JANOLI International Journal of
Computer Science and Engineering
(A Peer Review Journal)

journal homepage: https://janolijournals.com/index.php/jijcse

Volume 2, Issue 1, 2026 ISSN: 3107-4537
DOI: https://doi.org/10.64758/8ykg4954

Object Detection and Classification of Videos for Indexing &Retrieval Using Machine Learning

Shrinivas Amate
Department of CSE (AI&ML)
BLDEA'S VP Dr P G Halaktti College of Engineering and Technology, Vijayapur, Karanataka
aiml.srmate@bldeacet.ac.in
Dr. S R Biradar
Department of AI&ML
SDM College of Engineering and Technology, Dharwad, Karanataka
srbiradar@gmail.com

ARTICLE INFO
Keywords:

Intrusion Detection System (IDS), Internet of Things (IoT), Federated Learning, Edge Computing, Cyber Security, Anomaly Detection, Hybrid Model, Distributed Learning, Real-time Analysis, Adaptive Security.

Correspondence:
E-mail: srbiradar@gmail.com

ABSTRACT

In today's world, people have access to a huge amount of videos, both on the internet and television. But the amount of video data is large and it is increasing day by day as the technologies grow. So video data storage, classification for indexing, filtering,, and retrieval is a big issues. So it is very difficult for a human being to go through all the videos to search for a particular video and it is time-consuming, but the user wants a particular category of video or video genre within a short span. So that there is a need to classify the videos based on their genre, and subgenre, research has begun on automatically classifying videos.

Many works has been done for the classification of videos in certain categories or genre and sub-genre, by filling the semantic between low level features of video and high level concepts, so that user can find their specific interest of video within a narrow domain. There are different techniques have been developed for good understanding of video content and various video features have been recognized for best representation of videos. There are different techniques have been developed for a good understanding of video content and various video features have been recognized for the best representation of videos. Such as Gaussian Mixture Model (GMM), Neural Network (NN), Bayesian, Hidden Markov Model (HMM), and Support Vector Machine (SVM).

Object-based video classification method has been adopted in our work. The video classification system can be roughly divided into two major components: a module for extracting video features from key frames and another module to find feature similarities between video frames and the object from the database. Here the proposed project work is to classify the videos from the repository by the method of extracting regions of interest from each key frame of an input video clip. Then these regions of interest kept as features, are compared with features of objects for video recognition using the MSER algorithm. Finally, we retrieve the videos from the repository by understanding the constrained user text query is presented. The method is evaluated by experimentation over a dataset containing different types of videos i.e. cricket, football, cartoon, news, and movies.

Introduction

In today's world, Digital videos can be automatically classified into various categories or various genres & subgenre is an important issue in the field of video analysis. As the technology grows in video broadcasting control, video classification based on their content is very important for web multimedia administration; website can basically classify and filter a large amount of videos based on their content. It also takes advantages in HDTV (High Definition TV) and VOD (Video On Demand) applications.

There are various approaches for automatic video classifications have been attempted; approaches are divided into 4 groups.

i. Text based approach
ii. Audio based approach
iii. Visual based approach
iv. combined text, visual, and audio features

Most of the works reported on text based classification, but visual and audio based classification of videos work is less. So that our work is concentrated on classification of videos based on the content of input videos i.e. object based classification of video (object is a small Element in an image).

Content based video classification is continuously needed for many applications for example creating automatic video summarization, detecting specific action, for retrieving video sequence or activities in a video surveillance.

Due to the increasing semantic gap of videos, there is a need for computation tools to classify these videos into different genre. Accurately classifying the videos for good representation and efficiently retrieving the video data. So classification task is carried out effectively.

Need for Video Classification

In this world, today's generation have accessing huge amount of videos both on internet and television. But the amount of video data is large and it is increases day by day as the technology grows. So the data storage, classification for indexing, filtering and retrieval is big issue. So that it is very difficult for a human being to go through it all to search a particular video and it is time consuming, but the user want particular category of video or genre within a short of span. So that there is a need to classify the videos based on their genre, subgenre, research has begun on automatically classifying videos.

