iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** XII **Month of publication:** December 2019

DOI: http://doi.org/10.22214/ijraset.2019.12108

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue XII, Dec 2019- Available at www.ijraset.com

A Novel Approach to Detect Crimes and Assist Law Enforcement Agency using Deep Learning with CCTVs and Drones

Manish Pawar[1], Meet Dhanki[2], Sufyan Parkar[3], Chaitanya Dandekar[4], Balram Gupta[5]
[1]*Computer Engineering Dept,* [3, 4]*Electronics and Telecommunication Dept., Viva Institute of Technology*
[2]*Mechanical Engineering Dept., Dwarkadas J. Sanghvi College*
[5]*Computer Science, R.D National College*

Abstract: ***In India, crime has always been a serious issue and it is strenuous to tackle it due to lack of technology. The current security techniques are precarious and cannot be trusted, due to the negligence of the authorities and the policy setters, there's a monumental rise in the number of crimes taking place daily like kidnapping, abduction, molestation, etc. India ranks 133 out of 167 countries in Women, Peace and Security Index with Indore reporting the highest crime rate (769.1) among the megacities in India, followed by Bhopal (719.5) and Jaipur (597.1). Thus, an improved security system near bus-stops, schools, neighborhoods, and other public places is mandatory to detect violent or abnormal activities to avoid any casualties which could cause social or economic damage. The main focus should be to aid law enforcement agencies to avert the outbreak of violence by monitoring crime scenes and alerting the respective security officials. There is a need to instigate fear and perturbation in the minds of the criminals that they will get caught on committing any type of crime. Several image-based extraction techniques with smart methods of deep learning and artificial intelligence will help support the security officials to detect, monitor and track the perpetrator. Thus, we present a smart way to reduce violence, by criminal activity detection and recognition in a particular CCTV's field of view and make the drone identify, track and chase the criminal till he or she gets apprehended.***
Keywords: darknet19, yolo, deep learning, image segmentation, human activity recognition, violence detection, mavlink.

I. INTRODUCTION

Security has now become one of the preponderant concerns throughout the entire human-civilization due to the complex socio-economic structure of societies all over the world. With increasing crime rates, mendacious testimony and lack of significant evidence in hand, the perpetrator escape from getting into incarceration. Also due to weak security measures and lack of reliable resources the certainty of violableness increases. Video surveillance is thus a crucial monitoring tool that can provide profundity of the information and to detect a perpetrator's presence and study their actions which can possibly help to reduce crimes and thus have a derogatory effect on their actions. We thus need to exercise prudence in such paramountcy matters. In recent years, albeit a lot of work has been done in activity recognition of humans with little much focused on characterizing violence or detection of ferocious atrocity. But, evolution of automatic approaches through smart algorithms of deep learning and artificial intelligence, progress is being done as the demand for automated surveillance goes on increasing in various fields like public & home safety, military areas, etc. Recognizing fights and aggressive behavior in a video is an indispensable application area. The strategy used to detect violence considers the video sequence as a space-time volume, and using character recognition, object detection or other local feature extraction techniques, the crime scene is already been monitored and the data/information regarding each person passing through that spot is already recorded, thus having a plausibly substantiated and corroborated proof. This gives us a 24x7 surveillance of the vicinity, thereby aiding law-enforcement in a compelling way in critical cases. Although, a few difficulties arise in automatic violence detection due to its subjective nature which imposes some hindrance in deciding the exact point of violence.

Research has been done in detecting violence in a specific video by various state-of-the-art methods. Techniques like computer vision and deep learning have been applied to detect unusual activity or detect any weapons like guns, knives, etc. In the event of some political rally, religious gathering or some social get-together we may need some dynamic and efficient way to monitor and track casualties if any. But somewhere, the approach is partial and more work is needed to bring everything together to work effectively in real-life scenarios. In particular, we present a novel approach to reduce public violence by amalgamating various techniques to work together as a pipeline.

We basically put forward a flow to apply deep learning, computer vision, and image processing strategies to make surveillance at its

best and make use of an unmanned aerial vehicle to track the suspects. Initially, real-time violence or non-violence detection is done in a particular scene through CCTV's live coverage feed. After this, individual human recognition along with tracking takes place. This is done by fine-tuning pre-trained models in order to leverage the training and implementation of a newly generated model. Simultaneously, weapon detection takes place using a pre-trained Haar feature-based cascade classifier. The detected perpetrator's images are segmented from the entire scene and only those images are sent to the cloud for training, meanwhile sending an alert to a standby drone. Drone gets trained with the suspect's image as it arrives at the location and starts chasing the suspect until the cops take him or her into custody. In addition, a live video feed is also shared with the police so as to facilitate better navigation in order to intercept and apprehend the criminal. Then, some possible scenarios to extend our approach are manifested. Later, some drawbacks of the proposed methodology in various aspects are stated. Finally, this paper is concluded with further improvements to work on.

II. PREVIOUS WORK

Recent work by E. Bermejo, R. Sukthankar et al. [1] discussed how most of the studies have been focused more on pacifistic action detection and less on violent actions. They tested one of the most famous Bag-of-Words [11] framework for action recognition along with two best action descriptors like STIP [12] and MoSIFT [13]. The later part covers the research on violence detection on a database of 1000 sequences and is then segregated based on fight or non-fight and they were able to obtain an accuracy of around 90%. Reference [2] proposed by Amin Ullah et al. describes a triple-staged end-to-end deep learning violence detection framework. After detecting people in a video stream via a CNN model, a sequence of 16 frames is given to another CNN (3D-CNN) model where features of the sequences are extracted and given to the softmax classifier. Furthermore, the 3D-CNN model is optimized using an optimization toolkit (open visual interference and neural network toolkit) by intel, which converts the trained model for optimal execution of final stages for violent activity prediction. Anuja Naik et al. [3]in their survey mentioned various approaches for automated violence detection by using Optical Flow wherein they stated Violent flows (ViF) [14], Histogram of oriented optical flow (HOF) [15], Space-time interest point (STIP) [16] and Motion binary pattern, to detect abnormal behavior and other activities. A precis of various datasets for violence detection is also mentioned. Peipei Zhou et al. [4] introduced image acceleration field as a new input class, to extract motion attributes where each video is labelled as RGB images and the optimal flow is computed removing consecutive frames and acceleration field is obtained with respect to the optical flow field. Further, the model 'FightNet' is trained with three kinds of input modalities which are, RGB images for spatial network, optical flow images and lastly acceleration images. Lastly the output is determined by combining the three different input modalities.

