A MULTI-TEMPORAL FRAMEWORK FOR HUMAN VIOLENT EVENT ANALYSIS IN VIDEO SURVEILLANCE USING CNN

B.Pradeepa
Assistant Professor
Department of Electronics and Communication Engineering
Karpaga Vinayaga College of Engineering and Technology *Chengalpattu, TamilNadu, India*

A.Viji
Department of Electronics and Communication Engineering
Anna University (MIT Campus), *Chrompet, TamilNadu, India*

J.Joshan Anthesious
Department of Electronics and Communication Engineering
Anna University (MIT Campus), *Chrompet, TamilNadu, India*

***Abstract* —To detect violent behavior automatically in video surveillance is consider as one of the important event. A new approach is proposed to detect and analysis a novel framework of violent events which is called as High level activity (e.g., Fight, Kidnap, Gun Shoot etc..). The proposed methodology uses Deep Learning based CNN(Convolution Neural Network) Algorithm (Forward and Backward propagation Algorithm). In deep learning architecture, dense convolution layers are used for analyzing spatial and temporal information changes in the frames and used for analyzing the multidimensional changes in the image. Convolution layer is used for extracting features from the input video. Pooling layer is used to reduce the resolution of the image. All those layers are connected to fully connected layer to perform the classification task. By using BEHAVE DATASET and YouTube CCTV footage scenario experiment results achieves 90% of accuracy in violent behavior detection than the existing method available in the literature.**

Keywords — Violent Events, CNN Algorithm, Crowd Scene, Dense Convolution Layer, Behave Dataset .

I. INTRODUCTION

In Public places, Serious threat are produced by the high level activity which make public to be with personal security and social stability, nowadays security attendants leading to a huge pressure in public minds. detection, To detect and capture any variation in human body, discriminative and high level features [1] are difficult task in video surveillance. Existing method are obtained by analyzing violence related behaviors and characteristics they are blood, flame gunshots, explosions and car-braking etc. This type of detection belongs with lack of audio information surveillance.

To detect high level activity (violent behavior) in moving crowd is considered as difficult task because in moving crowd the velocity flow vector and magnitude of human direction can vary because of human moves from one place to another place frequently.

Cameras are considered as one part of our day to day life. In Areas Street, subways, train and bus station, airports, and sports stadiums many other public places everywhere these video surveillance used to monitor whatever happens within the frames. Particularly violence detection, system are needed to find humans fight and vandalism in many places.

Now focusing on important of violent events detection and recognition [2] in surveillance. Motion trajectory information and human orientation information and person-to-person fight detection are obtained by the proposed video surveillance method. Segmentation of frames is considered as a difficult task in real time videos.

The above mentioned drawbacks are overcome by increasing the intelligence human activities recognizing video surveillance and which will provide automatic alert to nearby police station in abnormal situation.

Human activities can be classified as two types based on complexity which is faced:

1. Low-level activity 2. High-level Activity

High level activity is analyzed by some changes in the body language of the human rather than normal behavior they are:

1. Group Violent Behavior 2. Two Persons Violent Behavior 3.Chase 4.Kidnap 5.Gun Shoot.

II.RELATED WORK

Many researchers shares important key points to analysis the abnormal activity of crowd . Donguhi Song et al [3] proposed the multi-layered temporal perception and last fusion of crowd and also author mentioned handling of high level activity is not easy. The main advantage is that function of utilizes early fusion among various features from various dataset.

Di Wu et al [4] Deep Dynamic Neural Networks (DDNN) for multi modal gesture recognition which is considered as a new method. At the same time for gesture segmentation and recognition author focus on Hidden Markov Model (HMM) and continuous labeling of cyclic video sequence. Within an HMM framework jogging, walking and running are classified by using deep learning based neural networks. Online segmentation and recognition are also obtained by this method.

Li Liu et al [5] proposed Spatial-Temporal Representations for Action Recognition. Recognizing human action obtaining robust features from video surveillance are considered as a difficult task. By using scale and different features both the color and efficient optical flow field are extracted by using SVM (Support vector machine)classifier. Proposed method developing discriminative spatiotemporal representations, which automatically divides the high level action color and motion information.

PiotrBilinski et al [6] proposed human high level action recognition by using Improved Fisher Vectors (IFV) violent actions are obtained. The proposed improvement provides accurate violent action reorganization (as compared to the standard IFV) and result in faster detection of violence behavior. Spatial information may contain useful information.For violence recognition SVM classifier is used. 4 Benchmark dataset, Hockey Fight dataset, violent flow dataset and movie fight dataset are used for evaluation.