Problem Definition & Proposed Approach

Video classification is a crucial task when classifying the videos from large repository. The video classification system can be roughly divided into two major components: a module for extracting video repository by understanding constrained user text query is presented. The method is evaluated by experimentation features from key frames and another module to find feature similarities between video frames and the object from the database. Here the proposed project work is to classify the videos from repository by the method of extracting regions of interest from each key frames of input video clip. Then from these regions of interest kept as features, are compared with features of objects for video recognition using MSER algorithm. Finally we retrieve the videos from over dataset containing different types of videos i.e. cricket, football, cartoon, news, and movies.

RELATED WORK

Many research studies have been done to develop efficient video classification system. Each system follows different method of implementation. **The survey seeks to provide a brief overview of researches done by many researchers on video** classification for indexing and retrieval **based on visual information such as graph matching, image, object, color, edge, motion etc.** This section summarizes several techniques to provide historical perspective.

MPEG-1 videos can be classified using Decision Tree classifier presented in [1]. But the decision tree is useful for optimal solutions only to the local, not to the global ones, thus the construction of decision tree is inaccuracy.

Gaussian Mixture Models (GMMs) that can be used to classify the videos into different categories & it also makes video classes according to their genre and PCA is used to reduce the dimensionality of video & audio features presented in [2], but it is not proper classify the videos.

A video classification model is built based on Hidden Markov model (HMM) presented in [3]. But this approach takes a comparatively low level simple features and it takes very large amount of training data for building a classification model, which includes heavy workload and time consuming. So that it is to improve the performance of the video classification algorithm, there is a need to re-examine the features of videos according to their video genres, and the classification policies must be compared to decrease complexities and improve the performance of video classification.

Bag of Visual Words as a model for identifying the local image features and improve the performance of image representation is presented in [4][5].But this approach has 2 disadvantages
First being ignored the spatial information of an image, so that the location information of an image block is ignored. Second one is Performance of the Accuracy of image classification is decreased, because each & every image block is represented by same visual vocabulary and it is not correct.
To segment and locate each & every moving objects in a video, an Efficient Motion Segmentation method is presented in [6].This method allows segmenting and locating on any number of moving objects in a each & every videos in an appropriate manner. And it also extracts the features of each tracked objects and then filtered & indexed in a database. But this system takes a long time for indexing if the video size is large. Because the each & every moving objects has to track in a video. Even though this system requires lot of information on specific objects for retrieving a particular video of interest.

To classify the videos using pre-defined class labels of text is presented in [7] [10]. This method is useful for searching the particular video clips, and it is some sort of similar to the traditional video search engines i.e. text-based. However, this system has advantages only when the images of a video clips, that have been already seen and it is attached properly with pre-defined class label. This system requires reprocessing of all the video clips again, when adding new class label, which is consumes lot of time and also extra space required.
Content based video retrieval approach is presented in [9] by using JSEG segmentation method.
In which every I/P frame is segmented into regions in that 10 largest segmented regions have been selected as objects, then color & texture features of tracked objects are used to describe the video clips. However this system is useful only when content based video retrieval. But this system includes extra step of segmentation of I/P frames. And it also takes extra time for processing of video clips which is provided as input for retrieval purpose, which is not efficient.

In today's world user also needs fastest video retrieval system, and they can search their video in narrow domain. But the existing popular video searching engines in [11] [12] are based on text annotations and descriptions of the video clip which is provided as a input by user. But these engines require extra time for matching user input keywords against largest database of text annotation. However, this system always requires intensive man power and extensive time to describe and to annotate those video clips.

A multimedia(videos, audio, image) data mining framework for extracting of events of soccer goal in soccer videos, by using decision tree and analysis of multimodal features is presented in [13] [15]. This system also proposed a common framework for filtering & indexing and also for summarizing sport videos (cricket, football, soccer, table tennis).

