International Journal on Applications in Information and Communication Engineering
Volume 2: Issue 2: Feburary 2016, pp 25-30. www.aetsjournal.com ISSN (Online) : 2394-6237

Graph Cut Models for Simultaneous Tracking and Recognitionusing the Neural Network Classifier

G.Karthic , B.Lalitha

***Abstract*—** A continuous video contains two important components such as tracks of the person in the video, and localization of the actions that are performed by the actors. The analysis of the activity is used for solving both the tracking, and recognition problems. In this paper, we have deployed an efficient activity analysis framework for determining the activity of the human in the video. Initially, the input video is obtained from the ULCA, and VIRAT datasets, then the video file is converted into multiple video frames named as frames.The information regarding each framesare obtained and further the frames are resized for preventing the memory from dumping.The noise present in each frame is filtered using the Gaussian filter. The Graph Cut technique is used for extracting the shape of the object from the background. The tracking of the video file is performed using the Bounding Box technique. The features from the resultant image are extracted using the Local Binary Pattern (LBP). Based on the features obtained the frames a pattern is generated. These feature values are grouped into activity segments using the Neural Network (NN) classifier. To validate the performance of the proposed NN classifier it is validated with the existing Support Vector Machine (SVM) classifier for the metrics such as accuracy, precision, and recall. The experimental results proved that the proposed NN classifier produced optimal results than the existing SVM.

Index Terms— Local Binary Pattern (LBP), Neural Network (NN), Support Vector Machine (SVM), Graph Cut, Bounding box technique, Gaussian filter.

I. INTRODUCTION

Object tracking is used in multiple applications such as robotics control, video retrieval, etc. The video tracking is the process of detecting an object in the image plane as it moves over the scene. The video tracking is preferred for various applications such as automated surveillance, video indexing, human-computer interaction, meteorology, and traffic management system. The key issue in the video tracking are motion estimation, and matching estimation. The motion estimation is used to predict the location of the region in the next video frame where the object would have been placed. The motion estimation information is very difficult to be determined, hence an effective mechanism for the determination of the fixed-size region is essential. In the matching estimation, an object is identified which is being tracked in the next video frame that is placed in the closed region of the next video frame. The motion estimation stage predicts the closed region. The location of the object of interest in the next frame is estimated in the matching estimation stage. The matching estimation algorithms incorporate a feature detection stage for performing the operations such as image classification, and segmentation. The object tracking algorithms implement feature detection for matching the pixels from the object being tracked between two consecutive video frames, then estimates the exact location of the object in the next frame. The existing techniques used for the activity recognition do not consider the tracks, location, and labels for determining the movement of the human in the scene. Hence, to overcome this issue, we have proposed an efficient activity analysis framework. Initially, the input video is converted into multiple video frames, then the noise present in all the frames are filtered using the Gaussian filter. The use of Gaussian filter prevents the edge from blurring. Further, they are computationally efficient. The filtered frames are provided as input to theHierarchical Markov Random Field-Sparce (HMRF-Sparce) technique. This technique, by comparing the intensity of the pixels, separates the background from the object. The object of interest is tracked using the bounding box technique, then the features present in the object of interest is substituted the Local Binary Pattern (LBP). The main advantage of using the LBP is increased accuracy, and stability. The extracted features are classified into various activity segments using the NN classifier. To validate the performance of the proposed NN classifierit is compared with the existing SVM classifier. The comparison results show that the proposed NN classifier provides higher accuracy than the SVM classifier. Further, the precision, and recall for the proposed activity detection framework is validated. The analysis results show that the suggested framework provides increased higher precision, and recall values for the different video input files.

The remainder of the paper is systematized as follows, Section II describes the literature review related to the existing human action recognition techniques. Section III illustrates the proposed human activity analysis framework, section IV describes the performance results of the proposed method, and Section V illustrates the conclusion of this paper.

G.Karthic , PG Student, Department of Computer science and Engineering, Sree Sowdambika College of Engineering, Virudhunagar, Tamil Nadu, India. (Email: karthicg64@gmail.com)

B.Lalitha , Assistant Professor, Department of Computer science and Engineering, Sree Sowdambika College of Engineering, Virudhunagar, Tamil Nadu, India.

