International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

MULTI SPECTRAL CLUSTER BASED VIDEO RETRIEVAL USING SPATIOTEMPORAL SALIENT OBJECTS

Renukadevi .S [1], Dr. S. Murugappan [2]

[1]*Assistant Professor, Department of Computer Science, KSIT, Bengaluru,*

[2]*Associate professor, Department of Computer Science,*

Tamil Nadu Open University,Tamil Nadu,(India)

ABSTRACT

Storage and retrieval of data is considered as a simple and straight forwarded task but found to be trivial when retrieval of information from video is concerned. In recent times there has been a significant increase in the digital content. Efficient retrieval from digital contents (i.e. video) provides competent communication solutions for several applications like video surveillance, educational purpose, monitoring terrorism and so on at an early stage and therefore improves the detection rate of culprit. In this work, a Multi Spectral clustered Spatiotemporal Feature with Graph-based Decision Tree (MSSF-GDT) indexing is performed for efficient video retrieval. A new graph-based data structure called decision tree is used that not only indexes but also organizes and retrieve similar videos from video data, reducing the computational time. With the indexed video data, Multi Spectral Clustering of spatiotemporal feature of the video is performed by applying the Largest Frequent Feature Identification (LFFI) algorithm that is independent of the bands of data (i.e. frame) and the size of the band. The LFFI algorithm extracts the key frames based on region of interest where retrieval is performed on the basis of high level semantic relationship. The performance of MSSF-GDT is evaluated with sports repositories data set using parameters such as similarity retrieval rate, multi spectral clustering accuracy, similarity retrieval time, indexing time with different videos.

Keywords— Multi Spectral Cluster, Spatiotemporal, Decision Tree, Largest Frequent Feature Identification, semantic relationship.

I.INTRODUCTION

With the recent technological advancement, data related to multimedia are designed straightforwardly resulting in vast data availability, detection of culprits on the web or in personal databases. Despite, early detection, the existing video retrieval approach to video surveillance makes it more difficult for detection of culprits with minimal response time. Different video retrieval mechanisms were designed to solve the difficulties, however the rate and the time at which the retrieval takes place has to be addressed.

An automatic Shot-based Key Frame Extraction (S-KFE) [1] for video indexing and retrieval used three phases, shot boundary detection, keyframe extraction and video indexing and retrieval. Initially, the frames were sequentially clustered into shots using shot boundary detection with the aid of edge based feature obtained from

the set of frames. With the obtained shots, relationship between consecutive frames was found using the block similarity based feature followed by which segmentation of shots was performed using dynamic clustering technique.

With the segmented shots, keyframe extraction was performed for visual content representation using grouped shot clusters resulting in the improvement of the average precision rate and recall. Though S-KFE method provided more meaningful representation of visual content and also enhanced the retrieval performance, less focus was emphasized on concept detection. Hence, retrieval performance with respect to concept detection remained unaddressed.

Bag of System Tree (BoS Tree) [2] constructed bottom-up hierarchy of codewords for efficient mapping of videos to the BoS codebook. A BoS Tree was constructed for fast-indexing of large BoS codebooks using bottom up hierarchy of codewords. The bottoms up hierarchy choose the most likely branches when traversing the tree, therefore reducing the computational cost.

Next, the BoS Tree was extended to codebook to handle spatiotemporal variations. The resultant form was experimented on different application, like video annotation, music annotation and retrieval, and video texture classification, ensuring minimum computational cost when compared to standard large codebook. Despite reduced computational cost to spatiotemporal variations, adapting approximate search using Decision Trees remained an open issue to be addressed.

Nowadays, many efficient methods for spatiotemporal object categorization lean on the visual contents and the construction of feature extraction in order to generate or extract efficient features. However, shot based key frame extraction cannot describe objectively and discriminatively the retrieval performance and neglects the spatial distribution of, although it allows significant distinctiveness of the representation through shot boundary detection.

Thus, our technique joins with this tendency in order to overcome these difficulties. We introduce a novel approach for spatiotemporal objects video retrieval called spatiotemporal object detection. It invests in the integration of the salient spatial object using the concept of locality of features and integrating temporal consistency of object based on visual content.

In addition, Graph-based Decision Tree indexing is performed with the detected spatiotemporal objects based on attribute selection namely, information gain, gain ratio and gini index. With the resultant indexed spatiotemporal objects, typicality criterion is applied to produce the inference rules that in turn help in minimizing the computational time. Finally, with the application of Largest Frequent Feature Identification (LFFI) algorithm, retrieval is performed on the basis of region of interest, thus improving the video retrieval rate and time.

The remaining of our paper is organized as follows: In Section 2, we presented state-of-the-art of the video retrieval using key frames. Section 3 described the proposed framework. In Section 4, extensive experiments are presented and discussed in detail in Section 5, followed by conclusions in Section 6.

II. RELATED WORKS

Recent efforts on developing video indexing and retrieval methods using spatiotemporal objects have mainly leveraged progression on analysis of individual frames, as well as progress on indexing and retrieval of spatiotemporal video data.

Adaptive Binary Tree-based Support Vector Machine (ABTSVM) [3] provided means for unified learning framework to retrieve content based image with better precision and recall. However, it did not address high dimensional dense features. To address this issue, a predictable hash code algorithm was designed in [4] with the aid of Expectation Maximization method resulting in the improvement of video-based face recognition.

There has been isolated research in the use of key frames for video indexing and retrieval. The Weber Binarized Statistical Image Features (WBSIF) [5] for video copy detection was presented, proved to be efficient in terms of precision, recall and accuracy. Yet another key frame extraction method for video using robust Principal Component Analysis was presented in [6] that selected most informative frames. A content-based retrieval method using Locality Sensing Hashing (LSH) was presented in [7] based on key frames for motion pattern recognition.

Recently new approaches were developed for semantic video indexing based on visual contents. In [8], a generic semantic video indexing scheme was investigated by exploiting fuzzy knowledge resulting in satisfactory performance. A survey of video indexing methods was presented in [9]. A key frame detection algorithm using image difference and classification was investigated in [10] to realize the real time recognition of dynamic sign language.

Beside the key frame detection approach, finding a subset of important data points for Graphics Processing Units (GPU) plays an important role in handling motion feature in many applications, mainly, surveillance videos. In [11], Motion Feature-based Key Frame extraction was presented improving the accuracy of motion information being retrieved with better precision and recall. A review of key-frame extraction methods were presented in [12].

