Safety Management and Human Factors, Vol. 151, 2024, 28–38
https://doi.org/10.54941/ahfe1005299

AI-Enhanced Ergonomics: Revolutionizing Industrial Safety Through Real-Time Posture Analysis and PPE Detection

Rafael Luque, José Ramón Vilanova, Gonzalo Díaz, and Eduardo Ferrera

Advanced Center for Aerospace Technologies (CATEC), C/ Wilbur y Orville Wright 19,
La Rinconada, 41309, Seville, Spain

ABSTRACT

Despite the continuous advancements in technology and safety regulations, professional accidents in the industry remain a persistent challenge. In the Industry 5.0 era, Artificial Intelligence and cutting-edge Computer Vision techniques are expected to have a transformative impact on industrial environments. In this context, Deep Learning applications can exhibit significant potential in both the primary detection of safety issues and the quick reaction in accidental scenarios. The system proposed in this publication uses tracking algorithms to identify human safety vulnerabilities and early detect people falling or requesting help. Specifically, the first component employs a Transfer Learning technique with YOLOv7 to efficiently determine and detect whether human Personal Protective Equipment is worn correctly. Additionally, the system utilizes YOLOv7 key points detection model to assess the human posture in real time, allowing machines to detect people falling or requesting help. The work concludes presenting experiments that scrutinize the algorithm's detection performance, under varied positions, evaluating the impact of GPUs and cameras and operator's distance to camera in a pilot aerospace experimental facility.

Keywords: Ergonomics, Deep learning, Computer vision, Aerospace, Safety management

INTRODUCTION

In the dynamic landscape of industrial operations, human workers are commonly exposed to an extraordinary number of risks, especially in the context of workplace injuries. Ergonomics, the science of designing environments and systems to optimize human well-being and performance, plays a pivotal role in mitigating workplace hazards and fostering optimal working conditions. The failure to adequately address ergonomic and safety become a burden on workers, resulting in injuries that compromise both the physical and the mindset of the operators.

Addressing these health concerns in industrial settings significantly influences the physical well-being and overall productivity of workers. Work-related injuries, whether caused by stress and strain or workplace accidents, can lead to chronic pain, limited mobility, and a decline in overall quality of life.

Tragically, some workplace accidents can lead to further consequences in sectors where the correct ergonomics or appropriate Personal Protective Equipment (PPE) uses could have mitigated risks and safeguard operators. Understanding the human cost of these challenges is crucial for motivating effective interventions. Moreover, industrial environments, characterized by elevated noise levels and operators' independent workstations, detecting falls or help calls represents a challenge, particularly in scenarios where rapid assistance is needed.

The integration of Artificial Intelligence (AI) and Deep Learning (DL) technologies to address these problems has the potential to mean a pivotal advancement in reshaping safety management within factories. The use of these cutting-edge technologies offers a transformative approach to identify and proactively address safety concerns, leading to a new era where prevention and reaction are equally addressed. Concretely, Computer Vision ease and prevent injuries through the implementation of real-time algorithms, which can be integrated in factories' security cameras. The possibilities remain outrageous, thanks to the implementation of tracking systems that can trace operators in real time, evaluating whether the posture is ergonomic or PPE are correctly worn.

This article provides an intersection of both AI and ergonomics by implementing an Artificial Vision system capable of real-time analysis of workers' gestures and the effective utilization of PPEs. By tackling these industrial challenges head-on, the objective is to pave the way for a safer, healthier, and more efficient working environment, thereby improving risk management and speeding up intervention in the event of injury.

More precisely, the contributions of this publication are:

- A brief review of AI-based ergonomics and how these problems are been addressed nowadays in the industry.
- A description of the Computer Vision system developed, aimed at detecting operator falls, gestures and PPE tracking in real-time.
- An experimentation of this system under different circumstances, regarding different GPUs, cameras and scenarios.
- A discussion of the results obtained including a roadmap of the next steps and opportunities of the solution.