A research by Maria Andersson, Joakim Rydell et al. [5] presented a Hidden Markov Model (HMM) approach to detect unusual behavior. The input gathering is done by distributed heterogenous sensory data such as CCTV, TIR (Thermal Infrared) cameras, radars or acoustic sensors. Here people are not viewed as each individual input rather they are viewed as a single entity. The HMM detects unusual behavior using a stochastic model and predicts output using various input streams. Sensors such as Thermal Infrared will give thermal images used to detect the various features improving the HMM. The weapon detection will be carried out by high quality Radar system, CCTV and TIR. these inputs sent to HMM determines the final output system. Michael S. Ryoo, J.K. Aggarwal et al. [6] stated Human Interaction Recognition from continuous videos using 10-fold leave- one-out cross validation. Activities like shaking, hugging, kicking, etc are recorded. The UT interaction dataset [16] was selected which had all 6 above mentioned activities covered over 20 videos divided among 2 sets composing of different weather conditions and subjects with different clothing. The model was trained on 9 sequences of activities out of 10 and tested on the remaining one. The subjects were required to depict the activity accurately. Irshad Ali et al. [7] presented an automatic approach on human detection and tracking using Viola and Jones AdaBoost cascade classifier, particle filter for tracking and color histograms. Integrating confirmation-by-classification method to recover from misses and reducing false detection along with tracking temporary occlusions, caused the results to improve. They experimented with these classifiers and algorithms and produce an output which tracks human heads through occlusions. Afshin Dehghan et al. [8] stated two methods for pedestrian tracking in videos with different crowd density. For videos with low density, each person is first detected using a part-based human detector. For detecting, they employed global data association method based on Generalized Graphs for tracking each individual. In videos with high crowd density, individuals are tracked using a scene structured force model or a crowd flow modelling.

A technical approach by Mikel Rodriguez et al. [9] concluded that achieved person density estimates can improve the tracking performance and person localization. Person detection task can be formulated as reduction of a joint energy function combining crowd density estimation and person localization over the constraints. Optimization of this energy function improving the performance of person detection and tracking in crowds is depicted. Roberto Olmos et al. [10] stated gun detection by a Faster R-

CNN [29] model trained on the available dataset and by using a sliding window and region proposal approach, the best classification model is assessed. Later on, a VGG-16 [17] based classification model is used that is pre-trained on the Imagenet dataset. A new metric of Alarm Activation per interval is implemented for a real-time automatic detection alarm system to prove the existence of a gun at the scene.

III. PROPOSED METHODOLOGY

Violence reduction is a core image processing and computer vision problem as signified by profusion of research papers. The proposed approach here is cleaved into various steps and Section *A* describing various methodologies and algorithms along with terminologies needed for each step while Section *B* elucidating the altogether flow to proffer the overall architecture.

A. System Overview

1) *Violence Detection:* The dataset used here [18] is collected from YouTube and CCTV videos comprising of 300 videos -150 fight and 150 non-fight sequences in it from various locations like street, bus, cafe, etc with each video being 2-3 seconds long approximately. Scenarios like kicking, fisting, wrestling, hitting with an object are included here. We followed a similar approach of [19] and used their model to train it on the above dataset. The model is illustrated in Fig. 1

