International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 4
ISSN: 2088-5334

SADY: Student Activity Detection Using YOLO-based Deep Learning Approach

Anagha Deshpande [a,*], Krishna Warhade [a]

[a] *School of Electronics and Communication Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India*
Corresponding author: [*]anagha.deshpande@mitwpu.edu.in

***Abstract*— Automating human activity recognition is one of computer vision's most appealing and pragmatic research areas. In this article, we have addressed the problem of video-based student activity detection. The student's activity detection using YOLO (SADY) aims to recognize the normal and abnormal student activities to ensure immediate intervention in case of any risk or necessity. We created our classroom data set of around 220 recordings depicting seven student classroom activities. The YOLOv4 Tiny model was retrained using 5000 labeled keyframes extracted from the train videos. The model was then tested for single or multiple activity detections. We presented the evaluated results for various values of hyperparameters like confidence threshold and Intersection Over Union (IoU) thresholds for the proposed model. The model assigns a unique confidence score and action label to each frame for the test videos by positioning recurrent activity labels. The proposed approach achieved a mean average precision (mAP) of 95% and a frame per second rate (FPS) of 45 for the student activity Class Room (CR) dataset and mAP of 95.18 % for the LIRIS dataset. The experimental findings using the Class Room recorded and LIRIS publicly accessible dataset show that our proposed approach outperforms existing approaches regarding recognition accuracy and speed. The comparable results obtained in this research work imply that the proposed framework could effectively monitor student's activities in schools, colleges, and universities.**

***Keywords*— Human activity recognition; convolution neural network; Class Room dataset; Yolov4Tiny; multi-action detection.**

Manuscript received 1 Nov. 2022; revised 18 Mar. 2023; accepted 13 May 2023. Date of publication 31 Aug. 2023.

I. INTRODUCTION

Video analytics has become an active research field in image and video processing [1]. Human Activity Recognition (HAR) applications are mainly categorized into active and assisted living, healthcare monitoring, surveillance systems, and tele-immersion. The key objectives of HAR systems are to monitor and analyze human activities efficiently and to interpret ongoing events successfully. The recognition of human activity using vision is subject to many challenges due to various factors such as change of view angle, blocking of view, the difference in execution rate, camera movement, and backdrop clutter. Presently, HAR has received much attention in the domain of video analysis [2-5]. Integrating artificial intelligence and analytics with closed-circuit surveillance, the ecosystem can shift from a reactive approach to a proactive one, potentially reducing criminal activity [6]. Growing elderly independent living demands specific smart assistants [7]. Despite ongoing research in HAR, complex human activity recognition is still considered a challenging task [8]. The traditional technique in HAR consists of three steps: preprocessing, feature engineering, and classification. Recently computer vision and machine learning-based techniques have evolved to develop better human recognition models [9]. After the successful application of convolutional neural networks (CNN) to image classification problems [10], researchers started using the potential of CNN for more advanced problems like image segmentation and object detection [11].

This work presents a visual human activity recognition method by analyzing video information retrieved from a single camera. The scenario of student activity in the classroom, considering the real classroom situation with many students attending classes. In this work, we explored the You Only Look Once (YOLO) version v4-tiny model. After extensive training from scratch for the Class Room student activity dataset, the testing video can show the activity class name, confidence value, and the bounding box of the local activity. In our method, various students' classroom activities that fall under the usual or unusual activity category are recognized. Our dataset has seven student activities: Hand Raise, Entry/Exit, Writing/Reading, Presentation, Throwing

objects, Mobile Conversation, and Head-Down. The prime focus of this work is to recognize multiple usual and unusual student activities in a classroom environment. The proposed work incorporated YOLOv4 tiny model as the inference time is faster than YOLOv4. The proposed model is also tested for human activity recognition for the LIRIS dataset, consisting of ten visually annotated human actions.[12]

The major contributions of this work are as below:

- We used the YOLOv4-tiny object detector model for the reorganization of multiple student activities for the first time.
- We prepared a large Class Room student activities dataset with ground truth labels.
- We developed and experimented with an effective method to train the YOLO model on the Class Room Students activity video dataset for high accuracy and better speed.
- We compared our approach with the recently used techniques by researchers and analyzed the effects of introducing the YOLOv4-tiny model for human activity recognition.

The rest of the sections of this study is arranged as follows. Section 2 presents the literature material and discusses the proposed model for activity recognition and dataset set collection. Section 3 represents implementation details, setup, experimental result, discussions, and comparison with existing work. Lastly, in Section 4, we summarize our work and conclude.

II. Materials and Methods

Object detection is a challenge in video analytics that needs both localization and classification of one or more elements inside an image. HAR is vital in computer vision, motivating many researchers for extensive study and experimentation. Human activity recognition involves a time sequence classification where there is a need to review a series of time steps to identify the action being performed correctly.

The literature survey presented in this paper relates to the scope of the paper. It focuses mainly on Vision-based HAR study in two categories: Machine Learning and Deep Learning approaches. In the Machine Learning and Deep Learning approach, expert knowledge is not required to get appropriate features, reducing feature extraction efforts. The network facilitates the automatic learning of the features. The Machine Learning approach can be well suited for problems addressing smaller data with lower-cost CPUs. A deep neural network may extract high-level features, making it suitable for complex and challenging tasks.