II. RELATED WORK

This section describes the various existing human action recognition techniques.*Brendel, et al*[1] proposed a volumetric-based approach for the activity recognition, and video parsing. Based on the sub activities, and hierarchical temporal, and spatial relations the suggested approach extracted the human activities. When compared to the traditional approaches, the proposed volumetric-based approach produced optimal results. *Wang, et al*[2] suggested a novel actionlet ensemble model for charactering the human actions. The suggested model prevented the noise, and successfully characterized both the human motion, and human-object interactions. Three datasets such as Kinect devices, multiview action recognition dataset that was captured using the Kinect device, and the dataset that was captured using the motion captures system were used for the evaluation. The experimental results proved that the suggested method produced optimal results than the state-of-the art algorithms. *Chaaraoui, et al*[3] proposed an evolutionary algorithm for determining the optimal subset of skeleton joints. As the suggested algorithm was based on the topological structure of the skeleton, the final success rate was optimal. When compared to the traditional RGB action recognition approach, the proposed evolutionary algorithm provided improved initial recognition rate, and optimal success rate for the MSR-Action 3D dataset.*Ofli, et al*[4]proposed the Sequence of the Most Informative Joints (SMIJ) representation for the human actions.The selection of the skeletal joints were automatic. The human actions were represented as a sequence of the most informative joints. When compared to the state-of-the art algorithms, the proposed SMIJ representation provided better performance.*Xia, et al*[5] proposed the Histograms of 3D Joint locations (HOJ3D) for representing the human postures. The action depth sequence from the HOJ3D was re-projectedusing the Linear Discriminant Analysis (LDA), and then clustered into visual words. The discrete Hidden Markov Models (HMMs) was used for modeling the temporal evolutions of the visual words. The suggested representation provided optimal results for the 3D action dataset.*Ji, et al*[6] proposed a novel 3D Convolutional Neural Network (CNN) model for the action recognition. The suggested model extracted the features from both the spatial, and the temporal dimensions. The suggested model produced multiple channels of information from the input frames. The information from all the channels were combined for representing the features. On applying the suggested model for the real-world environment, superior performance was achieved.*Tanays, et al*[7] analyzed the effectiveness of the sparse representation obtained from the context of the action recognition in videos. The human actions were modeled using three over complete dictionary learning frameworks. The over complete dictionary was constructed using the spatio-temporal descriptors. The suggested approach produced state-of-the art results on the public datasets. *Chen, et al*[8] proposed an efficient approach for unifying the activity categorization with the space-time localization. The upshot was the fastest method that evaluated the boarder space of the candidates. The suggested algorithm produced high speed, and accuracy than the existing search strategies.*Morariu, et al* [9] suggested a framework for the automatic recognition of complex multi-agent events. Based on the video analysis, the events were determined. The interval-based temporal reasoning was integrated with the probabilistic logical inference for preventing the combinatorial explosions. *Hoai, et al*[10] proposed the joint segmentation, and action recognition actions for preventing the limitations of the traditional methods. The suggested model was based on the discriminative temporal extension of the spatial bag-of-words model. The classification was performed using the multi-class SVM framework. When compared to the traditional methods, the proposed method produced optimal results for the honeybee, Weizmann, and Hollywood datasets.*Le, et al*[11] addressed the issue of building the high-level, class specific feature detectors from the unlabeled data. The feature detector was robust to the translation, scaling, and out-of-plane rotation. The network was trained to recognize 22,000 objects. When compared to the traditional approaches, the proposed trained network produced 70% better performance.*Oh, et al*[12] proposed a novel large-scale video dataset for validating the performance of the diverse visual event recognition algorithms. The suggested dataset had many outdoor scenes with the actions of the non-actors. Various types of evaluation modes were proposed for the visual recognition tasks. *Zhang, et al*[13] proposed an approach that efficiently identified the local, and long-range motion interactions. The suggested approach captured the combination of the hand movement of one person with the foot response of another person. The experimental results proved that the suggested approach effectively recognized awide variety of activity than the state-of-the art methods.*Yao, et al*[14] exploited the attributes, and parts for recognizing the human actions in the still images. The action attributes were described as the verbs. When compared to the traditional classification methods, the proposed method extracted the meaningful higher-order interactions.*Lara, et al*[15] proposed the centinela system for providing a highly accurate activity recognition. The suggested system identified the actions such as walking, running, sitting, ascending, and descending. A portable and unobtrusivereal-time data collection platform was included in the proposed system. The Centinela provided 100% accuracy for running, and sitting. Further, the classification accuracy for the ascending action was improved.