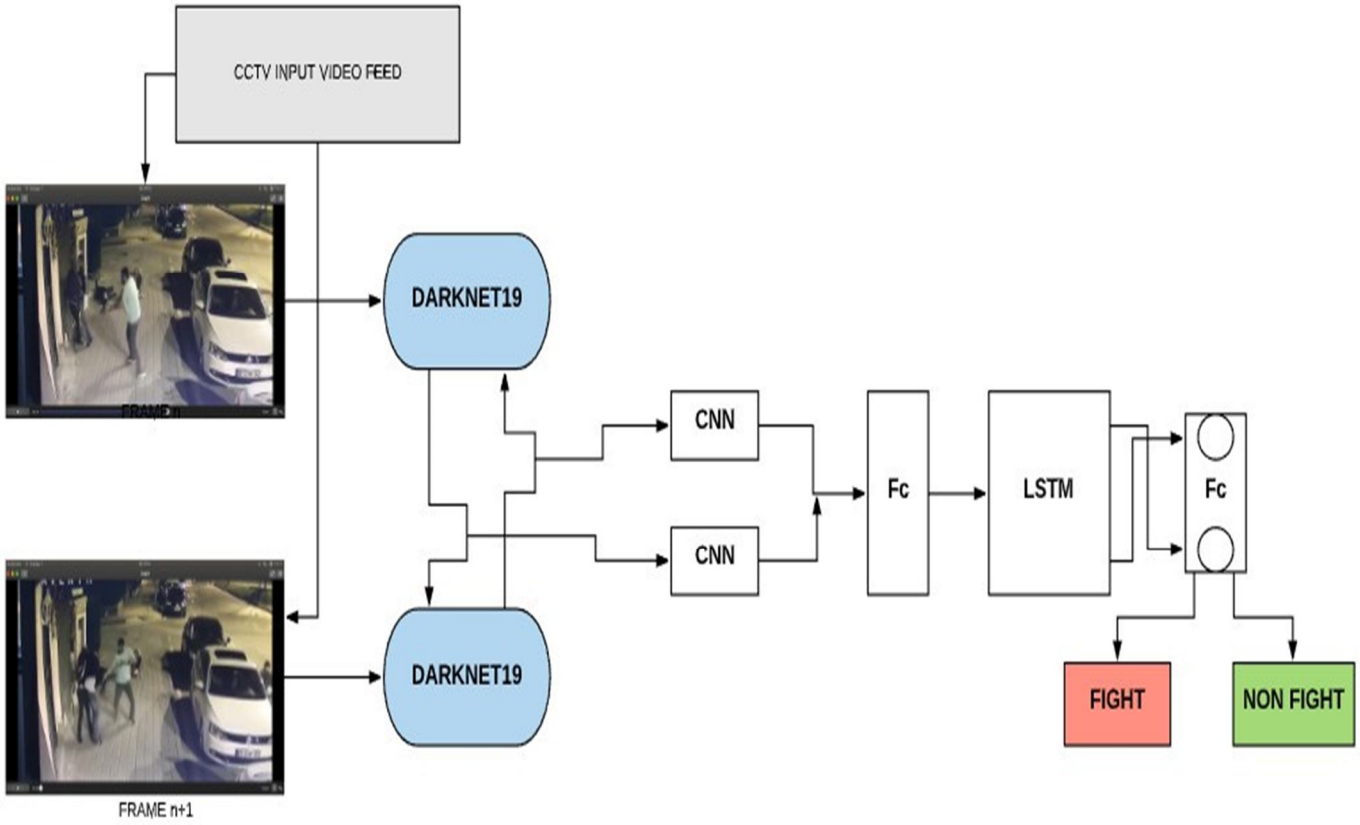


Fig 1 Fight/ Non-Fight Detection Model

The model takes 2 input frames which are processed by a pre-trained CNN, Darknet19 [20]. YOLO V1 [21] uses DarkNet framework trained on ImageNet-1000 [22] dataset as its feature extractor. The output of 2 frames from the bottom layer (for low-level feature learning) and top layer (for high-level feature learning) of model is merged and then fed to an additional CNN. These additional CNNs compare the 2 frames' features and try learning the appearance invariant features along with local motion features. Later, the CNNs outputs are passed to a fully connected layer and LSTM cell to learn global temporal features. At last, LSTM outputs are fed to a fully-connected layer which is a 2-neuron classifier representing fight and non- fight categories.

The pre-trained model is implemented by Darknet19 due to its accuracy on ImageNet and the above real-time prediction. To avoid the degradation issue, the additional CNN are implemented by the residual layers, since the darknet already has 19 convolutional layers.

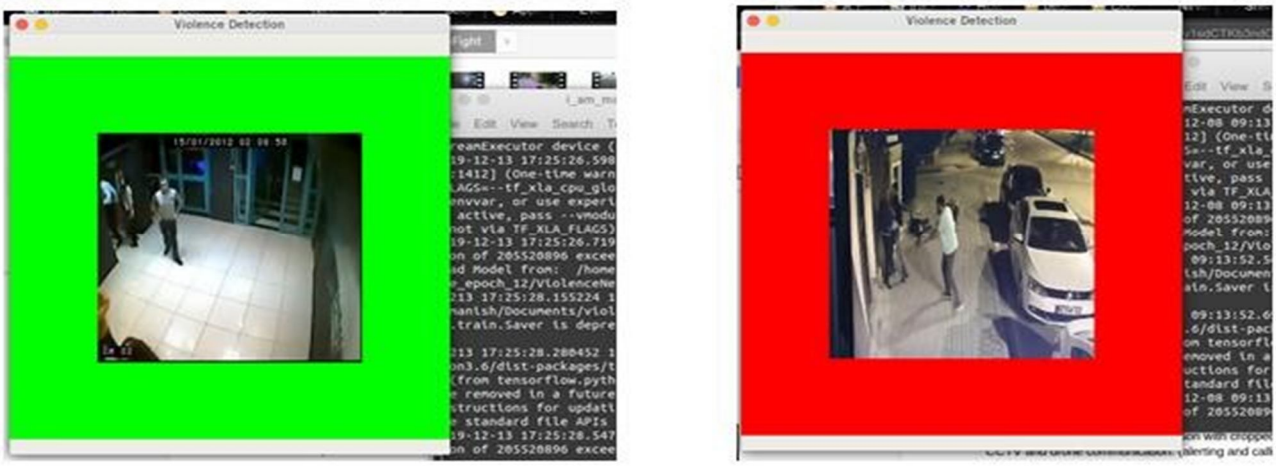


Fig 2. Green (Non-Fight) and Red (Fight) Detection

Thus, as shown in above screenshots (Fig. 2) where in a live CCTV feed, if violence is detected, its background turns red and remains green otherwise.

2) *Human Activity Recognition and Tracking:* Here, yolo along with deep sort is used to do person detection and to track them. YOLO v3 [23] uses a Darknet variant with a 106-layer fully convolutional architecture. Since, it makes detection at three different scales and applying 1x1 detection filters on features of 3 different places in a network, it is one of the faster object detection algorithms used where there's need of real-time processing along with less loss of accuracy.
Deep sort [24], on the other hand goes hand-in-hand with yolo to improve its performance. It's trained on millions of human images and extricates a 128-dimensional vector for each box, capturing the prominent features of each. Tracking is not just done by visualization of bounding boxes or distance or velocity, but also what the person looks like. It's useful in cases where people occlude or bounding boxes are too close. Deep sort computes deep features for each bounding box and uses similarity between them to factor the tracking.

Fig 3. Bounding Boxes for Person Detection

On feeding live CCTV input feed, the results were as expected and are shown in the Fig. 3
Further, to detect actions like sit, stand, walk, push, we referred another approach proposed by [25].
The proposed 3CDNN model gives predictions for all classes of events like sit, stand, push, fight, walk that could possibly occur in an input video. The model has 5 repetitive layers of convolution-3D and max-pooling with dropouts in middle. It uses softmax as its activation function. The model outputs the probability of set of images to represent each class of event. The event having highest probability is picked and is selected for labelling. A confusion matrix (shown in Fig.4) represents an event matrix.

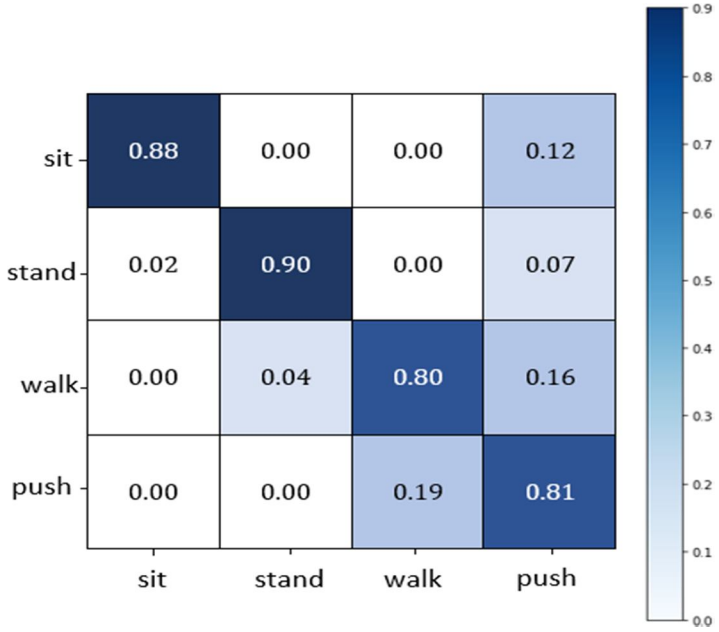


Fig 4 Event Matrix

First, the frames from video files are extracted. Individual person is detected along with coordinates of each using yolo v3 [23]. Coordinates of each person are tracked for all next frames. Then all the images are concatenated together which forms an input to the CNN model which predicts the probability of an event occurring. Finally, all frames with corresponding prediction label is stored in the same order in which they can be joined to form an output video.