Lixin Chen et al [7] proposed non-overlapping and visible views of human anomaly detection from different time dimension angle in cameras. Author used three steps they are 1) Real scene preprocessing using optical flow. 2) Real scenes based neighborhood weighted fuzzy c-means (NW-FCM) Algorithm were build. 3) To find time dependence data for analyzing the local and global path.. these method will gives better result among anomaly detection.

Zhongwei Cheng et al [8] proposed a reorganization of individual actions and pair interactions of human actions in video sequences and also motion trajectories and Gaussian processes are used to manage the human movements within the group. Visual appearance descriptions for group actions are succesfully proposed. The main advantage is that better group actions are obtained with better accuracy. Behave dataset are used here and give better result..

The proposed aim is to detect and analyze high level activity that is violent events in video scenes by using CNN. Some challenges faced to predict violent actions are:

- Tremendous variety of motion pattern.
- Specific human behaviours needed to be detected

- Distinguishing different actions still remains challenging task

Inorder to overcome the above mentioned drawback, Deep Learning based CNN Algorithms are used for effective result.

III. METHODOLOGY

Violent behaviour(High Level Activity) such as(Fight among the pedestrains, Kidnap, Following, chase) between the pedestrain is considered as one of the abnormal action. Anomaly activities are classified and identified by using Deep learning based CNN Algorithm. The Efficient framework of abnormal high level activity [9]is obtained by using CNN Algorithm.

Efficient and usefull framework of the proposed model are shown in Figure 1. Several potentially usefull framework for group behaviour classification are used and one of the standard dataset is used to classify the violent behaviour.

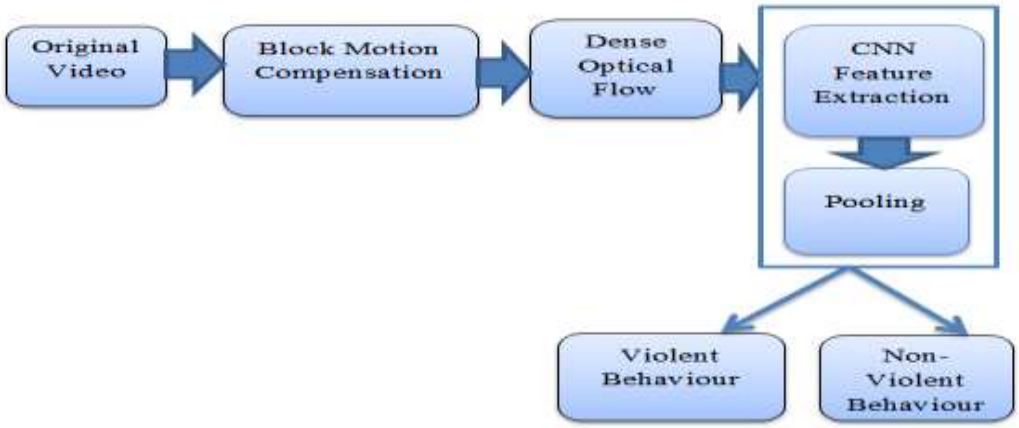


Figure 1.Efficient framework of violent event analysis

3.1 Dataset Description

Thus the dataset capture 25 frames per second with the resolution of 640×480, and consists of four video clips[10]: margaret#1, margaret#2, taku#1, and taku#2. BEHAVE Dataset consist of violencet and non-violence event because only high-level activities are concerned in the video.

EXAMPLE:

ID1	ID2	Start	End
[3,4]		;5826	;5926

Table 1: Dataset Labelled Description

Table 1 shows persons with 3 and 4 are in a group which has been described as "Walk Together" that frames starts at 5826 and ends at 5926.

BEHAVE video dataset shown in Figure 2 which consist of 4 video clips.

Figure 2. Behave Dataset

3.2 Block Motion Compensation

Motion Compensation implementation is easy, and can be widely used. These compensation divides the frames into square blocks which is shown in Figure 3. Each block used to analysis the motion and direction of the pedestrains . Motion Compensation [11] is used to predict a frame from a video, the main advantage of the compensation technique is previous and future frame are obtained by analysing the motion of the camera or object in the video, then frame is given to video compression for functing the video data encoding, generation of MPEG-2 files are considered as an main example.

