Vision-based Student Activity Observance System for the MCAST library

Keeron Spiteri
Institute of Information & Communication Technology
Malta College of Arts, Science & Technology
Corradino Hill
Paola PLA 9032
keeron.spiteri.g53977@mcast.edu.mt

***Abstract*—This paper presents a method for analyzing and recognizing student activity in the MCAST library through video surveillance cameras using machine learning methods, focusing on the YOLO (You Look Only Once) object detection family. The purpose of this research was to determine whether the use of computer vision can be used as a real-time student action observer in the library. The proposed system detects humans (student) and objects (book, laptop or phone), afterwards classifying the student's activity status to no_activity, using_book, using_laptop or using_phone. An activity hierarchy was implemented to determine the student's activity when multiple objects were detected in the vicinity. YOLOv8 was used as an object detector, DeepSORT for object tracking and a custom activity recognizer was used to classify student activities. Despite a small dataset of 198 annotated images, the system achieved an overall accuracy of 86%. False positives occurred when multiple objects were present or when object overlap did not truly reflect student engagement. The study demonstrates that computer vision offers a future foundation in real-time activity monitoring in educational environments.**

***Index Terms*—MCAST, Library, Human Activity Recognition, Surveillance System, Computer Vision**

I. Introduction

The MCAST (Malta College of Arts, Science & Technology) Library wants to provide students with a comfortable environment that includes the necessary resources to complete their academic studies. Library staff make decisions through data collected from online survey reports sent to students by email. Although most students check their school email, most ignore the survey report emails sent by the library, this leads to limited data for making informed decisions. To address this challenge, this study proposes a system that is capable of automatically tracking, classifying and analyzing student activities within the library. The data collected may provide invaluable insights for the library staff, allowing for more efficient informed decisions. By leveraging HAR (Human Activity Recognition) the system can provide analytical data in real-time without requiring manual input by students, thereby improving the volume and quality of data for the library staff. HAR has garnered significant attention due to its application in surveillance, healthcare, and human-computer interaction. Recent advancements have leveraged deep learning techniques, notably the YOLO object detection algorithm and the DeepSORT tracking algorithm, to enhance detection and tracking accuracy. This study demonstrates the unique challenges of implementing AI-based analytics in real-world environments, specifically utilizing existing CCTV infrastructure originally designed for security monitoring. This study addresses several challenges:

1) Which object detection algorithm can robustly detect users and activity objects from wide perspective angles?
2) How can a robust activity classification algorithm be developed for a library?
3) How can the algorithm determine the specific activity that the student is engaged in when the student has multiple different potential activities surrounding them?

II. Literature Review

A. Object Detection Using YOLO

In recent times the YOLO family has become widely used in real-time object detection applications due to its speed and time-efficiency [1], [2]. Unlike two-stage detectors such as R-CNN, Fast R-CNN, and Faster R-CNN [3], YOLO is a one-stage Convolutional Neural Network (CNN) detector, that simultaneously detects objects, their bounding boxes and classifies them by only requiring a single scan of an image [1], [4]. YOLO functions by splitting an image into a grid of S×S. Within each grid cell, multiple bounding boxes are predicted and each contains the coordinates, class probability score and confidence score for a potential detected object [3]. Following predictions, it's common for multiple bounding boxes to overlap resulting in the same object being predicted. YOLO employs Non-Maximum Suppression (NMS) to resolve this problem. NMS suppresses bounding boxes that have a low final score and low Intersection over Union (IoU) score. The final score reflects the likelihood that an object exists, and that it is of a certain class. Final Score = Confidence Score × Class Probability Score. The IoU score reflects the likelihood that the object's predicted bounding box is closer to the ground-truth bounding box [3], [4].

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{1}$$

B. Comparison of YOLO Iterations

1) Study One: YOLOv7 for Unsafe Human Actions: While conducting research on detecting five unsafe human actions, the researchers opted to use YOLOv7 since previous research had shown that compared to previous versions, the YOLOv7 version was the fastest and most accurate real-time object detection algorithm. Their model had obtained a result of precision 96%, recall of 95% and Mean Average Precision (mAP) of 99% at an IoU of 0.5 [1].

2) Study Two: YOLOv7 for Human Actions: While conducting research on detecting three human actions, the researchers opted to work with YOLOv5, YOLOv6 and YOLOv7. Previous research made a comparison from YOLOv1 to YOLOv4 and concluded that YOLOv4 obtained the best results. After training the models, YOLOv7 obtained a better result in performance and activity classification, receiving a precision score of 96% in standing, 94% in sitting and 97% in falling [4].