The structure of the paper is the following: Section 2 provides a literature review on systems addressing operator posture and PPE object detection; Section 3 details the Computer Vision-based solution, encompassing its architecture; Section 4 describes the experiments conducted within a pilot aerospace experimental facility, presenting the results obtained; finally, Section 5 concludes the paper and outlines a roadmap for the proposed solution.

LITERATURE REVIEW: HUMAN POSTURE AND PPE DETECTION

The proposed solution relies on the integration of two distinct types of Computer Vision techniques to accurately identify both PPE and human gestures: object detection and pose estimation algorithms.

In the realm of object detection models, significant advancements have been made in recent years, with noteworthy contributions from popular models such as SSD, R-CNN, SqueezeDet, MobileNet, and YOLO (Liu et al., 2020). These models are typically categorized into two-shot and one-shot detectors.

R-CNN serves as the banner of two-stage detection algorithms, employing a pioneering region-based approach. Nonetheless, it exhibits limited responsiveness for real-time predictions, even with the integration of the Fast R-CNN method.

The one-shot detectors are preferable for online applications due to their improved time of inference. SSD excels in real-time detection by predicting multiple classes using a single Deep Neural Network. On the other hand, SqueezeNet demonstrates lower performance compared to the other models mentioned, while MobileNet, operating as a single-shot multi-box detector, utilizes the Caffe Framework. YOLO ("You Only Look Once") family of algorithms stands as the more representative of the one-shot category, thanks to newer versions of YOLO, such as YOLOv7 and YOLOv8, that provide improved speed and accuracy for real-time object detection tasks (Cheng 2020).

The detection of Personal Protective Equipment through computer vision algorithms has transcended in several sectors. The PPE became popular during COVID-19 pandemic, where face mask was a must, enhancing the use of vision algorithms for its detection (Yunus, 2023). Other industries, such as construction sites have integrated this type of solutions due to the high risks of the sectors, by using CNNs (Nath et al., 2020); and YOLO, in its different versions, for detecting human and vests (Ferdous and Ahsan 2022) (Delhi et al., 2020). Similar approaches have been implemented with YOLOv5 in laboratories, addressing security issues for students (Ali et al., 2022). In the industrial sector, similar solutions have emerged, like the one presented by (Shi et al., 2023), using a customized version of YOLOv8 for the detection of vests, helmets and masks.

On the other hand, pose estimation algorithms cater and solve diverse computer vision needs that are helpful for gesture detection. Some of the most popular modules are OpenPose, PoseNet and YOLO Pose. OpenPose excels in real-time, multi-person pose estimation with high accuracy for the keypoint detection. PoseNet is a lightweight architecture for browsers or mobile devices. Finally, YOLOv7 has recently optimized pose models for real-time pose keypoint detection, showcasing versatility in various applications.

Pose estimation has arisen essential for signal recognition, which is the second cornerstone of this publication, so that human gestures can be recognized. The detection of falls has normally been solved by classical Computer Vision algorithms like the one implemented by (Luo, 2023). However, the rising of this pose estimation system has led to better performance of gesture detection, both for one and multiple person purposes (Debapriya Maji et al., 2022). One of the main uses of the pose estimation system is sign language recognition, thanks to the body pattern recognition (Zhou et al., 2023). These solutions have used both hand pose estimation, like (Damdoo and Gupta, 2022), or human body estimation, including the use of GCN

(Graph Convolutional Networks) (Dafnis et al., 2022), transformers (Woods and Rana, 2023) or Local Contrastive Learning (Hua et al., 2023).

CATEC's solution combines two components. The first component is set to retrain YOLOv7 Object Detection, known for its proven combination of speed and accuracy, to enhance the detection of PPE. The second component employs YOLOv7 Pose Estimation system for detecting human signals within an industrial environment. This approach will be furtherly presented in the methodology of this proceeding.

METHODOLOGY

This section introduces an AI-enhanced Ergonomics system, a sophisticated Computer Vision solution designed to assess the correct utilization of PPE and perform real-time detection and tracking of human posture. This capability facilitates early identification of individuals experiencing falls or requiring assistance.

