Communications on Applied Electronics (CAE) - ISSN : 2394 - 4714
Foundation of Computer Science FCS, New York, USA
Volume 5 - No. 2, May 2016 - www.caeaccess.org

A Framework for Event Classification from Video Sequences using Bayesian Neural Network

Putte Gowda D.
Research Scholar
P E T Research Center
Mandya, India

Padma M. C.
Professor and Head
Dept. of CSE, P E S College of Engineering
Mandya, India

ABSTRACT

Due to immense growth happened in multimedia field, research with videos and images have been received significant attention among the researchers. The automatic detection of the events presented in the video content may provide more useful information to the target audience. The motivation behind this approach is to design and develop a system for video event classification through video content analysis method using Bayesian neural network. Initially, the background from the video frames is estimated which is then segmented. Subsequently, features are extracted from the tracked objects and are classified to normal or abnormal event by applying Bayesian neural network classifier.UCSD Anomaly Detection Datasets is used for implementation. The performance of the proposed system is validated through the ROC curves and classification accuracy. From the comparative analysis made, the proposed technique obtained better results.

Keywords

Bayesian Neural Network, Feature extraction, Object Segmentation, Tracking, Video Classification

1. INTRODUCTION

Tremendous amounts of both image and video data has been continuesly uploaded to visual content databases of internet.The content for these databases is generally created in realistic settings from science and technology, social events, sports, news coverage and so on. In computer vision research in the recent years, this type of data has fast become the experimental data of choice because it encompasses large inter- and intra-class variability and presents very interesting challenges for problems like video event classification, detection and tracking, face recognition, human activity analysis, and so on [8].

Video event detection is a difficult problem in video content analysis and there are some challenges like bridging the semantic gap between high-level semantic features and low-level features. In general, based on the scene analysis technology, the event-based searching can be done. The scene analysis technology may be roughly categorized into the frame-based and the shot-based methods [7]. The task of activity recognition is to bridge the gap between a high-level abstract activity description and numerical pixel level data. A common approach involves the features of a moving object from image sequences are detected and tracked first. The goal of this step is to transform pixel level data into low-level features for activity analysis. From the tracked features, the type of moving objects and their spatio-temporal interaction are then analyzed. There are several challenges that need to be addressed to achieve this task [5]: Some methods make use of the multimodal features in classification of events by using temporal sequence representation. But, they suffer from a multiple-recognition problem.

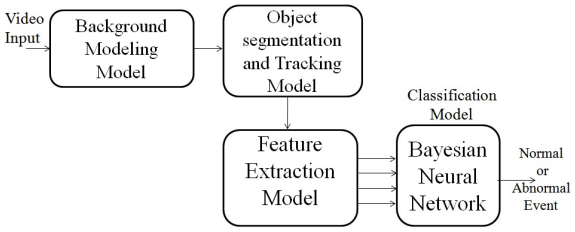

Fig. 1. overview of the proposed method

With probabilistic model and maximum likelihood decision making, it is hard to express the temporal relationship among multimodal features. Because of the temporal relationship is not fully utilized that leads to poor recognition performancesince since the feature models are independent of each other. [6].

Recently, many approaches have been proposed for event recognition which can be roughly categorise into two approaches namely , semantic approach and probabilistic approach. Semantic approaches are based on defining rules to model the events [3]. However, current approaches only describe a small portion of semantics, they do not indicate the appropriate recognition strategies and also they do not consider the event definitions and uncertainty inherent to low-level observations. But, compared to the semantic approaches, the probabilistic approaches have shown a superior performance [2]. They accurately achieved high precision within a domain and allowing an intrinsic uncertainty handling by learning the event models from training data. However, their usage is limited for different albeit related domains and they are not able to model complex relations. Therefore,in order to reslove these limitations, a combination of both approaches would be desirable. [4].
In this paper, Bayesian neural network for video event classification is proposed. The process consists of four modules like back ground modeling module, object segmentation and tracking module, feature extraction module and last module is classification module. The overview of the proposed method is shown in Fig. 1. The background from the video frames is estimated by the use of mode technique and shot segmentation technique in back ground modeling module. In object segmentation and tracking module, object is segmented using the FCM operator and each object in an image is tracked by searching for an object in subsequent image using neighborhood estimation among frames. Features are extracted from the tracked objects which include distance measure, size, number of pixels and histogram in the feature extraction module . In classification module, the final operation is carried out where the frames are classified to normal or anomaly using feed forward Bayesian Neural Network.

