers (means). The centers are in the same feature space as x.

a. Randomly choose k centers μ1, . . . , μk.
b. Repeat.
c. Assign x1 . . . xn to their nearest centers, respectively.
d. Update μi to the mean of the items assigned to it.
e. Until the clusters no longer change.

IV. FEATURE EXTRACTION

A. Video segmentation

Video segmentation is first step towards the text based video search aiming to segment moving objects in video sequences. Segmentation of video done with the help of step by step process of video segmentation, The complete video is first converted into scenes, then scenes are converted into shots and finally shots are converted into various frames.

Figure 5. JSEG segment results

B. Key frame extraction

Key frame extraction includes features color, texture, and shape as below mentioned.

1. Colors
For Colors, colors are defined on a selected color space. Color spaces shown to be closer to human perception and used widely in RBIR include, RGB, HSV (HSL), YCrCb and the hue-min-max-difference (HMMD). Common color, features or descriptors in RBIR systems include, color-covariance matrix, color histogram, color moments, and color coherence vector retrieval, scenery image retrieval, WWW image retrieval, etc.

2. Texture
The texture provides important information in image classification as it describes the content of many real-world images such as fruit skin, clouds, trees, bricks, and fabric.

3. Shape
Shape is a fairly well-defined concept. Shape features of general applicability include aspect ratio, circularity, Fourier descriptors, moment invariants, consecutive boundary segments

C. Distance calculation and similarity measurement

In terms of minimum distance corresponds to feature database and more similar images and video retrieved the distance calculation is measured. Euclidean distance is used as a similarity measure between two feature vectors and minimum.

Euclidean distance yields the best similarity. This similarity measure is used to give a distance between the query video and a candidate match from the feature data database.

The edge histogram descriptor (EHD) is found to be quite effective for representing natural images. It captures the spatial distribution of edges, somewhat in the same idea as the color layout descriptor.

D. Reducing the semantic gap

Reducing the semantic gap can be classified in different ways from different point of view. For example, by considering the application domain, they can be classified as those targeting at artwork.

1. Vocabulary
To find an 'ideal' vocabulary representing the rich semantics of images and text is not an easy task, psychophysical experiments are conducted to judge into the semantic categories that guide the human perception of image similarity. By analyzing the perceptual data, the most important 20 semantic categories (for example, portraits, crowds, cityscapes) in the perception of image similarity were established. Then 40 low-level features were discovered that best describe each category, such as number of regions, color, composition, number of edges, and the presence of the central object., another approach is to establish a so-called 'lexical basis functions' which contains 98 words to represent images.

2. WordNet
A 'WordNet' online lexical reference system is described. 'WordNet' organizes English words into synonym sets, each representing one underlying lexical concept. It is a 'dictionary' based on psycholinguistic principles so that searching can be done conceptually instead of alphabetically.

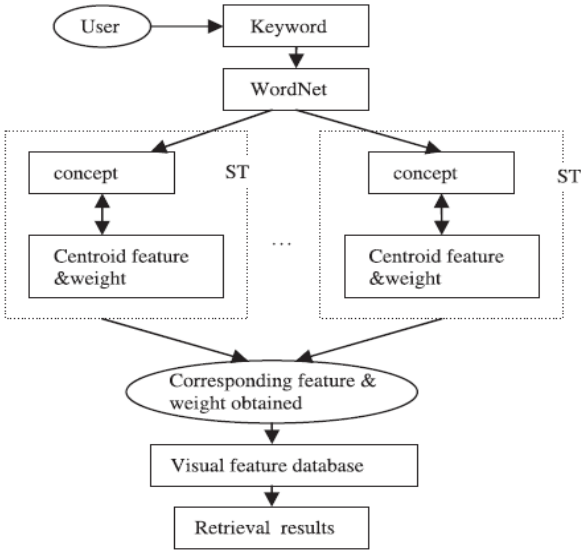


Figure 6. Image Retrieval Supported By Wordnet

V. DATA MODULES

A. Identify the unique frame

In the first module input will give in the form of video and it will convert them into the frames.The input will be given to the system in the form of a video. But instead of processing all frames, the first step is the identification of the unique frames from the video, that process is known as a preprocessing stage for reduction of number of processed frames. This process can done on a number of methods that invented some of them are SVM, OCR, DWT etc. which identify only the text appear image.

B. Text frame detection

After identifying the unique frames, next is to identify text frame based classification. The method needs to classify the image into text frame and non-text frame and identify only the text frame portion. The different methodology uses to classify that is OCR is used for identifies the text base and non-text based frame text. SVM use to identify the text based region. DWT is capture, region which consists of text based on the frequency transform. The objective of this first step is to remove the background from an input gray scale image where the background is interpreted as containing non-text scene contents.

