Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

A Modified Yolov3 Model (Yolov3sd) for Detection of Drinking and Smoking Actions

Sunaina[1], Dr. Dharamveer Sharma[2]
Punjabi University, Patiala, Punjab. 147002.
Email: [1]sunaina175@gmail.com, [2]dveer72@gmail.com

Abstract. *For many applications of intelligent systems, the ability to detect and recognize objects of interest from images is crucial. Precision and processing efficiency are two key considerations for such an object detection task, particularly for applications which need to operate in real-time. In this paper, a study is carried out to detect smoking and drinking actions from the images, which is very challenging due to various complexities exist in the background of image and must be detected in the real time so that disclaimer can be generated accordingly. To this end, first a dataset is generated by collecting images of smoking and drinking actions from the videos and movies. Then a variant of YOLOv3, a well-known object detector, is proposed named YOLOv3SD, which is the combination of DenseNet and YOLOv3. The DenseNet is combined to retain and reuse the features so that the model can detect actions in complex background with more precision. The experiment results shows that the proposed network has higher detection accuracy compared to the previous state-of-the-art networks such as Faster R-CNN, SSD, YOLOv2 and YOLOv3.*
Keywords*: Deep Learning, YOLOv3, DenseNet, object recognition*

1. Introduction

In recent time, object and action recognition has become a important field in computer vision. The aim of action recognition is to detect and classify actions occurring in images and videos and it is leading to many applications and specific fields such as vehicle and animal recognition on roads. Vision is much more than just seeing the picture, it is the knack to understand the context of the image, how many objects and subjects are present in the image and how they are related. To pace with the evolution of technology, it is required to incorporate this vision capability to computer. Due to the technological advancements in the field of mobile phones and cameras, the image and video data is continuously increasing. Now 48.33% of world population owns a smartphone [1], which is further increasing with the passage of time, and it is directly boosting the amount of image and video data created each day. YouTube which is the second most popular website after Google, witnessed the 500 hours of video upload every minute and 5 billion videos watched by people daily[2]. To tackle this huge amount of data, there is need to capture the content of this data automatically without the human intervention and understand what that data is representing. Earlier many traditional methods were used to identify the action in images and videos like histogram of gradient (HoG)[3], histogram of optical flow (HOF) [4] and scale invariant feature transform (SIFT) [5], but it was time and cost inept to apply these methods on large amount of multimedia data. The rise of neural networks and deep learning made it possible to reconnoiter the image and video data in less time and less cost comparatively. Deep learning proved to be so compliant from detecting objects to recognizing actions and everything that comes under computer vision that it unfastens the door of making the performance of computer vision equal to the human vision system.

Among different actions, there is specific significance of detecting drinking and smoking actions. The interest in these target action arises after the study and survey of different organizations who claimed that there is a direct impact of showing such scenes in the videos on the consumption behavior of viewers towards alcoholic products and drugs. Hence, to save the viewer and lessen this impact there is need to use the technical systems having the capability to detect such activities and show the warning message accordingly.

In deep learning, convolution neural network (CNN) is the most popular and robust network to apply in object detection and action recognition tasks. CNN is a combination of layers, which are responsible to extricate the important attributes from the dataset and based on that classification, detection and segmentation is performed. There are many algorithms for computer vision tasks which are using CNN as the base architecture, for example regions with CNN features (RCNN) [6], Fast RCNN [7], Faster RCNN [8],

Mask RCNN [9], you look only once (YOLO) [10], YOLOv2 [11], YOLOv3[12] etc. In this paper, modification in YOLOv3 algorithm named YOLOv3SD has proposed which is implemented on proposed dataset named 'interaction with cigarettes and liquor' (ICL), which consists of images of smoking and drinking actions, collected from videos and movies, to detect target actions i.e. smoking and drinking and comparison is shown of proposed model with other state of the art models on the ICL dataset.

The unexpended part of the paper is as follows. Next section includes the discussion on related work. The dataset section describes the dataset used in the proposed work and the augmentation techniques applied on the dataset. A brief introduction of the proposed techniques is given after the dataset section. Experiments Results and conclusion have been given after that.