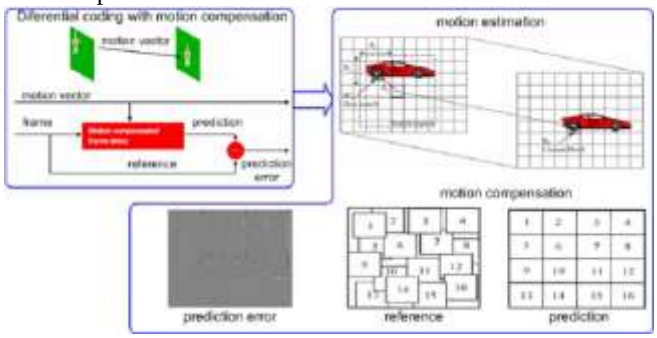


Figure 3. Block Motion Compensation

3.3 Optical Flow Field

Proposed work is based on motion segmentation algorithm which works on the basis of distribution of optical flow field. These extraction of optical flow algorithm is used to implemented on Opencv,The two consecutive gray-scale image are obtained from the x and y optical flow direction [12] . The $flow_x$ and $flow_y$ images and magnitude image(Mag) are obtained :

$$\text{Mag (i,j)} = \sqrt{flow_x\ (i,j) + flow_y (i,j)} \quad (1)$$

Where (i, j) represents pixel position, and $flow_x$ and $flow_y$ images denotes the x and y optical flow direction.

i) Difference of Movement

The mean value of the motion vector ratio M to Movement M are used to calculate group movement:

$$\bar{V}_x = \frac{1}{N_{blocks}} \sum_{i=1}^{N_{block}} V_{xi,} \quad \bar{V}_y = \frac{1}{N_{block}} \sum_{i=1}^{N_{block}} V_{yi,} \quad (2)$$

$$D = \frac{\bar{M}}{M} \quad (3)$$

These value consider the magnitude of the motion vectors and above motion vector directions represents the values of 0 to 1.

ii) *Violence*

The values are used as the coincidence indicator , motion in persons describe the violent events. The terminal indicator ,which is divided by the abnormal motion :

$$f_c = e^{-\lambda\sqrt{V_x^{-2} + V_y^{-2}}}$$
$$V = \frac{f_c}{D_M} \quad (4)$$

The value V of violence behaviour is used to discriminate behaviour of fight and walk among the people, which the term f_c cannot differentiate.

iii)*Difference of Direction* D_d

Directed motion vectors entropy values are:

$$D_d = \sum_{i=0}^{N_{angle}} p_i \log p_i \quad (5)$$

Where N_{angle} represent a resolution of direction and the Possibility p_i.

iv) *Speed*

The speed equation is used to analysis the movement of the crowd from one place to another place (i.e previous location (x_{t-1}, y_{t-1}) to the present location (x_{t,y_t})):

$$S = \sqrt{(x_t - x_{t-1})2 + (y_t - y_{t-1})2} \quad (6)$$

3.4 CNN Learn Hierarchical Features

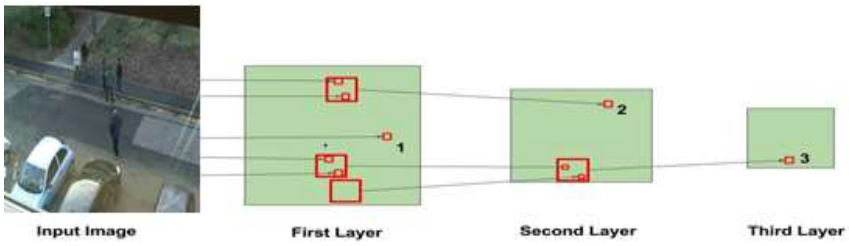


Figure 4. CNN Hierarchical Features

Hierarchical features of input images are shown in figure 4.

IV. EXPERIMENTAL RESULT

Thus the current invention for human violent behavior analysis which is already mentioned high level activity [13] are fight, kidnap, chase, follow etc are successfully obtained.The pixel positions of the image will belongs to image scaling which is image size. Here the image sizes are 640x480. Mapping can be computed from this information. Convolutional neural network (CNN) is the important technique for analysing multidimensional signals such as images.

The convolution layer is considered as the main function in CNN. The neurons in the layer look for specificfeatures. weighted sum of CNN [14] between two signals or function are mentioned. In image processing, convolution layer of term (x,y) are calculated, the function multiply the values element-by-element with the

convolution filter of (k × k) also single output are added. K is considered as kernal size. The convolution kernel is slid over the entire matrix to obtain an activation map.The convolution kernal is slide over the entire matrix to obtain an activation map.