To classify the videos & for efficient searching of frequent and rare video events, a robust GMM classifier is build is presented in [16]. This classifier identifies the real time incoming events for indoor surveillance.

The system presented in [17] [18] which employ a multimodal video indexing. It covers different aspects and issues in video indexing and the approaches that are used. But Multimedia data (video) does not have a single unique semantic, so this approach does not address this challenge.

Requirements for region detection

•For detecting the regions in a frame, translation, scaling, rotation i.e. transformations should be considered are illumination changes and full affine transform (i.e. from a different viewpoint a region should correspond to the same pre-image. Changing Viewpoint can be approximated locally by affine transform if assuming locally planar objects and orthographic camera, that is ignoring perspective effects).
•Region detection should be repeatable and stable, and capable to discriminate between different regions.

Fig. Detection of Regions

Maximally Stable Extremal Region

- MSER –It is a method for detecting distinguished regions in a gray scale image.
- It extracts a number of covariant regions called MSER's from the gray scale i.e. 2D image.
- These regions have taking through the same wide range of threshold i.e. intensity function of extremal property, inside the region and its outside the boundary.
- All pixels below the given threshold are marked as whiteand and above or equal are marked as black.
- If we are shown a threshold image 'I' with frame corresponding to the threshold 't', then black image will seen first ,after that whitespots related to the intensity of local, minima will appear then grows higher.
- These whitespot merges until & unless whole image is white.

- The connected of all components in a sequence is a bunch of extremal regions.
- Elliptical frames attached to the MSER's by finding ellipse to the regions. These regions descriptors are as features.
- The word extremal refers to the property of the pixels within the MSER's that have higher or lower intensity than the pixels on its outer boundary.
- This operation is carried out by performing first filtering all the pixels by the value of gray & then increasingly adding all pixels to the connected component by changing threshold, the area is monitored ,regions are varied wrt the minimal threshold are defined as maximally stable.
- let's make all the pixels below the threshold are marked as white and remaining are black.
- As we increasing the threshold, we seen sequence of threshold image sweeping from black to whitespots, we observe black image to image where whitespots appear will grows larger by merging up to the final image.
- Over the greater range of threshold the binarization is stable & will appear some invariance's to the affine transformation of image intensity and scaling.

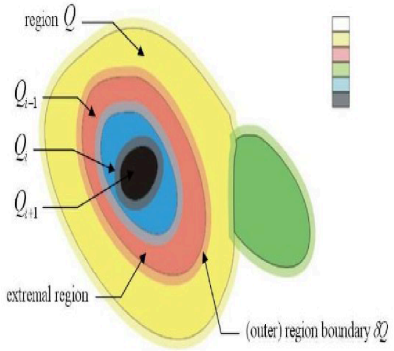
MSER Processing

The MSER extraction implements the following steps

1. Perform a simple luminance thresholding of the image for sweeping a threshold of intensity from black to white.
2. Then extracts all connected components i.e. regions of extremal.
3. Identifying the threshold when a regions of extremal becomes "Maximally Stable" i.e. region grows slowly after some range its growing stop and its region shape approximately become ellipse.
4. Those region descriptors are kept as features.

However it might be rejected, even if an extremal region becomes Maximally Stable if

i) It is too large (parameter MaxArea).
ii) It is too small (parameter MinArea).
iii) It is too variations i.e. unstable.
iv) It is similar to its parent MSER.