2. Related Work

Since past few years, many vision based deep learning methods have been proposed for object detection and action recognition in images. With the success of CNN in image classification [13], many networks are designed utilizing CNN as the base network. In 2014 Ross Girshilk et al. [6] combined region proposals with CNN which solve the CNN localization problem by using recognition within the regions. Further fast RCNN [7] is proposed to amplify speed of training and testing the model and detection accuracy is also improved. Faster RCNN [8] is the next variant of this family where region proposal network is introduced for generating cost free regions. J. Redmon et al. [10] transformed the problem of object detection from classification domain to regression domain and proposed an end-to-end network YOLO to localize the object in the image. Its next variant YOLO v2 [11] was able to detect the objects of different sizes as it was using multi scale training method. YOLO v3 [12] was the another variant which used darknet 53 network as feature extractor and is efficient in detecting small objects.

Many researchers of different fields have applied these models in distinct application like shuttlecock detection for badminton playing robot [14], densely connected electronic components etc [15]. X Sun et al. [16] applied the Faster RCNN for face recognition and also used multiple strategies like hard negative mining, feature concatenation, multi-scale training, model pre-training. A.K. Nsaif et al. [17] used the Faster RCNN for eye detection and combined it with Gabor filters and Naïve bayes methods to detect eyes under the condition of reflection from glasses. S. Albahli et al. [18] used Faster RCNN for numeric handwritten digits recognition. S. Yin et al. [19] used the faster RCNN to detect airports in the large scale remote sensing images and applied multi scale training to increase the robustness of the network. L. Wen et al. [20] used dual faster region based CNN for moving target detection from SAR images using one network for shadow detection and combining it with doppler energy detection in the doppler spectrum domain. Y. Xiao et al. [21] improved the object detection performance of faster RCNN by combining it with skip pooling and fusing contextual information. N. Zhang et al. [22] replaced feature extraction module of faster RCNN by Dense Net and used it in activity object detection.

S. Shinde et al. [23] used YOLO algorithm to perform detection, localization and recognition of actions in real time frames. Furhter, as most of the devices are non gpu devices, R Huang et al. [24] proposed YOLO-Lite to detect objects for non gpu computers by using shallow networks. W. Lan et al. [25] proposed the model R-YOLO by adding three passthrough layers i.e. route layer and Reorg layer to the YOLO network to pass shallow features and used for pedestrian detection. R Huang et al. [15] replaced the darknet53 architecture of YOLO v3 model with mobilenet and split the convolutional layer in depth-wise and point wise convolutions and used the model to recognize electronics components on integrated circuits. T. Ahmed et al. [26] replaced some layers of YOLO Model with Inception model and added a spatial Pyramid Pooling layer and used the modified YOLO model for object detection. Y Yin et al. [27] added random kernel convolution for extreme learning and proposed Faster YOLO for object detection. Z. Cao et al. [14] formulated the new loss function for tiny YOLOv2 and used the proposed M-YOLOv2 for shuttlecock detection for badminton robot. X. Wang et al. [28] proposed novel R-YOLO by expanding the detection branches of YOLOv4 so that small size text can be detected.

3. Dataset

The target actions of the presented work are novel to explore with deep learning models, therefore it

required collection of new dataset consisting of these actions.

The dataset is collected in the form of images and videos clips from movies of Bollywood, Hollywood and other cinemas. It provides more challenges than the datasets created for some specific actions in controlled environment with static background. The size of the images also varies. It ranges from 360x360 to 1024x1260. The dataset contains total 1200 images out of which 600 images contain drinking action and 600 images of smoking action.

After collecting the images and creating a dataset, the images are annotated by generating the bounding box around the target objects in all the images. The coordinates of the bounding boxes are calculated with the help of BBox-Label-Tool-master tool.

Sometimes the actions detection process suffers from various challenges such as occlusion, illumination changes, presence of multiple people and objects etc. These challenges become more prominent when the dataset is collected from the movies. To address these challenges, dataset includes the images where the objects are partially occluded, more than one instance of same object is present in etc. Fig 1 represents some of the sample images from dataset:

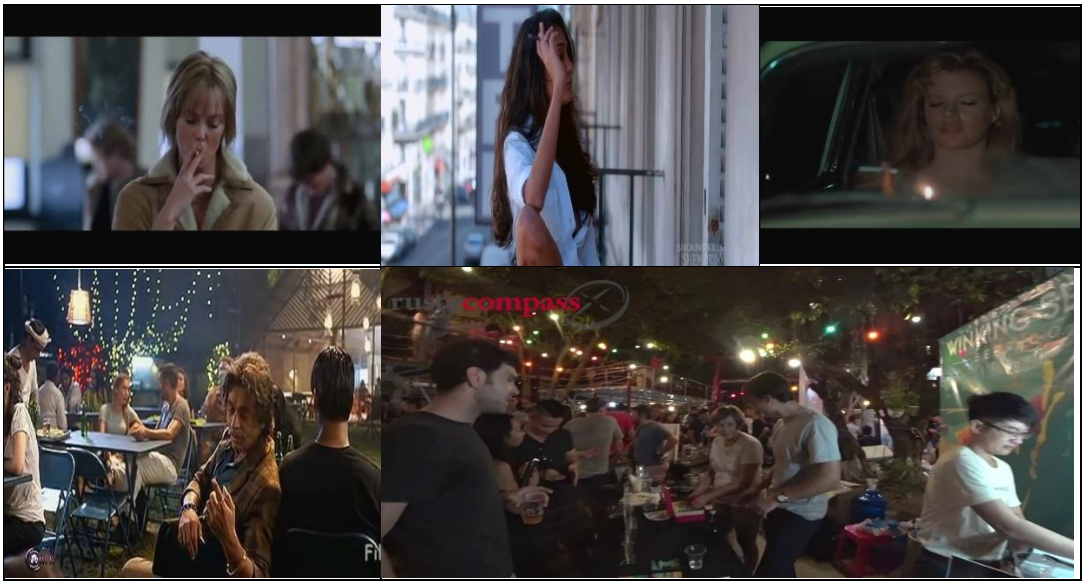

Fig 1: Sample images from the dataset

The above figure consists of some of the sample images from the dataset where image 1 and 2 (top to right) are suffering from occlusion, then next image shows the effect of illumination and image 4 and 5 show the presence of multiple objects and people in the scene while performing the actions.

3.1 Data Augmentation

When training the deep neural network, it requires the huge amount of data with diversity of samples. If the data is insufficient, then it may result in overfitting or underfitting which may cause reduction in the performance of the network on the unseen test data. Therefore, to increase the generalizability of the network, it is required to fed the model with expanded dataset, which have diversity in the samples. For this, various augmentation techniques are applied to smoking and drinking dataset to make the model more robust.

3.1.1 Brightness Augmentation:

Under varying lighting and imaging conditions, the human visual system can discern an object's surface's colour invariance; however, imaging equipment lack this colour invariance. Different lighting situations will cause an image's colour to deviate somewhat from reality. Three different brightness levels are applied on the images which are 0.5, 1.2, 1.7, where 0.5 increases the darkness of the image and brightness is increased by two levels i.e. 1.2 and 1.7. It helps to train the model to detect object in the presence of ambient light.(Fig 2(b)-2(d))

3.1.2 Augmentation by Flipping

The target object may appear in the image having different orientation. To make the model orientation invariant flipping is performed on the

dataset. By flipping the image, the orientation of the target object gets changed which makes the dataset orientation in location information. Flipping is again an in-place augmentation. Vertical flip flips the picture on the x axis, whereas horizontal flip flips it on the y axis.. Both Horizontal and vertical flipping is used to augment the data. A successive flipping is also applied to simulate the rotation of 180 degree. (Fig. 2(e)-2(g))

3.1.3 Blur Processing

While playing the video sometimes the picture quality deteriorates due to many possible reasons. To deal with this challenge, images of dataset are blurred and augmented with the base dataset. (Fig. 2(h))

Fig 2: Sample images after augmentation

Table 1: Image dataset generated by image augmentation method

	Base	**Brightness**	**Flipping**	**Blur**	**Total**
Smoking	600	1800	1800	600	**4800**
Drinking	600	1800	1800	600	**4800**

The table 1 summarizes the total base images and the images generated by augmentation techniques.

4. Methodologies

In our work a new model is proposed which is combination of YOLOv3 model and DenseNet Model. The brief of these models is discussed in this section. The suggested approach is then described using a diagram and the proposed model's layered structure.

4.1 YOLO v3

The YOLO (you only look once) network is an end-to-end object detection model that perform detection in single stage. It considers the object detection task as regression problem while in other algorithms like FasterRCNN consider this task as classification regression problem. The YOLO network divides each image in the training set into S * S grids. If the center point of an object's ground truth falls within the grid, it is the grid's responsibility to locate the target. The output of each grid contains the bounding box information having (x, y) as center point, (h, w) as height and width of bounding box and box confidence. Along with this information it also predicts the conditional class probability, which signifies the probability that the detected object is part of a certain class. When determining whether the grid contains items, confidence considers both the presence of objects and the precision of the predicted bounding box. When many bounding boxes identify the same target, non-maximum suppression (NMS) is used to select the optimal bounding box.

YOLO model is quick compared to Faster RCNN, however the error in detection is also higher. To address this issue, YOLOv2 introduced the "anchor" concept of Faster RCNN. Also, YOLOv2 improved the network architecture by using the convolution layer as output layer, rather than the fully connected layer. The features like direct location prediction, high resolution classifier, multi scale training and batch normalization which are implemented in YOLOv2 greatly increased the accuracy, when compared to YOLO model. But to detect multi scal e objects, YOLOv2 was still not the best solution.

To address the aforementioned issue, the next version of YOLO i.e. YOLOv3 is proposed in which object detection is performed using ResNet model and the feature pyramid networks (FPN). A residual model known as Darknet53 that had 53 convolutional layers served as the feature extractor for YOLOv3. It can be built more deeply from the viewpoint of the network structure, increasing the detection's precision. Another idea was to perform multi-scale prediction by leveraging FPN architecture, which improved the effectiveness of YOLOv3 over YOLOv2 in detecting small targets.

4.2 DenseNet

DenseNet is proposed by Huang et al. [29] to overcome the problem of information loss due to down sampling of the feature maps. In general, convolutional neural network, each layer is connected only with two layers that is immediate preceding and succeeding layers but each layer in DenseNet is connected to every other layer so that its output, or the feature map of all the preceding layers, is connected as its input, and its output serves as the input for all the succeeding levels.

$$Xl = Hl([x0,x1,x2\ldots\ldots\ldots xl-1]) \quad (1)$$

where XI is the output of lth layer and HL is the function performed by that layer.

The fundamental dense net model consists of four dense blocks, transition layers using convolution and pooling, and a classification layer at the last.

4.3 The proposed algorithm

In the presented algorithm, some of the components of the DarkNet 53 are replaced with the dense architecture. In fig 3, the original DarkNet53 framework which is used in YOLOv3 for feature extraction, and proposed framework containing some dense connection in few of the layers, responsible for feature reuse and fusion has shown. The suggested model contains two dense blocks. Each dense block has 16 convolutional layers, each of which takes input from all of the block's earlier layers. It encourages feature reuse and lessens feature loss. Fig 3 demonstrates the architecture of the proposed algorithm.

The model predicts the bounding box at three different scales like in YOLOv3. The first dense block is placed after 13th convolution layer having input of size 32 * 32. The second dense block is place after 29th layer having input size of 16 * 16. The careful placement of dense blocks is done to extract maximum features of target object. Fig 4 shows the detection framework of the YOLOv3SD model.

The specific network parameters used to train proposed YOLOv3SD model are shown in table 2:

Table 2: Network parameters of YOLOv3SD model

Input size	Batch Size	momentum	Learning rate	Training epochs
512 x 512	8	.9	.0003	200

For improved high resolution image processing the input size is taken as 512 x 512, which is then down sampled to 256 after second convolutional layer. The starting learning rate is 0.0003. The model is run for 200 epochs with a batch size of 8.

5. Experiments and discussion

In this study, extensive experiments with the trained YOLOv3SD model were conducted to demonstrate how the suggested model enhances the detection performance when compared to existing cutting-edge detectors. The YOLOv3SD model is built using Keras library. It is trained and tested on HP Z4 G4 workstation. The parameters of the workstation are: Intel Xeon (R) W-2195 (CPU), 64 GB RAM, Nvidia Quadro P4000 (GPU) and Windows 10 Pro for workstations (Operating System).

The following is a description of the some of the assessment criteria used to assess the model's performance.

The network Structure of YOLO V3

	Layer	Filter	Size	Output
	Conv	32	3x3	256x256
	Conv	64	3x3/2	128x128
1x	Conv	32	1x1	
	Conv	64	3x3	
	Residual			128x128
	Conv	128	3x3/2	64x64
2x	Conv	64	1x1	
	Conv	128	3x3	
	Residual			64x64
	Conv	256	3x3/2	32x32
8x	Conv	128	1x1	
	Conv	256	3x3	
	Residual			32x32
	Conv	512	3x3/2	16x16
8x	Conv	256	1x1	
	Conv	512	3x3	
	Residual			16x16
	Conv	1024	3x3/2	8x8
4x	Conv	512	1x1	
	Conv	1024	3x3	
	Residual			8x8
	AvgPool		Global	
	Connected		1000	
	Softmax			

convs convs convs

YOLO v3 detection

The Network Structure of YOLOv3SD

	Layer	Filter	Size	Output
	Conv	32	3x3	512x512
	Conv	64	3x3	512x512
	Conv	32	3x3/2	256x256
	Conv	64	3x3/2	128x128
1x	Conv	32	1x1	
	Conv	64	3x3	
	Residual			128x128
	Conv	128	3x3/2	64x64
2x	Conv	64	1x1	
	Conv	128	3x3	
	Residual			64x64
	Conv	256	3x3/2	32x32
8x	Conv	128	1x1	
	Conv	256	3x3	
	DenseNet			32x32
	Residual			32x32
	Conv	512	3x3/2	16x16
8x	Conv	256	1x1	
	Conv	512	3x3	
	DenseNet			16x16
	Residual			16x16
	Conv	1024	3x3/2	8x8
4x	Conv	512	1x1	
	Conv	1024	3x3	
	Residual			8x8
	AvgPool		Global	
	Connected		1000	
	Softmax			

convs convs convs

YOLOv3SD Detection

Fig 3: Layered structure of YOLOv3 and YOLOv3SD model

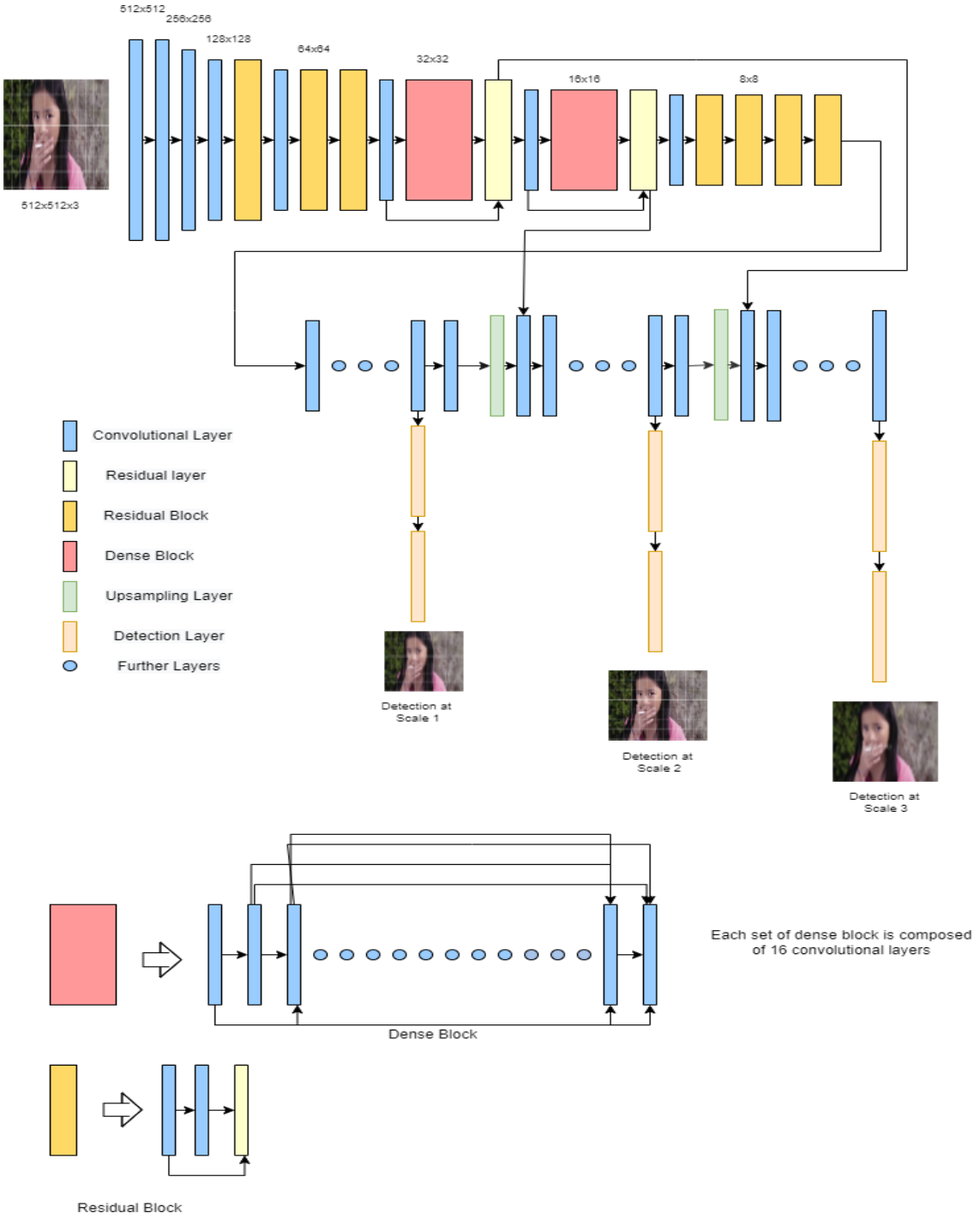

Fig 4: YOLOv3SD detection framework

Average Precision (AP):

Average precision of each class is calculated by taking different threshold value while calculating IoU. The threshold values are [0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.75,0.8,0.85,0.9]. The result of detection depends heavily on the threshold value. Taking the average of precision on different threshold values helps to make the results independent of threshold biasness.

AP= $\sum_{k=1}^{n-1}(Recall(k) - Recall(k+1)) * Precision(k)$ -----(2)

Where n is the total no. of threshold and Recall(k) is the recall value at k threshold value and Precision(k) is the precision value at k threshold value.

mAP

AP gives the results for each class. The mean Average Precision is calculated in order to assess and contrast the model's performance over the whole dataset, regardless of class.

mAP = $(\sum_{i=1}^{n} APi)/n$ ------(3)

5.1 Experiments:

The base dataset is divided into training and testing part in the ratio 5:1. As a result, after augmentation, the test data has 1600 images, whereas the training data contains 8000 images. The mAP presented in the results is calculated on the test dataset. Following are some experiments, where the suggested model is compared to other state-of-the-art models for both classes of the dataset. The experiments are done by varying the quantity of dataset and augmentation techniques also.

5.1.1 Influence of data category

The dataset consisting of both drinking and smoking actions was used to train the YOLOv3SD model. The average precision of the corresponding actions is shown in table 3.

Table 3: mAP comparison of different algorithms

Model	AP(Smoke)	AP(Drink)	mAP
FRCNN [8]	76.30	80.1	78.20
SSD [30]	77.55	76.45	77.0
YOLOv2 [11]	77.88	80.56	79.22
YOLOv3 [12]	81.44	79.74	80.59
YOLOv3SD (Proposed)	92.15	90.04	91.10

The table 3 shows that the overall results are ideal with the mAP value 85.34, but the detection of smoking action is better than the drinking action.

5.1.2 Comparison of different algorithms

In order to validate the model YOLOv3SD, presented in this paper, some other models like Faster RCNN, Single Shot Detector, YOLOv2 and YOLOv3 are trained on the same dataset. The mAP of all these models for the smoking and drinking dataset is shown in the table 3.

From these results it can be seen that the proposed model has significant improvement over the Faster RCNN, SSD and YOLOv2. The detection of Drinking action of YOLOv3 is very close to YOLOv3SD but for smoking action detection YOLOv3SD surpasses it.

5.1.3 Experiments by varying the quantity of dataset

This section examines how the size of the picture collection affects the performance of the YOLOv3SD model. To examine this, 2000, 4000, 6000, 8000 images are selected randomly from the image dataset to form corresponding dataset. Then the AP and mAP value of the model for respective datasets is obtained (Table 4).

Table 4: mAP results on different dataset sizes

No. of images	AP(Smoke)	AP(Drink)	mAP
2000	37.63	44.39	41.01
4000	65.71	51.68	58.70
6000	79.15	85.43	82.29
8000	92.15	90.04	91.10

These findings indicate that increasing the dataset size improves the YOLOv3SD model's ability to recognize objects.

5.1.4 Experiments by using different data augmentation methods

In this paper three augmentation technologies are used. To investigate how various augmentation techniques affect the model's performance, several experiments are performed. The whole dataset, the raw dataset, and individual augmentation technique data points are all used in the experiments. The corresponding mAP is shown in the table 5.

Table 5: mAP result on dataset with and without Augmentation

Dataset	mAP
Base Dataset	72.08
Full Augmented Dataset	91.10
Removing flipping from dataset	82.98

Removing Brightness processing from dataset	86.33
Removing Blur Processing from dataset	84.71

The results of the experiments shows that the model's performance was considerably enhanced through augmentation technology and the mAP improves from 72.08% to 85.34%. From these results, it can be inferred that the augmentation technologies could efficiently enhance the robustness and detection accuracy of the model.

5.2 Results

After training the model the weights are applied on the new images. Following are some sample results of the trained model. Along with each image the confidence score of the class is mentioned.

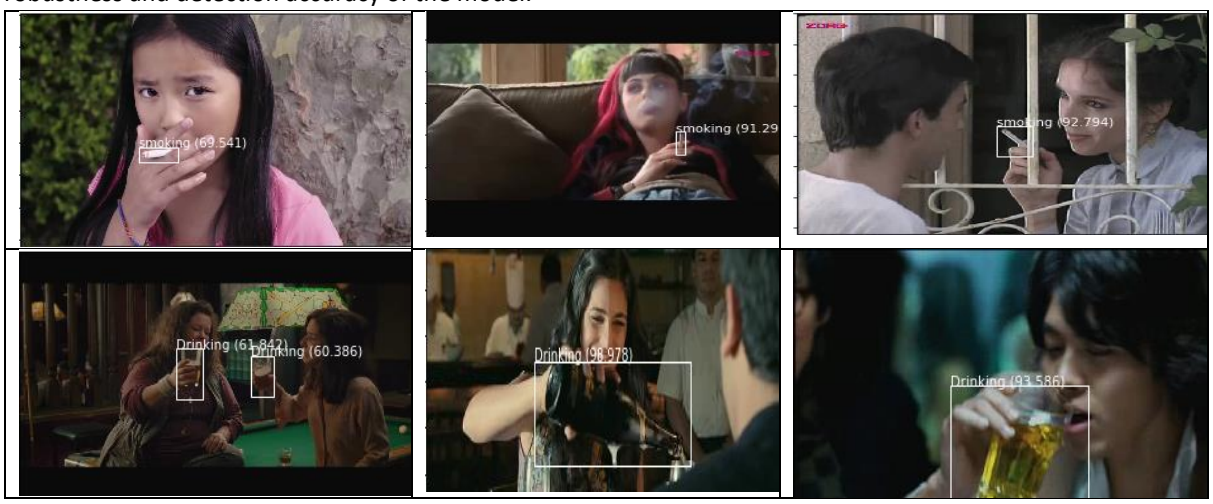

Fig 5: Sample results, First row shows the smoking detection and second row demonstrate the drinking action recognition.

From the images shown in fig 5, it can be concluded that the model is able to detect multiple objects in the same frame. Also, the bounding box prediction is very precise to the actual objects.

6. Conclusion

In this paper the improved YOLOv3SD model has been proposed to detect the drinking and smoking action. For dataset 1200 images are collected of smoking and drinking actions (600 each class), which are then augmented by flipping, blurring and brightness variations and the size of the dataset is increased by 8 times. The smoking and drinking actions in the images are labelled manually.

For better detection of the actions the dense blocks are added to the feature extraction part of the network. To demonstrate the efficacy of the presented model, its performance is compared with some of the other detection model. The comparison shows there is a substantial improvement with the suggested model in detecting target actions over other models like Faster RCNN, SSD etc.

YOLOv3SD model is well trained on detecting smoking and drinking actions, but real-time detection is still somewhat behind in terms of performance. Future work will focus on improving current models to recognize these activities in video and fulfil real-time needs.. To further increase the detection accuracy, the data augmentation methods will be refined also.

Statements and Declarations

a) Data Availability Statement

The dataset is created for the research purpose, which can be made available on receiving the reasonable request.

b) Compliance with Ethical Standards

All the work in the submitted manuscript follow the research ethics and there is no potential conflict of interest with any party if the manuscript is published.

c) Competing Interests

The writers did not receive funding from any organization for the work they submitted, and they have no material conflicts of interest, either financial or non-financial.

References

[1] Marcoux, T., Agarwal, N., Erol, R., Obadimu, A., & Hussain, M. N. (2021). Analyzing cyber influence campaigns on YouTube using YouTubeTracker. In Big Data and Social Media Analytics (pp. 101-111). Springer, Cham.

[2] https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/

[3] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). Ieee.

[4] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008, June). Learning realistic human actions from movies. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[5] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[6] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[7] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[8] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[9] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[11] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

[12] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[13] Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. Remote Sensing, 13(22), 4712.

[14] Cao, Z., Liao, T., Song, W., Chen, Z., & Li, C. (2021). Detecting the shuttlecock for a badminton robot: A YOLO based approach. Expert Systems with Applications, 164, 113833.

[15] Huang, R., Gu, J., Sun, X., Hou, Y., & Uddin, S. (2019). A rapid recognition method for electronic components based on the improved YOLO-V3 network. Electronics, 8(8), 825.

[16] Sun, X., Wu, P., & Hoi, S. C. (2018). Face detection using deep learning: An improved faster RCNN approach. Neurocomputing, 299, 42-50.

[17] Nsaif, A. K., Ali, S. H. M., Jassim, K. N., Nseaf, A. K., Sulaiman, R., Al-Qaraghuli, A., ... & Nayan, N. A. (2021). FRCNN-GNB: Cascade faster R-CNN with gabor filters and naïve Bayes for enhanced eye detection. IEEE Access, 9, 15708-15719.

[18] Albahli, S., Nawaz, M., Javed, A., & Irtaza, A. (2021). An improved faster-RCNN model for handwritten character recognition. Arabian Journal for Science and Engineering, 46(9), 8509-8523.

[19] Yin, S., Li, H., & Teng, L. (2020). Airport detection based on improved faster RCNN in large scale remote sensing images. Sensing and Imaging, 21(1), 1-13.

[20] Wen, L., Ding, J., & Loffeld, O. (2021). Video SAR moving target detection using dual faster R-CNN. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 2984-2994.

[21] Xiao, Y., Wang, X., Zhang, P., Meng, F., & Shao, F. (2020). Object detection based on faster R-CNN algorithm with skip pooling and fusion of contextual information. Sensors, 20(19), 5490.

[22] Zhang, N., Feng, Y., & Lee, E. J. (2021). Activity Object Detection Based on Improved Faster R-CNN. Journal of Korea Multimedia Society, 24(3), 416-422.

[23] Shinde, S., Kothari, A., & Gupta, V. (2018). YOLO based human action recognition and localization. Procedia computer science, 133, 831-838.

[24] Huang, R., Pedoeem, J., & Chen, C. (2018, December). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 2503-2510). IEEE.

[25] Lan, W., Dang, J., Wang, Y., & Wang, S. (2018, August). Pedestrian detection based on YOLO network model. In 2018 IEEE international conference on mechatronics and automation (ICMA) (pp. 1547-1551). IEEE.

[26] Ahmad, T., Ma, Y., Yahya, M., Ahmad, B., & Nazir, S. (2020). Object detection through modified YOLO neural network. Scientific Programming, 2020.

[27] Yin, Y., Li, H., & Fu, W. (2020). Faster-YOLO: An accurate and faster object detection method. Digital Signal Processing, 102, 102756.

[28] Wang, Xiqi, Shunyi Zheng, Ce Zhang, Rui Li, and Li Gui. "R-YOLO: A real-time text detector for natural scenes with arbitrary rotation." Sensors 21, no. 3 (2021): 888.

[29] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[30] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.