A video content representation method based on the recurring regions was presented in [13] with focus on its central visual elements, enabling efficient retrieval of video sequences. Video retrieval based on visual features was presented in [14] to re-find the shots taken during the procedure.

Content-Based Image Retrieval (CBIR) systems are available in public domain, all varied characteristics with respect to performance and features. In [15] image retrieval to support video indexing was presented from a web browser. A key frame extraction method based on unsupervised clustering and mutual comparison was presented in [16] with the aid of similarity index resulting in the improvement of concept detection rate. Feature extraction based on static and multi-resolution was presented in [17] resulting in the improvement of extraction of significant frames of interest.

A novel algorithm for content-based video indexing and retrieval was presented in [18] resulting in the improvement of average retrieval rate. Finding disturbing scenes in video was concentrated in [19].

The foremost contribution of this paper is to propose a video copy detection based on a new textural descriptor WBSIF, inspired from the former one namely WLD. Roughly speaking, the improvement here is the introduction of an efficient estimation of the local pixel contribution using BSIF. In this approach, video copy

detection uses key-frames generated from video by spatiotemporal transformation of original frames. Then, the vector feature is extracted using the proposed descriptor applied on these key frames.

III. MULTI SPECTRAL CLUSTERED SPATIOTEMPORAL FEATURE WITH GRAPH-BASED DECISION TREE INDEXING

We propose to detect spatiotemporal feature of the video based on both spectral cluster and visual content clustering. First, spectral cluster is applied on each selected video frame to extract spatial salient objects. Secondly, visual content clustering is used to validate the temporal consistency of these salient objects in adjacent video frames, thus the temporal boundary of each object is determined.

We use these spatiotemporal feature objects to describe the content of videos and index the video data using Graph-based Decision Tree Indexing model. Furthermore, Multi Spectral Clustering with the spatiotemporal feature of the video is performed by applying the Largest Frequent Feature Identification (LFFI) algorithm that is independent of the bands of data (i.e. frame) and the size of the band to describe the spatial relation of trajectories in each spatiotemporal object.

Spatiotemporal object detection

To guarantee the detection efficiency, the proposed MSSF-GDT indexing technique, shot keyframes are extracted based on the concept of locality of features and adopt spectral cluster to detect salient spatial object in each key frame. Figure shows the block diagram of spatiotemporal object detection.

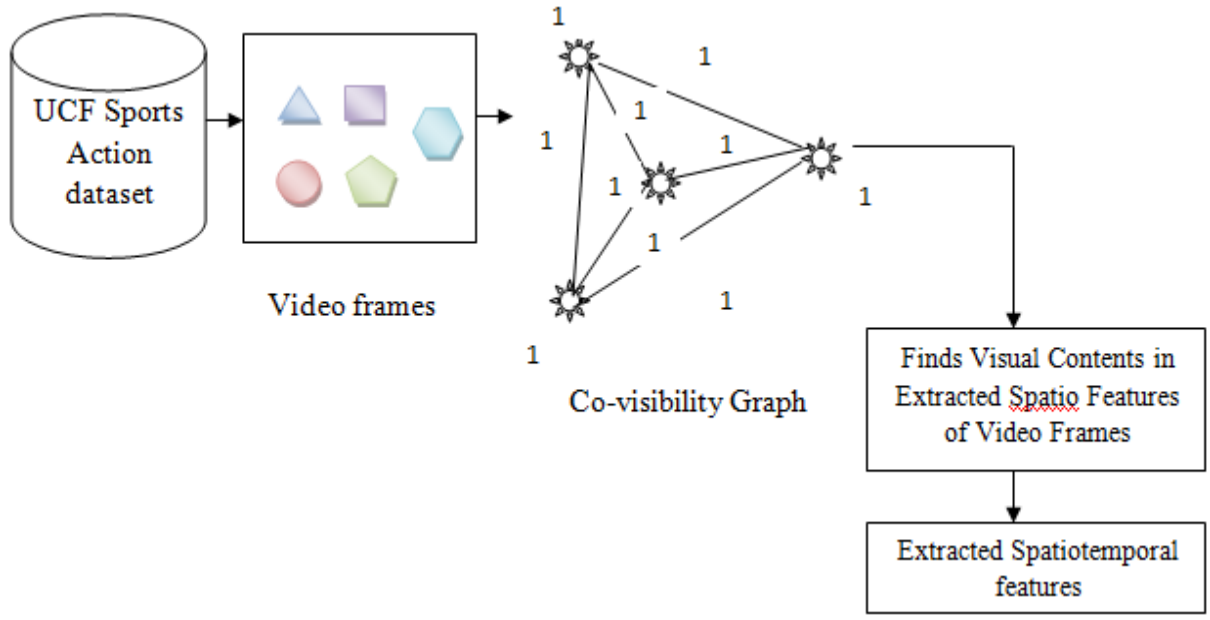

Fig.1. Block diagram of spatiotemporal object detection

As shown in the above figure 1, with the objective of extracting informative characteristics, an auxiliary graph is constructed during the scene exploration, called as the Co-visibility Graph (CovGraph). In this CovGraph, features are characterized as nodes and those features that have been observed in the same frame are associated by edges. The sets of nodes and edges are incrementally updated by including those features that were not detected in previous frames.

After detecting a set of spatial objects in the shot key frames, visual content clustering is utilized to validate the temporal consistency of these objects and generate spatiotemporal salient objects. Building visual content of points in videos is a recent topic for video content extraction and several feature extraction methods are applied to optimize the video trajectories. In this work, our method is based on visual content of video frames like, color, texture, shape and motion. It consists of extracting interest points based on these contents and tracking them in successive spatial salient regions to minimizing their distance between two points.

Graph-based Decision Tree Indexing

With the extracted spatiotemporal features, a new graph-based data structure called decision tree is used. With the resultant decision tree data structure, the proposed technique not only indexes but also retrieves similar frames or features from the video data, reducing the computational time involved in indexing. Hence, appropriate search is said to be ensured by applying Graph-based Decision Tree Indexing and therefore improvement in the video retrieval rate. Figure 2 shows the block diagram of Graph-based Decision Tree Indexing.