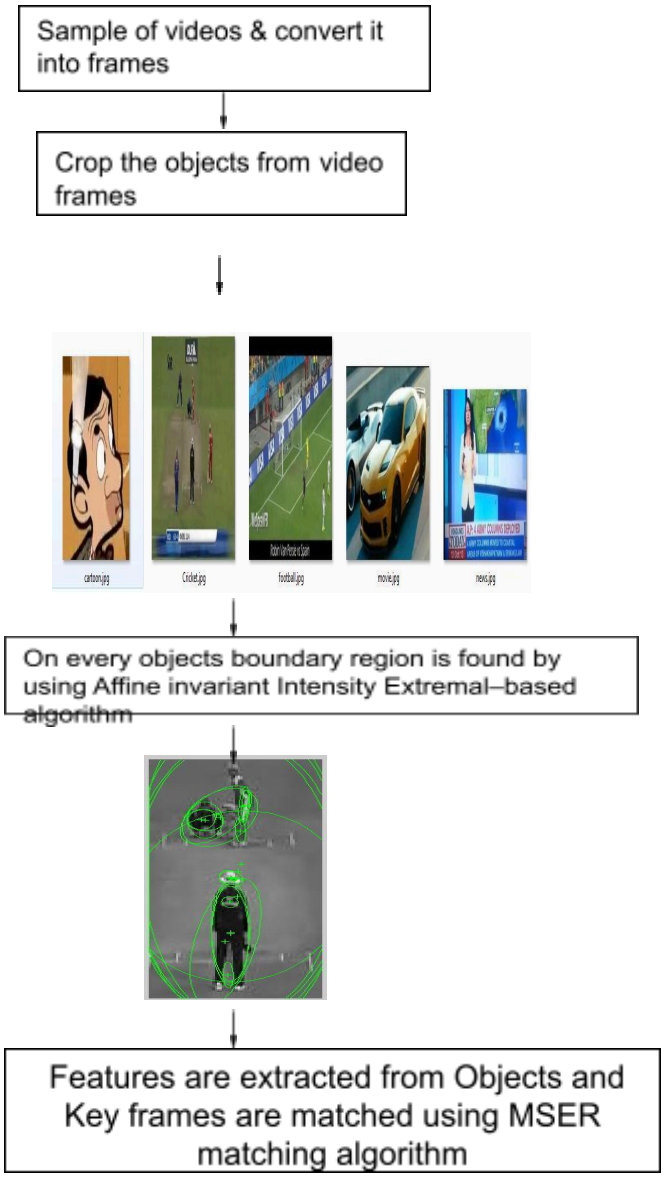


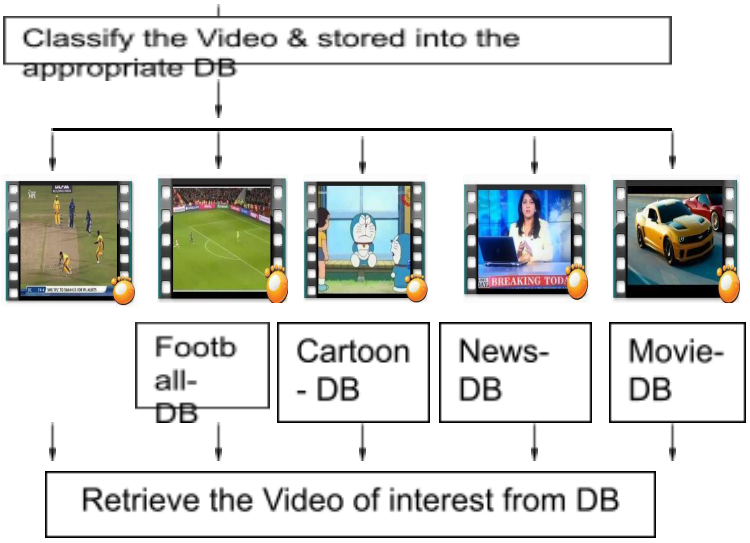
Margin = the number of thresholds for which the region is stable

Fig. Identification of internal & external regions

Proposed Algorithm

Step 1: Read the user input video archive.
Step 2**:** In the next step the videos from video archive are converted into frames (one video at a time is converted into number of frames).
Step 3: Each frame is then compared with the previous one and motion is detected. Frames with significant motion differences are separated as key frames**.**
Step 4: On every extracted key frame boundary region is detected by using Affine Invariant Intensity Extremal–based algorithm, and then the features are extracted around the boundary region.
Step 5: Crop the Objects from different video frames and these objects are stored in a database.
Step 6: Then the features of key frames are matched with the stored objects by using MSER matching algorithm. If it is matched then corresponding video is stored in the appropriate database.
Step 7: Otherwise it will matches with the next stored objects.
Step 8: User text input is then matched with video Cricke t- DB lentify which videos the user is asking for, if match is found retrieve the specified video.