III. PROPOSED METHOD

The overall flow of our proposed human activity analysis framework is depicted in the figure 1. The key components of our proposed framework are,

- Frame conversion
- Filtering
- Segmentation
- Video tracking
- Feature extraction

- Activity analysis

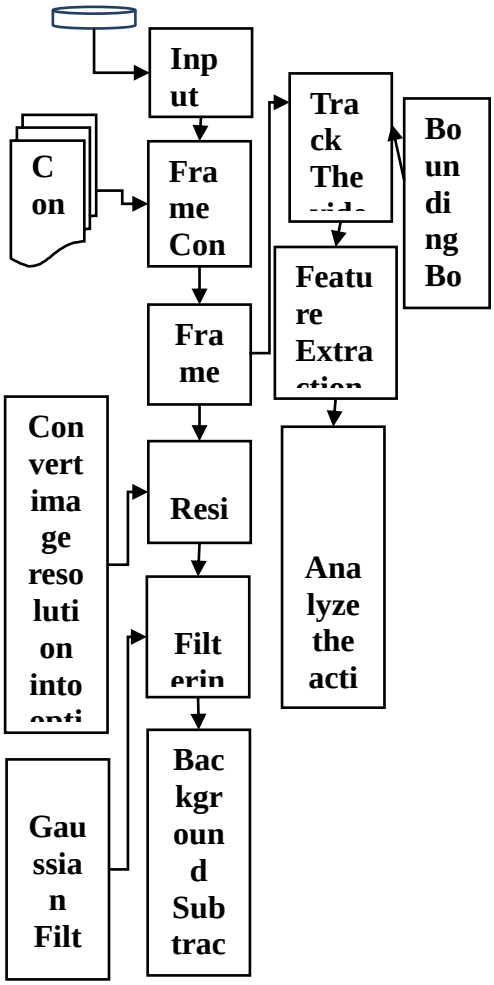
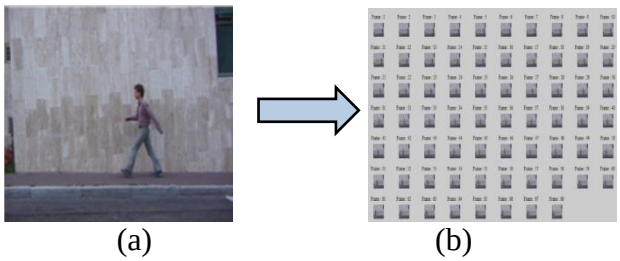


Fig.1. Overall flow of the proposed human activity analysis framework

A. Frame Conversion

The input datasets are obtained from [16], and [17]. The input video is converted into individual framesusing the frame conversion process. The conversion of the video file into frames is illustrated in the figure 2. The fig 2(a) depicts the input video file, and the fig 2(b) shows the converted frames.



(a) (b)

Fig. 2 (a) Input video file (b)Framess

The details of each frame such as number of frames, height, and width of the frames are collected after the frame conversion process. Further, to prevent the memory from dumping the large sized frames are resized to smaller frames. The resizing process of a single frame is illstrated in the figure 3. The same process is repeated for all the frames.