Fig 5. Human Activity Recognition

Thus, using the above approach in a live CCTV feed, person pushing, sitting, etc. is detected (Fig. 5) and thus it can successfully contribute towards improving violence detection in videos.

3) *Weapon Detection:* Two approaches were followed to detect weapon in a CCTV video feed. One is by using computer vision and the other is by traditional transfer learning. We compared the outputs from two of those and the latter one was found to be better. For the time being, we just focused on gun detection. A Haar cascade classifier was used to detect guns by computer vision approach.

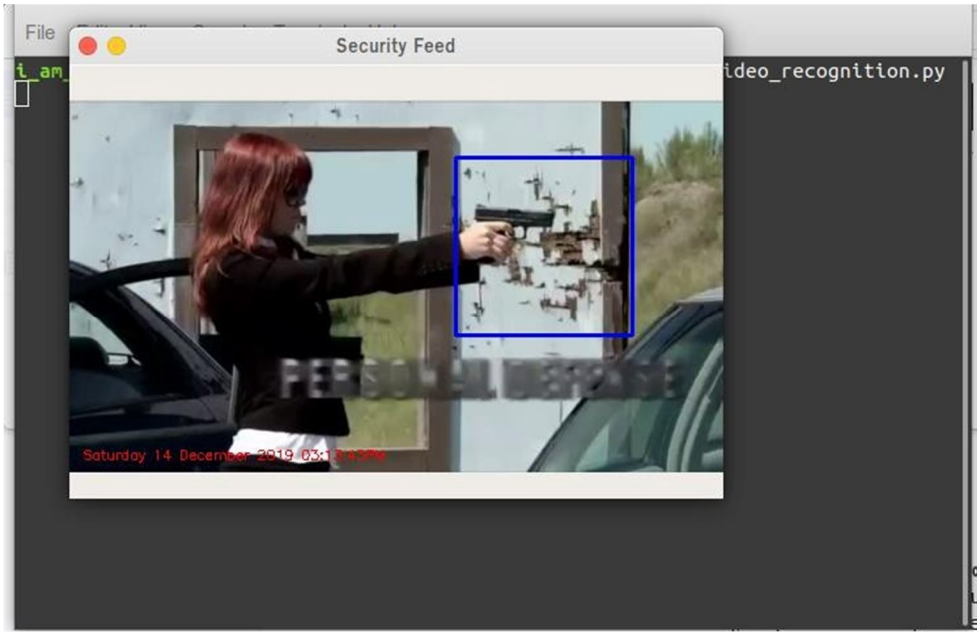


Fig 6. Gun Detection with Haar Cascade Classifier

As seen in Fig. 6, a gun is detected by a Haar cascade.

But it seemed to be vague when tested with different types of video feeds. So, we referred another approach by [26]. The dataset consisted of 1045 pistols and rifles images with an annotated JavaScript file. The model was trained on RetinaNet architecture [27] using fastai [28] library by Facebook. In object detection, there's extreme foreground and background class imbalance issue in case of crowded places or less contrast objects. In RetinaNet, loss focuses on hard samples since the lower loss is contributed by easy negative samples, ultimately improving the prediction. RetinaNet has backbone of ResNet+FPN for feature extraction along with two task-oriented subnetworks for bounding box and classification which makes it outperform Faster R-CNNs [29].

To test the performance, average precision score was used to evaluate the performance of each class.

Fig 7. Screenshots of Gun Detection using RetinaNet Architecture

As seen in the above Figure (Fig. 7), pistol is detected in an input CCTV feed with a certain accuracy level. The results produced were quite adaptable here. By taking datasets of knives, blades, etc. into consideration, we can have the model to detect other weapons too.

4) *Image Segmentation:* In a CCTVs field of view when a person comes into the frame window, then the person gets the bounding box with a tag and either of the two things will happen, the person in the frame is detected and if that person is treated as a new object then the images of the person are stored in a folder, this is done since the time that person entered the frame, till he or she goes out of the frame. Secondly, if the same person re-enters the frame then the model will recognize that person from the previous set of detected persons and it should give the same tag to the person. The algorithm for the image segmentation uses a re-identification neural network model [30]. This model identifies each individual person and thus produces results accordingly. If a total of M people is seen in the video, the output is a list of M folders, each belonging to one person. Each folder contains cropped images of that person from all frames in the video.