As shown in the figure, the block diagram of GDTI includes the extracted spatiotemporal features with three attribute selection namely, information gain, gain ratio and gini index. With extracted spatiotemporal features considered for experimentation, a decision tree is applied to the input spatiotemporal extracted video frames. A decision tree here includes, a non-leaf node representing an attribute, each branch represents an output frame of the test and each leaf node that holds a class label.

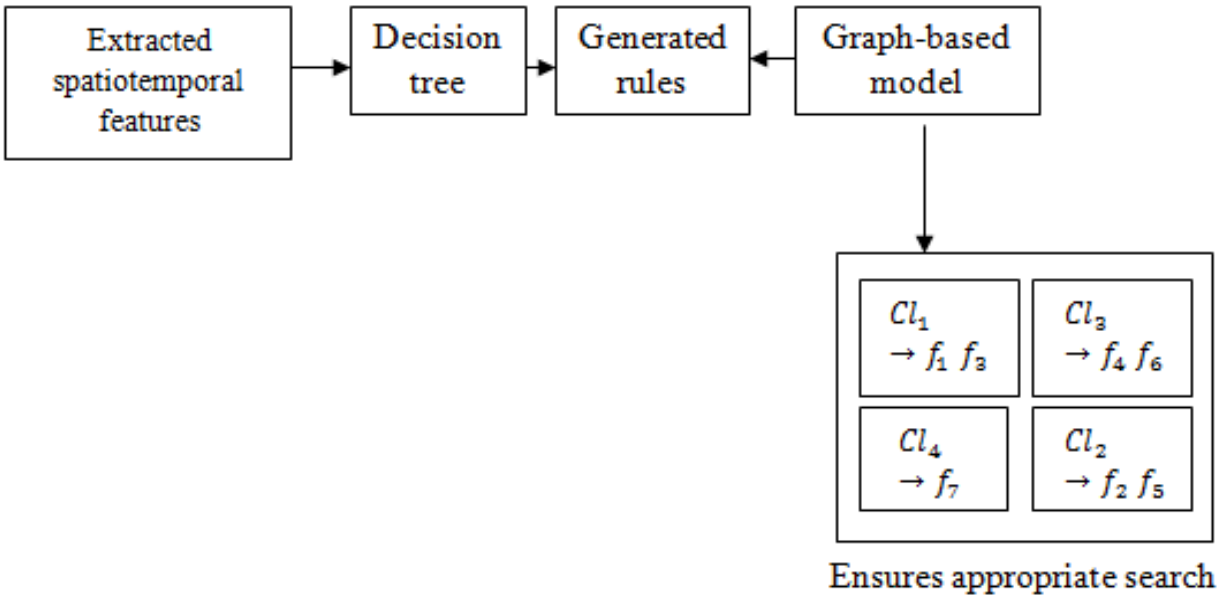

Fig.2. Block diagram of Graph-based Decision Tree Indexing

The pseudo code representation of Graph-based Decision Tree algorithm is given in algorithm 1.

Input: Video Samples ‘ ’, Classes ‘ ’, Video ‘ ’
Output: Indexed video data
1: **Begin**
2: **For** each Video Samples ‘ ’ with Video ‘ ’
3: Enumerate frames ‘ ’ in video ‘ ’
4: **End for**

5: **For** each typical frames ‘ ’ from overall frames ‘ ’
6: Measure Expected Information using eq. (1)
7: Measure Entropy using eq. (3)
8: Measure Split information using eq. (5)
9: Measure Information gain using eq. (6)
10: Select the typical frames ‘ ’ with maximal Information gain
11: Sort typical frames ‘ ’
12: **End for**
13: **End**

Algorithm 1Graph-based Decision Tree algorithm

Let us assume that ‘ ’ is the set of data samples or the video samples, the attributes of class label have ‘ ’ different value, and different classes ‘ ’. With graph-based model in MSSF-GDT technique, enumerate all the frames ‘ ’ in video ‘ ’. Select all typical frames ‘ ’ from overall frames ‘ ’ based on the typicality criterion (i.e. split information and gain ratio). Then, for a given video sample, the expected information ‘ ’ required for classification with ‘ ’ denoting the number of samples in class ‘ ’ is given by the following equation.

$$EI(vs_1, 1 \quad (1)$$

(2)

Set frame ‘ ’ with ‘ ’ different values ‘ ’.Then the video samples ‘ ’ could be divided into ‘ ’ subsets ‘ ’ by frame ‘ ’. Let ‘ ’ denote the number of the video sample of class ‘ ’ in a subset ‘ ’.The entropy ‘ ’ and information ‘ ’ expectations of the subsets divided by ‘ ’ are given by the following expression.

$$E(F) = \sum \frac{(vs_{1j}+1}{} \quad (3)$$

$$I(vs_{1j}, vs_{2j}, \dots, vs_{mj} \quad (4)$$

From (4), ‘ ’ represent the probability of video sample belonging to class ‘ ’. Then according to the split information ‘ ’ that is used to measure the frequency and uniformity of the split of video samples, the size of the information gain rate is compared in the process of the frame classification, present in each video. Here, the split information ‘ ’ and information gain ‘ ’ for each frame is expressed as given below.

(5)

(6)

The purpose of using decision tree in the MSSF-GDT technique is that to not only construct the tree but also produce the inference rules using the typicality criterion. With this, indexing is performed efficiently, therefore reducing the indexing time.

Multi Spectral Clustering

To the indexed video data, Multi Spectral Clustering is performed by applying the Largest Frequent Feature Identification (LFFI) algorithm that is independent of the bands of data (i.e. frame) and the size of the band. As the MSSF-GDT technique considers spatiotemporal objects of videos for which indexing is performed, multi spectral objects are said to exist. Hence, in the proposed work, Multi Spectral Clustering is performed, where each image of indexed object is considered as a band.

To entirely utilize the supplementary information which is present in multiple bands, the proposed work considers the video image on one multi-spectral video image rather than as a set of monochrome video images. For a video image with ‘ ’ bands, the brightness of each pixel is described as a point in a ‘ ’ space denoted by a vector of length ‘ ’. Hence, the proposed work is said to be independent of the bands of frame and size of frame.

The LFFI algorithm extracts the key frames based on region of interest and are annotated over the video. With the annotated video, comparison is made with similar frames for recognition of objects and therefore video retrieval. For example, to retrieve a key frame as belonging to one specific region, its intensities in different bands are said to form a ‘ ’ denoting its position in the ‘ ’ feature space.