C. Refinement of text regions

The OCR systems require text to be printed against a clean background for character recognition, a local background pixels contained within its interior. Once thresholds are determined for all candidate regions, thresholding operations based on an interactive selection method is performed in each candidate region to separate the text from its surroundings and from other extraneous thresholds are determined for all candidate regions, positive and negative images are computed, where the positive image contains region pixels whose gray level .Fig 8 gives an example of the defined modules.

VI. PERFORMANCE EVALUATION

The performance of the system is evaluated based on the Precision (PR) and Recall (Re) values. It shows the how much Precision and Recall is calculated for a given query video K.

Precision = No. of retrieved videos that are relevant to the query clip Total no. of retrieved videos Nr

Recall = No. of retrieved videos that are relevant to the query clip Total no. of relevant videos available in database Nt.

$$Re = Nr/Nt\ , \quad Pr = Nr/K. \tag{1}$$

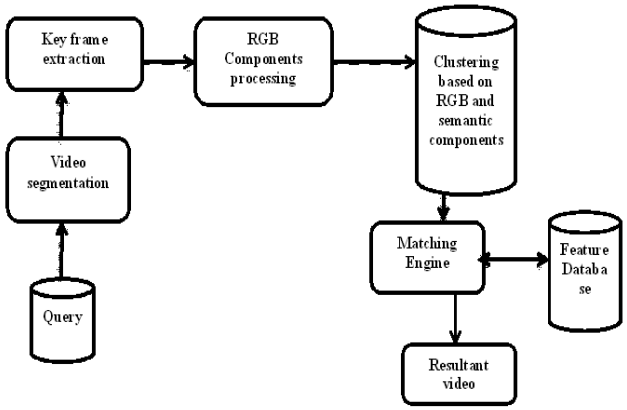


Figure 7: Text Based Video Retrieval System

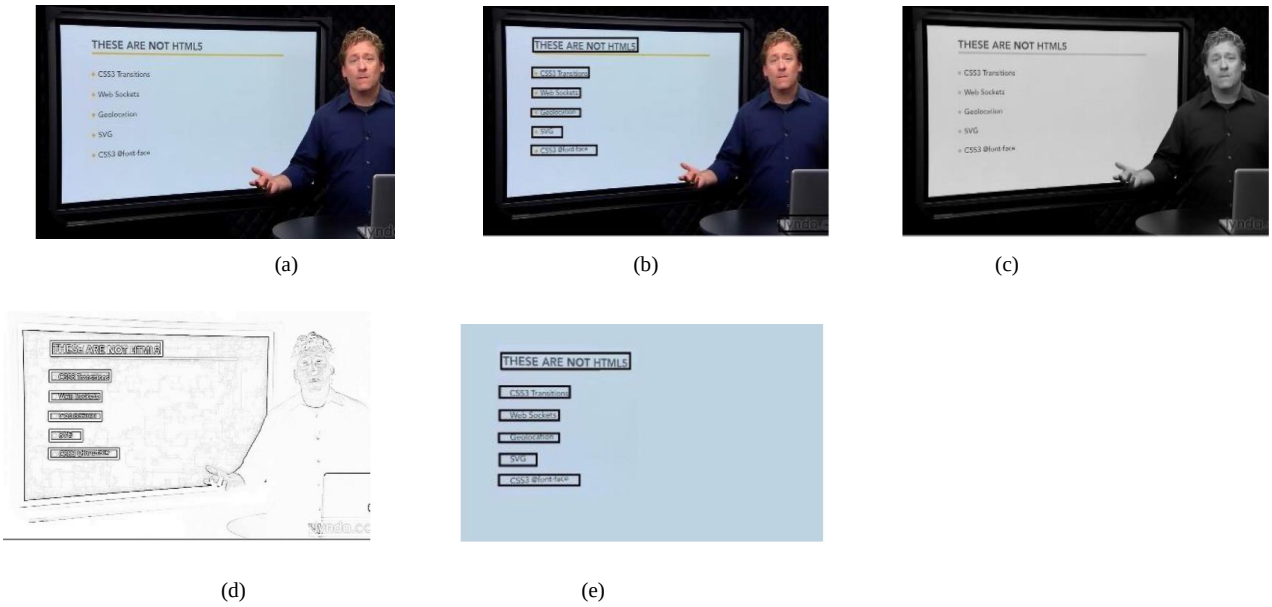
It is 'ideal' to have both high Pr and Re. Therefore, instead of using PR or Re individually, usually a joint PR (Re) curve is used to characterize the performance of an image retrieval system.

VII. CONCLUSIONS

Even with the considerable progress of academic research in video retrieval, there has been relatively little tremble of text, image, and video based video retrieval research on trading applications with some cubbyhole exceptions such as video segmentation. Also, choosing features that reflect real human interest remains an open issue. One reassuring approach is to use meta learning to automatically select or combine appropriate features. Another accessible is to develop an interactive user interface based on visually interpreting the data using a selected measure to facilitate the selection process. Spacious experiments comparing the results of features with actual human interest could be used as another method of analysis. Since user interactions are imperative in the determination of features, it is adorable to develop new methods, tools, and theories to facilitate the user's involvement.