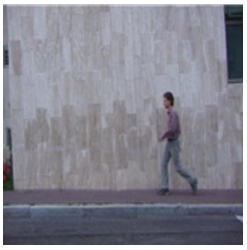


Fig. 3 Resized Frames

B. Filtering

The resized frames are then filtered using the Gaussian filter. The weights of the Gaussian filters are chosen based on the Gaussian functions. The Gaussian filter smoothens the image, and also prevents the Gaussian noises. Mathematically, the Gaussian filter is represented as follows,

$$G(a,b) = \left[\frac{1}{\sqrt{2\pi\sigma}} e \frac{a^2}{2\sigma^2}\right] \times \left[\frac{1}{2\pi\sigma} \frac{b^2}{e2\sigma^2}\right] \quad (1)$$

Where, σ^2 denotes the variance of the Gaussian filter.The figure 4 shows the resultant frames after the filtering process.

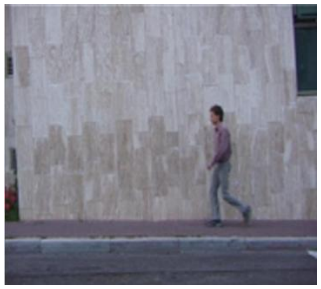


Fig. 4 Filtered frames

C. Segmentation

The segmentation is the process of extracting the shape of the object from the background. In this paper, Graph Cut technique is used for the extracting the human image from the background image. A graph cut approach makes use of efficient solutions of the maxflow/mincut problem between source and sink nodes in directed graphs. To take advantage of this we generate a s-t-graph as follows: The set of nodes is equal to the set of pixels in the image. Every pixel is connected with its d-neighbourhood(d=4,8). The figure 5 shows the process of extracting the shape of the human body from the background.

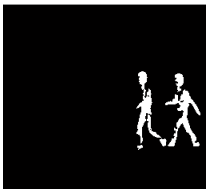


Fig. 5 Extraction of human body shape from the background

D. Video Tracking

The features present in multipleframes are collected and a pattern is generated based on the values of the features. These values are classified for detecting the human activity. The video tracking begins with the generation of the set of match hypothesis for the frame association and the set of tracks. Based on the features computed at the frame, the observation potential is computed for each frame. The classifiers such as Support Vector Machine (SVM), and Neural Network are usedfor grouping the frames into activity segments.

E. Feature Extraction

The node features and the edge features for the potential functions are computed for the training data using the Local Binary Pattern (LBP). The proposed LBP works in a 3x3 pixel block of an image. The pixels in the block are thresholded based on the center pixel value, multiplied by powers of two and then summed to obtain the center pixel value. As the number of neighborhood pixel is 8, a total of $2^8 = 256$different labels can be obtained depending on the relative gray values of the center and the pixels in the neighborhood. The figure 6 illustrates the feature extraction process using the LBP.

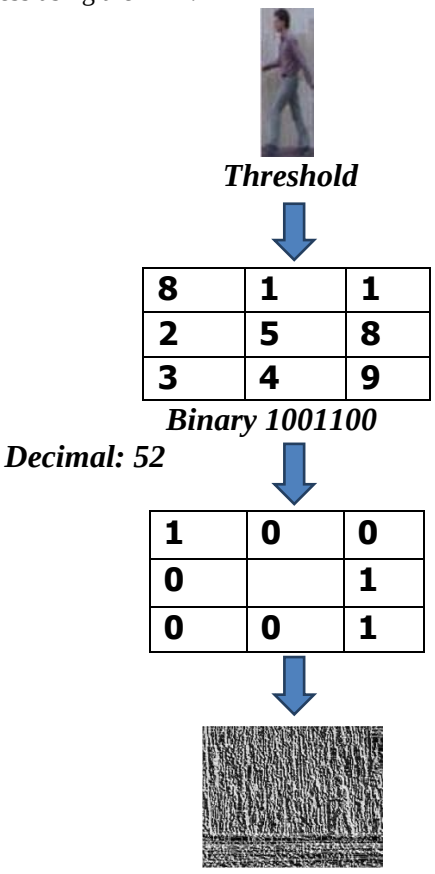


Fig. 6 Feature extraction using the Local Binary Pattern

F. Activity Analysis

Based on the extracted features, the activity being performed by human is analyzed. Here, the activity performed by the person in figure 6 is detected as walking.