For 32×32×3 input image and filter size of 3×3×3, and neurons corresponding to each location, then 30×30×1 output or activation of all neurons are considered as the activation maps. The activation map of one layer act as the input layer to the next layer. Initial layers are looking at smaller regions [15] of the image and learn simple features from image like features, edge,corners etc.

In deep Learning network, the neurons get information from the larger part of the image from various other neurons. Thus the neurons at the later layer can learn more complicated features from the human.

4.1 Performance Evaluation

By using precision – recall measures, accuracy performance evaluation is obtained. They are:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{F-Measure} = 2\frac{Precision \times Recall}{Precision + Recall}$$

Where,

True Positive (TP)

True Negative (TN)

False Positive (FP)

False Negative (FN)

True Positive tends to correctly predicted term (Positive Class) . Similarly True Negative tends to wrongly predicted term (negative Class)

False Positive tends to prediction of incorrect positive class, and false negative tends to prediction of incorrectly negative class.

TABLE3. COMPARISONS OF EXISTING AND PROPOSED METHOD

AUTHOR	RECOGNITION ALGORITHM	AVERAGE RECOGNITION
Donghui Song et al..(2018)	SVM(Support Vector Machine	70%
Sung-Kee Park(2018)	SGT(Situation Graph Tree)	78%
Proposed Method(2019)	CNN Algorithm	90%

The existing and proposed work comparison are shown in table 3, compared to existing method , 90% of accuracy is obtained by the proposed method .

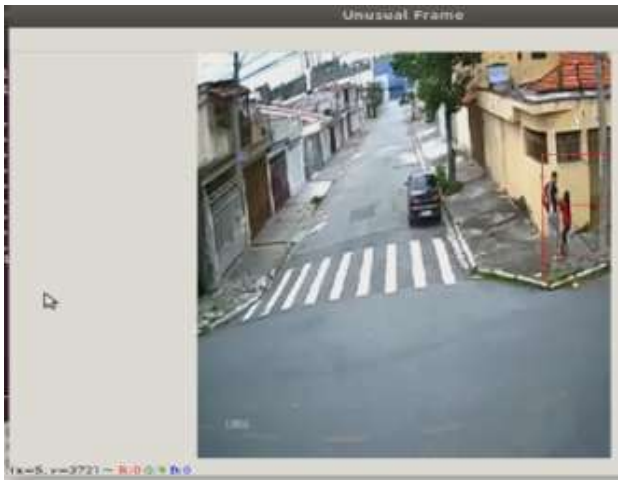
Figure 5. (BEHAVE DATASET) shows abnormal behaviors of humans in the videos. a) Two men violent behavior analysis. b) Group fight analysis. c) Chase Detection. d) Detection of CCTV Footage Kidnap e) Detection

of CCTV Footage Gun Shoot. Red color rectangular box correspond to the highly probable region. The Figure 5 denotes prediction of the anomalies in the Behave dataset[16]. Red colour rectangular box indicates violent behaviour like Group Fight, kidnap, Chase, Gunshoot.

(a)

(b)

(c)



(d)

(e)

Figure 5: (BEHAVE DATASET) a) Two men violent behavior analysis. b) Group fight analysis. c) Chase Detection. d) Detection of CCTV Footage Kidnap e) Detection of CCTV Footage Gun Shoot.

4.2 Performance Analysis in Graph

The proposed approach detect High level activity which is violent behavior activity(Fight, Kidnap, Gun shoot etc.,) in video surveillance by using Deep learning based CNN algorithm. Block Motion compensation is used to divides the frames into square block, so that according the motion and action of the pedestrian, feature extractions are done by CNN (Convolution Neural Network).The constant Threshold values and accurate action recognition in the scenes are used to obtain the highly accurate and perfect result.

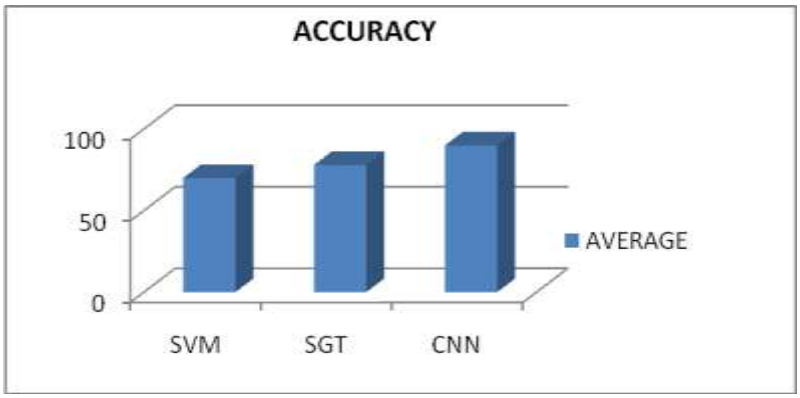


Figure 6. Performance Analyses in Graph

Figure 6 shows that compared to SVM (Support Vector Machine) and SGT (Situation Graph Tree) proposed CNN Method provides better Accuracy rate.

