An Approach to Resource-Efficient Optimization for Real-time Computer Vision-based OHS Monitoring on Resource-constrained Industrial Objects

Konstantin Kurochka
Information Systems and Technologies
Gomel State Technical University named after P. Sukhoj
Gomel, Belarus
kurochka@gstu.by

Konstantin Panarin
Department of Digitalization
Republican Unitary Enterprise «Production Association Belorusneft»
Gomel, Belarus
logran2@gmail.com

Daniil Karpenko
Information Systems and Technologies
Gomel State Technical University named after P. Sukhoj
Gomel, Belarus
protankist343@gmail.com

Abstract. **This paper proposes and investigates a methodology for the resource-efficient optimization of real-time AI-based computer vision technologies for occupational health and safety (OHS) monitoring tasks on industrial objects, specifically targeting deployment in resource-constrained systems. The core of the methodology enables the efficient use of CPU-based server devices by introducing two key improvements: a two-stage detection mechanism with dynamic Region of Interest (ROI) selection to significantly reduce the computational load on the primary object detection model, and a static background subtraction algorithm for pre-filtering the video stream and removing non-informative scene areas. A detailed analysis of the performance (in terms of processing speed for real-time capability) and accuracy of various configurations implementing this approach, based on models from the YOLO family, is conducted. A conceptual scheme for integrating the proposed optimization techniques into a typical video analytics pipeline is described, along with a methodology for creating and annotating a specialized dataset. The paper demonstrates that the developed methodology achieves an acceptable balance between violation detection accuracy and video stream processing speed necessary for real-time operation on CPUs, opening prospects for the wider adoption of intelligent safety systems in environments with limited computational resources.**

Keywords: **occupational safety, health and safety, personal protective equipment (PPE), computer vision, YOLO, object detection, neural network optimization, edge computing, industrial safety, background subtraction, focus of attention, optimization methodology, video analytics.**

I. INTRODUCTION

Ensuring a high level of occupational health and safety (OHS) is a critical task for all industries, especially for sectors with high production risks, such as oil and gas enterprises. Statistics on occupational injuries in these sectors, despite the measures taken, remain high, which requires the search for new, more effective approaches to control and prevent incidents [1, 2]. Non-compliance with established norms and rules for the use of personal protective equipment (PPE), as well as violation of safety procedures, are among the main causes of occupational injuries and emergency situations. Traditional control approaches, based on inspections and human observation, have significant drawbacks, including subjectivity, high cost, limited coverage, and the inability to ensure continuous real-time monitoring, which is crucial for timely intervention.

The development of computer vision and deep learning technologies provides opportunities for creating automated OHS monitoring tools. Modern neural network architectures, such as YOLO (You Only Look Once) [5], SSD (Single Shot MultiBox Detector), and Faster R-CNN [8], show high results in object detection tasks in images and videos. The effectiveness of using computer vision for industrial safety monitoring is confirmed by a number of studies. However, their widespread practical application for real-time monitoring at remote or mobile industrial objects often faces the problem of high demands on computational resources. Providing broadband internet for video stream transmission to a centralized cloud service is not always technically feasible, and the use of high-performance servers with GPUs at each facility is frequently not economically viable, necessitating resource-efficient solutions that can operate on existing or limited hardware.

The authors of this paper propose a methodology for the resource-efficient optimization of real-time computer vision-based OHS monitoring, adapted for effective operation on devices with limited computational capabilities, primarily CPUs, commonly found in resource-constrained systems at industrial sites. A key improvement proposed in the article is the integration of algorithmic components into a common processing pipeline designed to significantly reduce

computational load and enhance processing speed: a two-stage "focus of attention" mechanism and static background subtraction. This approach offers an effective, resource-efficient, and affordable way to enhance industrial safety, reduce injury risks, and improve the overall safety culture at enterprises through wider adoption of intelligent real-time monitoring technologies.

II. Existing approaches to object detection and methods for their optimization for CPU

A. Object Detector Architectures

The task of automatic object detection in images and video is one of the fundamental tasks in computer vision. Over the past decades, a number of successful approaches have been proposed, among which methods based on deep convolutional neural networks (CNNs) have taken a dominant position.

Modern CNN-based object detectors can be broadly divided into two main families: one-stage and two-stage detectors.

Two-stage detectors: A prominent representative of this family is the R-CNN (Regions with CNN features) architecture and its more advanced versions, such as Fast R-CNN and Faster R-CNN. Their operation is based on two main stages. In the first stage, a set of potential regions (region proposals) where objects might be located is generated. For this, Faster R-CNN uses a special Region Proposal Network (RPN) module. In the second stage, each of these regions is analyzed by a classifier to determine the object class and refine its boundaries (bounding box regression). Two-stage detectors generally demonstrate high detection accuracy, especially for small objects and in scenes with a large number of overlapping objects. However, their sequential architecture leads to higher computational costs and, consequently, lower processing speed compared to one-stage counterparts.

One-stage detectors: This family includes popular architectures such as YOLO (You Only Look Once) [10] and SSD (Single Shot MultiBox Detector) [5]. Unlike two-stage detectors, they predict object classes and coordinates in a single pass of the neural network through the image. The image is divided into a grid of cells, and for each cell, bounding boxes, objectness score, and class probabilities are predicted.

YOLO: The YOLO family has undergone significant evolution: from YOLOv1 to more modern versions such as YOLOv3, YOLOv4 [9], YOLOv5, and the newest YOLOv7, YOLOv8 [3]. Each new version brought improvements in architecture (e.g., using more efficient backbone networks like DarkNet, CSPDarkNet, or without them in recent versions), feature aggregation mechanisms (Feature Pyramid Network FPN [4], Path Aggregation Network PANet [4]), loss functions, and data augmentation techniques. YOLO models are known for their high processing speed, making them ideal for real-time tasks. However, earlier versions could be inferior to two-stage detectors in the accuracy of detecting small objects.

SSD: The SSD architecture uses a set of anchor boxes of various scales and aspect ratios on multiple feature maps of different resolutions. This allows for effective detection of objects of different sizes. SSD is also a fast detector, but its accuracy can depend on the correct selection of anchor box configurations for a specific task.

In the context of PPE monitoring on CPUs, one-stage detectors, especially modern YOLO versions, are of greatest interest due to their optimal balance of speed and accuracy, making them suitable candidates for resource-efficient real-time applications.

B. Methods for Optimizing Neural Network Models for CPU

For efficient operation of object detectors on CPUs with minimal compromise to accuracy, various optimization strategies are applied, such as quantization, knowledge distillation, and algorithmic optimizations.

Quantization: This method involves reducing the bit precision of the model's weights and/or activations [6]. Instead of 32-bit floating-point numbers (FP32), 16-bit or 8-bit integers are used. This leads to a reduction in model size, decreased memory bandwidth requirements, and accelerated computations.

Knowledge Distillation: This approach involves training a more compact "student" model using knowledge obtained from a larger and more accurate "teacher" model [7]. The "student" is trained not only on true labels but also on the "soft" predictions (class probabilities) of the "teacher". This allows the transfer of the generalization ability of the large model to the small one, often achieving better quality than when training the small model from scratch only on true labels.

Algorithmic Optimizations:

- Optimization of the data processing pipeline itself: minimizing data copying, efficient cache utilization, parallelization of independent operations.
- Use of specialized libraries and frameworks for CPU inference, such as ONNX Runtime. These tools provide optimized implementations of operations for various processor architectures,

support quantized models, and can automatically apply various graph optimizations.

The choice of a specific method or combination of optimization methods depends on the specifics of the task, requirements for accuracy and performance, as well as available tools and hardware platform. For the task of PPE monitoring at drilling sites, where real-time operation on a CPU is required, a combination of lightweight architectures, quantization, and algorithmic optimizations, such as Region of Interest Cropping and background subtraction, appears to be the most promising for achieving resource efficiency.

III. Proposed Optimization Methodology for OHS Monitoring

The proposed methodology aims to create an efficient video data processing pipeline for detecting OHS violations, which can be implemented on devices with limited computational resources. It is based on two main components: a "focus of attention" algorithm and a static background subtraction method, integrated into a common sequence of operations.

A. Static Background Subtraction

At many industrial monitoring sites, including drilling rigs, video cameras are often installed stationary. Under such conditions, a significant part of the observed scene remains unchanged over time (static background), while the changes of interest are primarily related to the movement of personnel and, to a lesser extent, the operation of some equipment. Effective separation of moving objects from the static background can significantly reduce the amount of data for subsequent analysis by neural network models.

We propose to use an adaptive background subtraction method based on Gaussian Mixture Models (GMM) [8]. The choice of GMM is due to its ability to effectively model a background that is not absolutely static, and its relative robustness to phenomena characteristic of an industrial environment, such as:

- Shadows: GMM can model changes in pixel intensity caused by moving shadows and not classify them as foreground objects.
- Lighting Changes: The adaptive nature of GMM allows it to gradually adjust to slow and moderate changes in the overall scene illumination level.
- Periodic movements of background elements: At drilling rigs, elements performing repetitive movements may be present (e.g., slight swaying of wires, rotor rotation at low speed if it is not the main object of interest). GMM can incorporate such regular changes into the background model, highlighting only more significant, atypical movements, such as people moving, as foreground.

Thus, the use of GMM allows filtering out a large part of irrelevant information (Figure), leaving for subsequent analysis by the neural network model only those areas of the frame where significant moving objects are present with high probability, primarily personnel.

Example of adaptive background subtraction

Integrating GMM as a preprocessing component can significantly reduce the amount of data fed to the input of subsequent detector stages, leading to an overall reduction in computational load.

B. Region of Interest Cropping

The main idea of this component is to intelligently reduce the volume of data fed to the input of the main, resource-intensive neural network detector model. Instead of processing each video frame entirely with a "large" high-accuracy recognition model, a two-stage process is proposed:

- Stage 1: Preliminary localization of regions of interest (ROI). At this stage, a fast and computationally "light" model (e.g., YOLOv8n) is used for coarse detection of key objects (people) that are highly likely to contain or be associated with the target monitoring objects. Due to its low complexity and small number of target classes, this stage requires insignificant resources. The result is a set of bounding boxes outlining potential ROIs.
- Stage 2: Detailed analysis of ROI. ROIs identified in the first stage (with some configurable padding to capture context and PPE) are cropped from the original frame. These fragments are passed to the input of the main, more accurate model (YOLOv8s) for detecting all target PPE classes. Due to the small ROI sizes, the main detector processes significantly fewer pixels, which reduces inference time while maintaining high accuracy.

IV. Experimental Study and Discussion of Results

To assess the effectiveness and practical applicability of the proposed optimization methodology, a comprehensive experimental study was conducted.

The key goals were a quantitative assessment of the performance gain (FPS) and an analysis of the impact of optimization components on the accuracy (mAP) of neural network detection of OHS violations.

A. Dataset Formation and Selection of Evaluation Metrics

For this study, a specialized dataset was created that maximally reflects the operating conditions at drilling rigs.

The basis of the dataset consisted of 12 video recordings obtained from surveillance cameras installed at various sections of active drilling platforms. The total duration of the source video material was about 8 hours. The recordings covered various work operations and times of day.

Individual frames were extracted from the video recordings. To ensure diversity and reduce data redundancy, a mixed approach was used: every 25th frame was extracted (at an original video FPS of 25 frames/sec, this gave approximately 1 frame per second), and additional frames containing significant changes in the scene or rare views of personnel and equipment were also selected. The total number of images extracted for annotation was about 11,500. The original video resolution varied, but for the dataset, images were resized to a standard Full HD (1920×1080 pixels) and HD (1280×720 pixels) resolution to assess the impact of resolution on performance and accuracy. Data annotation was performed manually using the CVAT (Computer Vision Annotation Tool). The following 8 object classes were defined: *helmet, no_helmet, vest, no_vest, person_front, person_back, gloves, no_gloves*.

During annotation, attention was paid to cases of partial object occlusion, different views, and scales. The class distribution was somewhat imbalanced: the *person_front* and *person_back* classes occurred more frequently than cases of explicit PPE absence, however, they were used in a separate model for fast ROI search, which let us avoided additional data balancing

The resulting dataset was randomly split into training (70 %), validation (15 %), and testing (15 %) sets, while maintaining class proportions in each set.

For quantitative evaluation of the effectiveness of the proposed solutions, commonly accepted metrics in object detection tasks were used: mAP (mean Average Precision), Precision, and Recall for each class, as well as FPS to assess the system's operating speed.

B. Analysis of Experimental Results

Experiments were conducted to evaluate the impact of the proposed optimization methods on performance and accuracy. The following main configurations were tested using the YOLOv8s model as the primary detector (input video resolution – HD, 1280x720):

1) Configuration A (Baseline): YOLOv8s processes the full video frame.

2) Configuration B (ROI Cropping): YOLOv8n for detecting people (ROI) + YOLOv8s for PPE analysis in the identified ROIs.

3) Configuration C (ROI Cropping + Background Subtraction): Sequential application: first background subtraction (GMM), then ROI detector (YOLOv8n), and then YOLOv8s on the final ROIs.

All measurements were performed on a test machine with an Intel Core i7-10700 CPU without using a discrete graphics card.

SUMMARY RESULTS OF PERFORMANCE AND ACCURACY

Type	***FPS***	***mAP@0.5***				
		Overall	***Helmet***	***Vest***	***No helmet***	***No vest***
A	8.5	0.812	0.953	0.930	0.821	0.791
B	16.2	0.863	0.972	0.955	0.881	0.845
C	***24.1***	0.835	0.965	0.942	0.843.	0.810

The data from Table clearly show that the proposed optimization methodology allows achieving a significant increase in performance (FPS) when working on a CPU. The baseline configuration (A) with full HD frame processing by YOLOv8s demonstrates 8.5 FPS. Background removal and applying the model to ROI allows accelerating frame processing to 24.1 FPS while increasing the overall mAP@0.5 accuracy metric from 0.812 to 0.835.

V. CONCLUSION

This paper proposed and experimentally validated a methodology for optimizing the application of neural network technologies for automated OHS monitoring, focused on effective functioning under limited computational resources on a CPU. The methodology, based on a combination of a dynamic ROI algorithm and static background subtraction, demonstrated the ability to significantly increase the video stream processing speed of YOLO models (up to 2.8 times) while maintaining an acceptable level of violation detection accuracy. This opens up opportunities for wider and more economically effective implementation of intelligent video monitoring systems for safety in industrial facilities.

REFERENCES

[1] F. M. Gafarov, G. F. Mingaleev, “Analysis of accidents at oil and gas industry facilities”, Siberian Fire and Rescue Bulletin, vol. 2 (29), 2023, pp. 202-207. (In Russian)

[2] E. A. Shapovalova, P. A. Battalova, “Analysis of the causes of injuries to oil and gas industry workers”, Young Scientist, vol. (4), 2023, pp. 32-34. (In Russian).

[3] N. Jegham, C. Y. Koh, M. Abdelatti, A. Hendawi, YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions, arXiv preprint arXiv:2411.00201v2, 2025.

[4] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, “Feature pyramid networks for object detection”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117-2125.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in Proc. European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[6] B. Jacob et al., "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 2704–2713.

[7] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis. (IJCV), vol. 129, no. 6, 2021, pp. 1789–1819.

[8] Z. Zivkovic, “Improved adaptive Gaussian mixture model for background subtraction”, In Proceedings of the 17th International Conference on Pattern Recognition (ICPR), 2004, vol. 2, 2004, pp. 28-31.

[9] A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. 2020

[10] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, “You only look once: Unified, real-time object detection”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.