Key frame identification/extraction

Key frame extraction and Video segmentation are the bases of video processing, for analysis & retrieval purpose. To provide a correct video summarization for genre and sub genre classification of videos for video indexing, filtering and retrieval, A Key frame extraction, is an essential part in video analysis. By using key frames which reduces the amount of data required in video classification for indexing & retrieval and for dealing with the content of video it also provides the framework.

Here the proposed method works on group of key frames extracted from a frames of video. It takes a list of key frames in particular order in which they will be extracted based on motion descriptor that specify whether two video frames are similar or dissimilar. It's main function is to choose smaller number of video representative of frames. It starts from 1st frame from sorted list of files. If consecutive frames are having same motion description, then two frames are similar. Repeat process till frames are similar, delete all similar frames & take 1st as key -frame. Start with next frame which is outside threshold & repeat the same procedure for all frames of video. The below figure shows the snap shot of key frames for cricket video.

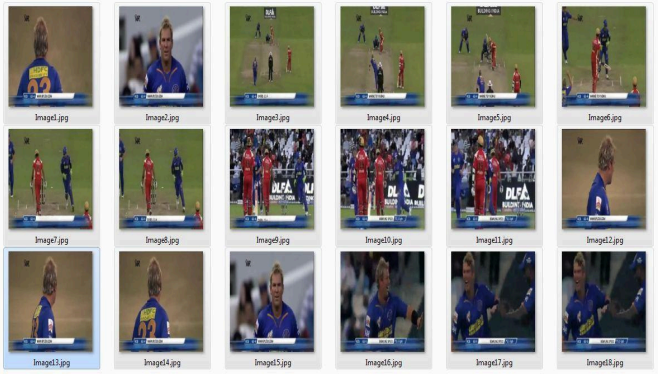


Fig.simple snapshot of Key frames of cricket video

IDENTIFICATION OF BOUNDARY REGIONS

- For detecting the regions in a frame, translation, scaling, rotation i.e. transformations should be considered are illumination changes and full affine transform.
- Region detection should be repeatable and stable, and capable to discriminate between different regions.
- Color image i.e. 3D image can be converted into gray image i.e. 2D image.
- Using MSER method for detecting distinguished regions in a gray scale image.
- It extracts a number of covariant regions called MSER's from a gray scale i.e. 2D image.
- These regions have taking through the same wide range of threshold i.e. intensity function of extremal property, inside the region and its outside the boundary.
- All pixels below the given threshold are marked as whiteand and above or equal are marked as black.
- If we are shown a threshold image 'I' with frame corresponding to the threshold 't', then black image will seen first ,after that whitespots related to the intensity of local, minima will looks, then grows higher.
- These whitespot merges until whole image is white.
- The set of all connected components in a sequence is a set of all extremal regions.
- Elliptical frames attached to the MSER's by finding ellipse to the regions. These regions descriptors are as features.

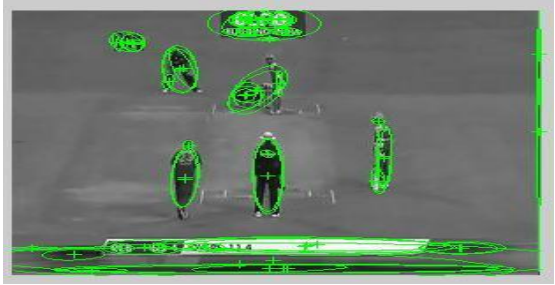


FIG .SIMPLE SNAPSHOT OF BOUNDARY REGION DETECTION

Feature Extraction

The feature extraction method implements the following steps using MSER algorithm
1. Perform a simple luminance thresholding of the image for sweeping a threshold of intensity from black to white.
2. Then extracts all connected components i.e. regions of extremal.
3. Detecting a threshold when a regions of extremal becomes "Maximally Stable" i.e. region grows slowly after some range its growing stop and its region shape approximately becomes ellipse
4. Those region descriptors are kept as features.