The system has been crafted to seamlessly operate in dynamic industrial environments, adept at adapting to background changes and ensuring safety under diverse conditions. To achieve this, it employs cutting-edge tracking algorithms that seamlessly integrate classical object tracking with key points tracking systems. This innovative solution serves as a valuable enhancement to existing safety measures within industrial settings, complementing and augmenting current alarm systems.

The presented system incorporates state-of-the-art DL models, capturing and analysing video images in real-time, providing crucial information on PPE detection, human falls, and call for help signals through security cameras. It is constructed using a combination of two distinct versions of the YOLOv7 algorithm: the first employing YOLO object detection for identifying various PPE items, such as safety vests and helmets, and the second utilizing YOLOv7 Pose for detecting key points and tracking human posture during videos.

The development process involves initiating from an open-source YOLO object detection pretrained model, followed by retraining to obtain new weights, crucial for tracking security vests and helmets on operators. This retraining necessitates the construction of a dataset for both PPE, ensuring accurate pose detection, and obtaining new weights for reliable tracking of these elements. Additionally, the system issues alerts when operators are not wearing the required equipment, mitigating potential safety risks. Specifically, the YOLOv7 model for PPE detection has been trained with a dataset comprising 3231 images.

On the other hand, a pretrained model is employed to recognize postures within industrial facilities, focusing on operator falls and "call for help" gestures. The YOLOv7 Pose model used is publicly available[1] and trained using the COCO-Pose Dataset, renowned for its reliability, supporting 17 key points for human figures and boasting over 200,000 labelled images for pose estimation tasks.

[1] https://github.com/WongKinYiu/yolov7?tab=readme-ov-file#pose-estimation

The approach of this research is concurrently executing both YOLO models in real-time, generating a comprehensive output comprising seven different bounding box-shaped predictions (see Figure 1): i) wearing helmet; ii) not wearing a helmet; iii) wearing reflective; iv) not wearing reflective; v) call for help gesture; vi) fall; and vii) stand. Notably, the key point model information is instrumental in deriving the remaining three classes related to safety operator posture.

An important detail to highlight is the system's offline deployment and testing on local hardware, ensuring independence from potential internet connection issues, thus safeguarding privacy and confidentiality. Finally, this innovative solution represents one of the outcomes of the 5R Cervera Network, a collaborative project involving five Spanish Research Centers, situated within the Industry 5.0 and Robotics paradigm. CATEC, an integral partner in this endeavour, spearheaded the development of this technology within the framework of a Pilot Aerospace Factory, primarily devoted to interactive monitoring and assistance for manual processes.

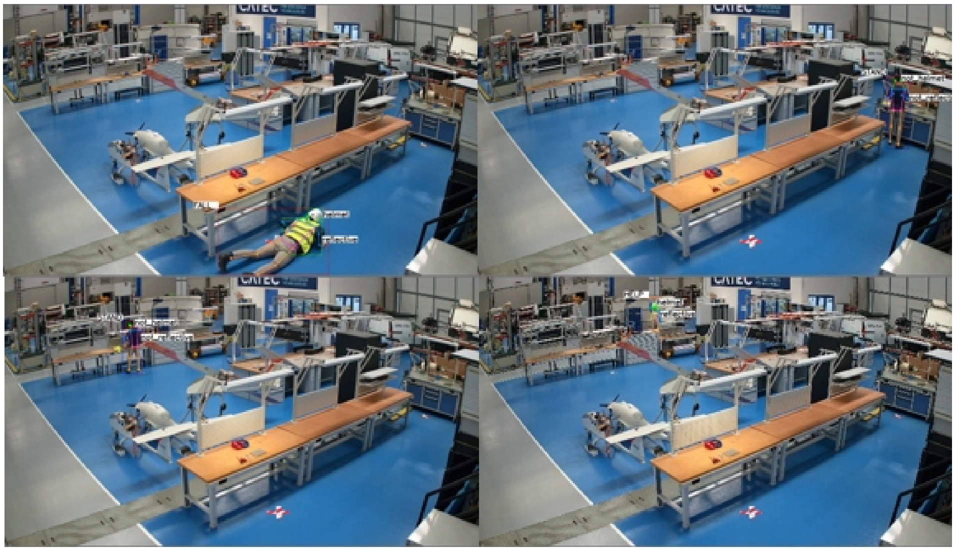

Figure 1: Experiments conducted within CATEC's pilot aerospace facility. The images showcase different model prediction outputs applied to various operator positions.

EXPERIMENTS

This section outlines the experiments undertaken to assess the system's performance across various scenarios, employing diverse cameras and GPUs. The objective is to thoroughly evaluate the model's effectiveness in terms of both inference time and accuracy.

Experimental Setup

The experiments have been conducted in the Pilot Aerospace Factory of CATEC as a part of the Project 5R Cervera Network. The CATEC facilities serve as a pertinent testing ground for evaluating the system's capabilities. The presence of advanced manufacturing equipment, such as a robotic cell and other manufacturing machinery, presented challenges that significantly impacted the model's performance.

Figure 2 illustrates the scenario employed to assess the system, comprising a blend of workstations and industrial corridors. Specifically, three distinct workstations were defined for use in this experimental set, supplemented by two additional positions designed to evaluate the system's performance under more intricate conditions:

- Position 1 (5 meters): It serves as an ideal proximity for the system's preliminary testing in a first iteration.
- Position 2 (10 meters): It exhibits a less favourable detection angle but is valuable for assessing the system's performance in complex environments.
- Position 3 (14 meters): It provides a superior angle compared to the second, though being more distant.
- Position 4 (21 meters): It presents a medium level of background noise.
- Position 5 (26 meters): Combines great distance with a high degree of background noise.

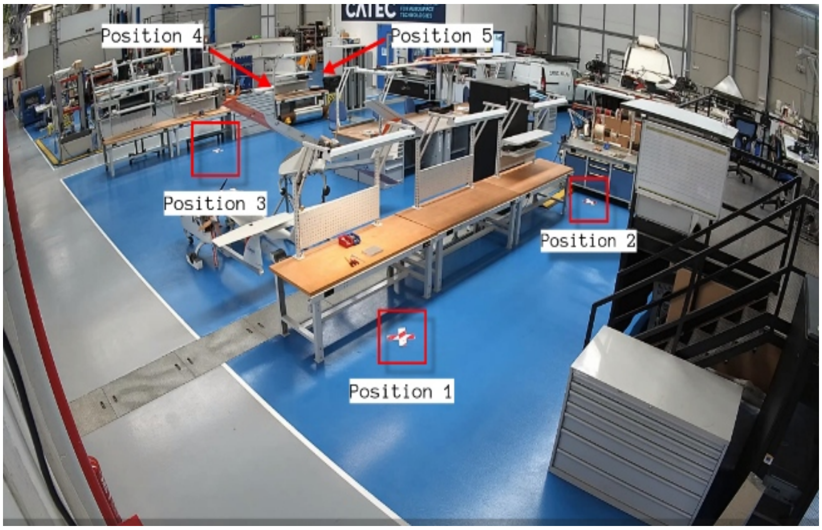

Figure 2: Working positions employed for testing the AI-enhanced ergonomics system.

In parallel to the different scenarios, the system's performance has also been evaluated considering two hardware factors: image quality due to cameras resolution and GPUs capabilities for inference speed. To comprehensively assess the impact of these devices on the model, a series of experiments has been carried out, examining the model across three distinct cameras models and GPUs, whose characteristics can be seen in Table 1 and Table 2, respectively.

Table 1. Cameras used during the experiments.

Camera	Price	Resolution (px)	Frames/Second
Logitech C920	Low	1280x720	10 fps
HIKVISION DS-2CD2T87G2-L	High	1920x1080	25 fps
ANNKE C500	Medium	2304x1296	25 fps

Table 2. GPUs used during the experiments.

GPU	Price	vRAM	GPU Clock Speed
Nvidia RTX A1000 Laptop	Low	4 GB	630 MHz
Nvidia RTX A3000 Laptop	Medium	12 GB	855 MHz
Nvidia Geforce RTX 4090	High	24 GB	2235 MHz

Experiment 1: Inference Time and Video Memory Allocated

The objective of this experiment was to gauge the influence of camera resolution and GPU attributes on inference speed and the associated video memory allocation. To achieve this objective, the study focused on evaluating the following two features:

- Video Memory allocated: This refers to the memory required to store the models while they are in operation.
- Inference Time: This is the duration that elapses between the execution of the code and the evaluation of the images.

Given that two models operate simultaneously, each with its distinct weights and structures, it is crucial to assess memory usage. This analysis is essential for comprehending the impact of GPU memory on computer vision systems. The results obtained for both metrics are presented in Figure 3.

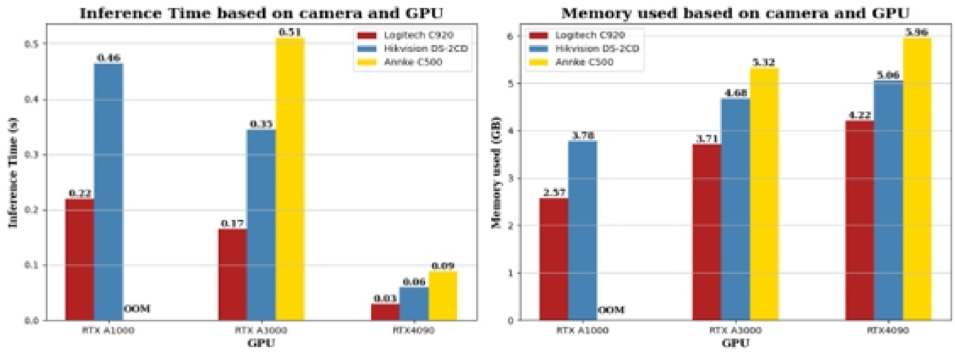

Figure 3: Experiment 1. Video memory allocated and inference time depending on camera resolution and GPU model (OOM means out of memory).

Experiment 2: Impact of Distance

This experiment is designed to demonstrate the model's efficacy across diverse detection scenarios and varying camera conditions, details of which are expounded upon in the preceding table. The primary objective is to discern seven classes: *helmet, reflective, not helmet, not reflective* (as dictated by the object detection model), and *help signal, stand,* and *fall* (attributed to the pose estimation model).

To ensure uniform conditions, the experiment has entailed video recording synchronization using the three different cameras. Manual ground truth labels have been intricately assigned to specific frames within each video (350 images in total), enabling a meticulous comparison with model predictions to derive performance metrics.

To assess the model's effectiveness, both Average Precision (AP) and mean Average Precision (mAP) metrics from the Pascal Method has been calculated using the methodology proposed by (Padilla et al., 2021). Average Precision, derived from the area under a pre-processed precision-recall curve, effectively captures the precision-recall trade-off and reflects the impact of confidence levels associated with predicted bounding boxes. Values are obtained for each class, with mAP representing the mean across all classes.

It is pertinent to note that prediction determination has supposed applying a confidence threshold for detection set at 0.5 and an Intersection over Union (IoU) threshold set at 0.65. These thresholds were instrumental in filtering the model predictions, discarding labels that did not meet the specified criteria.

Although specific frames from recorded videos were used for metric calculations, the comprehensive output predictions for the entire videos are accessible through the provided video links[2,3,4].

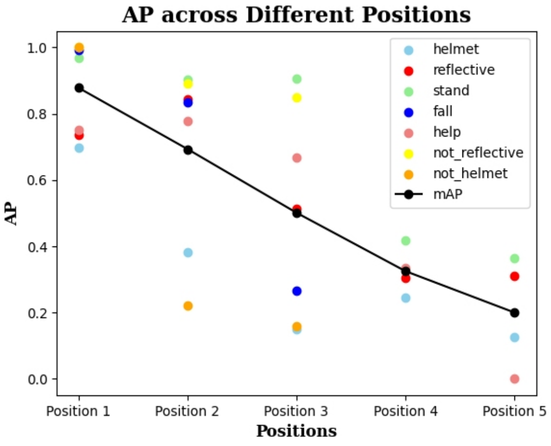

Figure 4: Experiment 2. AP for each class and mAP across different positions.

[2] https://www.youtube.com/watch?v=n232bVlP0cE

[3] https://www.youtube.com/watch?v=aiESqMOP4kk

[4] https://www.youtube.com/watch?v=9LTQGJgCNY0

RESULTS

Results from Experiment 1, illustrated in Figure 3, reveal significant performance disparities with more economical GPUs, notably influenced by camera resolution.

The evaluation of models underscores the computational complexity and memory demands. For instance, it is demonstrated that the NVIDIA RTX A1000 struggled to execute the model effectively for ANNKE camera due to insufficient memory capacity. Consequently, assessing this GPU's performance for this camera becomes unfeasible. Apart from that, the model runs a maximum of 5.96 GB, which may be assumable for a good number of GPUs.

Furthermore, the inference time is directly correlated with the pixels for analysis, presenting challenges that need prompt interventions. Emphasizing the criticality of detection speed, pricier GPUs, such as the NVIDIA GeForce RTX 4090, exhibit superior model performance, achieving rapid gesture recognition in less than 100 ms—particularly crucial during fall incidents.

In experiment 2 results', illustrated in Figure 4, the model offers a clear correlation between distance of the positions and cameras. The system is completely reliable at short distances, with a 0.87 mAP in the first position, and showing strength at recognizing all types of classes. Yet, the system offers acceptable results in intermediate distances, at 14 meters, with mAP = 0.5; while for longer distances, between 20 and 26 meters, the system's mAP values drop to 0.3249 and 0.2001 respectively.

Moreover, notable distinctions exist among different classes, such as *stand*, *fall*, and *not reflective*, consistently exhibiting high precision and recall across multiple positions. Contrary, classes like *helmet* and *reflective* show a great sensitivity to distance, struggling to maintain consistent performance for objects of these classes at positions 4 and 5. The observed trends emphasize the importance of optimizing the system for improved performance at greater distances, potentially through model fine-tuning or incorporating additional techniques for long-range detection.

In conclusion, understanding the distance-dependent variations in the system's performance is crucial for refining and enhancing its capabilities, especially in industrial environments where the deployment distances may vary. In consequence, it should be considered optimization strategies for improved object detection at greater distances, like: (i) evaluate the impact of environmental factors (e.g., lighting conditions) on model performance; (ii) consider the potential of incorporating additional sensors or adjusting model parameters to enhance performance in distant scenarios.

CONCLUSION

This publication navigates the intersection of AI and Ergonomics, presenting a Computer Vision system designed to enhance safety management in industrial environments. By addressing challenges related to PPE utilization, falls, and signals for help, the system offers a comprehensive solution for creating a safer, healthier, and more efficient working environment.

The key contributions of this publication include a concise review of AI-based ergonomics applications in industry, the detailed description of the developed Computer Vision system, experimentation with various cameras and GPUs under different conditions, and a discussion of the obtained results.

The results offer valuable insights into the system, thanks to the importance of selecting correct hardware devices, for both cameras and GPUs. Moreover, the system's dependency on the distance to camera is huge, decreasing its performance when operators are further.

The process of retraining YOLO models implies difficulties, due to a need for a high-quality dataset for better algorithms' performance in an industrial environment. Different backgrounds and similar colours are challenges to solve and need to be addressed when constructing the set of images, so techniques like data augmentation must be included. On the other hand, YOLOv7 pose detection algorithm's challenges are its configuration with a fixed set of 17 key points. Unfortunately, this setting lacks customization options, preventing users from adjusting the number of key points to better suit their specific needs. This rigidity poses a limitation, as having the flexibility to increase or decrease the number of key points could enhance the algorithm's utility. In case of the variation of the number of key points, only one model could be a solution for this system's detection of both PPE and human gesture; yet, for now, detecting a helmet or a vest with 17 key points is complicated and inefficient.

This investigation has offered important insights regarding future research lines that will enhance the performance of the solution, such as:

1. System fine-tuning to enhance performance at greater distances.
2. Exploring the inclusion of the model within different cameras point of view, integrating it in factory's security alarms.
3. Addressing the limitations of pose estimation system, customizing the number of key points detected.

In summary, the developed AI-enhanced Ergonomics system presents a promising avenue for improving safety and efficiency in industrial settings. Future endeavours should focus on refining and optimizing the system to address specific challenges observed in real-world environments.

ACKNOWLEDGMENT

This study has been funded by the project "5R, Cervera Network in robotic technologies for intelligent manufacturing", contract number CER-20211007 under "Centros Tecnológicos de Excelencia Cervera"; funded by "The Centre for the Development of Industrial Technology (CDTI)".

REFERENCES

Ali, L. et al. (2022). 'Development of YOLOv5-Based Real-Time Smart Monitoring System for Increasing Lab Safety Awareness in Educational Institutions'. Sensors, 22(22).

Cheng, R. (2020). 'A survey: Comparison between Convolutional Neural Network and YOLO in image identification'. Journal of Physics: Conference Series, 1453, p. 12139.

Dafnis, K. M. et al. (2022). ‘Bidirectional Skeleton-Based Isolated Sign Recognition using Graph Convolutional Networks’. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7328–7338, Marseille, France. European Language Resources Association.

Damdoo, R. and Gupta, A. (2022). ‘Gesture controlled interaction using hand pose model’. International journal of health sciences, pp. 10417–10427.

Debapriya Maji et al. (2022). ‘YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss’. CVPR Workshops 2022: 2636–2645.

Delhi, V. S. K., Sankarlal, R. and Thomas, A. (2020). ‘Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques’. Frontiers in Built Environment, 6.

Egi, Yunus. (2023). ‘YOLO V7 and Computer Vision-Based Mask-Wearing Warning System for Congested Public Areas’. Journal of the Institute of Science and Technology. 13. 22–32.

Ferdous, M. and Ahsan, S. M. M. (2022). ‘PPE detector: A YOLO-based architecture to detect personal protective equipment (PPE) for construction sites’. PeerJ Computer Science, 8.

Hua, Y. et al. (2023). ‘Part Aware Contrastive Learning for Self-Supervised Action Recognition’. International Joint Conference on Artificial Intelligence.

Liu, L. et al. (2020). ‘Deep Learning for Generic Object Detection: A Survey’. International Journal of Computer Vision, 128(2), pp. 261–318.

Luo, B. (2023). ‘Human Fall Detection for Smart Home Caring using Yolo Networks’. International Journal of Advanced Computer Science and Applications, 14(4).

Nath, Nipun & Behzadan, Amir & Paal, Stephanie. (2020). ‘Deep learning for site safety: Real-time detection of personal protective equipment’. Automation in Construction. 112. 103085.

Padilla, R. et al. (2021). ‘A comparative analysis of object detection metrics with a companion open-source toolkit’. Electronics (Switzerland), 10(3), pp. 1–28.

Shi, C. et al. (2023). ‘GBSG-YOLOv8n: A Model for Enhanced Personal Protective Equipment Detection in Industrial Environments’. Electronics (Switzerland), 12(22).

Woods, L. T. and Rana, Z. A. (2023). ‘Constraints on Optimising Encoder-Only Transformers for Modelling Sign Language with Human Pose Estimation Keypoint Data’. Journal of Imaging, 9(11).

Zhou, L. et al. (2023). ‘Human Pose-based Estimation, Tracking and Action Recognition with Deep Learning: A Survey’.