The rest of the paper is structured as follows: Section 2 gives the Related Works. Section 3 gives the proposed approach. Section 4 gives Results and Discussion. Conclusion is summed up in Section 5.

2. RELATED WORKS

Despite numerous efforts in video content analysis, it remains a major challenge in terms of effectively integrating the multiple physical features to deduce the semantic events due to the well-known semantic gap. In response to this matter, some efforts in the research have been directed to extend the basic content analysis methods with the facilitation of more supervised approaches like heuristic method [9], E-R model [10], and Hidden Markov Model (HMM) [11]. In [9], a set of fixed rules is derived on the basis of the multimodal cues. But, the derivation process becomes infeasible with the increment of the number of multimodal features. In addition to this, the fixed thresholds adopted in the rules are not general enough for a large number of video samples. In [10], Tovinkere and Qian proposed a hierarchical E-R model on the basis of 3D data of the locations of players and ball. Here, it is trying to model the domain knowledge and semantic meaning for the soccer games. Thereafter, a set of rules are generated to determine the occurrence of the event. However, the 3D information is not available in the video data, the generalization of this work is highly limited. In [11], a method Hidden Markov Model (HMM) was proposed to detect and recognize soccer highlights. Here, each model is trained separately for each type of event. this method as mentioned in the preliminary results can detect and recognize penalty and free kick events. However, it has the problem to deal with long video sequences.

Almost all the classification methods adopted model-based approaches in the video event detection field, which present some good qualities when the video data are with a high level of consistency. Therefore, A generalized framework for video event classification proposed which gives better performance compared to other methods.

3. PROPOSED APPROACH

This section describes the proposed a tecnique that involves the various modules to classify the video event by analyzing the video content. The technique consists of four modules, namely back ground modeling module, object segmentation and tracking module, feature extraction module and classification module. The main objective of the technique is to classify the video/frame as anomaly or normal.

3.1 Background Modeling Module

In this module, the background from the video frames is estimated by the use of mode technique. The video clip is initially split into frames and the important frames are found out using the shot segmentation technique. Wavelet Transform and distance measure based shot segmentation is employed in this paper. Subsequently, back ground is extracted by taking the mode of pixel value series at each image location which forms the background. Block diagram of the background modeling module is given in Fig. 2.

Back ground is extracted by taking the mode of pixel value series at each image location which forms the background after the shot segmentation is performed,. Mode at any image location is the pixel value that occurs most frequently at that location. Since the background is motionless, during the entire analysis time the color values of this pixel would approximately be the same. For the foreground, the moving vehicles occupying the pixel may be in different colors and shapes at different times. So, it should

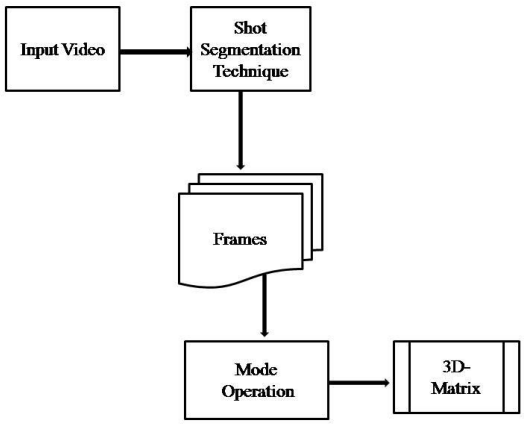

Fig. 2. Block diagram of background modeling module

be the background-pixel color vector that occurs the most frequently in the frame. The obtained pixels are stored in the form of 3- dimensional matrix.