Hence, a particular class is selected via an upper threshold ‘ ’ and lower threshold ‘ ’ for each band. In this way, ‘ ’ ‘ ’ is said to be determined in the feature space. Only if the feature vector of a frame points to a position within this cube, is the video classified as belonging to this class. Hence, it is said to be largest frequent feature identification. The pseudo code representation of Largest Frequent Feature Identification (LFFI) algorithm is given in algorithm 2.

Input: vector length ‘ ’, upper threshold ‘ ’, lower threshold ‘ ’, typical frames ‘ ’, triplets ‘ ’, ‘ ’ ‘ ’,
Output: video retrieval
1: **Begin** 2: **For** vector length ‘ ’ with upper threshold ‘ ’ and lower threshold ‘ ’ 3: **For** each typical frames ‘ ’ 4: **For** inclination type corresponds to an arrangement, ‘ ’ 5: **Repeat** 6: Measure raw correlation value using eq. (7) 7 Measure neighborhood motion using eq. (8) 8: **Until** (all video frames are processed) 9: **End for** 10: **End for** 11: **End for** 12: **End**

Algorithm 2 Largest Frequent Feature Identification (LFFI) algorithm

As provided in the above algorithm, the proposed technique works on group of frames extracted from a video. Key frames are identified using regions of interest where retrieval is performed on the basis of high level semantic relationship using inclination detection. Inclination detection involves a process of detecting a change in position of an object (i.e. frame) relative to its neighborhood region or change in neighbor region relative to an object.

The calculation of the inclination motion score is as follows. An inclination type corresponds to an arrangement, ' ' which is a set of spatiotemporal feature in a specific relative position. In the proposed work, the arrangement is represented as a set of triplets ' ', in which each of the ' ', ' ' and ' ' are integers and ' ' denotes the number of elements in the arrangement.

The raw correlation value for the frame ' ' at the position '( )' is defined as a product that involves all offsets contained in the frame. This is mathematically formulated as given below.

$$RawCorr(a, b, t; T) = [F(a + a_i, b + b_i, t + t_i) - F'_{sh}$$

(7)

From (7), ' ' represents the luminance of the image at the position ' ' and ' ' denotes the average luminance across the shot. An arrangement that is transposed along the ' ', which is represented as ' ', represents the arrangement in which each triplet ' ' of ' ' is replaced by ' ', where ' ' is the distance of the arrangement in the ' '. Finally, the neighborhood score at position ' ' for arrangement ' ', in direction ' ' is mathematically formulated as given below.

$$NeighMotion\ (a, b, t; F; X) = RawCorr(a, b, t; F) - RawCorr\ (a, b, t; F^x) - RawCorr$$

(8)

From (8), the neighborhood motion of the frame ' ' in ' ' for the corresponding templates '( )' is obtained using the raw correlation templates with respect to the transposed values of the frames ' '. The main function of the proposed technique is to select smaller number of key frames. If consecutive frames, lie within the threshold, (upper and lower threshold), then two frames are said to be similar and retrieve all similar frames and therefore video. The above said process is repeated till frames are similar. The process is then started with the next frame which is outside of the threshold & the above said steps are repeated for the all video frames.

IV. EXPERIMENTAL SETTINGS

In order to implement the proposed technique, UCF Sports Action Data Set [20] is used. The experimental dataset comprises a total of 150 sequences with the resolution of 720 x 480. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. The performance of MSSF-GT is evaluated using JAVA language with sports repositories data set using parameters such as, multi spectral clustering accuracy, multi spectral clustering time, true positive rate for video retrieval and video retrieval time for different user requests.

The dataset includes the following 10 actions, namely, Diving (14 videos), Golf Swing (18 videos), Kicking (20 videos), Lifting (6 videos), Riding Horse (12 videos), Running (13 videos), SkateBoarding (12 videos), Swing-Bench (20 videos), Swing-Side (13 videos) and Walking (22 videos).

In order to evaluate the retrieval performance of the proposed technique in a qualitative and quantitative manner, different sub-sets of query images representing different video samples for the entire experimental database and a number of query images in each sub-set varies such that 10, 20, 30,. . ., 100 is chosen. Table 1 summarizes the characteristics of the dataset.

Table 1 Summary of the characteristics of UCF sports dataset

Actions	10
Clips	150
Mean clip length	6.39s
Minimum clip length	2.20s
Maximum clip length	14.40s
Total duration	958s
Frame rate	10fps
Resolution	720 * 480
Maximum no of clips/class	22
Maximum no of clips/class	6

The experimental work is compared against the existing shot based keyframe extraction (S-KFE) [1] for video indexing and retrieval and Bag of System Tree (BoS Tree) [2] to identify the effectiveness of MSSF-GT technique. The performance of the MSSF-GT technique is measured in terms of multispectral clustering accuracy, multispectral clustering time, true positive rate for video retrieval, video retrieval time with respect to total number of video samples and size of video.

V. DISCUSSION

The performance of Multi Spectral clustered Spatiotemporal Feature with Graph-based Decision Tree (MSSF-GDT) for video retrieval is compared with the existing Shot based Key Frame Extraction (S-KFE) [1] for video indexing and retrieval and Bag of System Tree (BoS Tree) [2]. The performance is evaluated according to the following metrics.

Impact of multispectral clustering accuracy

This section discuss about the performance measure of multispectral clustering accuracy and comparison made with the existing methods Shot based Key Frame Extraction (S-KFE) [1] for video indexing and retrieval and Bag of System Tree (BoS Tree) [2]. Table 2 shows the result of multispectral clustering accuracy versus the varying video samples. To better perceive the efficacy of the proposed MSSF-GDT technique substantial experimental results are illustrated in Figure 3 and compared against the existing S-KFE [1] and BoS Tree [2].

Accuracy in the proposed technique is arrived at in terms of multispectral clustering accuracy. In the proposed work typicality criterion is used to measure multiple spectral clustering. Hence, multi spectral clustering is the percentage of ratio of number correctly clustered videos based on typicality criterion to the total number of video samples considered for experimentation. Here, typicality criterion involves features involving both split information and gain ratio

$$MC_{acc} = \frac{No.of\ correctly\ clustered\ videos\ (}{vs} \quad (9)$$

From (9), multispectral clustering accuracy ‘ ’ is obtained using the total number of video samples ‘ ’ and typical frames ‘ ’ respectively.