Jagadeesh et al. [13] employed an activity recognition architecture that included Optical Flow Estimation, Scale-Invariant Feature Transform (SIFT) feature extraction, and Support Vector Machine (SVM) Classifier classification. Deshpande and Warhade [14], the author proposes that the HOG and PCA features are embedded and inputted into the ANN and an optimized SVM classifier. The methodology was tested on a benchmark KTH dataset, and recognition accuracy improved to 99.21%. Agarwal et al. [15] applied the R transform technique in combination with a Principal Component Analysis (PCA) and independent component analysis. The Hidden Markov Model (HMM) was employed further for activity recognition.

In recent years, researchers started focusing on Deep learning-based approaches for addressing the HAR problems. Comprehensive surveys are provided in some previous studies [16]-[18]. Human activity recognition problems are efficiently addressed using Convolution Neural Networks (CNN) [19], [20]. The key advantage is that CNN can be used to learn an entire process from start to end, including feature extraction and classification. CNNs' superior feature learning necessitates meticulous regularization and ample labeled data.

Wang et al. [21] proposed CNN with two streams; CNN is trained using separate information from spatial and temporal using the PKU-MMD dataset that contains 51 activities. In the end, the two pieces of information are combined to get the inferences. Almaadeed et al. [22] analyzed each sequence from the scene to recognize the related actions by 3D convolutional neural networks (3DCNNs). The author claims that the human activity approach provides accurate multi-human action recognition.

Hamdy Ali et al. [23] modeled the activities as 3D forms created by layering 2D silhouettes in a Spatiotemporal volume. In Liu et al. [24], the MHI and VGG-16 neural networks collect spatial and temporal features, whereas the Kalman filter and Faster R-CNN deep model capture static data and detect the target position. In Arifoglu and Bouchachia [25], the authors use variants of RNNs (e.g., GRUs and LSTMs) to recognize routine activities and detect unusual activities of old people distressed by dementia. The comparison models show that RNNs beat other ML models at most measured metrics like accuracy, precision, and recall, and LSTMs performed best of the investigated RNN models.

Using transfer learning techniques, various pre-trained models can be deployed for activity recognition [26], [27]. The author Jadhav and Begampure [28] used deep neural architecture for human activity detection. The transfer learning method using pre-trained Inception-V3 cascaded with LSTM. The transfer learning approach benefits in the reduction of training duration as it employs learned weights.

In Shinde and Kothari [29], the YOLO model is employed for human activity recognition for the LIRIS dataset. The LIRIS dataset comprises ten human activities that have been visually annotated, including categories such as interactions between humans and objects and between humans. Activity labels and confidence values are assigned to every frame of the video. Intermittent frames from the video sequence are treated instead of considering the full video stream. The model claims the real-time recognition speed as 15-16 FPS (Frame per second).

The literature review depicts that video-based human activity recognition provides promising results regarding recognition accuracy, but they suffer from false recognition. Issues when working with real-world deployment. More specifically, the challenges are intra-class variation, the disparity in human appearance, camera viewpoint change, backdrop clutter, occlusion, and illumination conditions.

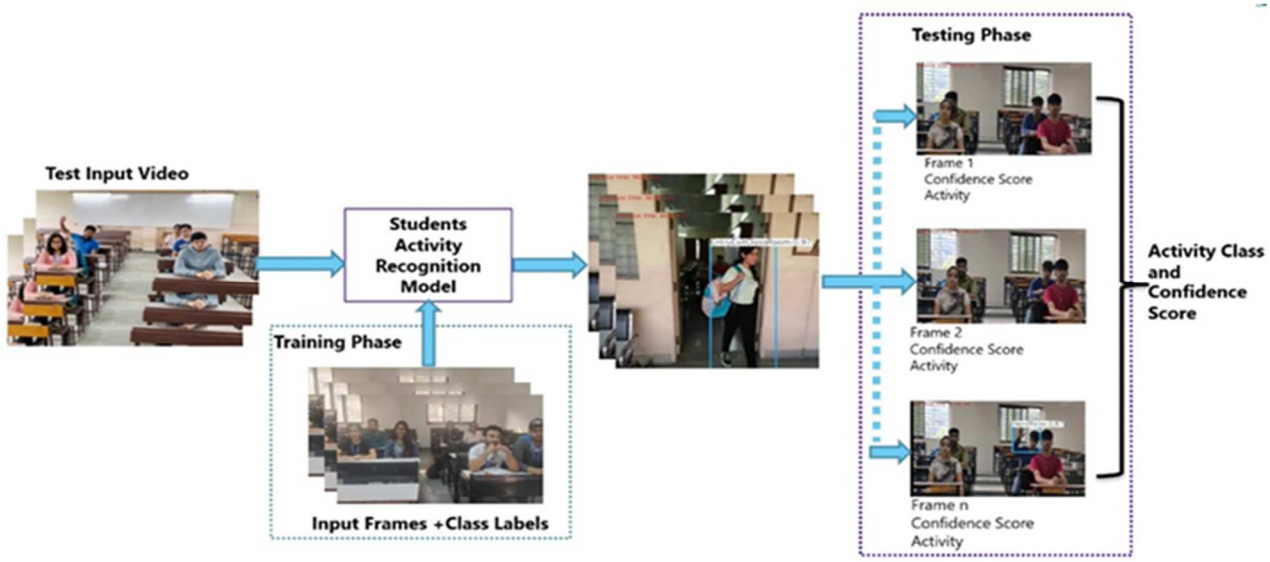

Fig. 1 Workflow diagram of the proposed model

In our approach, we chose the YOLO model to recognize and localize activity from video without using optical flow information from frames. Instead of using CNN for classification, in this work, we chose CNN as the detector model, YOLOv4Tiny. Human activity recognition using object detection techniques was significantly more robust to occlusion, complex scenes, and challenging illumination. In the present work, we have removed the redundant frames, thus reducing the computational time.

The workflow diagram in Figure 1 illustrates the proposed technique in this paper. In the training stage, firstly, we recorded a custom video dataset comprising a set of usual and unusual student activities. For the present work, we focused on seven activities comprising usual and unusual student activities in classroom environments. The technique implemented for training and testing the model is the Tiny Version 4 YOLO Model [30], which will be discussed in-depth in the next section.

The testing is carried out by setting different values of the IoU (Intersection over Union) and confidence thresholds. In the final step, assessment parameters like precision, recall, F1-score, and mAP are calculated to evaluate the Smart student's Activity Detection performance using YOLO (SADY).

A. Object detection method

YOLO Model offers significant benefits compared to systems that are based on classifiers. YOLO looks at the whole image and the predictions made by the image's global context. YOLO makes predictions using a solo network, contrasting approaches like R-CNN, which needs thousands of networks to make a single picture prediction. YOLOv4 is a popular single-stage object detector published in April 2020. YOLO model splits the object detection task into two steps, first Regression is used to determine object position using bounding boxes, and then classification to YOLOv4 achieved outstanding performance on the COCO dataset for speed and accuracy.

YOLOv4 in Backbone uses the Darknet architecture (darknet-53), which has stacked 53 layers, resulting in a 106-layer fully convolution architecture for automatic object identification.

B. YOLOv4-Tiny Model

The real-time implementation of human activity detection is affected by the inference time in two-stage detection systems like CNN, RCNN, Faster RCNN, etc. In anomalous or suspicious activity, recognition inference time plays a vital role. The YOLOv4-tiny model is selected for the work to tradeoff between recognition accuracy and speed. Another advantage of using YOLOv4-tiny is a lighter model with fewer convolution layers than YOLOv4.

The YOLO model takes the entire image of size $640X480$ and divides it into SXS grds ($20X15$). For each grid cell, it estimates B bounding boxes and their confidence scores. Each box is responsible for giving conditional class probability C_n. Figure 2 shows the attributes of bounding boxes where b_x , b_y indicate the middle location point of the box with height b_h and width b_w. Table 1 shows the sample bounding box values for the activity classes from the annotation process.

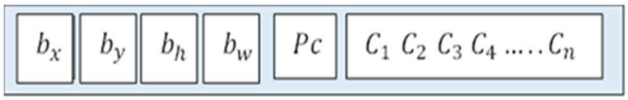

Fig. 2 The Bounding Boxes Attributes of YOLOv4

TABLE I
SAMPLE BOUNDING BOX VALUES FOR ACTIVITY CLASSES

Class Index	X Coordinate	Y Coordinate	Width	Height
0	0.4398523	0.4557086	0.4103321	0.3631889
1	0.5859778	0.4104330	0.2693726	0.5570869
2	0.5276752	0.4862204	0.3092250	0.2076771
3	0.5380073	0.6250000	0.217712	0.1269685

C. Datasets

1) Class Room (CR) Dataset: The dataset recording tasks comprise human activities that include multiple students in the video simultaneously, activities including multiple people, and human-object interactions. The purpose of this

dataset is to capture students' activities in the classroom environment which can be used to recognize the usual and unusual behavior of students in the classroom. The dataset was created by getting the help of twenty student volunteers who participated in seven different activities in the various classrooms. The dataset has been shot with one moving mobile camera delivering color videos in MP4 format. Each video sequence is 4–5 seconds long and captured using frames per second rate of 30 FPS. The Class Room dataset comprises the videos, a collection of RGB frames with a resolution of 640X480. The Class Room dataset requires a disk memory size of up to 100MB.

Many of the video datasets available fall into the following categories:

- Simple repetitive actions like jumping and clapping (e.g., KTH [31], Weizmann dataset [32])
- A real-time dataset for human-to-human or human-to-object interactions (e.g., UCF101 [33])
- Actions on YouTube, daily activities videos (e.g., UCF YouTube [34], Google Ava [35])
- RGB-D dataset encompassing depth of knowledge in the video (e.g., MSRC-12 [36])

The datasets mentioned above are with simple or general human activities. The Class Room dataset was recorded with the specific application in context to students' activities in classrooms at schools, colleges, and universities. The application of this dataset, along with the activity classification, can also help detect abnormal or infrequent student activities in classrooms. It needs to extract motion information from video, separate it from color and texture data, and the characteristics of human behavior. The dataset was recorded considering the real-time challenges as follows:

- Camera View Angle Change
- Multiple People Present in a Frame
- Illumination Change and Scale differences

Fig. 3 illustrates the specimen frames from the Class Room (CR) dataset for various student activities. Table 2 provides a summary of the recorded classroom dataset and Table 3 shows the abbreviations used for students' activities from the Class Room recorded dataset.

TABLE II
BRIEF DESCRIPTION OF RECORDED CLASSROOM DATASET

Sr. No	Student's Activity Class	Total No. of Videos	Total No. of Frames
1	Hand Raise	22	2640
2	Writing/ Reading	22	2640
3	Presentation	22	2640
4	Entry-Exit	22	2640
5	Head Down	22	2640
6	Mobile Conversation	22	2640
7	Throwing Object	22	2640
		154	**18480**

TABLE III
THE ACTIVITY CATEGORIES IN THE CLASSROOM DATASET

Sr. No.	Activities	Abbreviations
1	Hand Raise by Students	HR
2	Writing on Notebook / Reading Book	WR
3	Presenting Seminar, Project topics	PS
4	Enter/Exit a classroom (pass through a door)	EECR
5	Sitting with Head Down	HD
6	Mobile conversation	MC
7	Throwing Object	TO

2) LIRIS Dataset: The LIRIS human activities dataset includes videos that depict individuals carrying out everyday activities, such as conversing, making phone calls, and exchanging objects. The dataset is fully annotated, with XML annotations indicating the bounding boxes of the activities. The videos were captured using two different cameras: a mobile robot-mounted camera that produced VGA resolution grayscale videos and depth images using an MS Kinect depth camera and a consumer camcorder that generated DVD resolution color videos.

Fig. 3 Sample frames in Class Room Dataset

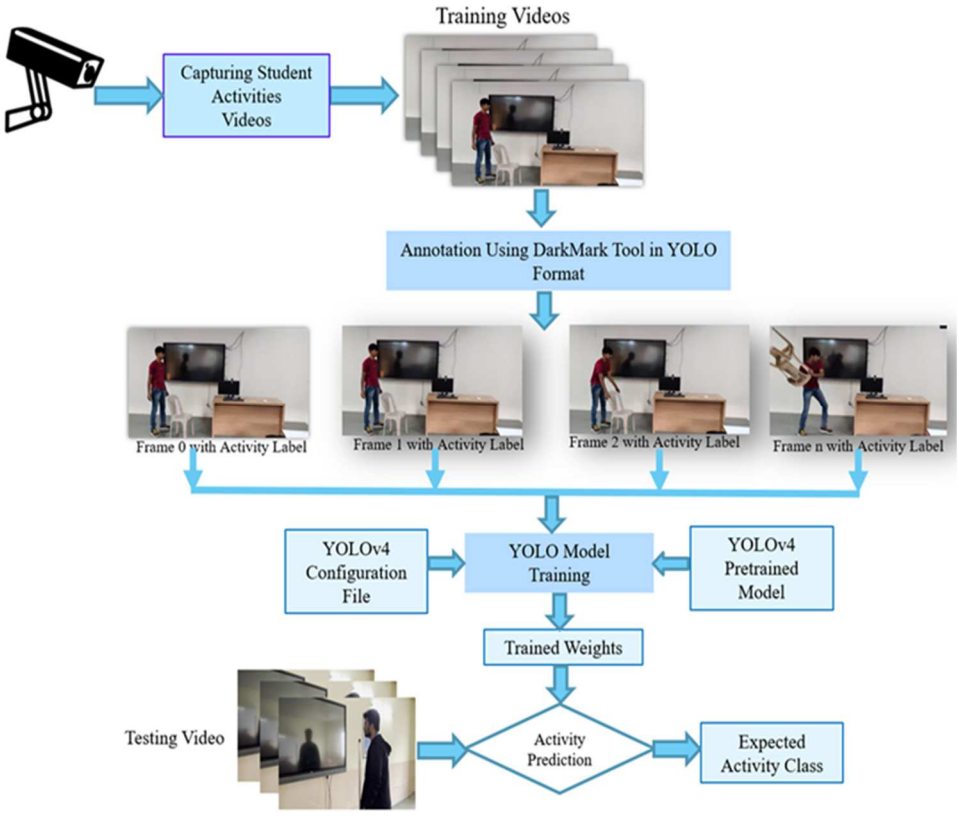

Fig. 4 YOLOv4 Training Flow

D. Model Training

The YOLO-based student activity recognition model training flow diagram is described in Fig. 4. Firstly, each video is segmented into 30 frames per video from the Class Room dataset. To eliminate data redundancy, frames with intervals of five were selected for annotations. Each video was segmented into 30 frames and labeled as one of the seven activity groups using the dark-mark tool. YOLO requires the below-listed files to start the training: The train and test text files contain a file path to train and test images.

- The data file contains the path to train, test, cfg (configuration), and weights backup files.
- The names file contains class names according to line numbers starting from zero.

$$Filter\ Size = (Number\ of\ Classes + 5) * 3 \quad (1)$$

$$Max_{batches} = 2000 * number\ of\ classes[\,minimum] \quad (2)$$

The filter values in the configuration file of YOLO (.cfg file) for the convolution layer before each YOLO layer is calculated by equation 1, and the maximum batch size is calculated by equation 2[37]. Table 4 shows a few of the training parameters set to train the model on the Class Room dataset. In this work training the model from scratch was necessary as the existing trained models are trained from a completely different category of classes (COCO dataset). The model was trained for 20000 iterations with a batch size of 64 and 8 subdivisions. The average loss found was around 0.04. Fig. 5 displays the training progress graph for loss versus the numb of epochs.

TABLE IV
TRAINING PARAMETERS FOR DEEP TRAIN MODEL

Training Parameters used (train.cfg)			
Batch:	64	Hue:	0.1
Subdivisions:	8	Learning_rate:	0.002610
Width:	640	Burn_in:	1000
Height:	480	Max_batches:	20000
Channels:	3	Policy:	steps
Momentum:	0.9	Steps:	16000,18000
Decay:	0.0005	Scales:	0.1,0.1
Angle:	0	Cutmix:	0
Saturation:	1.5	Flip:	1
Exposure:	1.5	Filter Size:	36

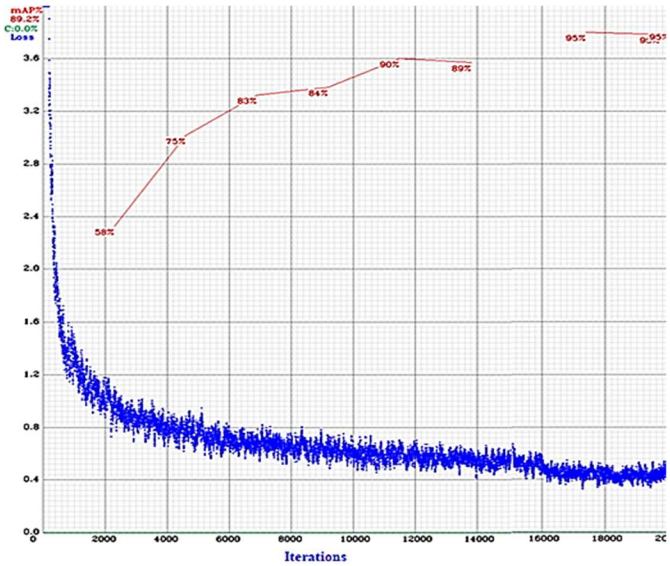

Fig. 5 Graph for training loss vs. epochs

TABLE V
AVERAGE PRECISION VALUES FOR EACH ACTIVITY CLASS

Sr. No	Class Abbreviation	Class Name	AP (%) (IoU Threshold: 0.25)	AP (%) (IoU Threshold: 0.50)	AP (%) (IoU Threshold: 0.75)
1	**EECR**	Entry-Exit Class Room (Class 0)	76.11	76.11	22.00
2	**HR**	Hand Raise (Class 1)	99.54	99.54	99.54
3	**HD**	Head Down (Class2)	99.93	99.93	99.93
4	**MC**	Mobile Conversation (Class 3)	100.00	100.00	98.48
5	**PS**	Presenting Seminar (Class 4)	99.79	99.79	99.79
6	**TO**	Throwing Object (Class 5)	93.33	93.33	73.81
7	**WR**	Writing Reading (Class	100.00	97.06	80.17

III. RESULTS AND DISCUSSIONS

This section describes the performance evaluation of the presented study for students' activity recognition. The experimentation was completed on a Linux (64-bit) operating system with an Intel Core i5-8250U CPU running at 1.60 GHz, a GPU (GeForce GTX 1650), and 16 GB of graphics card RAM. We executed the proposed student activity detection system using Python with CUDA 11.0, cuDNN 7.6.5, and OpenCV 3.2. Darknet is a C/CUDA neural network framework for computer vision tasks like object detection and image classification. To get the CPU and GPU computation support, an open-source Darknet framework is used for training the YOLO [38]. Table 6 shows the evaluation parameters used to test the model performance for the student's activity dataset.

A. Quantitative Results

The average precision is the primary evaluation parameter for measuring the model's performance. Intersection over Union is a statistic for evaluating the accuracy of an object detector on a specific dataset.

$$IoU = Area\,(B\_p \cap B\_(gt))/Area\,(B\_p \cup B\_(gt)) \quad (3)$$

In equation 3,B_p is the predicted bounding box and B_gt is the ground truth bounding box. The confidence score indicates the chance that an anchor box has an object defined by equation 4. In equation 4 $Pr(object)$ indicates the likelihood that the detection region comprises an object. The IoU threshold is the degree of overlap between the ground truth and prediction boxes that must be present for the prediction to be measured as a true positive. The confidence score and the IoU are used to evaluate whether a finding is true or false. To identify the True Positives (TP) IoU threshold and Confidence threshold values play a significant role.

$$Confidence\ Score = P_{r(Object)} \times IoU \quad (4)$$

In YOLO (You Only Look Once), IoU (Intersection over Union) is used as a threshold to determine whether a predicted bounding box should be considered a true positive detection. The standard threshold value for YOLO varies depending on the specific implementation and the application's needs. Typically, the threshold value is set between 0.1 and 0.5. The experimentation is carried out to calculate Average Precision (AP) values for each class in the dataset for different IoU threshold values.

Table 5 shows the percentage values for average precision for each activity class. Empirically, the IoU threshold value of 0.50 is a good choice for opting for greater precision for all the classes. Fig. 6 shows the graphical representation of the average precision results.

$$AP = \sum_{i=1}^{n-1}(r_{i+1} - r_i)\, p_{interp}(r_{i+1}) \quad (5)$$

where $r1, r2, \ldots., rn$is the recall levels

$$mAP = \sum_{i=1}^{K} AP_i/K \quad (6)$$

Quantitative parameters such as precision, recall, F1-score, and mean average precision is used to test the classification performance of the implemented model. Precision in the classification results for each class indicates how many data items expected to belong to that class are accurately classified. Recall indicates the percentage of data in that class that is correctly predicted. The F1 score is the harmonic mean of Precision and Recall.

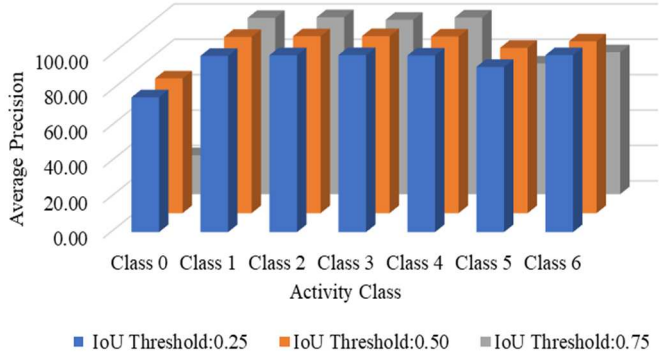

Fig. 6 Average Precision results for each activity class

TABLE VI
EVALUATION PARAMETERS FOR DIFFERENT CONFIDENCE THRESHOLDS

Sr. No.	Confidence Threshold	Precision (%)	Recall (%)	F1-score (%)	mAP(%)
1	0.25	95	97	96	95.11
2	0.5	97	96	96	95.11
3	0.75	98	93	95	95.11

The model is tested for the LIRIS dataset [12] comprising the videos in RGB frames. The ten visual annotations are available in XML format. The LIRIS dataset comprises handshaking, picking an object, object hand covering, entering, and exiting the room, etc. The dataset of 167 videos is divided into 109 training and 58 testing videos. The proposed model is trained up to 20000 epochs and the training mean average precision was 95.18 % for IoU threshold 0.5. The mean average precision from the methodology proposed in this paper for the randomly selected 22 test videos is 97.63%.

Fig. 7 Qualitative results on test videos

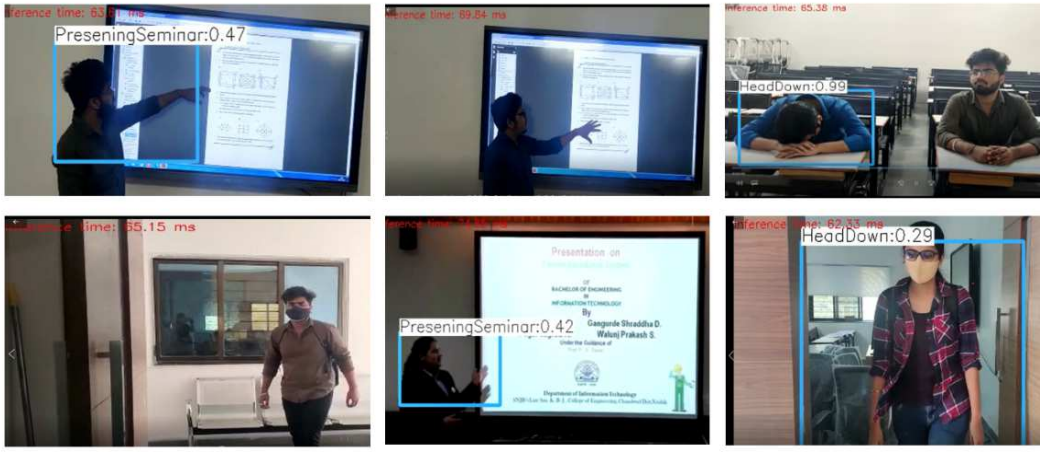

Fig. 8 Qualitative results on real-time videos

B. Qualitative Results

A qualitative performance evaluation of a proposed method is achieved by visual inspection and assessing its accuracy. Figure 7 shows the qualitative result, frames shot from test videos accurately depicting single and multi-activities detected at a time. The algorithm is also tested on a few unseen real-time captured videos from different activity classes. Fig. 8 shows the resulting frames with class names and scores. All the activity classes mentioned in the dataset are detected with 97% accuracy except the class of entry-exit from the classroom, which needs improvement in better understanding of features and good annotation methods.

C. Comparison with Previous Work

After experimentation, we compared model performance with the previously reported publications. Precision, recall, and F1 score are metrics used to evaluate the performance of a classification model. Mmereki et al. [40] in the study deployed YOLOv3 to detect and recognize human actions in aerial footage, achieving an average precision of 82.30% and an F1 score of 88.10%. In the YOLO patient monitoring system, abnormal patient activities like falling, heart pain, fainting, and vomiting are classified with an F1-Score of 89.2%.[39]. F1- score is more useful than accuracy for imbalanced class distribution.

Shinde and Kothari [29] employed activity detection using the LIRIS dataset comprising varied activities like handshaking, picking an object, hand-over, entering, and exiting the room, etc., and quoted precision of 89.88%. The comparison of the proposed model results with the state of art techniques published in the literature is shown in Table 7. The comparison chart depicts that the proposed models considerably outperform the other models.

TABLE VII
COMPARISON OF RESULTS FROM THE PROPOSED MODEL WITH THE STATE-OF-THE-ART TECHNIQUES

Sr. No.	Methods	Precision (%)	Recall (%)	F1-score (%)
1	W. Mmereki, et al. [40]	82.30	-	88.10
2	G. Malik, et al. [39]	90.134	88.421	89.269
3	Shinde, et al. [29]	89.881	88.083	88.358
4	Proposed Model	**97**	**96**	**96**

IV. CONCLUSION

The major contribution of this work is developing a lightweight, fast trainable, speedy recognizing model and creating a custom dataset of student activities. Video frames are processed individually at specific intervals to reduce temporal redundancy and processing time for fast and easy

computation. We chose the YOLOv4-tiny model as the backbone network for CNN to achieve a faster inference time. After being implemented using the computer machine mentioned above, the proposed model can process the student's classroom dataset images at 42–45 frames per second faster than the former technique [30]. The evaluation parameters stated in Table 8 show that Student's Activity Detection using YOLO is significantly more robust to occlusions and illumination changes and appropriate for real-time detection. The limitation of this approach for large datasets is image annotation, as the image annotation task is a little expensive and time-consuming. Furthermore, we intend to extend our work to more challenging datasets.

ACKNOWLEDGMENT

The authors thank Dr. Vishwanath Karad, MIT World Peace University, Pune, Maharashtra, India, for the support in dataset recording. The authors also appreciate the efforts and support of B.Tech. students of MIT World Peace University, Pune, India.

REFERENCES

[1] F. Hidayat, F. Hamami, I. A. Dahlan, S. H. Supangkat, A. Fadillah, and A. Hidayatuloh, "Real Time Video Analytics Based on Deep Learning and Big Data for Smart Station", *Journal of Physics: Conference Series,* vol. 1577, no. 1, July 2020, doi:10.1088/1742-6596/1577/1/012019.

[2] H. Amanullah, S. Letchmunan, M Zia, U. Butt, H. Fadratul, "Analysis of Deep Neural Networks for Human Activity Recognition in Videos – A Systematic Literature Review", *IEEE Access*, vol. 99, pp 1-1, 2021, doi: 10.1109/Access.2021.3110610.

[3] R. Mondal, D. Mukherjee, P. K. Singh, V. Bhateja and R. Sarkar, "A New Framework for Smartphone Sensor-Based Human Activity Recognition Using Graph Neural Network," in *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11461-11468, 15 May, 2021, doi: 10.1109/JSEN.2020.3015726.

[4] M. Bendali-Braham, J. Weber, G. Forestier, Lhassane Idoumghar, P Alain Muller, "Recent trends in crowd analysis: A review", *Machine Learning with Applications*, vol. 4, June 2021, 100023, ISSN 2666-8270, doi:10.1016/ jmlwa.2021.100023.

[5] M. R. Bhuiyan., J. Abdullah, N. Hashim, F. Farid, "Video analytics using deep learning for crowd analysis: a review", *Journal Multimedia Tools Applications*, vol. 81, pp. 27895-27922, March 2022, doi:10.1007/s11042-022-12833.

[6] S. Bhalla, K. Singh, "Exploration of Crime Detection Using Deep Learning", Innovations in Cyber-Physical Systems. *Lecture Notes in Electrical Engineering*, vol. 788, pp. 297-304, September 2021.

[7] A. Hayat, F. Morgado-Dias, B.P. Bhuyan, R. Tomar, "Human Activity Recognition for Elderly People Using Machine and Deep Learning Approaches", MDPI *Journal Information*, vol. 13, issue 6, pp. 275, 2022, doi: 10.3390/info13060275.

[8] C. Jobanputra, J. Bavishi, N. Doshi, "Human Activity Recognition: A Survey, *Procedia Computer Science*, vol. 155, pp. 698-703, 2019, ISSN 1877-0509, doi: 10.1016/j.procs.2019.08.100.

[9] A. M. F and S. Singh, "Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications," *Proc 3rd International Conference on Signal Processing and Communication* (ICPSC), pp. 110-114, 2021, doi: 10.1109/ICSPC51351.2021.9451736.

[10] S. S. Yadav, S.M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis", *Journal of Big Data*, Vol.6, 113, December 2019, doi:10.1186/s40537-019-0276-2.

[11] A. Ullah, M. Khan, W. Ding, V. Palade, Ijaz Ul Haq, S. W. Baik, "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications," *Applied Soft Computing*, vol. 103, 107102, May 2021, doi: 10.1016/j.asoc.20 21.107102.

[12] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C. E. Bichot, C. Garcia, B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements", *Computer Vision and Image Understanding*, vol. 127, pp.14-30, October 2014.

[13] B. Jagadeesh, & C M Patil, "Video Based Human Activity Detection, Recognition and Classification of actions using SVM", *Transactions on Machine Learning and Artificial Intelligence*, vol. 6, no. 6, January 2019, doi: 10.14738/tmlai.66.5287.

[14] A. Deshpande, K. K. Warhade, "An Improved Model for Human Activity Recognition by Integrated feature Approach and Optimized SVM", *Proc. International Conference on Emerging Smart Computing and Informatics* (ESCI), April 2021, pp. 571-576.

[15] A. Agarwal, A. Sharma, A. Gupta, V. Goel, "Human Movement Recognition System using R", *International Journal of Engineering and Advanced Technology* (IJEAT) ISSN: 2249-8958, vol. 8, Issue 5, pp. 560-566, June 2019.

[16] M. F Abdul, S. Singh, "Computer Vision-based Survey on Human Activity Recognition System", Challenges and Applications, *Proc 3rd International Conference on Signal Processing and Communication*, 2021, pp.110-114.

[17] D. R Beddiar., Nini B., Sabokrou M. et.al, "Vision-based human activity recognition: A survey", *Multimedia Tools and Applications*, vol. 79, pp. 30509–30555, August 2020, doi: 10.1007/s11042-020-09004-3

[18] H-B Zhang, Y-X Zhang, B. Zhong, Qing Lei, L. Yang, Ji-Xiang Du, and D-S Chen. "A Comprehensive Survey of Vision-Based Human Action Recognition Methods", *Sensors,* vol. 19 no. 5, 1005, February 2019, doi: 10.3390/s19051005.

[19] Sarnaik, Neha, "Human Activity Recognition using CNN", *International Journal of Scientific and Research Publications* (IJSRP), vol 10, issue 2, February 2020, pp 9804, doi:10.29322/IJSRP.10.02.2020.

[20] N. Junagade and S. Kulkarni, "Human Activity Identification using CNN," *Proc Fourth International Conference on I-SMAC* (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), 2020, pp. 1058-1062, doi: 10.1109/I-SMAC49090.2020.9243477.

[21] K. Wang, Xuejing Li, Jianhua Yang, Jun Wu, Ruifeng Li, "Temporal action detection based on two-stream You Only Look Once network for elderly care service robot "*International Journal of Advanced Robotic Systems*, vol 18, issue 4, July 2021.

[22] N. Almaadeed, O. Elharrouss, S. Al-Maadeed, A. Bouridane, A. Beghdadi "A novel approach for robust multi-human action recognition and summarization based on 3D convolutional neural networks", March 2021, doi:10.48550/arXiv.1907.11272.

[23] H. Hamdy Ali, H. M. Moftah, A. Youssif, "Depth-based human activity recognition: A comparative perspective study on feature extraction", *Future Computing and Informatics Journal*, vol. 3, issue 1, pp 51-68, 2018, doi: 10.1016/j.fcij.2017.11.002.

[24] C. Liu, Y. J. Yang, H. Haima, Yang X., Hu J. L., "Improved human action recognition approach based on two-stream convolutional neural network model", *The Visual Computer,* vol. 37, pp. 1327–1341, June 2021, doi: 10.1007/s00371-020-01868-8.

[25] D. Arifoglu, A. Bouchachia, "Activity recognition and abnormal behavior detection with recurrent neural networks", *Procedia Computer Science*, vol. 110, pp.86–93, 2017.

[26] S. Chakraborty, R. Mondal, P. K. Singh, R. Sarkar, and D. Bhattacharjee, "Transfer learning with fine tuning for human action recognition from still images", *Multimedia Tools Applications* 80, vol. 13, pp. 20547–20578, May 2021, doi:10.1007/s11042-021-10753-y.

[27] Oh S., Ashiquzzama A., Lee D., Kim Y., Kim J., "Study on Human Activity Recognition Using Semi-Supervised Active Transfer Learning", *Sensors,* Basel, Switzerland, vol. 21, no. 8, April 2021, doi: 10.3390/s21082760.

[28] P. M. Jadhav, S. Begampure, "Intelligent video analytics for human action detection: a deep learning approach with transfer learning", *International Journal of Computing and Digital Systems*, vol .11 no.1, pp. 64–71, July 2021, doi:10.12785/ijcds/110105.

[29] S. Shinde, A. Kothari, G. V, "YOLO based human action recognition and localization", *Procedia Computer Sci*, vol. 133, pp. 831–838, 2018.

[30] J. Zicong, L. Zhao, S. Li, and Y. Jia, "Real-time object detection method based on improved YOLOv4-tiny" *Journal of Network Intelligence*, vol. 7, no.1, February 2022, doi: 10.48550/arXiv.2011.04244.

[31] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," *Proceedings of the 17th International Conference on Pattern Recognition 2004*, September 2004, vol.3, pp. 32-36, doi: 10.1109/ICPR.2004.1334462.

[32] L. Zelnik-Manor, &M. Irani, "Event-based analysis of video", *Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, December 2001, vol. 2, pp. II-II.

[33] K. Soomro, A.R Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild” *CRCV-TR-12-01,*2012, http://arxiv.org/abs/1212.0402.

[34] J. Liu, J. Luo, and M. Shah, “Recognizing Realistic Actions from Videos in the Wild", *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), August* 2009, pp. 1996-2003, doi: 10.1109/CVPR.2009.5206744.

[35] G. Chunhui, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, et al. "Ava: A video dataset of spatiotemporal localized atomic visual actions", *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, December 2018, pp 6047-6056, doi: 10.1109/CVPR.2018.00633.

[36] P. Barmpoutis, T. Stathaki, and S. Camarinopoulos, "Skeleton-Based Human Action Recognition through Third-Order Tensor Representation and Spatio-Temporal Analysis", *Inventions,* vol. 4, no.9, February 2019, doi: 10.3390/inventions4010009

[37] H. Hendry, Rung-Ching Chen, “Automatic License Plate Recognition via sliding-window darknet- YOLO deep learning”, *Image and Vision Computing*, vol. 87, pp. 47-56, ISSN 0262-885, July 2019, doi: 10.1016/j.imavis.2019.04.007.

[38] A. Bochkovskiy, C. Y. Wang, H. Yuan, M. Liao., “YOLOv4: optimal speed and accuracy of object detection”, April 2020, https://arxiv.org/abs/2004.10934

[39] G. Malik, Muhammad H., Yousaf, Shah Nawaz, Zakaur Rehman, Hyung Won Kim, “Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture”, *Electronics*, vol 9, no. 12, November 2020, doi: 10.3390/electronics9121993.

[40] W. Mmereki, R. S. Jamisola, D, T. Mpoeleng, Petso, "YOLOv3-Based Human Activity Recognition as Viewed from a Moving High-Altitude Aerial Camera", *Proc 7th International Conference on Automation, Robotics and Application*s (ICARA), Feb 2021, pp. 241– 246.