DATASETS

We have collecting 100 of videos from YouTube website (www.youtube.com) and that will be used as datasets for performing experiments on it. These videos contain the different categories of objects such as cricket, football, cartoon, news and movies. The sample snapshot for the input videos of the proposed system is given in Figure.

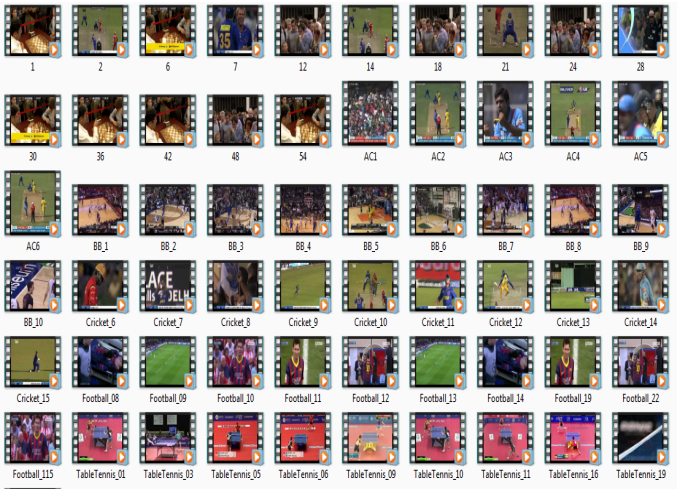


Fig. Sample Snapshot of User Input Dataset

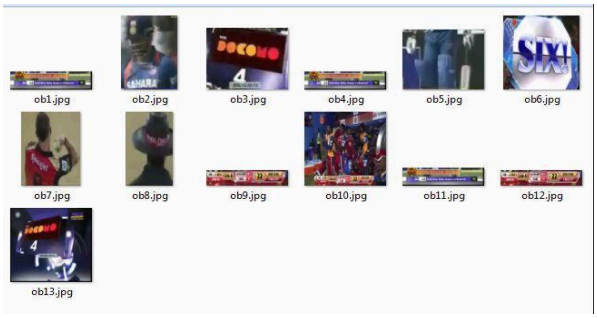


Fig 5.2 Simple Snapshot of Cricket Video Objects

Performance Analysis

The performance of the proposed system is measured on the input dataset using the accuracy rate. Fig 5.5 shows the graph of accuracy rate, for performance analysis, videos from each category are given to the proposed system and results are evaluated which are shown in table 5.1 as follows.
It is also important to know the accuracy of the proposed system which is measured as

$$\text{AccuracyRate} = \frac{\text{No. of times correctly videos classified}}{\text{No. of tests conducted}}$$

NO. OF TESTS CONDUCTED

TABLE 5.1 EXPERIMENTAL RESULTS

VIDEOS	ACCURACY
Cricket	96
Football	80
Cartoon	92
News	90
Movie	86

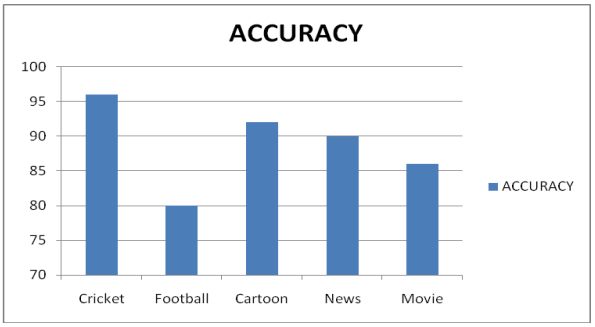


Fig. Graph showing experimental accuracy rate.