Table 2 Tabulation of multispectral accuracy

Video samples	*Multispectral clustering accuracy (%)*		
	MSSF-GDT	**S-KFE**	**BoS Tree**
10	82.15	77.10	71.11
20	88.90	83.85	77.86
30	93.16	88.11	82.12
40	96.78	91.73	85.74
50	85.18	80.13	74.14
60	90.24	85.19	79.20
70	93.15	88.10	82.11
80	81.76	76.71	70.72
90	88.98	83.93	77.94
100	95.14	90.09	84.10

Results are presented for different number of video samples with differing resolution for video retrieval. The multispectral clustering accuracy on several video samples with varying resolutions is shown below. The results reported here confirm that with the increase in the number of video samples, multispectral clustering accuracy is not found to be linear. It first increases for 40 video samples, followed by a fall in curve and so on. The process is repeated for 10 different video samples. The observance of non-linearity is due to the size different in different video samples and also the presence of noise.

In order to investigate the multispectral clustering accuracy by video files to perform video retrieval while keeping up with video samples, we simulated verifying both S-KFE and BoS Tree for different implementation runs. As illustrated in Figure 3, the proposed MSSF-GDT technique performs relatively well when compared to two other methods S-KFE [1] and BoS Tree [2]. The multispectral clustering accuracy is improved in the proposed MSSF-GDT technique by extracting the object based on spatiotemporal characteristics. By detecting the object using spatiotemporal characteristics, locality of features and visual content of points in videos are considered. By minimizing the distance between two points for a given video file using feature locality and visual content, in turn results in the improvement of detection, and therefore the accuracy rate

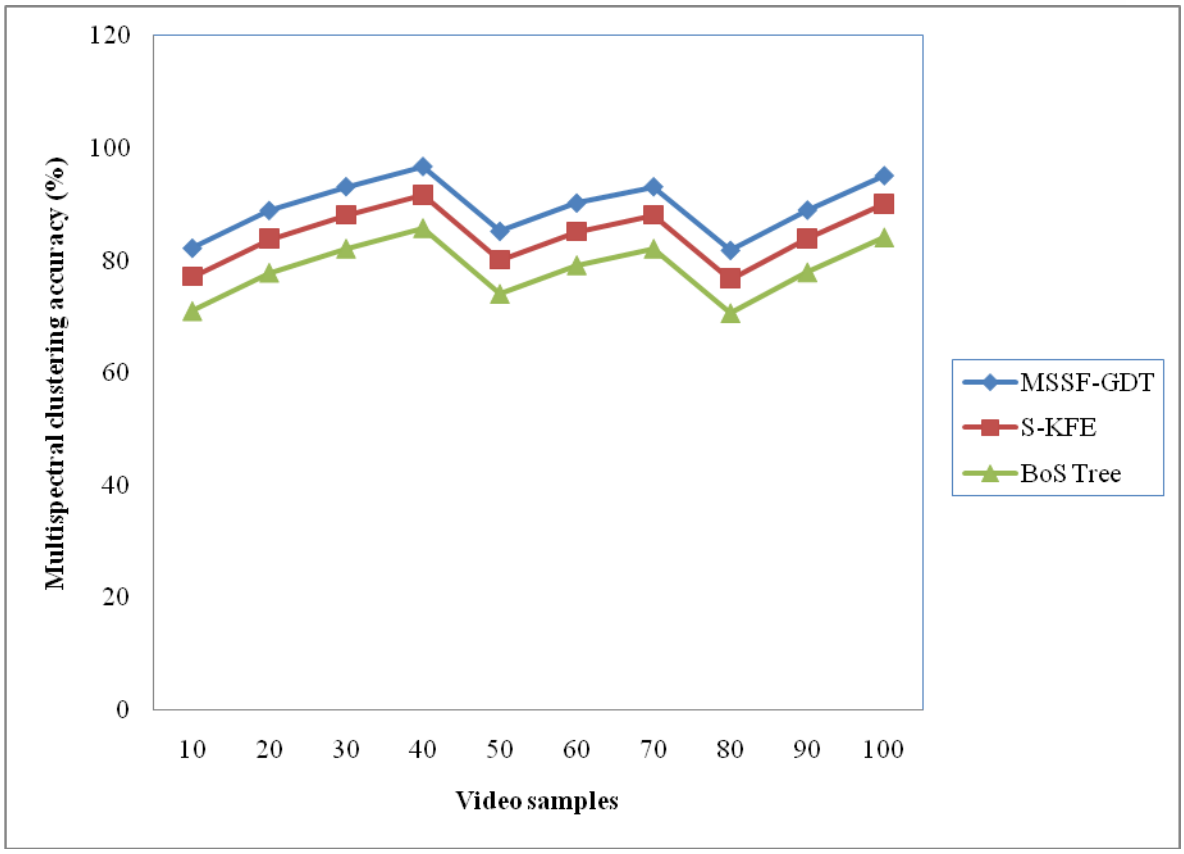

Fig.3. Measure of multispectral clustering accuracy

In order to investigate the multispectral clustering accuracy by video files to perform video retrieval while keeping up with video samples, we simulated verifying both S-KFE and BoS Tree for different implementation runs. As illustrated in Figure 3, the proposed MSSF-GDT technique performs relatively well when compared to two other methods S-KFE [1] and BoS Tree [2]. The multispectral clustering accuracy is improved in the proposed MSSF-GDT technique by extracting the object based on spatiotemporal characteristics. By detecting the object using spatiotemporal characteristics, locality of features and visual content of points in videos are considered. By minimizing the distance between two points for a given video file using feature locality and visual content, in turn results in the improvement of detection, and therefore the accuracy rate.

Moreover, the multispectral clustering accuracy in MSSF-GDT technique is improved by removing the dissimilar features with the application of Co-visibility Graph which incrementally updates those features that were detected in previous frames. Hence, the rate at which the clustering is performed is said to be improved using MSSF-GDT method by 6% compared to S-KFE [1]. Moreover, in MSSF-GDT method using visual content clustering, according to the visual content of video frames, the spatiotemporal object detection detects the frames resulting in improving accuracy by 14% compared to BoS Tree.

Impact of multispectral clustering time

In order to reduce the complexity during clustering and measure the efficiency of multispectral clustering accuracy for different video samples towards video retrieval, the time taken to perform multispectral clustering

using distinguished video samples is considered. In the experimental setup the total number of video samples considered ranges from 10 to 100 is provided in Table 3. The complexity on clustering time using the technique MSSF-GDT provides comparable values than the state-of-the-art methods.

The multispectral clustering time in other words refer to the time taken to perform multispectral clustering with respect to the total number of video samples ‘ ’ considered for experimentation. In the proposed technique, multispectral clustering for each video frame is conducted using LFFI algorithm based on the neighborhood motion ‘ ’. Therefore, multispectral clustering time in the proposed method is measured as given below.

(10)

From (10), lower the multispectral clustering time ‘ ’ with respect to the number of video sample considered ‘ ’, more efficient the method is said to be.

Table 3 Tabulation of multispectral clustering time

Video samples	*Multispectral clustering time (ms)*		
	MSSF-GDT	*S-KFE*	*BoS Tree*
10	6.09	10.12	17.11
20	8.23	22.43	33.32
30	12.53	30.23	46.54
40	16.21	36.64	60.35
50	14.74	30.32	55.33
60	25.12	50.64	86.12
70	32.32	53.23	93.65
80	36.54	62.12	95.43
90	40.12	67.43	99.21
100	45.75	73.22	106.27

A comparative analysis for multispectral clustering time with respect to different video samples was performed and compared with the existing S-KFE and BoS Tree is shown in Figure 4.

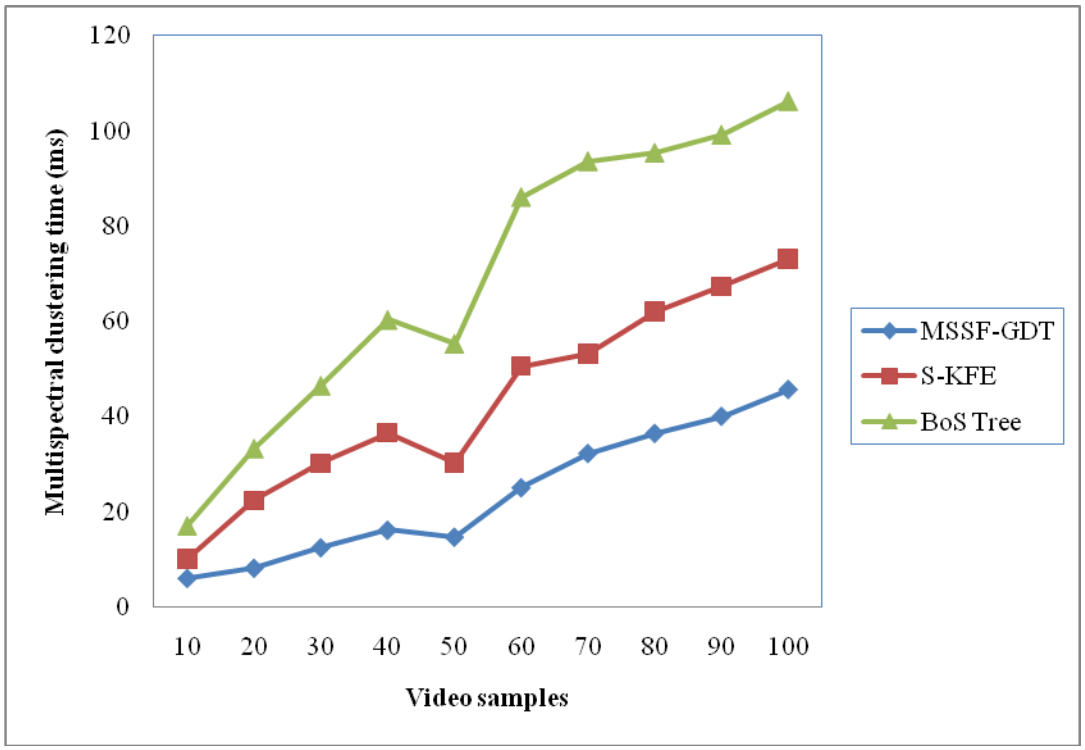

Fig.4. Measure of multispectral clustering time

The increasing video samples of 10 to 100 are considered for experimental purpose. As illustrated in figure 4, comparatively while considering video samples of increasing order, multispectral clustering time also increases. However, with the video samples of 50, minimum change in clustering time was observed, though betterment achieved using the proposed technique MSSF-GDT. This is because of the size of video considered is different for different videos.

The targeting results of distinguished video samples to measure the multispectral clustering time using MSSF-GDT technique is presented. This is compared against the two state-of-the-art methods S-KFE [1] and BoS Tree [2] and provided in figure 4 is presented for visual comparison based on different video samples of different actions. Our technique differs from the S-KFE and BoS Tree in that we have incorporated inclusion of graph-based data structure to the extracted spatiotemporal objects. The graph-based data structure used in the proposed technique provides management of database like indexing with the help of decision tree. With the aid of the decision tree, the proposed technique only indexes similar frames which in turn reduces the multispectral clustering time using MSSF-GDT method by 48% compared to S-KFE and 67% compared to BoS Tree respectively.

Impact of true positive rate for video retrieval

Table 4 below shows the true positive rate for video retrieval for MSSF-GDT technique, S-KFE and BoS Tree versus ten different video samples. The true positive rate for video retrieval over S-KFE and BoS Tree increases gradually though not linear for differing video files. To measure the true positive rate, three measures are required, namely, number of hits, number of missed hits and number of false hit. A correctly detected shot is called a hit, a not detected shot is called a missed hit and a falsely detected shot is called a false hit. With this, the true positive rate for video retrieval is measured mathematically as given below.

(11)

As given above (11), the true positive rate ‘ ’ is measured in terms of percentage (%).

Table 4 Tabulation for true positive rate

Video samples	*True positive rate for video retrieval (%)*		
	MSSF-GDT	*S-KFE*	*BoS Tree*
10	90.43	83.43	80.40
20	91.12	83.62	81.09
30	91.88	85.16	81.85
40	83.49	76.64	73.46
50	93.96	89.13	85.92
60	94.23	90.02	84.19
70	85.81	81.32	75.76
80	96.44	91.95	86.40
90	96.88	92.20	87.84
100	97.60	93.82	87.56

Table 4 shows the true positive rate for MSSF-GDT method, S-KFE and BoS Tree versus ten different video samples using different video images.

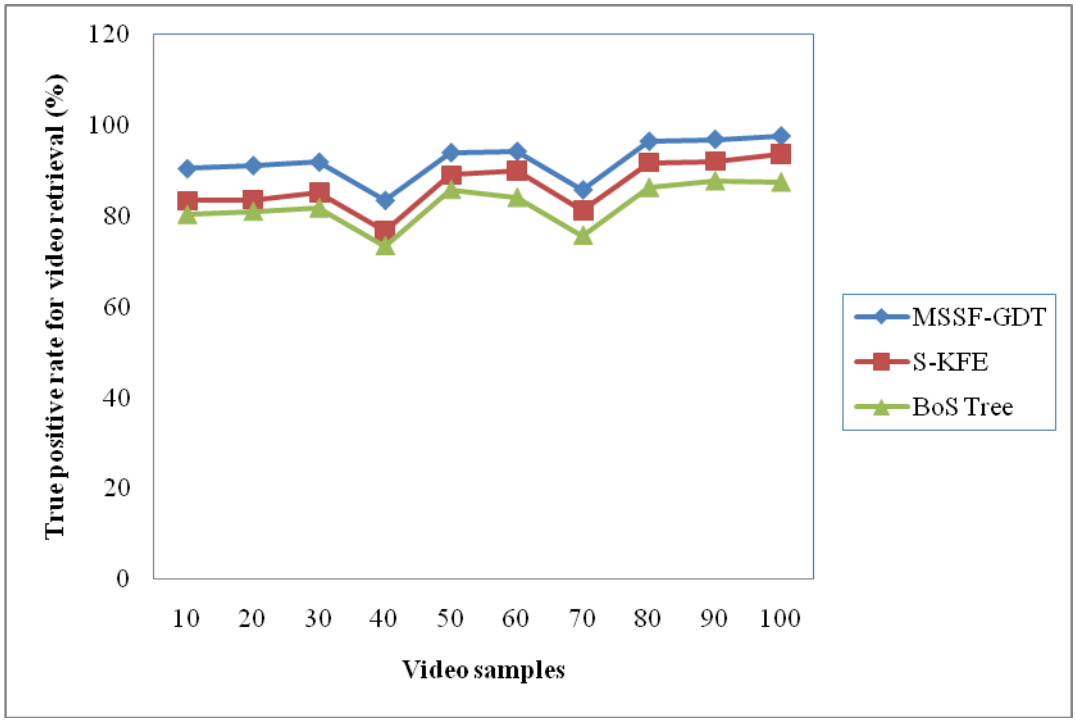

Fig.5. Measure of true positive rate

From figure 5, it is illustrative that the true positive rate is improved using the proposed technique MSSF-GDT. For example, when the number of video samples was 50, the true positive rate was 93.96% using MSSF-GDT, 51 percent compared to S-KFE and 73 percent compared to BoS Tree. Also with 90 video samples, the true

positive rate for video retrieval was 41% better compared to S-KFE and 60% compared to BoS Tree respectively.

By observing the dense video frame behavior with differing actions of video, the true positive rate for video retrieval is improved. This is because with the application of Graph-based Decision Tree Indexing algorithm selects the typical frame from the overall frame based on the typicality criterion. The typicality criterion considered in the proposed technique is split information and information gain for each frame. This typicality criterion used in the proposed technique constructs the tree reducing the search space and also produces inference rules. As a result, the true positive rate for video retrieval is said to be improved by 6% compared to S-KFE and 12% compared to BoS Tree respectively.

Impact of video retrieval time

The time taken to retrieve the video plays a main role in video surveillance, monitoring terrorism and so on. Lower the time taken to retrieve the video, more efficient and effective the method is said to be.

$$VR_{time} \quad (12)$$

From (12), the video retrieval time ‘ ’ is measured using the video size ‘ ’. Table 5 given below shows the tabulation for video retrieval time. In this section to check the efficiency of MSSF-GDT technique, the metric video retrieval time is evaluated and compared with the state-of-the-art methods, S-KFE [1] and BoS Tree [2] and is measured in terms of milliseconds (ms).

Table 5 Tabulation for video retrieval time

Video size (MB)	*Video retrieval time (ms)*		
	MSSF-GDT	*S-KFE*	*BoS Tree*
113.6	4.13	9.52	14.12
323.7	8.32	11.22	14.26
349.5	10.43	14.81	18.39
454.5	11.12	16.35	23.45
635.2	15.33	18.53	30.56
905.3	18.22	22.76	33.69
936.2	20.66	24.67	37.73
970.6	22.12	28.48	41.88
1000.1	24.23	32.92	44.96
1040.7	27.54	39.83	51.03

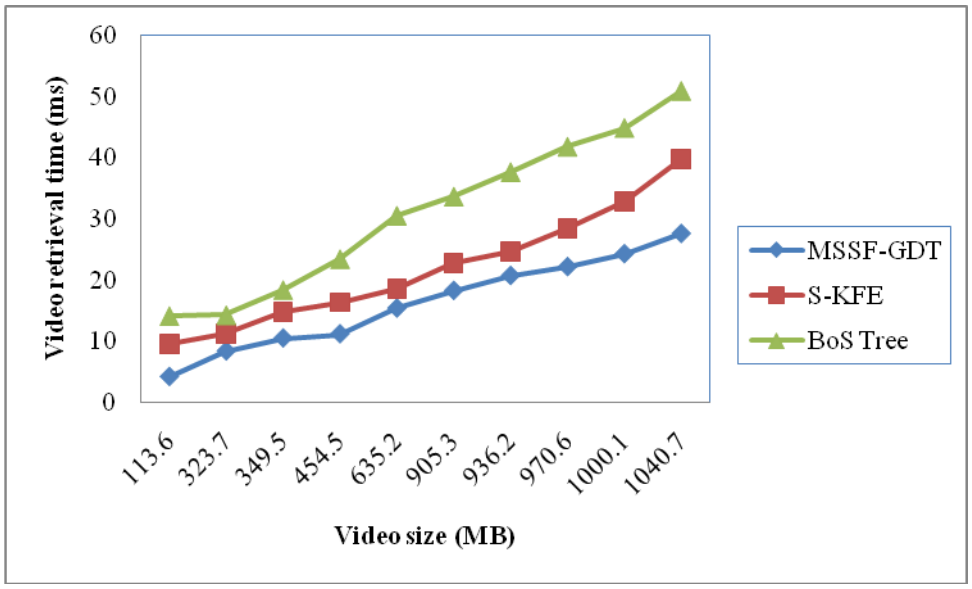

Fig.6. Measure of video retrieval time

From Figure 6 it is clear that the MSSF-GDT technique performs better than S-KFE [1] and BoS Tree [2]. In MSSF-GDT technique, with an increase in video size, the video retrieval time also increases. With the construction of Largest Frequent Feature Identification (LFFI) algorithm, regions of interest with high level semantic relationship are considered whenever key frames have to be identified. Inclination detection using the algorithm is measured on the basis of neighborhood region or neighborhood motion of the frame. This in turn helps in improving the video retrieval time by 28% compared to S-KFE. In addition, smaller number of key frames is selected where frames lie within the lower and upper threshold. As a result, better performance is provided and therefore the video retrieval time is improved by 49% compared to BoS Tree. The advantage of the Largest Frequent Feature Identification algorithm is that its computational complexity is independent of the bands of data (i.e. frame) and the size of the band.

VI. CONCLUSION

Storage and retrieval of video data has become an important paradigm for video surveillance, terrorism and so on. Currently, there are many video retrieval methods that offer different methods with different performance attributes. With the growing number of video retrieval methods, it has also become challenging to apply it while considering the time and rate at which the retrieval is said to take place. Therefore, Multi Spectral Clustered Spatiotemporal Feature with Graph-based Decision Tree (MSSF-GDT) indexing is investigated to improve the video retrieval rate and time based on spatiotemporal video objects. In this context, this work presents the Graph-based Decision Tree Indexing to systematically measure the indexing rate and time using a novel Graph-based Decision Tree algorithm. The method also addresses key challenges related to computational time involved in retrieval of video files based on the co-visibility graph and visual contents. An algorithm, Largest Frequent Feature Identification (LFFI) is presented to improve the true positive rate of video retrieval based on the neighborhood motion of the frame. Experiments conducted using the UCF sports dataset shows that the MSSF-GDT outperforms in terms of video retrieval rate and time when compared to the state-of-the-art methods.

REFERENCES

[1] G.G. Lakshmi Priya, S. Domnic, "Shot based keyframe extraction for ecological video indexing and retrieval", Ecological Informatics, Elsevier, Sep 2013

[2] Adeel Mumtaz, Emanuele Coviello, Gert R.G. Lanckriet, and Antoni B. Chan, "A Scalable and Accurate Descriptor for Dynamic Textures Using Bag of System Trees", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 37, NO. 4, APRIL 2015

[3] K. Seetharaman, S. Sathiamoorthy, "A unified learning framework for content based medical image retrieval using a statistical model", Journal of King Saud University – Computer and Information Sciences, Elsevier, May 2015

[4] Ran He, YinghaoCai, TieniuTan, LarryDavis, "Learning predictable binary codes for face indexing", Pattern Recognition, Elsevier, April 2015

[5] Aissa Boukhari, Amina Serir, "Weber Binarized Statistical Image Features (WBSIF) based video copy Detection", Journal of Visual Communication and Image Representation, Elsevier, Oct 2015

[6] Chinh Dang and Hayder Radha, "RPCA-KFE: Key Frame Extraction for Video using Robust Principal Component Analysis", IEEE Transactions on Image Processing (Volume: 24, Issue: 11, Nov. 2015)

[7] Gregory Casta˜n´on, Mohamed Elgharib, Venkatesh Saligrama, and Pierre-Marc Jodoin, "Retrieval in Long Surveillance Videos using User-Described Motion & Object Attributes", IEEE Transactions on Circuits and Systems for Video Technology (Volume: 26, Issue: 12, Dec. 2016)

[8] Nizar Elleuch & Anis Ben Ammar & Adel M. Alimi, "A generic framework for semantic video indexing based on visual concepts/contexts detection", Multimedia Tools and Applications, Springer, Apr 2014

[9] Muhammad Nabeel Asghar, Fiaz Hussain, Rob Manton, "Video Indexing: A Survey", International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 03 – Issue 01, January 2014

[10] Li Shurong, Huang Yuanyuan, Hu Zuojin and Dai Qun, "Key Frame Detection Algorithm based on Dynamic Sign Language Video for the Non Specific Population", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.8, No.12 (2015), pp.135-148

[11] Ran Zheng, Chuanwei Yao, Hai Jin, Lei Zhu, Qin Zhang, Wei Deng, "Parallel Key Frame Extraction for Surveillance Video Service in a Smart City", PLOS ONE | DOI:10.1371/journal.pone.0135694 August 18, 2015

[12] Lino Ferreira, Luis A. da Silva Cruz and Pedro Assuncao, "Towards key-frame extraction methods for 3D video: a review", EURASIP Journal on Image and Video Processing, May 2016

[13] Lukas Diem and Maia Zaharieva," Video Content Representation Using Recurring Regions Detection", Springer, May 2016

[14] Jennifer Roldan Carlos, Mathias Lux, Xavier Giro-i-Nietoy, Pia Munoz and Nektarios Anagnostopoulos, "Visual Information Retrieval in Endoscopic Video Archives", International Workshop on Content-based Multimedia Indexing, July 2015

[15] Gabriel de Oliveira Barra, Mathias Lux, Xavier Giro-i-Nieto, "Large Scale Content-Based Video Retrieval with LIvRE", International Workshop on Content-based Multimedia Indexing, June 2016

[16] Nitin J. Janwe, Kishor K. Bhoyar, "Video Key-Frame Extraction using Unsupervised Clustering and Mutual Comparison", International Journal of Image Processing (IJIP), Volume (10) : Issue (2) : 2016

[17] J. Kavitha, Dr.P.Arockia Jansi Rani, "Static and Multiresolution Feature Extraction for Video Summarization", Procedia Computer Science 47 (2015) 292 – 300

[18] M.Ravinder and T.Venugopal, "Content-Based Video Indexing and Retrieval using Key frames Texture, Edge and Motion Features", International Journal of Current Engineering and Technology, Vol.6, No.2 (April 2016)

[19] Florian Eyben, Felix Weninger, Nicolas Lehment, Bjorn Schuller, Gerhard Rigoll, "Affective Video Retrieval: Violence Detection in Hollywood Movies by Large-Scale Segmental Feature Extraction", PLOS ONE, December 2013 | Volume 8 | Issue 12 | e78506

[20] Khurram Soomro and Amir R. Zamir, "Action Recognition in Realistic Sports Videos, Computer Vision in Sports", Springer International Publishing, 2014