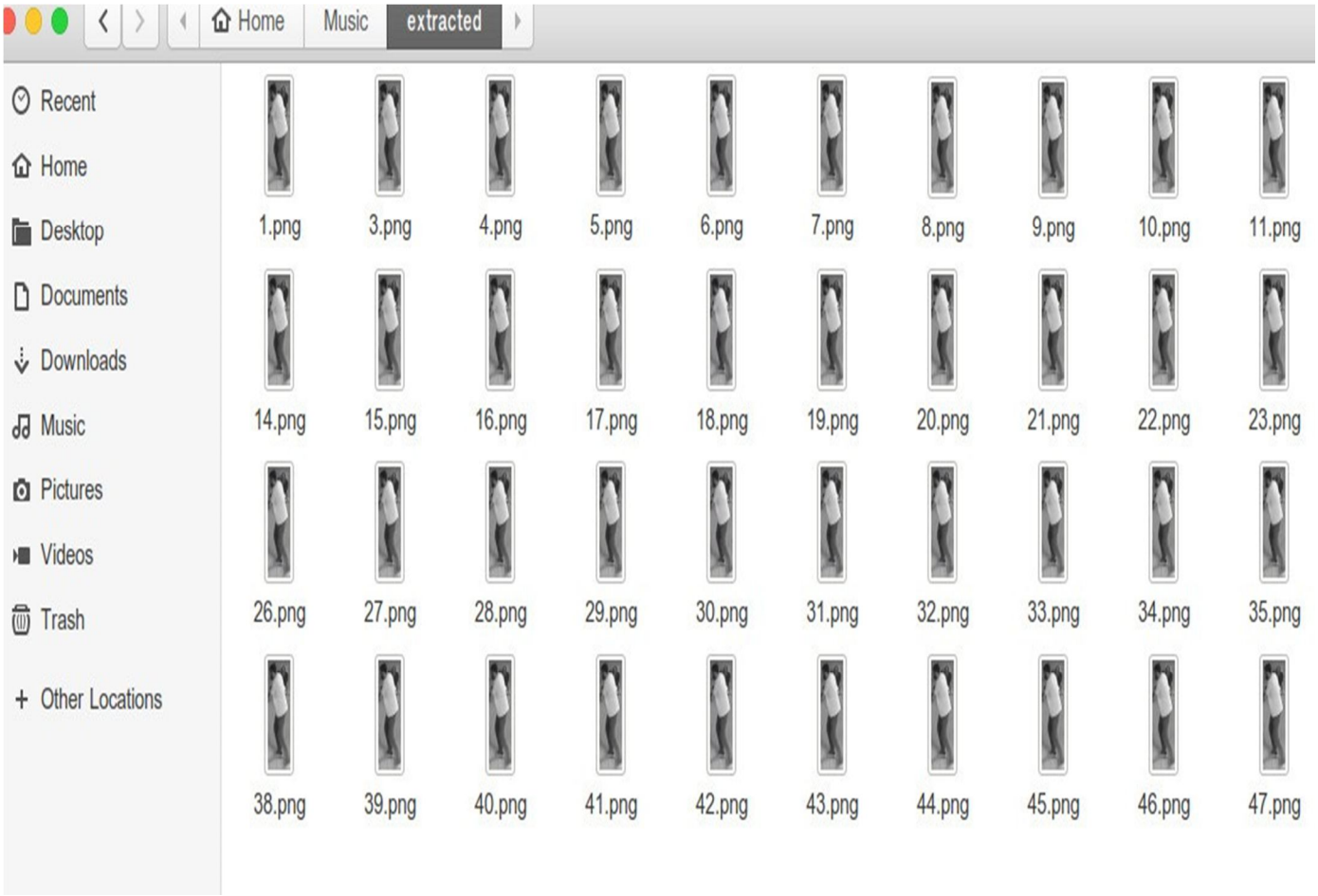


Fig 8. Images of identified suspect

The above screenshot (Fig. 8) shows a cropped image of person detected in the footage.

5) *Drone tracking/ Chasing:* For the drone to be able to track the person, depth prediction can be used. But, since it can be difficult to obtain, some tricks with the bounding box can be applied here. Once the drone detects the required person or the suspect, there's a bounding box around it. It has four numbers describing the box, it's bottom-right corner (x1, y1) and its top left corner (x, y). Mentioned those, area of the rectangular box along with its center can be computed. Area is width*height
i.e. (x1-x) * (y1-y) and as for the center, it's (x+x1)/2 and (y+y1)/2.
This information can be useful for the drone to track any detected person. The center of rectangle tells whether the person is in the center of the picture or if in left or right. By this, drone is instructed to turn left or right seeing its vertical axis, aiming to bring the person in the center of the shot. If the person is detected in the upper image part, then drone is commanded to raise up or down. Using the area of the rectangle, closeness of the person can be roughly estimated. A tiny rectangle would indicate that the person is far away, while a bigger rectangle indicates the person is very close. Equally, drone can be commanded to move ahead or backward to bring it to the desired distance to the person.

The pseudocode is framed as follows.

def drone track ((x, y), (x1, y1)):

#variables

area, center = calculate_area_and_center ((x, y), (x1, y1)) #obtain c center and y center between 0.0 and 1.0 normalized_center[c] = center[c]/image.width normalized_center[c1] = center[c1] / image.width

if normalized_center[c] > 0.6:

"turn_right"

elif normalized_center[c] < 0.4:

"turn_left"

if normalized_center[c1] > 0.6:

"raise upwards”

elif normalized_center[c1] < 0.4:

"raise downwards"

#if the area is too big move backwards if area is too big:

"move backwards" elif:

"move ahead" Return

Fig 9. Drone follows the suspect [31]

As seen in the above image from the drone’s perspective, drone tracks the person based on bounding box trick as mentioned above.

6) *Communication with Drones:* The Micro Air Vehicle communication protocol (MAVLink) [32] is a point-to-point communication protocol allowing entities to share information which can be used as a lightweight messaging protocol for bidirectional communication with drones and between on-board drone components with Ground Control Station (GCS) in between. It is designed as a header-only message marshaling library and can be used to transmit the orientation of the vehicle, its GPS location and speed. CCTV and drone use MAVlink protocol for communication and the drone can reach the CCTV’s location by using simple Google Maps API.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue XII, Dec 2019- Available at www.ijraset.com

B. Framework and Illustrations

All the above techniques and methodologies are piped into a flow to propose a way to reduce violence in a particular scenario. The Fig. 10 below gives a detailed flow of the entire architecture.

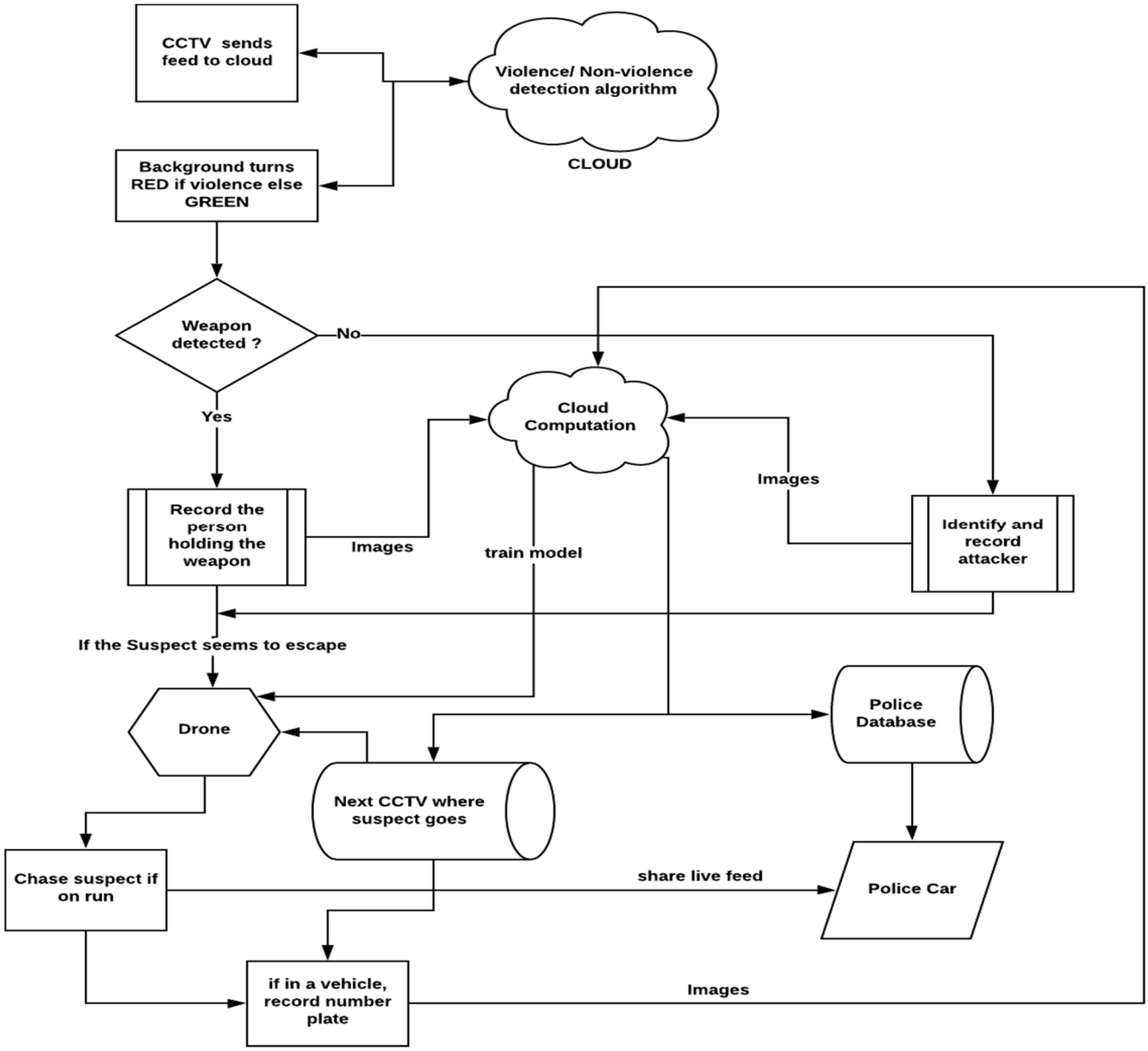


Fig 10. Flow-Chart explaining the entire architecture

Initially, consider a 24x7 monitoring CCTV that constantly updates it video feed to cloud. The cloud, which is a hybrid type of cloud, uses Fig.1 model to detect violence or non-violence i.e. fight or non-fight in the CCTV feed. The background turns red when violence is detected and remains green otherwise. Then, person recognition and simultaneously human activity recognition like sit, stand, push, fight, etc is detected over the captured red violent scene.

Next, an additional weapon i.e. gun detection is implemented over it which recognizes malicious objects in the video. Now when a weapon is detected, it will start to capture only the holder of the weapon regardless of him drawing the weapon using image segmentation, and will upload the captured images on the cloud. These images will be given to a basic convolutional neural network (CNN) model where it will be trained on these images to detect the perpetrator. The cloud's database will be accessible to the authoritative station preferably the police headquarters. Now, the CCTV alerts the drone through mavlink protocol [32] and drone gets ready from standby mode to reach at given location through maps API.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue XII, Dec 2019- Available at www.ijraset.com

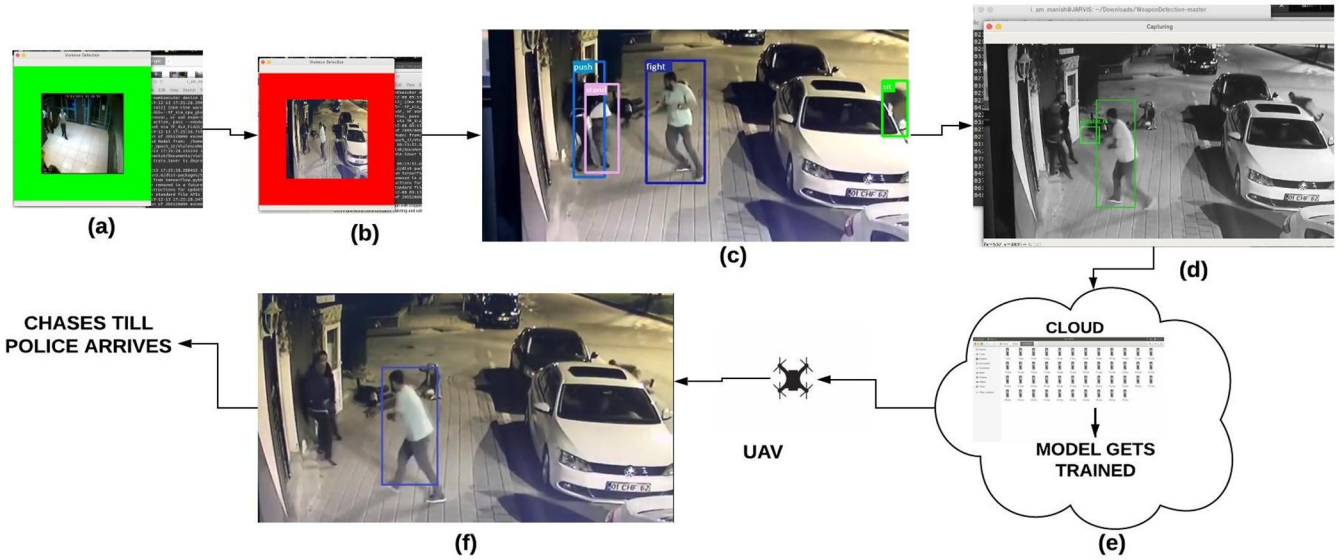


Fig 11 (a). CCTV is green normally and continuously sending data to cloud (b). the cloud detects violent scene through CCTV's perspective (c). human activity recognition (d). weapon (gun) detection with suspect (e). suspect's segmented images sent to cloud for model training (f). suspect detected by the drone.

All CCTV's in the vicinity will send the respective suspect images to cloud. The model gets trained with all suspect images sent from all CCTV coverages.

The following representation (Fig. 12) illustrates the network architecture and how each drone is assigned to each regional area. The Area_1 is magnified with CCTVs detecting the crime and uploading it to the cloud.

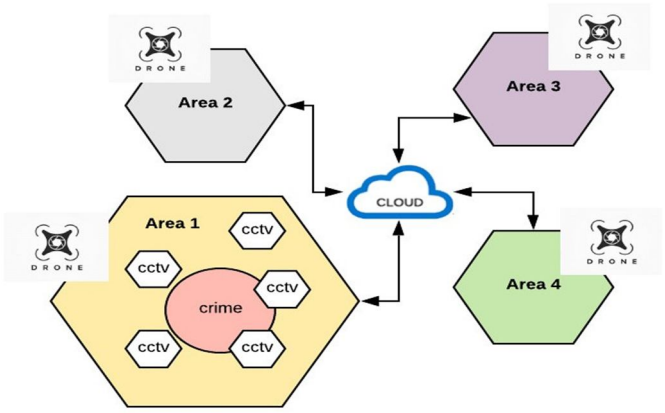


Fig 12 Representation of the entire network architecture, with one drone per region.

Now, there is a huge chance that the perpetrator leaves the field of view of the CCTV it was captured on, however the neighboring CCTV's locations will be on the cloud. So, when the perpetrator tries to escape, the neighboring CCTV will already be provided with the dataset of that person and if that same person is detected in that CCTV, it'll act accordingly and thus, this will help in continuous tracking of the suspect.

If the perpetrator tries to avoid the field of view of the CCTV coverage, the drone can be used to pursue the target and try bringing the suspect in the center of the bounding box, maintaining a safe distance. The live feed via the cloud will be accessible to the central authority which can take action depending on the situation.

There may be a possibility that the perpetrator tries to escape in a vehicle. In this scenario a license plate detection algorithm [33] can be implemented in the cloud and thus drone will be able to track the detected license plate of that vehicle.

If the perpetrator succeeds in leaving the drone's range, then using the scatternet like architecture the drone will contact the nearby drone to continue the chase, thereby maintaining a constant eye on the criminal, and broadcasting a live video feed that will be used by the authority.

C. Limitations

A high level of accuracy is expected by the subjects in depicting the activities, there are always chances of violence going undetected when these activities aren't captured properly. The CCTV camera identifies the subject by its basic appearance (like clothing, hair colors and facial features), the issue arises when the appearance of an innocent person matches with the suspect, it may cause an error and yield an inaccurate prediction thereby sending wrong information to the central authority. However, causality cannot be assumed since the data that is obtained is correlational.

Better prediction rate of the entire architecture is pertinent with high end CCTV camera, better drone capabilities and better processors for deep learning model computations. Recommended hardware specifications to be used is an NVIDIA GeForce 1080Gi processor with a frame rate of 23ms/Frame. Selecting a DJI Phantom 3 Standard as a drone can remarkably increase the flight time to 25 mins and range up to 0.5 miles (1 kilometer) within the network. If in cold areas then it'll affect battery life as per conditions. Moreover, network gets congested due to overload of heterogeneous data (i.e., text, images, videos etc.) and repetition of queries causing slow access time of drone.

However, the cameras which are installed by the government are not up to the mark and may hamper the accuracy of the model. The highest available resolution is 700 TVL (Television lines) for CCTV camera. A trade-off between efficiency of the model and the costing must be minimized.

In CCTV, a hard disk failure or a corrupted memory card can result in loss of consequential evidence as they were never backed up or printed. In case of internet shutdown, these cameras will lose their connection from servers and the cloud may not be able to detect the current situation or perform any computational activities.

IV. CONCLUSION AND FUTURE SCOPE

This work elucidates an entire architecture of a Smart Surveillance system, from the detection of violence till stopping the criminals from getting escaped which includes a combination of state-of-the-art deep learning and image processing techniques with drones. It detects one or more subjects involved in treacherousness with the help of CCTV cameras which interprets the information faster and more efficiently than any human observer. This architecture can be customized for serving a specific purpose in a dynamically flexible environment. This structure will be feasible in areas where police or security intervention is less, especially in countries like India, Bangladesh, etc. This scalable and effective approach can be useful in reducing violent activities and can come to aid in range of fields from public areas like bus-stops, schools, shopping malls, parks, events like Kumbh mela to military, traffic flow monitoring and disaster management too.

The overall architecture can be improved by increasing the amount of data required for training the models. Use of other state- of-the-art models or faster image processing methods can help increase accuracy without much less loss. Accurate depiction of human activities using high end cameras can improve the accuracy of the overall architecture. Better drone equipment's, enhancing its mobility and maneuvering can ameliorate the performance and almost reduce the chance of suspect getting escaped. Also, powerful cloud infrastructure along with a greater number of CCTVs can make the model more robust thus improving its performance.

REFERENCES

[1] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques", Computer Analysis of Images and Patterns-CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Heidelberg

[2] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq and Sung Wook Baik," Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network ", Sensors (ISSN 1424-8220; CODEN: SENSC9)

[3] Anuja Jana Naik, M.T. Gopalakrishna, "Violence Detection in Surveillance Video-A survey", International Journal of Latest Research in Engineering and Technology (IJLRET) ISSN: 2454-5031

[4] P. Zhou, Q. Ding, H. Luo, and X. Hou, ''Violent interaction detection in video based on deep learning,'' J. Phys., Conf. Ser., vol. 844, no. 1, 2017, Art. no. 12044.

[5] Andersson M., and J. Rydell, "Estimation of crowd behavior using sensor networks and sensor fusion", FUSION'09, 2009, pp. 396-403.

[6] R. Vezzani, D. Baltieri, and R. Cucchiara. "HMM based action recognition with projection histogram features", ICPR, 2010

[7] Matthew N. Dailey, Irshad Ali, "Multiple human tracking in high-density crowds", Image and Vision Computing, vol. 30, pp. 966–977, 2012.

[8] A. Dehghan, H. Idrees, A. R. Zamir, M. Shah, "Automatic detection and tracking of pedestrians in videos with various crowd densities", Pedestrian and Evacuation Dynamics, Springer, 2012, pp. 3–19.

[9] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert., "Density-aware person detection and tracking in crowds", ICCV, 2011

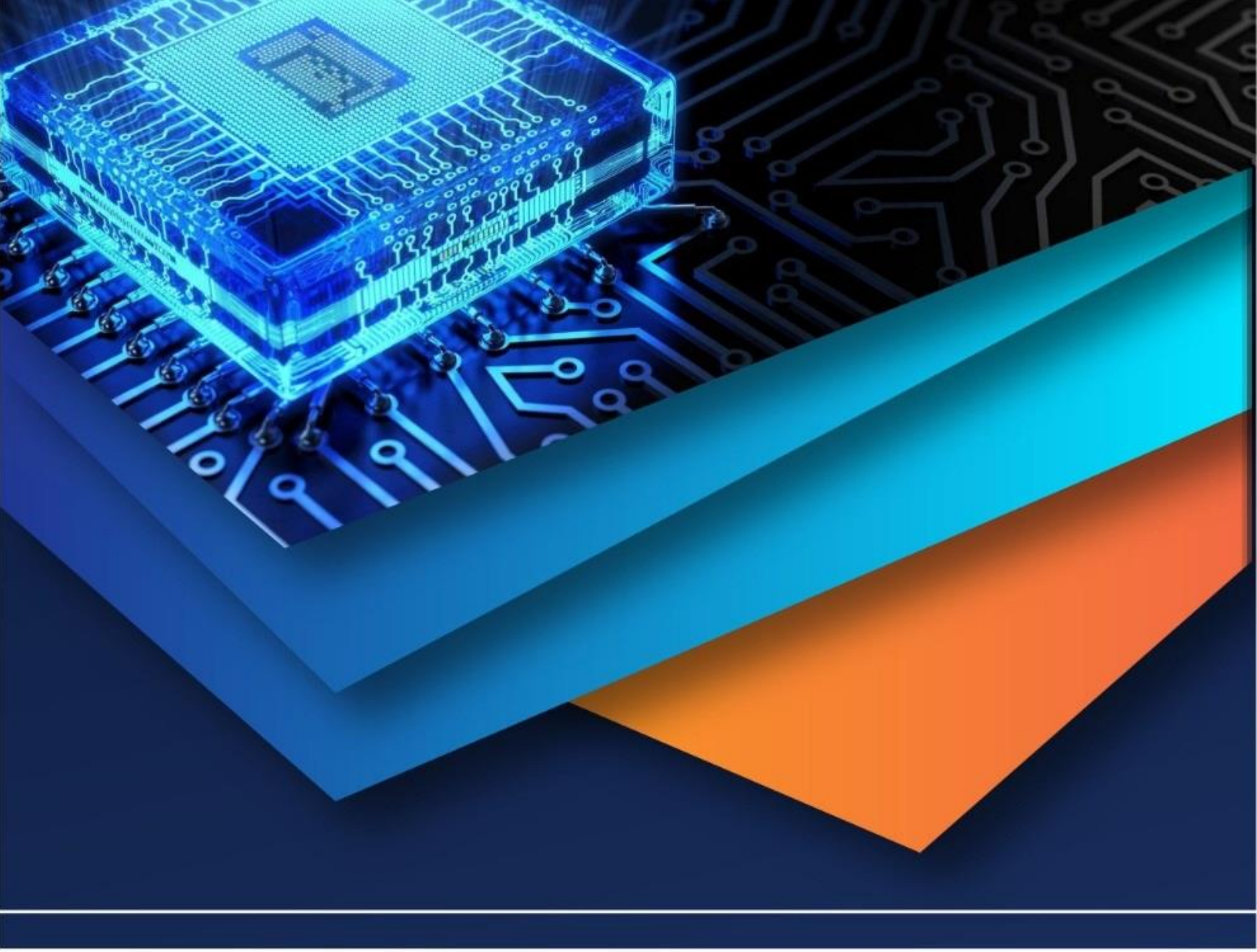
[10] Roberto Olmos, Siham Tabik and Francisco Herrera, "Automatic handgun detection alarm in videos using deep learning", Neurocomputing, volume 275, pp. 66-72, 31 January 2018

[11] Lewis, D.: Naive Bayes at Forty: "The independence assumption in information retrieval", European Conference on Machine Learning. pp. 4–15 (1998)

[12] Laptev, I., "On space-time interest points", International Journal of Computer Vision. vol. 64, pp. 107–123 (2005)

[13] Chen, M., Hauptmann, A., “MoSIFT: Recognizing human actions in surveillance videos”, Tech. rep., Carnegie Mellon University, Pittsburgh, USA (2009)
[14] Tal Hassnar, Y. Itcher, O. Kliper-Gross, Violent Flows, “Real-Time Detection Of Violent Crowd Behavior”, Computer Vision on Pattern Recognition Workshops (CVPRW), 2012, pp. 1-6.
[15] Yan Chen, Ling Zhang, Biyi Lin, Yong Xu, XiaoboRen, “Fighting Detection Based on Optical Flow Context Histogram”, Second International Conference on Innovations in Bio-inspired Computing and Applications, IEEE-2011, pp:95-98.
[16] Ryoo, M. S. and Aggarwal, J. K., "Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities”, IEEE International Conference on Computer Vision (ICCV), 2009, Kyoto, Japan,
[17] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”, ICLR, 2015.
[18] Ş. Aktı, G.A. Tataroğlu, H.K. Ekenel, “Vision-based Fight Detection from Surveillance Cameras”, IEEE/EURASIP 9th International Conference on Image Processing Theory, Tools and Applications, Istanbul, Turkey, November 2019.
[19] Violence Detection by CNN + LSTM, http://joshua-p-r-pan.blogspot.com/2018/05/violence-detection-by-cnn-lstm.html

[20] J. Redmon and A. Farhadi., "Yolo9000: Better, faster, stronger", Computer Vision and Pattern Recognition, arXiv:1612.08242
[21] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi., "You Only Look Once: Unified, Real-Time Object Detection", arXiv:1506.02640 [cs.CV]
[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei., "ImageNet: A Large-Scale Hierarchical Image Database", CVPR 09, 2009
[23] J. Redmon and A. Farhadi., "Yolov3: An incremental improvement", arXiv preprint arXiv:1804.02767, 2018.
[24] Nicolai Wojke, Alex Bewley, Dietrich Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric", arXiv preprint arXiv:1703.07402 [cs.CV]
[25] Shiv-Kumar-Yadav, Event-Detection-In-Classroom, https://github.com/Shiv-Kumar-Yadav9/Event-Detection-In-Classroom
[26] Rajas Kakodkar, Ethan Alberto, Anikta Shirodkar, Ashton Rodrigues, "Real Time Gun Detection Classifier", Department of Computer Engineering, Padre Conceicao College of Engineering
[27] Tsung-Yi Lin, Priya Goyal, "Focal Loss for Dense Object Detection", arXiv preprint arXiv:1708.02002 [cs.CV]
[28] 'fast.ai' by Facebook, https://www.fast.ai/
[29] Ross Girshick., "Fast r-cnn", IEEE International Conference on Computer Vision, pages 1440–1448, 2015.
[30] Liang Zheng, Yi Yang, "Person Re-identification: Past, Present and Future", arXiv:1610.02984 [cs.CV]
[31] Sarthak Jain, nanonet blog, https://nanonets.com/blog/how-to-add-person-tracking-to-a-drone-using-deep-learning-and-nanonets/
[32] mavlink guide, https://mavlink.io/en/,
[33] Cheng-Hung Lin, Yong-Sin Lin, "An efficient license plate recognition system using convolution neural networks", ICASI-2018, doi: 10.1109/ICASI.2018.8394573



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY