iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53958

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: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Abnormal Activity Recognition in Private Places Using Deep Learning: A Survey

Anjali Suthar[1], Prof. Jayandrath Mangrolia[2], Prof. Ravi Patel[3]
[1]*M. Tech. Student, Department of Artificial Intelligence, Charutar Vidya Mandal University, Anand, Gujarat, India*
[2, 3]*Assistant Professor, Department of Information Technology, Charutar Vidya Mandal University, Anand, Gujarat, India*

Abstract: ***The method of analysing human motion using computer and machine vision technologies is known as "human activity recognition," or HAR. One of the applications of human activity recognition in security systems is anomaly detection. Surveillance cameras have been widely placed as the foundation for video analysis as the demand for security has grown. Identifying aberrant behaviour necessitates considerable human effort, which is one of the major challenges in surveillance video analysis. It is important to set up video recording in order to detect unusual activity automatically. Our intelligent video surveillance system can detect an abnormality in a video using deep learning technologies. Real-time detection of activities is also conceivable, and these video frames will be saved in the system as images for the user to study. The proposed Abnormal Activity Recognition system was designed with the purpose of finding and detecting anomalies in the financial industry, especially in an ATM context, using a live stream. The first part of the research focuses on the use of image deep learning algorithms to recognise different products and detect anomalous behaviour utilizing ATM monitoring systems***
Keywords: YOLOv5, YOLOv4, YOLOv3 Convolution Neural Network (CNN), Object detection

I. INTRODUCTION

Real-time prediction of the presence of one or more objects, along with their classes and bounding boxes, is the task of computer vision that has taken the industry by storm. Object detection can use a neural network to classify and localize an object in the image. Benefitting from this capability, there is a tremendous amount of work that is being done in the different streams of life from facial recognition to autonomous driving cars, security applications and robotics[2].Modern detectors have been in the development to identify the objects in higher frame rate.

In this paper, proposes a deep learning-based system for detecting suspicious events in a bank-ATM context in real time. Bounding boxes, which functioned as classes in this case, are utilized to detect tagged items. This is then used to categories labels in video and forecast whether the occurrences are normal or abnormal. that result is calculated using the Motion representation Depth data is derived from the classes' bounding boxes. Then multi-stream CNNs are used to distinguish constituents and actions. The choosing of an appropriate algorithm for a certain job.

The results of this implementation were pretty remarkable, since the maximum accuracy and speed were seen. In contrast to traditional object detection utilising static pictures, video object detection detects things using video data. Autonomous driving and video surveillance are two applications that have played a significant influence in the advancement of video object detection.

Detecting objects in video required conducting object detection on each picture frame. Object detection techniques may be divided into two categories: (1) one-stage detectors and (2) two-stage detectors. One-stage detectors are frequently more computationally efficient than two-stage detectors. However, two-stage detectors have been found to provide greater accuracies than one-stage detectors.

However, using object detection on each image frame does not take into consideration the following attributes in video data: (1) Since there exist both spatial and temporal correlations between image frames, there are feature extraction redundancies between adjacent frames. Detecting features in each frame leads to computational inefficiency. (2) In a long video stream, some frames may have poor quality due to motion blur, video defocus, occlusion, and pose changes. Detecting objects from poor quality frames leads to low accuracies. Approaches for video object recognition make an effort to solve the aforementioned problems. Some methods, such feature fusion on several layers, employ the spatial-temporal information to increase accuracy. Other strategies concentrate on enhancing detection effectiveness and eliminating information redundancy. Moving forward, the YOLOv5, YOLO-6 and YOLOv7 will be talked about respectively.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

However, there is always a trade-off between speed and accuracy among these methods., YOLO versions were developed and in each version there was a speed accuracy trade off. Furthermore, methodology improvements and structure of YOLO will be discussed. Afterwards, we will compare the performance of all three models to analyze which one is the most accurate.

II. RELATED WORK

The YOLO algorithm uses convolutional neural networks (CNN) to quickly identify objects. The approach just needs one forward propagation through a neural network, as the name would imply, to detect objects.

A. Object Detection with Deep Learning

Deep learning has been widely employed in artificial intelligence object identification, which is the process of identifying and finding objects in digital photos or videos. Deep learning neural networks for object detection are trained on huge datasets of labeled photos, where the algorithms learn to recognize things by extracting features such as edges, corners, textures, and colors from the images. These traits are then utilized to forecast the existence and placement of items in previously unseen images.

Furthermore, detecting items that take up between 2% and 60% of an image's area is an area where object detection excels. It is also very efficient at detecting items with distinct borders Additionally, it detects groups of objects as a single item and performs object localization at high speed (>15fps).

Furthermore, object detection is becoming more prevalent in a range of industries, with applications ranging from company productivity to personal security. Convolutional Neural Networks (CNNs), for example, have shown remarkable success in achieving high accuracy and real-time performance in a wide range of applications, including autonomous cars, surveillance systems, and face recognition systems. Because they can automatically learn and adapt to different item classes, orientations, sizes, and lighting conditions, these algorithms are particularly successful at recognising objects in complicated and dynamic situations.

B. Introduction to yolo family

1) YOLOv1

RCNN models were the most popular models for object detection at the time. Although the RCNN family of models was accurate, it was very sluggish due to the multi-step process of locating the recommended region for the bounding box, classification on these regions, and lastly post-processing to enhance the result.

YOLO was established with the purpose of eliminating multistage detection and doing object detection in a single stage, hence improving inference time.

a) Performance

YOLOv1 sported a 63.4 mAP with an inference speed of 45 frames per second (22ms per image). At that time, it was a huge improvement of speed over the RCNN family for which inference rates ranged from 143ms to 20 seconds.

b) Technical Improvements

The YOLO model's primary operation is based on its unified detection approach, which combines several components of object identification into a single feed neural network.

The model breaks an input picture into several grids and assesses the likelihood that an object will be found within each grid. This is repeated for all of the grids into which the image is split. The programme then aggregates neighbouring high-value probability grids into a single item. Low-value predictions are eliminated using a method known as Non-Max Suppression (NMS)[23].

The model is trained in a similar fashion where the centre of each object detected is compared with the ground truth. In order to check whether the model is correct or not and adjust the weights accordingly.

2) YOLOv2 – Better, Faster, Stronger

YOLOv2 was capable of detecting over 9000 different item types. This version improved on the previous version YOLOV1.

a) Performance

On the VOC 2012 dataset, YOLOv2 achieved a performance of 78.6 mAP. The table below shows that it outperformed other item detection algorithms on the VOC 2012 dataset.

b) Technical Improvements

YOLOv2 version introduced the concept of anchor boxes. Anchor boxes are nothing but predefined areas for an image that illustrates the idealized position of the objects to be detected. We calculate the ratio of overlap over union (IoU) of the predicted bounding box and the pre-defined anchor box. The IoU value acts as a threshold to decide whether the probability of the detected object is sufficient to make a prediction or not.

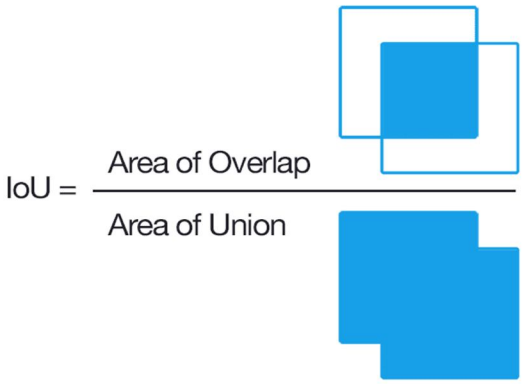
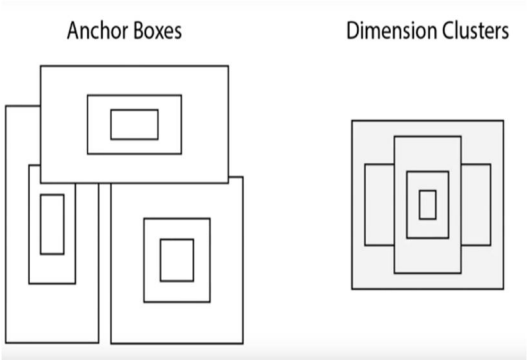


Fig 1..Graphical illustration of intersection over union (IoU) metric[23]

In the case of YOLO, anchor boxes are not computed randomly. Instead, the YOLO algorithm examines the training data and performs clustering on it (dimension clusters). All this is done in order to ensure that the anchor boxes that are used represent the data on which we will be training our model. This helps in enhancing the accuracy a lot.



.Fig 2. Anchor boxes converted to dimension clusters[23]

c) Additional Improvements

- In order to adapt to different aspect ratios, the YOLOv2 model is randomly resized throughout the training process (this is called multi-scale training).
- For the model to be robust the YOLOv2 model was trained on a combination of the COCO dataset (80 classes with bounding boxes) and the ImageNet dataset (22k classes without bounding boxes). When the model processes an image with labels the detection and classification error is calculated. Whereas when the model sees a label-less image it backpropagates the classification error only. This structure is called the Word Tree.
- Inference speeds of up to 200 FPS and mAP of 75.3 were achieved using a classification network architecture called darknet19 (the backbone of YOLO).

3) YOLOv3: An Incremental Improvement

This model was a little bigger than the earlier ones but more accurate and yet was fast enough.

a) Performance

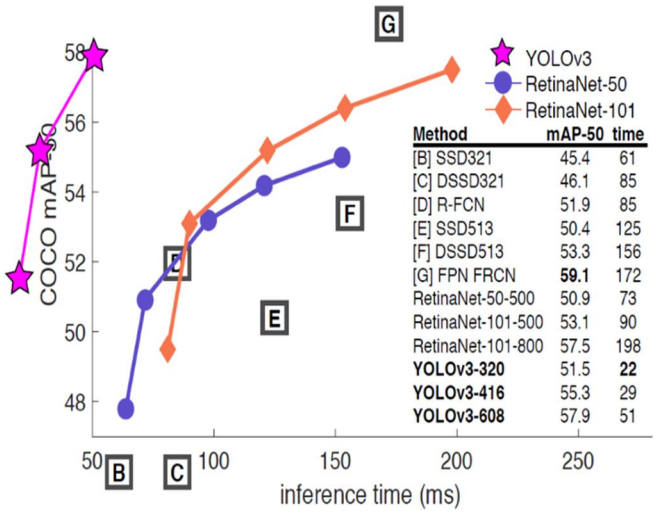
YOLOv3-320 has an mAP of 28.2 with an inference time of 22 milliseconds. (On the COCO dataset). This is 3 times fast than the SSD object detection technique yet with similar accuracy Comparisons

b) Technical Improvements

YOLOv3 consisted of 75 convolutional layers without using fully connected or pooling layers which greatly reduced the model size and weight. It provided the best of both worlds i.e. using residual models (from the ResNet model) for multiple feature learning with feature pyramid network(FPN) while maintaining minimal inference times.

A feature pyramid network is a feature extractor that extracts different types/forms/sizes of features for a single image. It concatenates all the features so that the model can learn local and general features.

By employing the use of logistic classifiers and activations the class predictions for the YOLOv3 goes above and beyond RetinaNet-50 and 101 in terms of accuracy.As the backbone, the YOLOv3 model uses the Darknet53 architecture.



Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

Fig 3. mAP versus Inference time[23]

4) YOLOv4 – Optimal Speed and Accuracy of Object Detection

a) Performance

The YOLOv4 model outperforms other detection methods such as efficientDet and ResNext50. It is powered by the Darknet53 backbone (the same as the YOLOv3).

b) Technical Advancements

The bag of freebies (techniques that improve model performance without raising inference cost) and the bag of specials (techniques that improve accuracy while increasing computation cost) were added in YOLOv4.

On the COCO dataset, it has a frame rate of 62 frames per second and a mAP of 43.5 percent.

c) Bag of Freebies (BOF)

- Data augmentation techniques: Cutmix (cut and mix numerous photos containing items to be detected), Mixup (random image mixing), Cutout, Mosaic data augmentation.
- Bounding box regression loss: Experiment with various bounding box regression types. MSE, IoU, CIoU, and DIoU are some examples.
- Regularisation: There are several regularisation strategies such as Dropout, DropPath, Spatial Dropout, and DropBlock.
- Normalisation: Added cross mini-batch normalisation, which has been shown to improve accuracy. In addition to approaches such as iteration-batch normalisation and GPU normalisation.

d) Bag of Specials BOS

- Spatial attention modules (SAM): Uses the inter-spacial feature connection to generate feature maps. Increase accuracy while increasing training time.
- Non-max suppression (NMS): When we group objects together, we receive several bounding boxes as predictions. Non-max suppression minimises the number of false/excess boxes.

- Non-linear activation functions: The YOLOv4 model was used to examine several types of activation functions. For instance, ReLU, SELU, Leaky, Swish, and Mish.
- Skip-Connections, such as weighted residual connections (WRC) and cross-stage partial connections (CSP).

5) YOLOv5: Latest YOLO?

YOLOv5 is rumoured to be the next member of the YOLO family to be launched in 2020 by Ultranytics, only a few days after YOLOv4. No paper has been produced, and there is some dispute in the community over whether the use of the YOLO trademark is justified given that it is only the PyTorch implementation of YOLOv3.

a) Performance

Because there is no official document yet, the legitimacy of the performance cannot be verified. It achieves the same, if not higher, accuracy (mAP of 55.6) than the other YOLO models while using less computer resources.

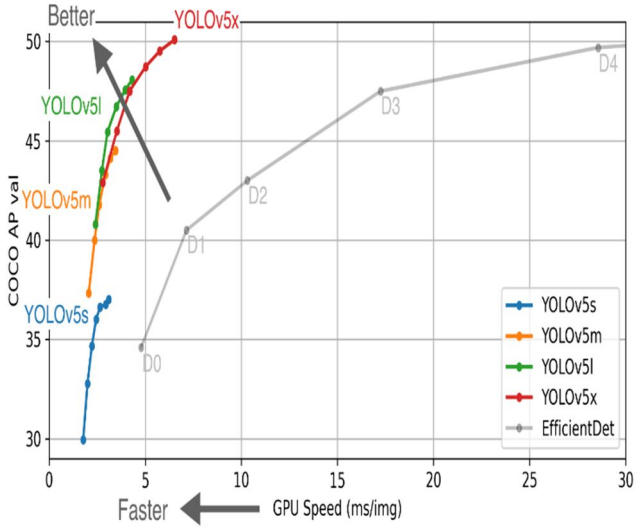


Fig 4.Accuracy comparison of different models[23]

b) Technical Improvements

- Improved data augmentation and loss calculations (now that the model's foundation has changed from C to PyTorch) • Auto learning of anchor boxes (they no longer need to be manually inserted)
- Backbone use of cross-stage partial connections (CSP).
- Path aggregation (PAN) network is used in the model's neck.
- Easier to train and test framework (PyTorch).
- Simple to use and install.
- The new version supports YAML files instead of CFG files, which considerably improves the form and readability of model configuration files.

III. METHOD USED

In this section, we will study about architecture of yolov5 and find out why yolov5 is capable for object detection. The inference time and accuracy of the model was carefully observed and then was compared with other models.

A. Improvements in Yolov5

Any computer vision enthusiast has surely heard of YOLO models for object detection. Ever since the first YOLOv1 was introduced in 2015, it garnered too much popularity within the computer vision community. Subsequently, multiple versions of YOLOv2, YOLOv3, YOLOv4, and YOLOv5 have been released albeit by different people. In this article, we will give a brief background about all the object detection models of the YOLO family from YOLOv1 to YOLOv5.

B. Basic Working of YOLO Object Detector Models

1) Precision and recall are critical for deducing and judging the correctness and robustness of any ML-based model. As a result, the author of YOLO continued to try to develop an object identification model that maximises mAP (mean average precision). The ratio of genuine positives to total positive predictions (right or inaccurate) is referred to as recall.

2) Precision is defined as the proportion of genuine positives to ground truth positives (total right predictions).

3) Mean average precisions (mAP) are the sum of all average precisions.

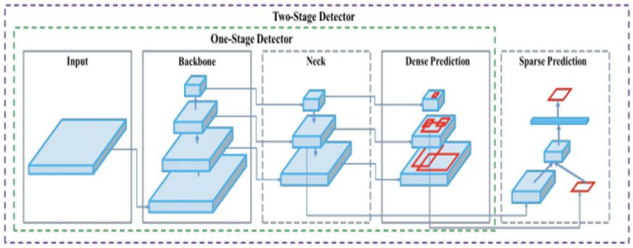


Fig 5. One-stage detector (input, backbone, neck, and dense prediction), two-stage detector (input plus sparse prediction)[24]

Aside from that, the architecture of all YOLO models follows a similar pattern of components, as seen below -

a) *Backbone:* A convolutional neural network that collects and generates visual features of various shapes and sizes. As feature extractors, classification models such as ResNet, VGG, and EfficientNet are employed.

b) *Neck:* A group of layers that combine and mix properties before passing them on to the prediction layer. Examples include the feature pyramid network (FPN), the path aggregation network (PAN), and the Bi-FPN.

c) *Head:* Includes neck characteristics as well as bounding box forecasts. To finish the detection process, performs classification and regression on the features and bounding box coordinates. Outputs four values, usually x and y coordinates along with width and height.

C. Data Augmentation in YOLOv5

YOLOv5 runs training data through a data loader, which augments data online, with each training batch. The data loader performs three types of augmentations:

1) Scaling.

2) Colour space changes.

3) Mosaic enhancement.

The most unique is mosaic data augmentation, which mixes four photos into four random ratio tiles.

Mosaic augmentation is particularly effective for the widely used COCO object identification benchmark, assisting the model in learning to overcome the well-known "small object problem," in which little things are not spotted as reliably as bigger ones.

It is worth mentioning that experimenting with your own set of augmentations to maximise performance on your particular work is worthwhile.

Here's a screenshot of enhanced training pictures in YOLOv5.

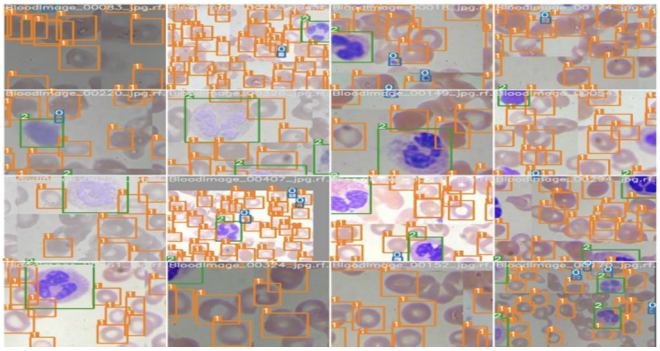


Fig 6.A grid of 16 blood cell images to which image augmentations have been applied[23]

D. Bounding Box Anchors That Learn On Their Own

The YOLOv5 network predicts bounding boxes as deviations from a set of anchor box dimensions in order to produce box predictions.

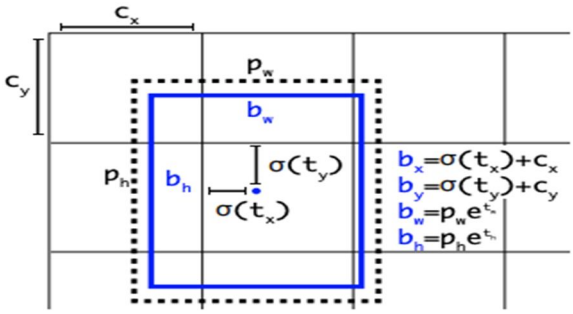


Fig 7.Derivation of bounding box[23]

The most severe disparity in anchor boxes may arise while attempting to identify giraffes that are very tall and narrow, or manta rays that are very wide and flat. When you enter custom data into YOLOv5, all YOLO anchor boxes are auto-learned.

IV. RESULTS AND COMPARISION

When applied to a sample image, YOLO algorithms. Please check the Supplementary Material at the conclusion of Section 4 for additional photographs and a video. Figure 5 also depicts the performance of YOLO algorithms in both PC and CC. Table A1 also has detailed findings, which demonstrate the average precision of the three YOLO algorithms for all labels. Furthermore, Table 3 displays the accuracy and recall of the algorithms; YOLOv3 has a high precision but a poor recall, indicating that the model has to be improved. For an algorithm to be considered efficient in our work, it must strike a balance between precision and recall, which is represented in the method's F1 score. Precision and recall are balanced in YOLOv4 and YOLOv5l, as shown. As a result, YOLOv4 and YOLOv5l have greater F1 scores than YOLOv3, while having higher accuracy. The models in YOLOv4 and YOLv5 have balanced precision and recall, resulting in a good F1 score.

Fig 7. Object detection with YOLOv3[23]

Fig 8. Object detection with YOLOv4[23]

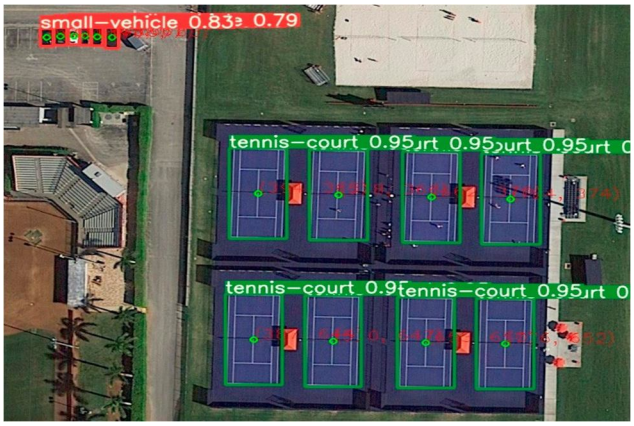


Fig 9. Object detection with YOLOv5[23]

V. FUTURE WORK

In this study, we conducted a broad literature review on object identification algorithms, their many versions, and their diverse needs. Based on the results of the survey, we discovered the following concerns with real-time object identification and tracking: 1. The majority of existing algorithms are image/video-based. It loses some information while extracting features from images/videos. As a result, detection and tracking are difficult. 2. Some detection and tracking algorithms can identify and track several objects while also dealing with occlusion. However, greater computational and memory requirements are required.

Based on the comparison of yolov model, the yolov5 version is best suitable for proposes a deep learning-based system for detecting suspicious events in a bank-ATM context in real time. Bounding boxes, which functioned as classes in this case, are utilised to detect tagged items. This is then used to categorise labels in video and forecast whether the occurrences are normal or abnormal. that result is calculated using the Motion representation Depth data is derived from the classes' bounding boxes. Then multi-stream CNNs are used to distinguish constituents and actions. The choosing of an appropriate algorithm for a certain job.

VI. CONCLUSION

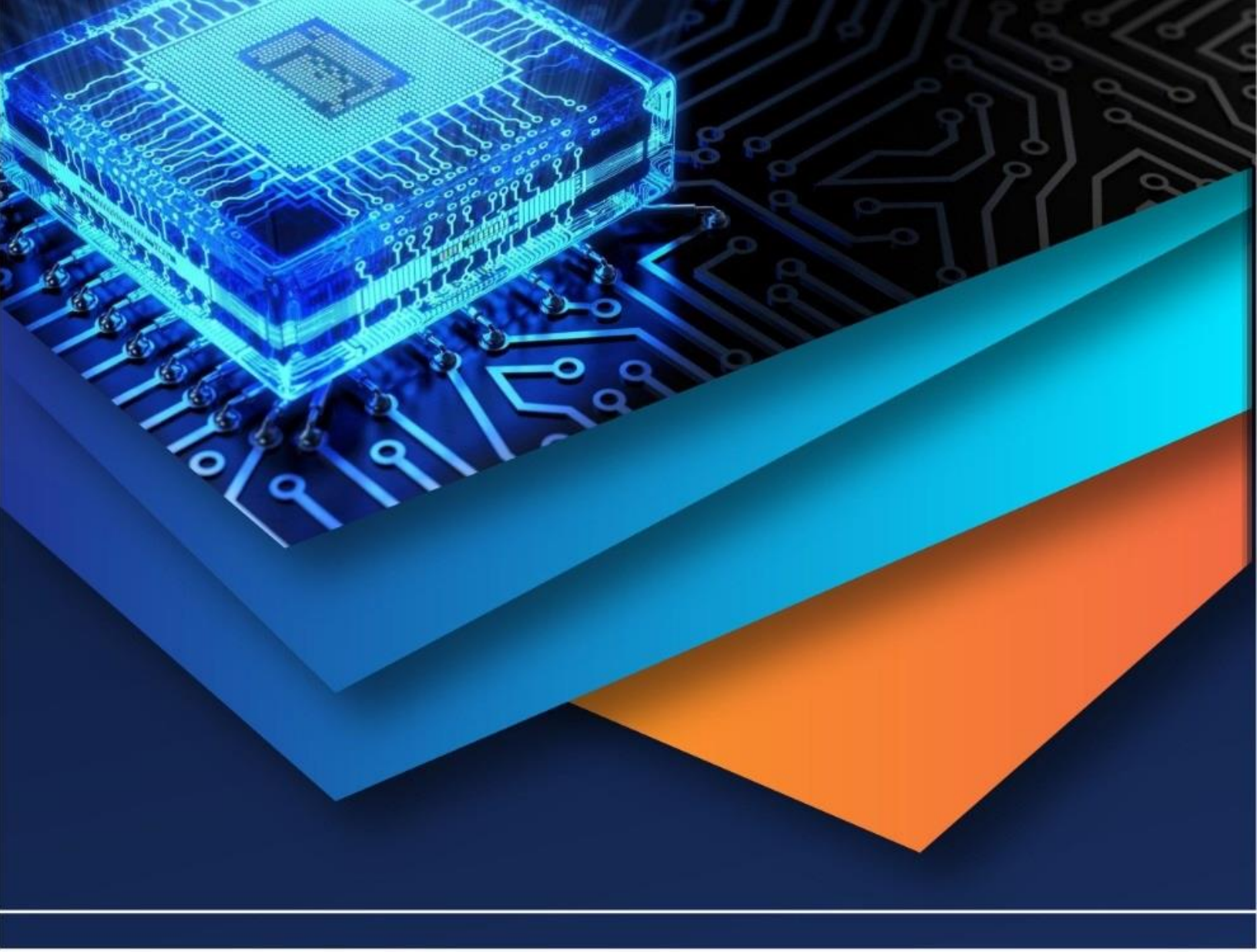
While working on an object identification module that would determine favorable and poor landing places in real-time. Based on prior relevant work, we were unable to determine the object identification technique that performs best in this application while meeting the desired safety requirements. As a result, we chose YOLOv3, YOLOv4, and YOLOv5 because of their high detection speed and accuracy in real-time applications, and we compared their accuracy and speed to see which method works best for emergency landing place identification.

Based on the findings of our research, as shown in Fig 7,8,9 we select the algorithm with the best accuracy, YOLOv5. In the future work, an establish the viability of combining YOLOv5 with the ATM -I dataset for fast and accurate object recognition with multiple object and actions.YOLOv5 models are best approach for object detection and tracking.

REFERENCES

[1] Vikas Tripathi; Hindawi Publishing Corporation, "Robust Abnormal Event Recognition via Motion and Shape," Journal of Electrical and Computer Engineering, pp. 1-11, 2015.
[2] Pushpajit A. Khaire and Praveen Kumar, "RGB+D and deep learning based real time detection of suspicious," Springer; Journal of Real-Time Image Processing, pp. 1-13, 2021.
[3] P. A. Khaire, "RGB+D and deep learning based real time detection of suspicious," Journal of Real-Time Image Processing, pp. 1-13, 21.
[4] C. Shiranthika, "Human Activity Recognition Using CNN & LSTM," IEEE, 2021.
[5] T. S. Bora, "HUMAN SUSPICIOUS ACTIVITY DETECTION SYSTEM USING CNN MODEL FOR VIDEO SURVEILLANCE," IJARIIE, 2021.
[6] R. Vrskova, "A New Approach for Abnormal Human Activities Recognition," Sensor, 2022.
[7] S. Sabbu, "LSTM-Based Neural Network to Recognize Human Activities," Hindawi, pp. 1-8, 2022.
[8] Rajeshwari S, Vismitha G, Sumalatha G and Safura Aliya, "Unusual Event Detection for Enhancing ATM Security," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, pp. 1-6, 2021.
[9] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," SIGKDD Explor. Newsl., vol. 12, no. 2, pp. 74–82, Mar. 2011, doi: 10.1145/1964897.1964918.
[10] A. Murad and J.-Y. Pyun, "Deep Recurrent Neural Networks for Human Activity Recognition," Sensors, vol. 17, no. 11, p. 2556, Nov. 2017, doi: 10.3390/s17112556

[11] P. Kuppusamy and C. Harika, "Human Action Recognition using CNN and LSTM-RNN with Attention Model" International Journal od Innovative Technology and Exploring Engineering(IJITEE), vol.8,Issue 8, pp.1639-1643, 201
[12] https://www.analyticsvidhya.com/blog/2022/03/basics-of-cnn-in-deep-learning
[13] Y. Chen, K. Zhong, J. Zhang, Q. Sun, and X. Zhao, "LSTM Networks for Mobile Human Activity Recognition," presented at the 2016 International Conference on Artificial Intelligence: Technologies and Applications, Bangkok, Thailand, 2016, doi: 10.2991/icaita- 16.2016.13
[14] https://ieeexplore.ieee.org/document/904397
[15] https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939
[16] C. Jobanputra, J. Bavishi, and N. Doshi, "Human Activity Recognition: A Survey," Procedia Computer Science, vol. 155, pp. 698–703, 2019, doi: 10.1016/j.procs.2019.08.100
[17] https://deepai.org/publication/evaluating-two-stream-cnn-for-video-classificatio
[18] https://www.codeproject.com/Articles/1366433/Using-Modified-Inception-V3-CNN-for-Video-Processin
[19] https://www.kaggle.com/datasets/mehantkammakomati/atm-anomaly-video-dataset-atma
[20] A. Murad and J.-Y. Pyun, "Deep Recurrent Neural Networks for Human Activity Recognition," Sensors, vol. 17, no. 11, p. 2556, Nov. 2017, doi: 10.3390/s17112556
[21] T. Zebin, M. Sperrin, N. Peek, and A. J. Casson, "Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, Jul. 2018, pp. 1–4, doi: 10.1109/EMBC.2018.8513115.
[22] https://github.com/pjreddie/darknet/blob/master/data/coco.names
[23] https://machinelearningknowledge.ai/a-brief-history-of-yolo-object-detection-models
[24] https://www.irjet.net/archives/V8/i4/IRJET-V8I4809.pdf
[25] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowdedness," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2015.
[26] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab ," in Proceedings of the IEEEinternational conference on computer vision, 2013.
[27] Lu, S. (2019). Deep learning for object detection in video Journal of Physics Conference Series, 1176.
[28] Simonyan, K., Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos.
[29]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY