Suspicious Activity Detection And Tracking In Surveillance Videos

S. Adarsh
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Giridhar Kannan
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

S. Poorvaja
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

B. S. Vidhyasagar
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

J. Arunnehru
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

***Abstract*— Recently, the utilization of security cameras for crime prevention and early detection of emergencies worldwide has been increased. The expansion in the use of surveillance cameras has aided in crime detection, captures and crime prevention. However, in many cases, it will be recognized and resolved after the occurrence of the crime and concerning continuous surveillance, the weight on the surveillance side is overwhelming and there are situations where suspicious activity may go unnoticed. To overcome this obstacle, a surveillance system that employs Human Activity Recognition techniques which can efficiently decide if the objective individual is an ordinary individual or a suspicious individual can be deployed. It is likewise expected that establishing detection systems can act as a hindrance against crime. This paper proposes a surveillance system that utilizes YOLO and ResNet for detecting suspicious individuals and activities.**

Keywords—Surveillance, Crime Detection, Suspicious Activity, Human Activity Recognition, YOLO, ResNet.

I. Introduction

With the increase in globalization and technological revolution, the necessity of the hour is especially being focused around security generally including physical, cyber, intellectual, etc. the most priority of protection obviously being the human life, has become crucial with the rise in hostile behaviour in recent times. So as to detect these behaviour and to avoid casualties before things get out of hands, an efficient system to research human behaviour and to predict whether or not they are a threat or not is very essential to cope up with the radically growing society. Behaviour detection is a technology related to computer vision that deals with detecting instances of humans in certain scenarios. It has applications in many areas of video surveillance.

As of now, acts of crime are increasing in a global scale and are proven to be unpredictable and planned in such a way to throw off security officials. to assist combat this a predictive method which will transcribe a given set of actions/gestures so as to foresee any upcoming potential threats in order that they will be addressed beforehand. Prevention is always better than cure thus preventing a threat before it gets administered can reduce the amount of casualties and reduce the general negative impact created.

This paper presents an approach which uses You Only Look Once (YOLO) [1], a Convolutional Neural Network (CNN) Model for object detection and a Residual Network (ResNet-34) [2] trained with UT-Interaction dataset [3] which detects suspicious individuals and hostile behaviour.

II. LITERATURE SURVEY

There are numerous methods and approaches for detecting abnormal behaviour in a surveillance system. Generally, the existing techniques can be categorized into three learning methods.

The first technique is the supervised learning method. In this method, the labels and parameters of the different behaviour are given as the input. The test data and training data are compared and the behaviour is classified accordingly. Ko et al. [4] proposes a method which assumes that human behaviour comprised of sequences of body postures with respect to their temporal characteristics, contributes to the recognition of the displayed behaviour. YOLO is used to detect and differentiate the humans detected in the frame. The VGG16 CNN model then classifies the behaviour of the detected humans. Long Short-Term Memory (LSTM) network then detects the abnormal behaviour.

The second one is the unsupervised learning method. In this method, the network forms the given data without any labels. To cluster the data into their respective behaviour, the data that have identical characteristics are formed as a cluster group. Zhang et al. [5] proposes a three-phase approach which uses Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) for creating classifiers. In the next phase, the irregular events are recognized by an ensemble learning algorithm. In the final phase, abnormal behaviour models are derived from normal behaviour model to reduce the rate of false positives, which are the incorrectly classified outputs.

The third one is the semi-supervised learning method. This method requires a combination of both labelled and unlabelled data. This method can prove to both advantageous and disadvantageous. Insufficient labels won't drastically affect the training of the model where the unlabelled data can be used. Li et al. [6] present a four-step approach to identify the irregularity which can construct normal behaviour models that can efficiently avoid encountering the over-fitting problem by using hidden Makarov Models (HMM) and Maximum a posteriori (MAP).

III. Proposed Methodology

In this paper, an approach that uses two CNN models for detecting and classifying the exhibited behaviour of the detected objects is proposed.

The proposed system ideally uses the input frames from the footage recorded in the security camera, further which it is processed by a YOLO model pre-trained on the COCO [7] dataset, where the individuals present in the footage are detected. After which the frames are then sent to the Residual network (ResNet-34) model. The ResNet– 34 is trained with the UT-Interaction dataset in order to classify the actions present in the input frames. Finally, once the classification is done, the behaviour is labelled on the screen accordingly. The dynamic training algorithms used in the ResNet-34 provide more scope on detecting a wider range of actions in order to improve the efficiency of threat detection.

The workflow presented in this paper is represented in Figure 1.

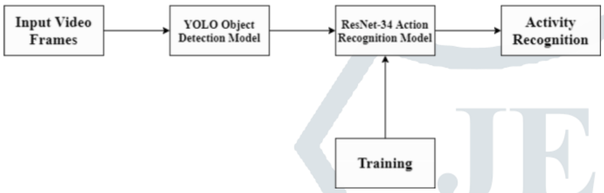

Input Video Frames
YOLO Object Detection Model
ResNet-34 Action Recognition Model
Activity Recognition
Training

Fig. 1: Workflow of the proposed system

IV. YOLO (You Only Look Once)

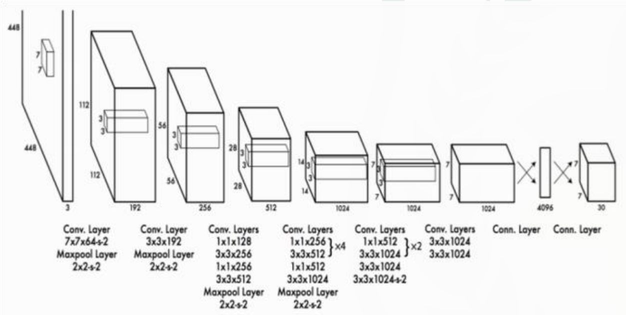


Fig. 2: Architecture of YOLO [1]

A. *Object Detection*

YOLO [1] is a progressive approach which employs advanced techniques to detect objects wherein it applies a single CNN to the entire video frame which then divides the frame into a grid. Bounding boxes represent the rectangle that covers an object. Each grid is then evaluated by measuring the confidence score of each bounding box against the threshold scores.

The YOLO model is faster and more accurate than previous models. For changing speed and accuracy of the model, there is no need to retrain the model. The trade-off between speed and accuracy can easily be achieved by changing the size of the model.

YOLO resizes the assessed image to 448 x 448 pixels. The image is then processed using the CNN and is output in the resolution 7 x 7 x 30 tensor. Tensor gives the data about the bounding box rectangle and the confidence score distribution of all the system trained classes.

B. *Advantages over other detectors*

Region-Convoluted Neural Network (R-CNN) [8] is a popular method of anchored object detection. It is similar to YOLO in the fact that the datasets are fully end-to-end trained however; the steps are much more involved.

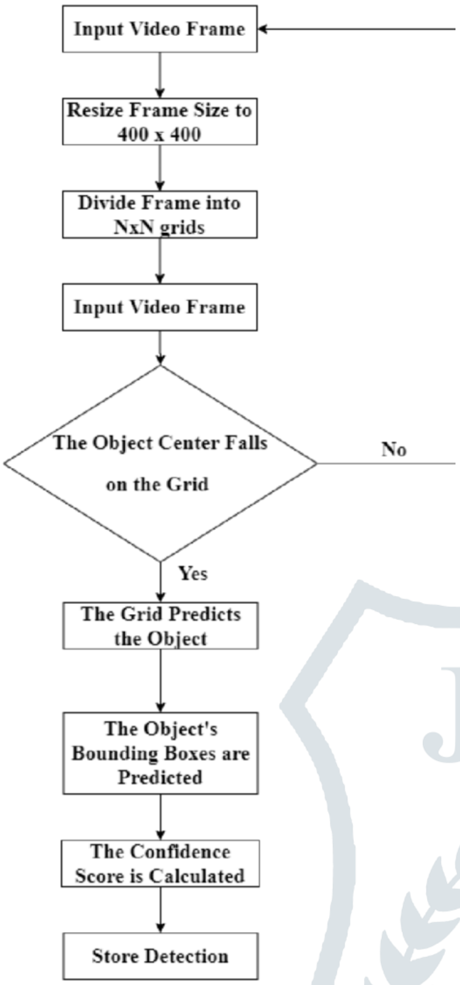
Single Shot Detector (SSD) [9] is a fairly new object detection system. It finds a healthy balance between accuracy and quickness. It runs a CNN over the input image one time and creates a feature map. SSD also uses anchor boxes of various aspect ratios, similar to Faster RCNN [10] and learns the off-dataset to a certain extent than learning the box. In order to hold the scale, SSD predicts bounding boxes after multiple convolutional layers. Since every CNN layer works at diverse levels, it is able to detect a number of objects of different classes.

Compared to Faster R-CNN, YOLO object detection has certain advantages,

- It runs a lot faster than Faster R-CNN because it has a simpler architecture and is trained to do bounding box regression and image classification at the same time. Due to its simplistic architecture, it is about 10 times faster than the Faster R-CNN.
- YOLO has superior image processing power, where it can process images at up to 90 frames per second.
- The image processing can be done with negligible latency over a few milliseconds. This means that live footage can be processed in real time which leads to better efficiency in real-life scenarios.
- Object classification and localization is done using a single CNN rather than using two-step methods.

V. ResNet-34 (Residual Network)

Fig. 3: Flowchart of YOLO model for object detection


Input Video Frame
Resize Frame Size to 400 x 400
Divide Frame into NxN grids
Input Video Frame
The Object Center Falls on the Grid
No
Yes
The Grid Predicts the Object
The Object's Bounding Boxes are Predicted
The Confidence Score is Calculated
Store Detection

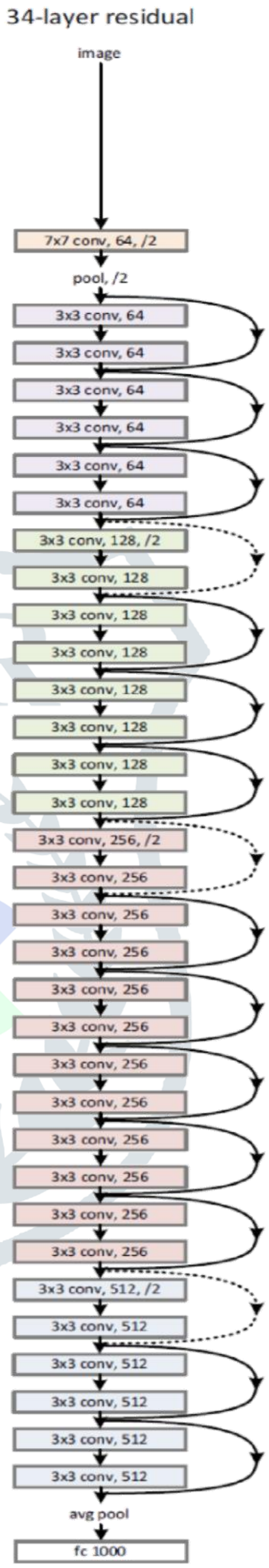


Fig. 4: Architecture of ResNet-34 [2]

A. Training and Classification

Resnet-34 is a Deep Neural Network which has 34 layers. It's unique features allow it to contain multiple layers whilst keeping the training of the neural networks simple and easy. This allows the network to deconstruct the data in the form of input frames and thoroughly analyze it to give more accurate outputs. Skip connections are present between various layers of the Resnet which allows adding the outputs of previous layers to the present layer. This enables us to train deep neural networks with better performance than previously possible. This results in extremely low error rate (~3.6%) in properly trained datasets. Resnet is used to accurately determine the actions of the various elements in real time. Due to its architecture, it is deployed to analyze video footage in real time and to deduce the actions performed.

UT-Interaction dataset is used for training the network. Set 2 of the dataset is given as the training data. The action labels and configuration files are accessed by the network and the model is trained on it. Set 1 of the dataset is given the testing data for the model. The dataset has four action labels which may potentially be considered as hostile. If the network classifies the behaviour as any one of those four action labels, the behaviour is classified and labeled as hostile behaviour.

Fig. 5: Workflow of the ResNet-34 model

VI. UT-Interaction Dataset

This dataset [3] consists of video clips of human to human interactions. The time interval and the bounding boxes along with the truth labels are provided. It is an efficient dataset for training behaviour recognition models used for surveillance and security systems. The complex actions are performed by multiple actors in two different settings. The dataset is divided into two sets. The first set consists of video clips showcasing two actors performing different activities in a fairly static environment with a different zoom level and minimal jitter. The next set consists of video clips showcasing multiple actors performing the same activities but in a slight dynamic environment with wind and more jitters.

The actions performed in this dataset are represented in Figure 6.

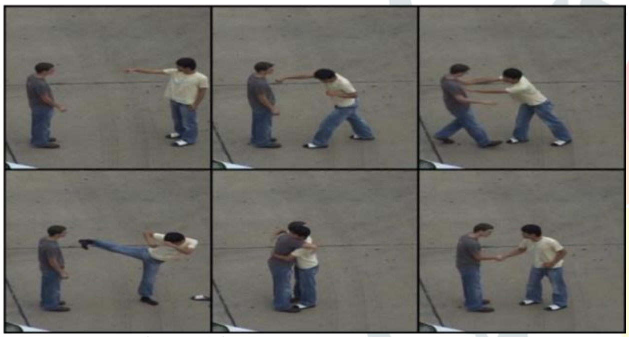


Fig. 6: Sample frames of UT-Interaction dataset containing action labels punching, hugging, hand shaking, pointing, pushing and kicking

VII. Result

A. Specifications

The methodology presented in this paper is developed and implemented on the system having the following specifications:

- OS: Windows 10 1903 (64 bit)
- CPU: Intel® Core™ i5-7300HQ CPU 3.50GHz
- RAM: 8 GB
- GPU: GTX 1060 Max-Q
- GPU RAM: 6 GB

B. Quantitative Results

The fundamental assessment metric for activity classification in this paper is the accuracy of the test data. A confusion matrix is plotted to calculate the accuracy of the model. The rows show the actual actions whereas the columns show the predicted actions in the confusion matrix.