3.2 Object Segmentation and Tracking Module

In this module, object is segmented using the FCM operator. Initially, foreground is obtained which is then sharpened and black and white converted. Subsequently FCM operation is carried out and areas lesser than a threshold is removed. Thereafter, each object in an image is tracked by searching for an object in subsequent image of the video clip that overlaps most with the given object using neighbourhood estimation among frames. The block diagram of the object segmentation and tracking module is given in Fig. 3.

By subtracting the back ground from each image in the video clip the foreground is obtained. The logic in the approach is that of detecting the moving objects from the difference between the current frame and a reference frame, also called background image. In this case, frame differencing as the foreground detection technique is used. Let the input video sequence be defined as $Vid[x, y, t]$ where x and y defines the pixel position and t gives the time. Then the frame difference at a time $t + 1$ is given by:

$$F_{dis}(t+1) = |Vid[x, y, t+1] - Vid[x, y, t]| \qquad (1)$$

The background is assumed to be the frame at time t. After obtaining the foreground, it is sharpened by the use of laplacian filter. Subsequently, it is converted to gray scale format.
There exist many clustering algorithms in data mining and Fuzzy C-means (FCM) is one among them. FCM gives more accurate clustering results with the inclusion of fuzzy concept when compared to K-means. Minimization objective function of FCM is defined by:

$$S_i = \sum_{i=1}^{N} \sum_{j=1}^{cen} \mu_{ij}^{n} ||z_i - cen_j||^2 \qquad (2)$$

Here, μ_{ij} is the membership degree function of i^{th} D-dimensional measured data (D_i) in cluster j . Centroid of j^{th} cluster is represented by cen_j. The minimization function of FCM is iteratively optimized for clustering by updating of membership function μ_{ij} and centroid cen_j on every iteration. The updating equations are given by:

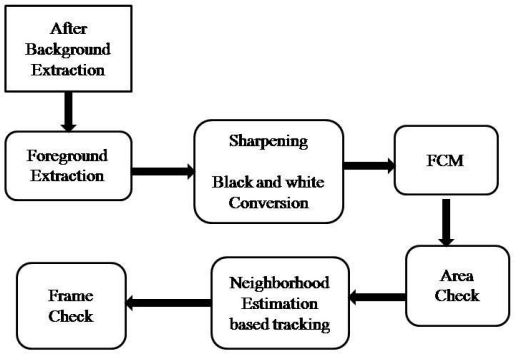

Fig. 3. Block diagram of object segmentation and tracking module

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{cen} \left(\left\| \frac{z_i - cen_j}{z_i - cen_k} \right\| \right)^{\frac{2}{m-1}}} \quad (3)$$

Where, z_i is the input data, is the centroid cen and m is a positive constant. The centroid under investigation is cen_j and other centroids are represented by cen_k. The centroid updating equation is given by:

$$cen_j = \frac{\sum_{i=1}^{N} \mu_{ij} z_i}{\sum_{i=1}^{N} \mu_{ij}} \quad (4)$$

3.3 Feature Extraction Module

In this module, features are extracted from the tracked objects. The extracted features include distance measure, size, number of pixels and histogram. Distance measure takes the distance between centroid of objects in subsequent frames (or images) in clip in x and y directions. Here, initially centroids of the objects are found out in each frame. Let the objects in the frame be represented by Ob. Let the pixels inside the i^{th} object are represented by poi_j where $0 < j < n$ and the respective centroid is calculated as:

$$cen_i = \frac{\sum_{j=1}^{n} poi_j}{n} \quad (5)$$

After finding out the centroid for each object for each frame, Euclidean distance is taken between centroid of objects in subsequent frames. This distance found out forms a feature.

Another feature taken is the size of bounding box of object in x and y directions. The minimum or smallest bounding or enclosing box is a term used in geometry. In this case, the bounding box for two dimensions (x and y) is found out.

Histogram value is the fourth feature taken. Histogram is a graphical representation of the distribution of image pixels. Histogram peak value is the maximum histogram value in the taken interval. Histogram value is the histogram peak value divided by total number of pixels in object image after removing background pixels.

