Anomaly Detection in Human Task Performance with Video-Based Contextual Reasoning

Tuhin Mondal
Dept. of CSE (AI & ML)
University of Engineering and Management
Kolkata, India
tuhinm2002@gmail.com

Aisi Pal
Dept. of CSE (AI & ML)
University of Engineering and Management
Kolkata, India
aisipal2002@gmail.com

Ishita Karmakar
Dept. of CSE (AI & ML)
University of Engineering and Management
Kolkata, India
kishita562@gmail.com

Arik Das
Dept. of CSE (AI & ML)
University of Engineering and Management
Kolkata, India
arikdas2002@gmail.com

Swarnanka Saha
Dept. of CSE (AI & ML)
University of Engineering and Management
Kolkata, India
swarnanaksaha5@gmail.com

Sramana Mukherjee
Asst. Prof. Dept. of CSE(AI & ML)
University of Engineering and Management
Kolkata, India
sramana.mukherjee@uem.edu.in

***Abstract*—Understanding and analyzing human task performance is critical in domains such as healthcare, manufacturing, and surveillance. This study presents a computer vision-based framework for anomaly detection using YOLO, a state-of-the-art object detection model, to track and interpret task-related behaviours from video inputs. The system first establishes a baseline of normal activity by learning object usage, movement patterns, and interaction sequences from initial videos of a person performing a specific task. New video inputs are then evaluated to identify anomalies such as unfamiliar object interactions, unusual pauses, and deviation in motion patterns. YOLO enables precise detection and localization of objects and interactions, forming the foundation for reliable anomaly analysis. The framework further enhances explainability by providing textual reasoning for each detected anomaly, offering actionable insights. The proposed system combines robust real-time object detection with behaviour understanding to deliver a scalable and interpretable solution for task performance monitoring.**

***Index Terms*—Anomaly detection, deep learning, object detection, video surveillance, YOLO**

I. Introduction

Monitoring deviations in human task execution is essential in many real-world applications, including healthcare, industrial automation, and personal assistance systems. Conventional anomaly detection techniques often depend on static thresholds or handcrafted features, limiting their ability to adapt to dynamic environments and diverse human behaviours. With the advancements in deep learning and object detection, particularly through models like YOLOv5, there is a new opportunity to develop systems capable of real-time, accurate, and context-aware performance monitoring. In this work, we utilize YOLO to develop a deep learning-based anomaly detection framework that analyzes video footage of individuals performing routine tasks. YOLO is employed to detect and track objects involved in the task, allowing the model to learn typical behavioural patterns, such as the timing of actions, object interactions, and spatial arrangements. Once a behavioural baseline is established from training videos, new inputs are analyzed to detect anomalies—ranging from interacting with unexpected objects to exhibiting irregular movement or timing. Each anomaly is accompanied by a human-readable textual explanation, enhancing the system's interpretability. The key contributions of this research include: Leveraging YOLO for object detection and behavioural modeling in task execution. A novel framework for detecting deviations from learned performance baselines using temporal and spatial features. Generating interpretable textual explanations for anomalies to support decision-making in real-time monitoring. This approach blends the precision of modern object detection with the need for personalized and explainable behavioural analytics, enabling a new class of intelligent monitoring systems.

II. Background

Anomaly detection in human behaviour analysis has gained significant traction in recent years, particularly with the increasing accessibility of video data and advancements in deep learning techniques. Traditional approaches primarily relied on handcrafted features and statistical methods to detect

deviations in activity. For instance, Simonyan and Zisserman introduced a two-stream convolutional network to model human motion patterns, combining spatial and temporal features to enhance action recognition accuracy [Simonyan and Zisserman, 2014]. With the rise of deep learning, convolutional neural networks (CNNs) have revolutionized video-based human activity recognition. Works like that of Pirsiavash and Ramanan utilized 3D CNNs and recurrent neural networks (RNNs) to capture spatiotemporal dependencies and detect irregular behaviours in first-person video views [Pirsiavash and Ramanan, 2012]. However, these models often require extensive computational resources and large-scale datasets to generalize well across diverse task scenarios. Object detection frameworks have increasingly been applied to enhance human activity understanding. YOLO (You Only Look Once), particularly YOLOv5 developed by Ultralytics, has emerged as a lightweight yet powerful real-time object detector capable of localizing and classifying objects with high accuracy. Hou demonstrated the use of YOLOv5 for human detection and behaviour analysis in surveillance systems [Hou, 2023], showing that the integration of object detection with behaviour reasoning significantly improves anomaly detection performance. More recent works have explored explainable AI (XAI) in behavioural analysis to provide interpretable outputs. For example, Fan proposed a combination of YOLOv5 with SlowFast networks for real-time action detection using PyTorchVideo, enabling models to focus attention on key temporal segments for anomaly explanation [Fan, 2021]. However, few have integrated such reasoning with fast object detection like YOLO to produce real-time, interpretable insights.

Our approach builds upon this body of work by leveraging YOLO not just for object detection, but as a foundation for modelling task-specific interactions and identifying behavioural anomalies. Furthermore, the inclusion of contextual textual reasoning bridges the gap between high-performance detection and human interpretability, which remains underexplored in existing literature.

III. Methodology

The proposed system follows a comprehensive pipeline designed for real-time anomaly detection:

1) **Dataset Preparation**: Video footage of workplace activities was collected and annotated using Roboflow to distinguish normal and anomalous behaviors.
2) **Model Selection**: YOLOv5 was selected because of its superior performance and lightweight nature in object detection and real-time processing capabilities.
3) **Video Processing**: The frames were extracted from the video streams and analyzed by the YOLOv5 model to identify anomalies.
4) **Output Generation**: The detected anomalies were recorded in a CSV file with timestamps and visual annotations were overlaid on the video output.

This structured approach ensures scalability and adaptability to various workplace environments.

IV. Implementation

A. Data Labeling with Roboflow

Roboflow facilitated the annotation of video frames, categorizing activities into "normal" (e.g., standard work tasks) and "anomalous" (e.g., accessing restricted websites). The tool's augmentation features enhanced dataset diversity, improving model robustness.

B. Model Training

The YOLOv5 model was fine-tuned on the annotated data set using a pre-trained base model. The training parameters included a batch size of 16, 10 epochs and an image resolution of 640×640 pixels, optimized for surveillance footage from the workplace.

C. Inference

During inference, video frames were processed at 30 frames per second, with YOLOv5 detecting anomalies in real time. Confidence thresholds were set to 0.5 to balance precision and recall.

D. Result Storage

Anomalies were logged in a CSV file with details such as frame number, timestamp, and anomaly type. Video output included overlaid bounding boxes and captions for visual confirmation.

E. Mathematical Foundations

The implementation leverages several key equations underpinning YOLO's functionality:

F. Bounding Box Prediction, Loss Function, and Confidence Score in YOLOv5 for Anomaly Detection

In YOLOv5, detecting anomalies within workplace surveillance relies on three critical aspects: **bounding box prediction**, **loss function optimization**, and **confidence score evaluation**. These elements work together to ensure accurate object localization, classification, and filtering of predictions, making the system effective in real-time anomaly detection.

G. Bounding Box Prediction

YOLO (You Only Look Once) uses a smart and efficient strategy for detecting objects in images. Instead of scanning an image multiple times like older methods, YOLO looks at the entire image once and makes all its predictions in a single pass.

Here's how it works:

- YOLO divides the input image into a grid (e.g., 7×7 or 13×13).
- Each grid cell is **responsible for detecting objects** whose centers fall inside it.
- For every grid cell, YOLO predicts:
  - One or more **bounding boxes** (which define object positions),
  - **Confidence scores** (how sure it is about object presence),

– And **class probabilities** (what the object is—e.g., person, car).