V.CONCLUSION

Experimental results shows better result with improved accuracy and also which is compared to practical violence detection in a surveillance environment. By using BEHAVE dataset and YouTube CCTV footage Scenarios proposed approach gives better result .In future work different dataset can be used to get a better accuracy.

REFERENCES

[1] Vahora, S. A., and N. C. Chauhan. "Deep neural network model for group activity recognition using contextual relationship." Engineering Science and Technology, an International Journal 22.1 (2019): 47-54.

[2] Zhou, Peipei, et al. "Violence detection in surveillance video using low-level features." PLoS one 13.10 (2018): e0203668.

[3] Song, Donghui, Chansu Kim, and Sung-Kee Park. "A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance." Information Sciences 447 (2018): 83-103.

[4] Wu, Di, et al. "Deep dynamic neural networks for multimodal gesture segmentation and recognition." IEEE transactions on pattern analysis and machine intelligence 38.8 (2016): 1583-1597.

[5] Bilinski, Piotr, and Francois Bremond. "Human violence recognition and detection in surveillance videos." 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2016.

[6] Liu, Li, et al. "Learning spatio-temporal representations for action recognition: A genetic programming approach." IEEE transactions on cybernetics 46.1 (2016): 158-170.

[7] Lixin Chen ,Huiwen Guo .,' Detecting Anomaly Based on Time Dependence for Large Scenes' IEEE International conference on Information and Automation , China , August 2016.

[8] Cheng, Zhongwei, et al. "Recognizing human group action by layered model with multiple cues." Neurocomputing 136 (2014): 124-135.

[9] Burghouts, Gertjan J., et al. "Complex threat detection: Learning vs. rules, using a hierarchy of features." 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2014.

[10] http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/

[11] http://www.reasearchgate.net/figure/predictive-source-coding-with-motion-compensation_fig1_229028758

[12] Burghouts, Gertjan J., and Klamer Schutte. "Spatio-temporal layout of human actions for improved bag-of-words action detection." Pattern Recognition Letters 34.15 (2013): 1861-1869.

[13] Blunsden, Scott, and R. B. Fisher. "The BEHAVE video dataset: ground truthed video for multi-person behavior classification." Annals of the BMVA 4.1-12 (2010): 4.

[14] Ni, Bingbing, Shuicheng Yan, and Ashraf Kassim. "Recognizing human group activities with localized causalities." 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.

[15] Albanese, Massimiliano, et al. "A constrained probabilistic petri net framework for human activity detection in video." IEEE Transactions on Multimedia 10.8 (2008): 1429-1443.

[16] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[17] https://docs.opencv.org/2.4/modules/video/doc/motion_analysis_and_object_tracking.html.

[18] Ramin Mehran†, Brian E. Moore‡, Mubarak Shah†A 'Streak line Representation of Flow in Crowded Scenes' Computer Vision Lab, ‡Department of Mathematics University of Central Florida

[19] Kai-Wen Cheng and Yie-Tarng Chen .., ' Video Anomaly Detection and Localization Using Hierchical Feature Representation and Gaussian Process Regression'.., IEEE Conference on computer Vision pattern ,2015.

[20] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali and V. Murino, "Analyzing Tracklets for the Detection of Abnormal Crowd Behavior," *2015 IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, 2015, pp. 148-155.

[21] Lixin Chen,Huiwen Guo.,' Detecting Anomaly Based on Time Dependence for Large Scenes' IEEE International conference on Information and Automation , China , August 2016

[22] Navneet Dalal and Bill Triggs .., 'Histogram of Oriented Gradients for Humans Detection' Proc., IEEE Conference on 2016

[23] HajarYouselfi ,AzadehNazemi .., 'Locally Anomaly Detection in Crowded Scenes Using Locality Constrained Linear Coding' Artificial Intelligence and Signal Processing Conference.