3.4 Classification Module

In this module, using the features extracted in the previous model, the final operation is carried out where the frames are classified to normal or anomaly. The classification is carried out using feed forward Bayesian Neural Network. The clips are divided into 2;80% for training and 20% for testing in the neural networks.

Let the net input to i^{th} unit in $(m+1)^{th}$ layer be represented as:

$$inp^{k+1}(m) = \sum_{i=1}^{sk} \varpi^{m+1}(m,i) y^k(i) + x^{k+1}(m) \quad (6)$$

The output of the m^{th} unit can be defined by:

$$x^{k+1}(m) = G^{k+1}(inp^{k+1}(m)) \quad (7)$$

The main duty of the network is to learn associations between a specified set of input-output pairs $\{(g1, \tau_1), (g2, \tau_2), ...(gn, \tau_n)\}$. The performance index for the network is given by:

$$F = \frac{1}{2} \sum_{p=1}^{p} (\tau_p - y_p^{N_s})^T (\tau_p - y_p^{N_s}) = \frac{1}{2} \sum_{p=1}^{p} er_p^T er_p \quad (8)$$

Here, $y_p^{N_s}$ is the output of for p^{th} input and $er_p = \tau_p - y_p^{N_s}$ is the error for p^{th} input. In back propagation, the performance index is approximated by the steepest descent rule and can be defined as:

$$F = \frac{1}{2} \sum_{p=1}^{p} er_p^T er_p \quad (9)$$

Total sum of squares in steepest decent approximation, is replaced by the squared errors for a single input-output pair. Hence,

$$\Delta \varpi^k(m,i) = -\alpha \frac{\partial F}{\partial \varpi^k(m,i)} = -\alpha \gamma^k(m) y^{k-1} \quad (10)$$

$$\Delta y^k(m) = -\alpha \frac{\partial F}{\partial z^k(m)} = -\alpha \gamma^k(m) \quad (11)$$

Where,

$$\gamma^k(m) = \frac{\partial F}{\partial inp^k(m)} \quad (12)$$

Here,α is the learning rate and γ is the sensitivity of the performance index. It can also be shown that sensitivities satisfy the relation:

$$\gamma^k = G^k(inp^k) \varpi^{k+1^T} \gamma^{k+1} \quad (13)$$

The learning is carried out for about 80% of the frames and rest 20% is used for testing. In both learning and testing, the initial process of three modules is carried out as pre-processing to the networks. After the learning stage, the network is fed with the test data frames to classify as normal or anomaly.

4. RESULT AND DISCUSSION

In this section, the results achieved by the proposed technique is analysed.