Most of the results are fixed along the diagonal of the confusion matrix. The accuracy achieved by the proposed methodology is found to be 82%.

Table I: Classification accuracy of the UT-Interaction dataset shown by the confusion matrix.

	Pointing	Punching	Pushing	Kicking	Hugging	Hand Shaking
Pointing	0.95	0.04	0.0288	0.22	0.00	0.04
Punching	0.04	0.87	0.01	0.00	0.01	0.00
Pushing	0.01	0.03	0.85	0.01	0.22	0.05
Kicking	0.00	0.00	0.05	0.80	0.00	0.04
Hugging	0.00	0.02	0.22	0.00	0.75	0.00
Hand Shaking	0.04	0.00	0.02	0.00	0..00	0.70

Table 2: Performance comparison with other classification techniques.

Number	Method	Performance Score (%)
1	S. Saha [11]	58.10
2	M. S. Ryoo et al. [12]	70
3	S. Mukherjee et al. [13]	81.27
4	Presented Method	82

VIII. Conclusion and Future Work

This paper presents a methodology that uses two neural networks for the detection of humans and classification of their displayed behaviour. The video frames are fed to the YOLO network which detects the humans in the frame. These frames are then delivered to the ResNet-34 which then recognizes the activity displayed by the detected humans. If the activity displayed by the human is identified to be suspicious and recognized as one of the four hostile actions present in the action labels of the dataset, the activity is classified as hostile behaviour and a message is displayed. By comparing the accuracy of the presented model with that of the existing models, we can observe that the result accuracy of the proposed model is higher. In future, the detection algorithms can be further calibrated to need fewer frames to detect and classify the activity. It can be improved to detect more complex behaviors under different and more complex environments.

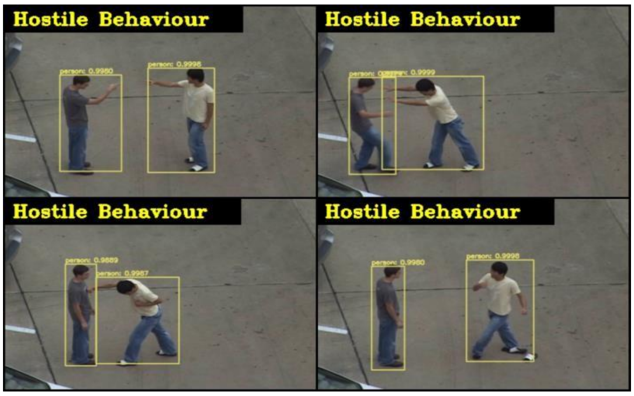


Fig. 7: Recognition of displayed behaviour in the frame

ACKNOWLEDGMENT

The research work presented in this paper is supported by SRM Institute of Science and Technology. We gratefully acknowledge the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai for providing the computing power and all the guiding faculties for lending their support for this project.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 779-788.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770778.

[3] Ryoo, M. S. and Aggarwal, J.K.: "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)", http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

[4] Ko, K., and Sim K.: "Deep convolutional framework for abnormal behaviour detection in a smart surveillance system," Engineering Applications of Artificial Intelligence, 67, 226-234. Doi:10.1016/j.engappai.2017.10.001. (2018).

[5] Hu, Derek & Zhang, X.P. & Yin, Jie & Zheng, Vincent & Yang, Qiang. (2009). Abnormal activity recognition based on HDP-HMM models. 1715-1720.

[6] Wang, Yongxiong & Li, Xuan & Ding, Xueming. (2016). Probabilistic Framework of Visual Anomaly Detection for Unbalanced Data. Neurocomputing. 201. 10.1016/j.neucom.2016.03.038.

[7] Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context." *ECCV* (2014).

[8] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587.

[9] Liu, Wei & Anguelov, Dragomir & Erhan, Dumitru & Szegedy, Christian & Reed, Scott & Fu, Cheng-Yang & Berg, Alexander. (2016). SSD: Single Shot MultiBox Detector. 9905. 21-37. 10.1007/978-3-319-46448-0_2.

[10] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[11] Singh, Gurkirt & Saha, Suman & Sapienza, Michael & Torr, Philip & Cuzzolin, Fabio. (2016). DEEP LEARNING FOR DETECTING MULTIPLE SPACE-TIME ACTION TUBES IN VIDEOS. 10.13140/RG.2.1.5129.6248.

[12] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2009, pp. 1593-1600.

[13] Mukherjee, Snehasis & Mallik, Apurbaa & Mukherjee, Dipti. (2015). human action recognition by dominant motion pattern. 10.1007/978-3-319-20904-3_43.