ACKNOWLEDGMENT

I hereby take this opportunity to express my heartfelt gratitude towards the people whose help was very useful



(a) (b) (c)

(d) (e)

Figure 8. An Example of edge detection and text extraction from text contained video (a) Original Image (b) Edge Detection (c) Converting Grayscale (d) Text Frame Extraction (e) Text Detection

for the completion of my research work on the topic of "A Text Based Video Retrieval Using Semantic And Visual Approach "It is my privilege to express sincerest regards to the project Guide constructive criticism throughout the duration of my Prof.A.N.Bhute, for his valuable inputs, able guidance, encouragement, whole-hearted cooperation and project work.

REFERENCES

[1] Di Zhong and Shih-Fu Chang, 1999, "An Integrated Approach for Content-Based Video Object Segmentation and Retrieval", IEEE Transactions On Circuits And Systems For Video Technology, Vol.9,No.8, pp.1259-1268.

[2] Oh J.H., and Bandi B., 2002, "Multimedia Data Mining Framework For Raw Video Sequences", in Proceedings ACM International Workshop Multimedia Data Mining, Edmonton, AB, Canada, pp.18–35.

[3] Yuan J., Wang H., Xiao L., Zheng W., Li J., Lin F., and Zhang B.,2007, "A Formal Study Of Shot Boundary Detection, IEEE Transactions Circuits Systems Video Technology", Vol.17, No.2, pp.168–186.

[4] Divakaran A., Radhakrishnan R., and Peker K.A., 2002, "Motion Activity Based Extraction Of Key-Frames From Video Shots",Proc. IEEE International Conference of Image Process., Vol.1, Rochester, NY, pp.932–935.

[5] Hanjalic A., Lagendijk R. L., and Biemond J., 1999, "Automated Highlevel Movie Segmentation For Advanced Video-Retrieval Systems", IEEE Transaction Circuits System Video Technology, Vol.9, No.4, pp.580–588.

[6] Changsheng Xu, Yi-Fan Zhang, Guangyu Zhu, Yong Rui, Hanqing Lu and Qingming Huang, 2008, "Using Webcast Text for Semantic Event Detection in Broadcast Sports Video", IEEE Transactions On Multimedia, Vol.10, No.7, pp.1342-1355.

[7] Noboru Babaguchi, Yoshihi koKawai, and Tadahiro Kitahashi, 2002,"Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration", IEEE Transactions On Multimedia, Vol.4, No.1, pp.68-75.

[8] Yao Wang, Zhu Liu, and Jin-Cheng Huang, "Multimedia Content Anaylsis Using Both Audio and Visual Clues", IEEE Signal Processing Magazine, pp.12-36,2000

[9] Zhao L., Qi W., Wang Y.J., Yang S.Q., and Zhang H.J., 2001, "Video Shot Grouping Using Best First Model Merging", in Proc. Storage Retrieval Media Database, pp.262–269

[10] Li H.P., and Doermann D., 2002, "Video Indexing and Retrieval Based on Recognized Text", in Proceedings IEEE Workshop Multimedia Signal Process, pp.245–248.

[11] Yoshitaka A., Hosoda Y., Hirakawa M., and Ichikawa T., 1998,"Content-Based Retrieval Of Video Data Based On Spatio temporal Correlation Of Objects", in Proceedings IEEE Multimedia Computing and Systems, pp.208-213.

[12] Dr. Sunitha Abburu "Multi Level Semantic Extraction For Cricket Video By Text Processing**"** Professor & Director, Department of Computer Applications Adhiyamaan College of Engineering, Hosur,pin-635109, Tamilnadu,India, International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5377-5384

[13] Alexander Haubold , Apostol (Paul) Natsev, Milind R. Naphade "Semantic Multimedia Retrieval Using Lexical Query Expansion And Model-Based Reranking**"** Department of Computer Science Columbia University, New York, NY 10027, IBM Thomas J. Watson Research Center Hawthorne, NY 10532

[14] Rashmi M, Roshan Fernandes "Video Retrieval Using Fusion of Visual Features and Latent Semantic Indexing" Dept. of Computer Science and Engineering, Nitte Mahalinga Adyanthaya Memorial Institute of Technology, Nitte, Udupi Dist., Karnataka, India, International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 2, Issue 5, May 2014

[15] Stéphane Clinchant, Julien Ah-Pine, Gabriela Csurka "Semantic Combination of Textual and Visual Information in Multimedia Retrieval**"** Xerox Research Centre **,**Europe chemin de Maupertuis 38240 Meylan, France

[16] B V Patel and B B Meshram, "Content Based Video Retrieval Systems" Shah & Anchor Kutchhi Polytechnic, Chembur, Mumbai, INDIA Computer Technology Department, Veermata Jijabai Technological Institute, Matunga, Mumbai, INDIA, International Journal of Ubi Comp (IJU), Vol.3, No.2, April 2012