IV. PERFORMANCE ANALYSIS

The performance of the proposed Neural Network (NN) classifier is validated against the existing SVM classifier for the metrics such as,

- Accuracy
- Precision
- Recall

A. Accuracy

The accuracy defines the proximity of the measurement results to the true value.The accuracy of the proposed NN classifier is computed using the following equation,

$$Accuracy = \frac{(TP+TN)}{n} \quad (2)$$

Where,
TP is the number of true positives
TN denotes the number of true negatives
n represents the total population.

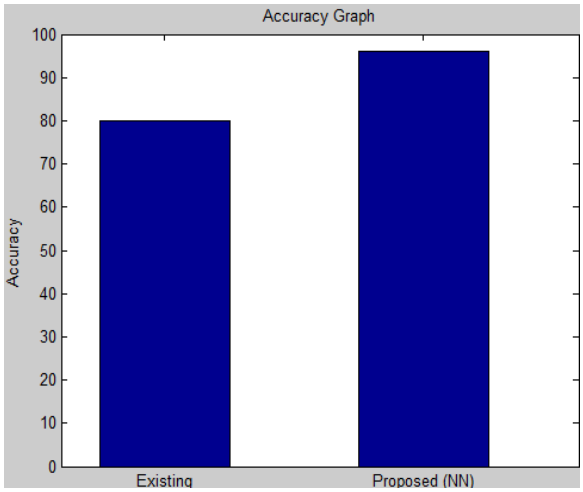


Fig. 7 Comparison of accuracy for the existing, and the proposed method

The figure 7 shows the comparison of accuracy for the proposed NN, and SVM classifier. From the figure it is analyzed that the proposed NN classifier provides higher accuracy than the existing SVM classifier.

B. Precision

The precision is defined as the ratio of the true positives and the sum of true positive, and true negative values. It is computed using the following equation,

$$Precison = \frac{TP}{TP+FP} \quad (3)$$

The figure 8 shows the comparison of the precision for the various video files. Each iteration in the graph represents anindividual video file. From the graph it is concluded that the precision value is high for all the iterations.

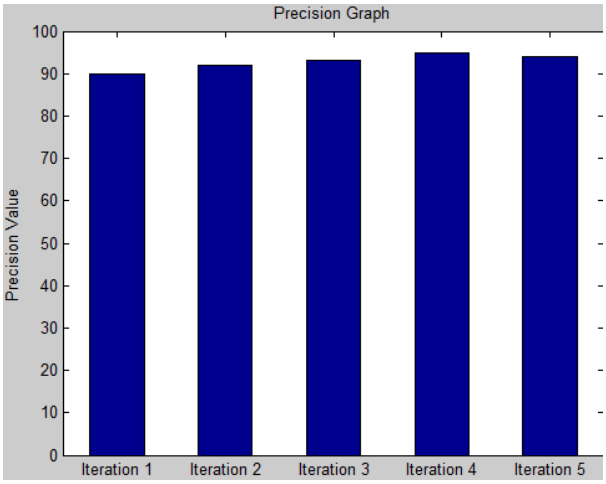


Fig. 8 Comparison of precision for multiple iterations

C. Recall

The recall is defined as the ratio between the True Positive,and the sum of True Positive and True Negative values. It is computed using the following equation,

$$Precison = \frac{TP}{TP+FN} \quad (4)$$

The figure 9 shows the comparison of the recall for the various video files. From the graph it is concluded that the recall value is high for all the iterations.