3) Study Three: YOLOv8 for Customer Activity: While conducting research on a customer activity detection system in fast food courts, the researchers opted to use YOLOv8 for object detection due to its state-of-the-art (SOTA) performance and real-time processing capabilities. The system's obtained precision score for each activity was: 72% eating, 78% drinking, 75% sitting, 70% eating and drinking, and 70% empty table [2].

4) Study Four: YOLOv7 for Critical Safety Concerns: While conducting research on detecting three critical safety concerns in real-time, the researchers opted to use YOLOv7 as it was the most recent version, and it solved issues that previous YOLO iterations had. YOLOv7 obtained an average precision of 37.20% at an IoU of 0.5, which is higher than that of YOLOv5 when recognizing objects. YOLOv7 is faster than previous iterations since a Graphics Processing Unit (GPU) can be utilized. YOLOv7 makes use of CSPDarknet53 which is a backbone network that is deeper and more potent than Darknet53 used in YOLOv5 [3].

C. YOLO Pre-processing Techniques

Resizing the data to the same resolution size is important as the model would be inaccurate on different resolutions. In these studies, the images were resized to 640x640 [1], [4].

Pixel normalization is applied to the data by scaling the pixel values to the range [0,1] [1], [3].

Image smoothing and contrast enhancement is applied to data to increase the model's accuracy and reduce noise [3].

D. YOLO Data Augmentation Techniques

Data augmentation is used to enhance the training quality by modifying the training data using several techniques

- Abbad and Ibraheem had used several techniques to generate three modified versions of each image:
  - 50% probability horizontal flipping
  - Random cropping (0%–20%)
  - Random rotation (-10° to +10°)
  - Random brightness adjustment (-25% to +25%)
  - Random exposure adjustment (-25% to +25%)
- Phatangare et al. had used a few techniques, but there was never a mention of the values:
  - Random rotation
  - Horizontal flipping
  - Random brightness modification

E. Object Tracking Using DeepSORT

Research was conducted on object tracking methodologies, analyzing and reviewing both object detection and tracking techniques to identify which algorithms are obtaining high accuracy in real-time detection. Object tracking involves a number of phases, including object detection, object classification, assigning unique identifiers for discovered objects and tracking the object. Object tracking may be done in a variety of ways. Three popular techniques are the target tracking based on Mosse with Kernelized Correlation Filters (KCF), target tracking based on the Siam algorithm, and DeepSORT based on YOLO. Target tracking based on Mosse with KCF obtains high speeds and requires minimal hardware needs, however it is easy to lose the targeted object since the tracking frame scale is constant and cannot follow the target's changing scale. DeepSORT is built upon the Simple Online and Realtime Tracking (SORT) algorithm, created to track multiple objects simultaneously. DeepSORT enhances SORT's object identification, since it would sometimes mistakenly assign a new or incorrect ID to an object. DeepSORT based on YOLO uses the YOLO algorithm to detect objects, and after detection DeepSORT would be used for object tracking. DeepSORT starts by employing appearance feature extraction using the Residual Network (ResNet) which is a pre-trained CNN. Extracted features are used in future frame to re-identify objects by using the Hungarian algorithm. Motion prediction would be applied using Kalman filters to predict the future positions of the detected objects based on the objects' previous states. Kalman filters function by predicting the object's movement by tracking its relative consistent speed and its previous predictable movement pattern. Certain situations might result in an object not being clearly visible, Kalman filters are used to predict the object's position, bridging the gap until the object becomes clearly detectable again [5].

F. Activity Recognition Methods

Two studies were analyzed from Abbad and Ibraheem, and Garcia-Garcia and Pinto-Elias. Both studies used YOLO for object detection and classification of activities [1], [4]. Although this solution is a success for both studies, in other environments a result with false positives can be obtained since multiple potential activities can be detected simultaneously around a person, and with no confident way of knowing whether a person is participating in the detected activity. A customer activity detection system was developed by utilizing a restaurant's surveillance system. Four polygonal zones were drawn, each corresponding to a table's seating area. Activities at each table were determined based on the presence of objects in the table's spatial zone. A hierarchical rule set was created to prioritize and classify activities when multiple activities were detected simultaneously [2]. S. Phatangare, S. Kate, D. Khandelwal, A. Khandetod, and A. Kharade, used YOLO for object detection and classification, while activity recognition was handled by rule-based logic [3].