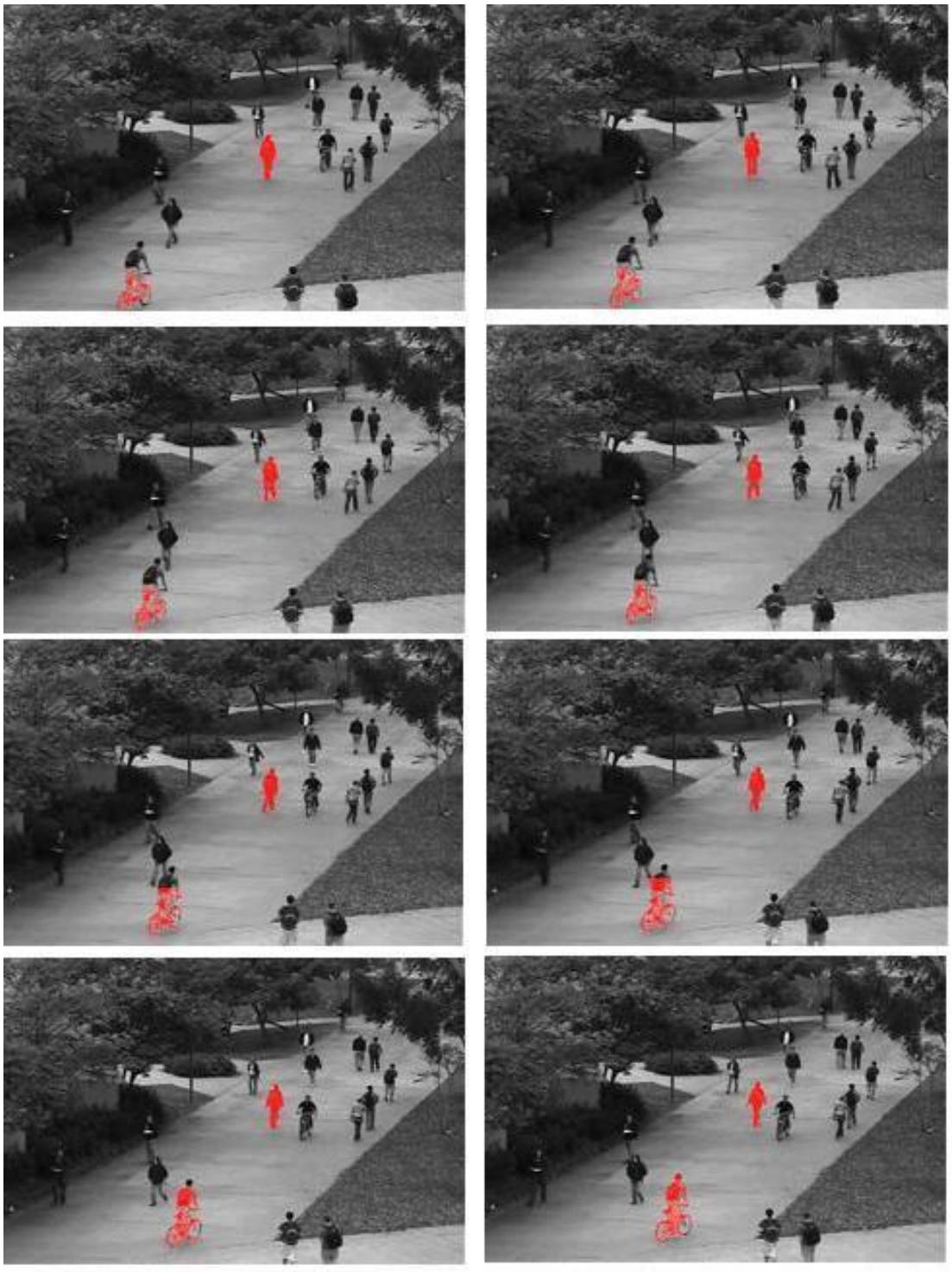

Fig. 4. Simulation results

4.1 Database Used

The database used includes UCSD Anomaly Detection Datasets [1]. The UCSD Anomaly Detection Dataset is used and it was acquired with a stationary camera mounted at an elevation, overlooking pedestrian walkways. The crowd density was variable in the walkways , ranging from sparse to very crowded. In the normal setting, the video contains only pedestrians. Abnormal events are due to either non-pedestrian entities circulation in the walkways or due to anomalous pedestrian motion patterns. The anomalies commonly occurring like bikers, skaters, small carts, and people walking across a walkway or in the grass that surrounds it.

The data was split into 2 subsets namely Peds1 and Peds2, each corresponding to a different scene. The recorded video footage from each scene was split into various clips with 200 frames. The first Peds1 subset include clips of groups of people walking towards and away from the camera and some amount of per-

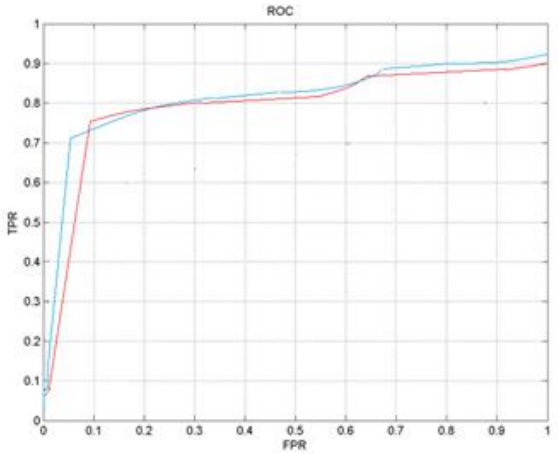

Fig. 5. ROC Curves

spective distortion. It consists of 34 training video samples and 36 testing video samples. Peds2 include scenes with pedestrian movement parallel to the camera plane. It contains 16 training video samples and 12 testing video samples.