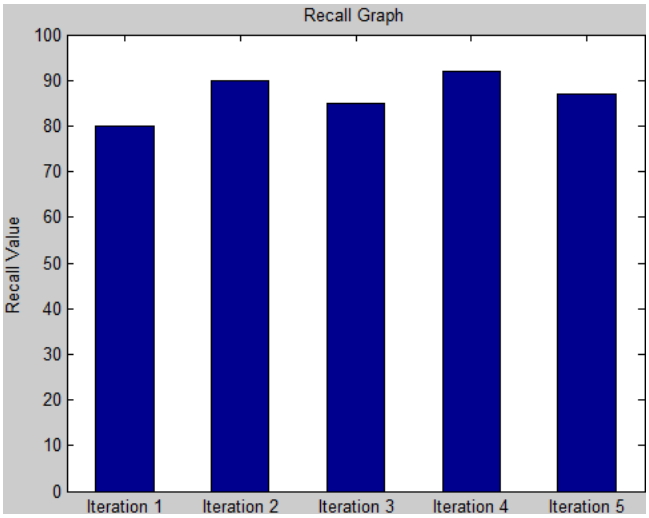


Fig. 9 Comparison of recall for multiple iterations

D. Comparison of Precision Values

The performance of the proposed Hierarchical Markov Random Field-Sparce (HMRF-Sparce) is compared with the existing segmentation techniques such as Bag of Word (BOW), HierarchicalMarkov Random Field -Dense (HMRF-Dense), Morphological, Zhu. The figure 9 shows that the precision of the proposed HMRF-Sparce is higher than the existing methods.

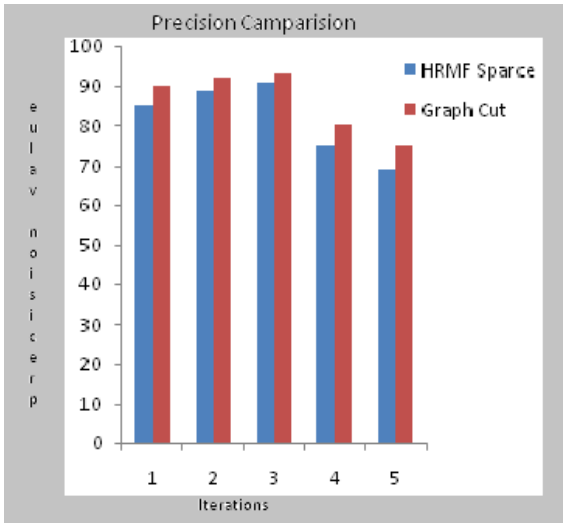


Fig. 10 Comparison of Precision Value for the existing and the proposed segmentation techniques

V. CONCLUSION

An efficient NN based classifier is used for tracking the human activity. During the frame conversion process, the input video file is converted into multiple frames. The noise present in the resized frames are removed using the Gaussian filter. To extract the shape of the human from the background, the background subtraction technique is used. The bounding box technique is used for tracking the background subtracted frame. The features from the frames are obtained using the LBP. With the extracted features, the activity being performed by the human is analyzed using the NN classifier. The performance of the NN classifier is validated against the SVM classifier. The experimental results prove that the suggested NN classifier produceoptimal performance in terms of accuracy, precision, and recall than the existing SVM. Further, when compared to the existing segmentation techniques such as BOW, HMRF Dense, Morphological, and Zhu techniques, the proposed HMRF-Sparce technique provides higher precision value.

REFERENCES

[1] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 778-785.

[2] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 914-927, 2014.

[3] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *Expert Systems with Applications*, vol. 41, pp. 786-794, 2014.

[4] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, pp. 24-38, 2014.

[5] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 20-27.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221-231, 2013.

[7] a. R. K. W. Tanaya Guha, "Learning Sparse Representations for Human Action Recognition," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, pp. 1-14, 2011.

[8] C.-Y. Chen and K. Grauman, "Efficient [14] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1331-1338.

[9] Ó. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data," *Pervasive and Mobile Computing*, vol. 8, pp. 717-729, 10// 2012.

[10] V. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3289-3296.

[11] *ULCA Department of Statistics*. Available: http://statistics.ucla.edu/
"VIRAT Video Dataset."activity detection with max-subgraph search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1274-1281.
[10] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3265-3272.

[12] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8595-8598.

[13] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3153-3160.

[14] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *Computer Vision–ECCV 2012*, ed: Springer, 2012, pp. 707-721.