TABLE I
COMPARISON OF RELATED STUDIES. (N/S = NOT SPECIFIED)

Author(s), Year, Ref.	**Topic**	**Techniques Used**	**Datasets Used**	**Train**	**Test**
Abbad et al. 2023 [1]	Unsafe Actions Detection for Humans using YOLOv7	YOLOv7	UT-Interaction, ISR-UoL 3D Social Activity	465	N/S
Ismail et al. 2024 [2]	Customer Activity Detection using YOLOv8 and Status Order Algorithm	YOLOv8, DeepSORT	Custom	N/S	N/S
Phatangare et al. 2023 [3]	Real-Time Human Activity Detection using YOLOv7	YOLOv7	COCO, Custom	N/S	N/S
Garcia-Garcia et al. 2022 [4]	Human Activity Recognition implementing YOLO models	YOLOv5, YOLOv6, YOLOv7	UR Fall Detection, Custom	70%	15%
Pujara et al. 2022 [5]	Real-Time Multi-Object Detection and Tracking with YOLO and TensorFlow	YOLOv7, DeepSORT	COCO, Custom	N/S	N/S

III. RESEARCH METHODOLOGY

The MCAST library staff are currently making informed decisions with intent of improving library resources by analyzing data obtained through survey reports sent to students by email. A high percentage of students do not bother to submit a survey report, resulting in a limited data pool. The objective of this prototype is to obtain analytical data from student activities performed in the MCAST library. The data provided would be of higher quality and quantity, resulting in better informed decisions being made by the library staff. The system is designed to operate on the existing CCTV infrastructure without requiring specialized hardware or manual input from students.

The system consists of four primary stages: video input processing, object detection, object tracking, and activity classification. Each stage is modular and can be tuned or replaced independently, allowing flexibility and scalability.

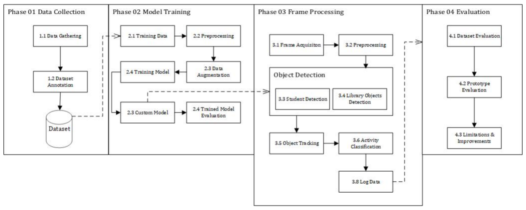

Fig. 1. Research Pipeline

A. Data Collection

An online dataset of students performing activities in a classroom was used to train, validate and test the model. The dataset was split to 70% training, 15% validation and 15% testing. Pre-processing steps included auto-orientation and frame resizing to 640x640 pixels to prepare the image.

B. Object Detection

The YOLO object detector was used due to previous research indicating that it is known for its real-time performance, efficiency and high detection accuracy, which can be seen in Section II-A. The YOLOv8s pre-trained model was used for training, as it requires low performance to run, resulting in a good trade-off between speed and precision. The model was trained on humans (students) and objects (book, laptop and phone) using 100 epochs, batch size of 32 and an image resolution of 640x640. To enhance the quality of our training dataset, data augmentation techniques were applied resulting in 3 augmented versions of each image:

- 50% probability horizontal flipping
- Random cropping (0%–20%)
- Random rotation (-15° to +15°)
- Random Hue adjustment (-15° to +15°)
- Random Saturation adjustment (-25% to +25%)
- Random brightness adjustment (-15% to +15%)
- Random exposure adjustment (-15% to +15%)
- Random noise adjustment (0% to +0.10%)

C. Object Tracking

The DeepSORT object tracker was used to track humans across consecutive frames, assigning a unique identifier to each object, ensuring temporal consistency and analysis over time.

D. Activity Classifier

The classifier algorithm was developed to assign activities based on the presence of activity objects near a student. An activity was classified once a student's bounding box had a minimum of 50% intersection overlap with an activity object's bounding box. The IoA (Intersection over Area) formula was used to determine the intersection between the student and object's bounding boxes.

$$\text{IoA} = \frac{\text{Area of Intersection}}{\text{Area of Outer Object}} \tag{2}$$

To determine an activity when multiple potential activity objects are in the vicinity of a student, an activity hierarchy was introduced following the provided structure:

1) *"using_book"*: score 1
2) *"using_laptop"*: score 2
3) *"using_phone"*: score 3

- *"no_activity"*: was assigned to students with no presence of an activity object nearby.

The order reflects increasing levels of objects that are potentially more distracting.

E. Tools and Environment

This study was performed using the Google Colab Python environment notebook, providing an efficient environment for intensive tasks such as training and testing. Several Python libraries were employed for various tasks. The official YOLOv8 framework from the Ultralytics team was chosen for object detection. DeepSORT using PyTorch was chosen for object tracking due to its long research background and stable tracking. OpenCV handled image processing, frame extraction and video writing.

IV. Findings & Discussion of Results

A. Data Collection Challenges

The dataset was obtained from Roboflow and required annotations to be applied. Due to time limitations 198 images were annotated and used, resulting in a small dataset and the trained model being overfitted. Additionally the dataset was not captured from the desired angle.

B. Activity Classifier Results

The activity classifier algorithm was developed to assign an activity to a student when an activity object had been detected being 50% overlapping within the student's bounding box. This algorithm was successful at recognizing activities but presented negatives:

1) Single Object False Positives: The algorithm would only assign an activity when bounding boxes have a minimum of 50% overlap, resulting in times where false positives get assigned since the algorithm cannot verify whether the student is participating in the activity.

2) Multiple Objects False Positives: An activity hierarchy was introduced to assist in assigning a valid activity when multiple potential activity objects are all within the minimum 50% overlap of the student's bounding box. While this system improved the reliability, a decision to make the system prioritize objects that are more distracting in an academic environment was taken, as presented in Section III-D. This has resulted in false positives being assigned since the algorithm cannot verify what the student is participating in.

TABLE II
Activity Classifier Performance Metrics

Activity Class	**Precision**	**Recall**	**F1-Score**
no_activity	0.84	0.92	0.88
using_book	0.53	0.54	0.53
using_laptop	1.00	0.89	0.94
using_phone	0.94	0.83	0.88

Figure 2 presents the Confusion Matrix, visualizing the algorithm's recognition. Due to a small dataset the *using_book* class was often predicted as *no_activity* which is wrong. The *using_laptop* class obtained a very high F1-Score of 0.94, however this result is due to the dataset being small and only containing a single instance of a laptop.

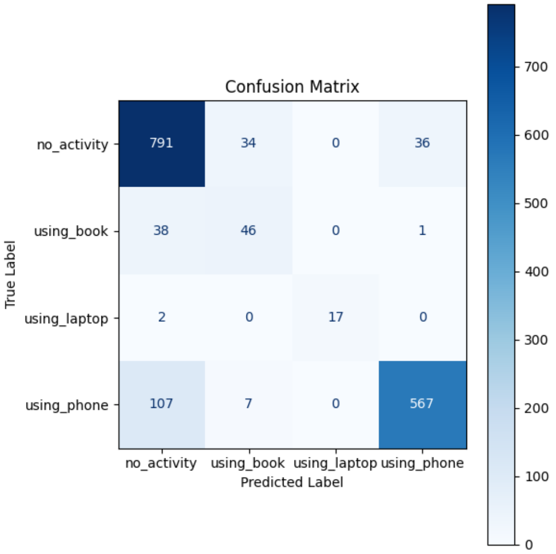

Fig. 2. Activity Classifier Confusion Matrix

The activity classifier achieved an overall accuracy of 86%, indicating reliable activity recognition. Notably, the *no_activity* class yielded the highest recall at 92%, which forms a significant portion of the observed data. This result suggests that the system is capable of distinguishing between active and inactive behavior. Although the system is capable of providing more data, the data is not to be dependent upon as the activity classifier can provide a fairly large false positive pool.

V. Conclusion

In this study a proposed solution to obtain analytical data of student activities using computer vision in a library was explored. Despite the challenges posed by a limited and imbalanced dataset, the proposed system achieved an overall accuracy of 86%, demonstrating promising performance in identifying activities. The design of the activity classifier, although successful still presented false positives in multiple scenarios. The activity hierarchy while simple and effective at assigning an activity to a student when multiple potential objects are in the vicinity, still presented false positives as it had to be tuned to prioritize activities that are most distracting in an academic environment. The current activity classifier will never be fully right in determining the activity the student is doing, since it cannot track the student's hands and eyes.

Future work should focus on using a larger dataset that includes data from a wide-perspective angle, potentially from the MCAST library. The YOLO object detector provided few false detections but future work may explore a more recent YOLO model. Additionally, exploring an activity classifier with pose estimation can be a way forward in reducing false positives by providing more precise student behavior data.

Overall, this study contributes a foundation for academic behavior analysis, offering valuable insights into student engagement in educational environments.

References

[1] M. F. Abbad and I. N. Ibraheem, "Unsafe actions detection for humans using yolov7," in *Proceeding - 2023 International Conference on Artificial Intelligence Robotics, Signal and Image Processing, AIRoSIP 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 122–127.

[2] M. Ismail, A. Zakaria, A. S. A. Yeon, S. M. M. S. Zakaria, L. M. Kamarudin, M. R. Z. Abidin, R. Visvanathan, X. Mao, and N. M. Yusof, "Customer activity detection using yolov8 and status order algorithm," in *Proceedings - 2024 International Conference on Cyberworlds, CW 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 301–307.

[3] S. Phatangare, S. Kate, D. Khandelwal, A. Khandetod, and A. Kharade, "Real time human activity detection using yolov7," in *7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2023 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1069–1076.

[4] S. Garcia-Garcia and R. Pinto-Elías, "Human activity recognition implenting the yolo models," in *Proceedings - 2022 International Conference on Mechatronics, Electronics and Automotive Engineering, ICMEAE 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 127–132.

[5] A. Pujara and M. Bhamare, "Deepsort: Real time & multi-object detection and tracking with yolo and tensorflow," in *Proceedings - International Conference on Augmented Intelligence and Sustainable Systems, ICAISS 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 456–460.