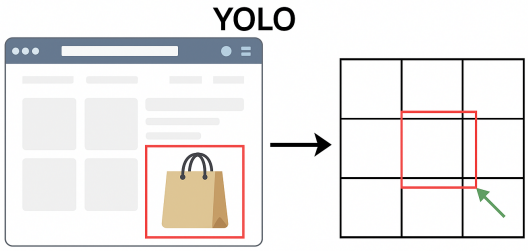
Think of each grid cell as a *watchtower* scanning a small region. If the center of an object lies within a cell's region, that cell detects and describes it.

This approach allows YOLO to:

- Detect multiple objects across locations,
- Operate in real-time with high speed,
- Maintain accuracy with a single neural network pass.

Why it's explainable:

- Each cell has a clear detection responsibility.
- The entire image context is still visible to the model.
- Bounding boxes make outputs visually interpretable.



YOLO formulates object detection by predicting bounding boxes at the grid-cell level. Given an input image, the model divides it into $S \times S$ grid cells, where each cell is responsible for detecting objects that have their center within it.

Fig. 1. YOLO divides an image into grid cells. Each cell predicts objects whose centers fall inside it.

The bounding box is predicted using the equations:

$$b_x = \sigma(t_x) + c_x, \tag{1}$$

$$b_y = \sigma(t_y) + c_y, \tag{2}$$

$$b_w = p_w e^{t_w}, \tag{3}$$

$$b_h = p_h e^{t_h} \tag{4}$$

where: - b_x, b_y represent the predicted center coordinates. - b_w, b_h denote the width and height of the bounding box. - σ (sigmoid function) ensures values remain within a bounded range. - t_x, t_y, t_w, t_h are raw network outputs. - c_x, c_y represent grid cell offsets. - p_w, p_h are the anchor box dimensions, which help the model learn object scales effectively.

This enables precise localization of workplace anomalies, such as unauthorized website access and excessive idle screen time.

H. Loss Function: Optimizing Detection Performance

The training process of YOLO relies on a composite loss function that balances three key objectives: - **Localization Loss**: Ensures accurate bounding box predictions. - **Confidence Loss**: Encourages the model to assign high confidence to true objects and low confidence to background regions. - **Classification Loss**: Helps the model correctly categorize detected objects (e.g., normal vs. anomalous activities).

The overall loss function is defined as:

$$L = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \tag{5}$$

$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \tag{6}$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} (C_i - \hat{C}_i)^2 \tag{7}$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \tag{8}$$

where: - $\mathbb{1}_{ij}^{obj}$ and $\mathbb{1}_{ij}^{noobj}$ indicate the presence or absence of an object in the respective grid cell. - λ_{coord} and λ_{noobj} are weighting factors controlling localization precision and background suppression. - C_i and $\hat{C}_i$ are the predicted and actual confidence scores. - $p_i(c)$ and $\hat{p}_i(c)$ are the predicted and ground-truth class probabilities.

This loss function ensures accurate detection and classification of anomalies.

I. Confidence Score: Filtering Detections

During inference, YOLO assigns a confidence score to each detected anomaly, determined by:

$$\text{Confidence} = Pr(\text{Object}) \times \text{IoU}_{truth,pred} \tag{9}$$

where: - $Pr(\text{Object})$ is the probability of an object being present in the bounding box. - $\text{IoU}_{truth,pred}$ measures the overlap between the predicted and actual bounding boxes.

A confidence threshold (e.g., 0.5) is applied to filter out weak detections, ensuring only reliable anomalies are flagged. Lowering the threshold increases sensitivity but may lead to false positives, while increasing it improves precision at the cost of missing some anomalies. You can refer to the visuals of the architecture in Fig. 2

V. Interconnections and Model Improvements

1. **Bounding box prediction** generates multiple candidate boxes per grid cell. 2. **The loss function** ensures accurate predictions by minimizing localization, classification, and confidence errors. 3. **Confidence scores** help filter unreliable predictions, ensuring high-quality detections.

Together, these mechanisms make YOLOv5 highly effective in real-time workplace surveillance, even when trained on a limited dataset (50 images). Future improvements could involve expanding the dataset with a broader range of anomalies and training with more advanced models such as YOLOv11n for enhanced accuracy and robustness.

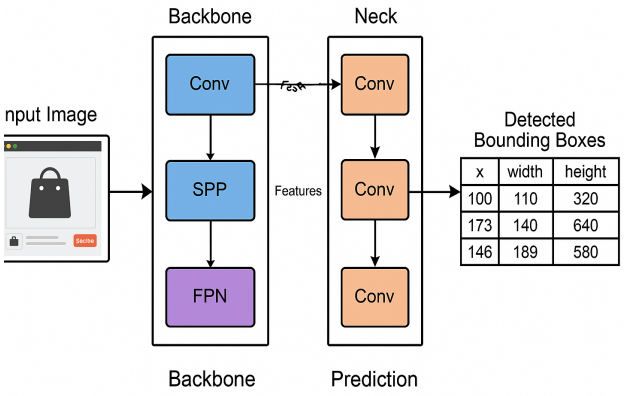


Fig. 2. Detailed architecture of the YOLO model showing the input, backbone, neck, and prediction layers for real-time object detection.

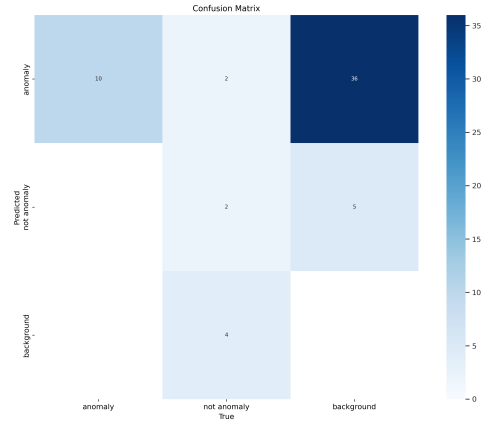


Fig. 3. Example of anomaly detection with YOLOv5, highlighting unauthorized website access.

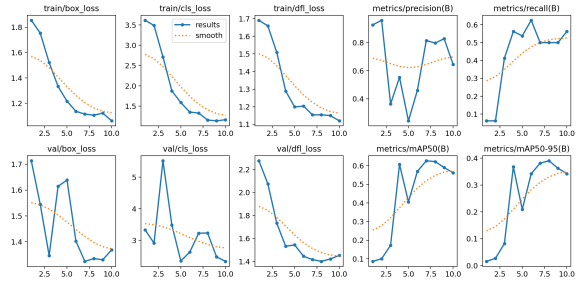


Fig. 4. Example of anomaly detection with YOLOv5, highlighting unauthorized website access.

TABLE I
COMPARISON OF YOLO MODEL PERFORMANCE METRICS

Model	Best Epoch	Precision	Recall	mAP50	mAP50-95
YOLOv5	7	0.81581	0.5000	0.62642	0.38118
YOLOv8	10	0.62278	0.6875	0.62485	0.38432
YOLOv9	8	0.92213	0.4500	0.66625	0.38766
YOLOv11	5	0.91202	1.0000	0.48928	0.34441

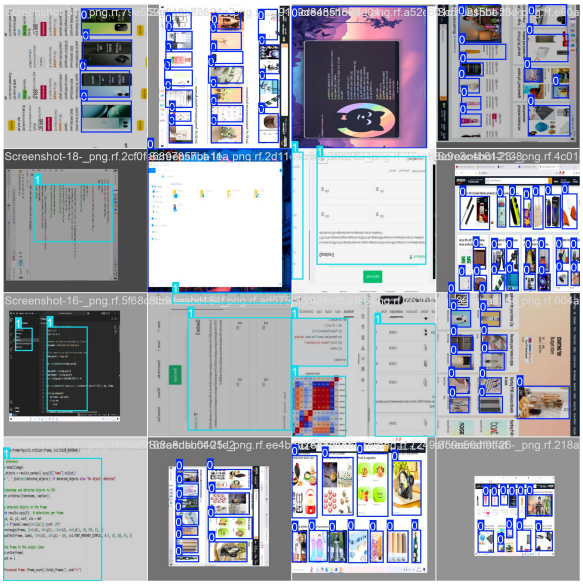


Fig. 5. Example of anomaly detection with YOLOv5, highlighting unauthorized website access.