4.2 Simulation Results

Fig. 4. gives the simulation results. The figure gives the first eight images of the Peds1 dataset. The objects in red give the found out non pedestrian entities in the walkways.

Table 1. Comparative table for accuracy(%)

Threshold set	Proposed Approach	Base Paper
0.1	24.7	21.6
0.2	28.2	29.4
0.3	53.8	46.9
0.4	74.6	67.8
0.5	88.6	87.2
0.6	90.1	89.3
0.7	93.7	89.4
0.8	86.7	85.2
0.9	82.1	76.7
1.0	79.7	69.7

4.3 Comparative Analysis

In this section, the proposed technique is compared with the other technique [12]. The proposed technique average accuracy and for the other technique [12] is given in table 1. The values are given for different threshold levels set.

From the table, it can infer that the proposed technique has performed well obtaining better accuracy than the other technique [12]. The evaluation metrics employed are accuracy and ROC (Region of Convergence). In a ROC curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. Fig.5. gives the ROC curves obtained for proposed technique and the other [12] technique. Blue line indicates the average ROC curve for proposed technique and red gives the average ROC for the other [12] technique. From the figure, it can infer that the proposed technique has achieved better ROC curve.

5. CONCLUSION

In this paper, The video event classification by analyzing the video content using Bayesian neural network is proposed. The main objective of the technique is to classify the video/frame as abnormal or normal. For the implementation, UCSD Anomaly Detection Datasets is used. The performance of the proposed approach has been illustrated using UCSD Anomaly Detection Dataset and it is validated through the ROC curves and classification accuracy. From the comparative analysis made, it can observe that the proposed technique obtained better results.

6. REFERENCES

[1] UCSD Anomaly Detection Dataset from http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[2] G. Lavee, E. Rivlin, M. Rudzsky, Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev., vol. 39, no. 5, pp. 489504, 2009.

[3] P. Turaga, R. Chellappa, V. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, pp. 14731488, 2008.

[4] Juan C. SanMiguel, Jos M. Martnez, "A semantic-based probabilistic approach for real-time video event recognition", Computer Vision and Image Understanding, vol. 116, pp.937952, 2012.

[5] Somboon Hongeng, Ram Nevatia, Francois Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods", Computer Vision and Image Understanding, vol. 96, pp. 129162, 2004.

[6] Hsuan-Sheng Chen, Wen-Jiin Tsai, "A framework for video event classification by modeling temporal context of multimodal features using HMM", J. Vis. Commun. Image R., vol. 25, pp. 285295, 2014.

[7] Cheng-Chang Lien, Chiu-Lung Chiang, Chang-Hsing Lee, "Scene-based event detection for baseball videos", J. Vis. Commun. Image R., vol. 18, pp. 114, 2007.

[8] Gaurav Srivastava, Josiah A. Yoder, Johnny Park, Avinash C. Kak, "Using objective ground-truth labels created by multiple annotators for improved video classification: A comparative study", Computer Vision and Image Understanding, vol. 117, pp. 13841399, 2013.

[9] Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng and Harpreet Sawhney, "Video event recognition using concept attributes, IEEE Workshop on Applications of Computer Vision , pp.339-346, 2013.

[10] Xiang Ma,Schonfeld, D. ; Khokhar, A.A.,"Video Event Classification and Image Segmentation Based on Noncausal Multidimensional Hidden Markov Models,"IEEE Transactions on Image Processing,Vol.18,No.6,2009.

[11] Mao-Hsiung Hung and Chaur-Heh Hsieh, Event Detection of Broadcast Baseball Videos, Circuits and Systems for Video Technology, IEEE Transactions on, Vol.18 ,no. 12,pp. 1713 1726,2008.

[12] Weixin Li, Vijay Mahadevan, Member, and Nuno Vasconcelos,"Anomaly Detection and Localizationin Crowded Scenes, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 35,pp. 1-15, 2013.