The above calculated accuracy rate may differ as the number of videos in the dataset increases. In the proposed work user query text is case sensitive.

Conclusion

In this proposed system, videos are classified into different categories or genre that can be used for indexing and retrieval purpose. First sample of video can be converted into frames, Each frame is then compared with the previous one and motion is detected. Frames with significant motion differences are separated as key frames**.**

On every extracted key frame boundary region is detected by using Affine Invariant Intensity Extremal–based algorithm, and then the features are extracted around the boundary region. Crop the Objects from different video frames and these objects are stored in a database. Then the features of key frames are matched with the stored objects by using MSER matching algorithm. If it is matched then corresponding video is stored in the appropriate database. Otherwise it will matches with the next stored objects. User text input is then matched with video name to identify which videos the user is asking for, if match is found retrieve the specified video.

Experimental results show that this algorithm performs well in the object based video classification.

References

1. Ba Tu Truong, Svetha Venkatesh, and Chitra Dorai. Automatic Genre Identication for Content-Based Video Categorization. In Int. Conference on Pattern Recognition, 2000,
volume 4: 230-233.

2. Xu, L.Q., Li, Y.. Video classification using spatial-temporal features and PCA. Proceedings of the 2003 International Conference on Multimedia and Expo (ICME '03), 2003,
volume 3: 485 – 488

3. Geetha, M.K., Palanivel, S. HMM Based Automatic Video Classification Using Static and Dynamic Features. International Conference on Computational Intelligence and Multimedia
Applications, 2007. Volume: 3: 277-281

4.Van Gemert J C, Veenman C J, Smeulders A W M., et al, “Visual word ambiguity”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271-1283, 2010.

5.Wu L, Hoi S, Yu N, “Semantics-preserving bag-of-words models and applications”, *IEEE Transactions on Image Processing,* vol. 19, no. 7, pp. 1908-1920, 2010.

6.R. H. Y. Chung, F. Y. L. Chin, K.-Y. K. Wong, K. P. Chow, T. Luo, and H. S. K. Fung. Efficient block-based motion segmentation method using motion vector consistency. In *Proc. IAPR Conference on Machine Vision Applications*, pages 550–553, Tsukuba Science City, Japan, May 2008.

7.S. Fan, X.Q. Zhu, A.K Elmagarmid, W.G. Aref, and L. Wu. Classview: Hierarchical video shot classification, indexing, and accessing. *IEEE Transactions on Multimedia*, 6(1):70–86, Febrary 2012.
9.M. Smith and A. Khotanzad. An object-based approach for digital video retrieval. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004)*, volume 1, pages 456–459, Los Alamitos, CA, USA, April 2010.
10.B.S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7, Multimedia Content Description Interface*. Wiley, 2010.
11.http://video.google.com.
12.http://www.youtube.com.
13. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based video retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell., 22(12):1349{1380}, August 2010.
14.S. C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A decision tree-based multimodal data mining framework for soccer goal detection," in Proceedings of IEEE International Conference on Multimedia and Expo (ICME), pp. 265–268, IEEE, June 2004.
15.B. Li, J. H. Errico, H. Pan, and I. Sezan, "Bridging the semantic gap in sports video retrieval and summarization," Journal of Visual Communication and Image Representation, vol. 15, no. 3, pp. 393–424, 2004.
16.V. A. Petrushin, "Mining rare and frequent events in multi-camera surveillance video using self-organizing maps," in Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '05), pp. 794–800, ACM, August 2005.
17.C. Snoek and M.Worring: " A Review of multimodal indexing"; Intelligent Sensory Information System, University of Amsterdam, 2009.
18. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based video retrieval IEEE Trans. Pattern Anal. Mach. Intell., 22(12):1349{1380}, August 2010.