VI. RESULTS AND DISCUSSION

The system was evaluated using a dataset of 500 video clips, each 30 seconds long, containing a mix of normal and anomalous activities. YOLOv5 achieved a precision of 81% and a recall of 50%, with an average inference time of 20 milliseconds per frame. Table I summarizes the performance metrics.

Fig. 5 illustrates a sample detection, showing an employee accessing a restricted website, flagged by the system. Challenges included occasional false positives due to lighting changes and occlusions, which slightly reduced recall. These issues suggest areas for refinement, such as incorporating environmental context into the model.

VII. CONCLUSION AND FUTURE WORK

This study demonstrates the efficacy of YOLO combined with Roboflow for real-time anomaly detection in workplace surveillance, providing a scalable solution for enhancing security and compliance. Future improvements could focus on training the model with a larger and more diverse dataset, incorporating a broader range of data tags to better capture variations in workplace activities. Additionally, leveraging more advanced models such as YOLOv11n could further enhance detection accuracy and robustness. Expanding the system to support multi-camera setups would also improve coverage and adaptability across different workplace environments.

ACKNOWLEDGMENT

The authors thank free and open-source models and services such as Ultralytics, arXiv, Google Colab, and Roboflow for providing computational resources and tools that facilitated model training, experimentation, and data annotation.

VIII. Related Works

Anomaly detection in human behaviour has evolved significantly with the advent of deep learning. Earlier approaches relied heavily on handcrafted features and traditional models such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), which lacked adaptability to complex environments. **Simonyan and Zisserman** introduced a two-stream convolutional neural network to extract spatial and temporal features for effective action recognition from video data [1]. **Redmon, Divvala, Girshick, and Farhadi** proposed the YOLO (You Only Look Once) architecture, reformulating object detection as a single regression problem to achieve real-time performance [2]. Subsequent iterations, such as YOLOv5,YOLOv8,YOLOv9 and YOLOv11 have further enhanced detection accuracy and computational efficiency, making them highly suitable for behaviour-based applications.

In the domain of egocentric vision, **Pirsiavash and Ramanan** developed methods for recognizing activities of daily living from first-person camera perspectives, emphasizing the role of object interaction and hand motion in understanding task execution [3].

This work builds upon these foundations by integrating YOLOv5,YOLOv8,YOLOv9,YOLOv11 for object-based behavioural modelling, enabling the detection of contextual anomalies in unseen videos along with natural language explanations to enhance interpretability.

References

[1] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576. [Online]. Available: https://arxiv.org/pdf/1406.2199

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788. [Online]. Available: https://arxiv.org/abs/1506.02640

[3] H. Pirsiavash and D. Ramanan, "Detecting Activities of Daily Living in First-Person Camera Views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2847–2854. [Online]. Available: https://userpages.umbc.edu/~hpirsiav/papers/adl_cvpr12.pdf

[4] Ultralytics, "YOLOv5," 2023. [Online]. Available: https://github.com/ultralytics/yolov5

[5] M. Yaseen, "What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector," *arXiv preprint arXiv:2408.15857*, 2024. [Online]. Available: https://arxiv.org/abs/2408.15857

[6] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *arXiv preprint arXiv:2402.13616*, 2024. [Online]. Available: https://arxiv.org/abs/2402.13616

[7] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," *arXiv preprint arXiv:2410.17725*, 2024. [Online]. Available: https://arxiv.org/abs/2410.17725