Advances in Science and Technology
Research Journal

Advances in Science and Technology Research Journal, 2026, 20(2), 284–297
https://doi.org/10.12913/22998624/211569
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.08.10
Accepted: 2025.12.15
Published: 2026.02.01

Utilisation of a stereo camera and artificial intelligence methods for automatic employee motion tracking in manufacturing enterprises

Andrzej Chmielowiec[1*], Leszek Klich[1], Adam Błachowicz[2], Weronika Woś[1], Sylwia Sikorska-Czupryna[1]

[1] Faculty of Mechanics and Technology, Rzeszow University of Technology, 37-450 Stalowa Wola, Poland
[2] Piklington Automotive Poland, 27-600 Sandomierz, Poland
* Corresponding author's e-mail: achmie@prz.edu.pl

ABSTRACT

The article explores the application of stereo images and neural networks for tracking designated manufacturing employees, with a focus on optimizing production processes. The primary objective of this study was to present a novel approach for the automatic construction of movement trajectories of employees, an issue of significant importance for improving production efficiency. By analysing these trajectories, valuable insights can be gained regarding the optimal arrangement of workstations, facilitating adjustments to their positions in alignment with the actual workflows. This approach aligns closely with the principles of lean manufacturing, offering a method to enhance operational efficiency. The authors proposed a solution based on U-Net type neural networks for classifying objects within stereo images, alongside the integration of stereoscopic imaging with regression techniques for accurate 3D localization of objects. A thorough analysis of the proposed AI models was presented, accompanied by the results of practical tests conducted under varying configuration parameters of the image processing system. The study highlighted the novelty of the approach, contributing to the advancement of automated monitoring and optimization in manufacturing environments.

Keywords: artificial intelligence, neural networks, U-Net, computer vision, motion tracking, lean manufacturing.

INTRODUCTION

The idea of worker movement tracking is one of the Lean Manufacturing methods, which allows process engineers or specialists to improve workplace organization. Of course, in modern industrial factories, production lines are designed according to Lean rules. Standard tools like Value stream mapping or Single piece flow are used to preliminary optimize workplaces, during project design stage. Nevertheless, later, in real industrial environment, it is necessary to adapt and improve workplace organisation, to be in line with expected performance. The problem of workstation design is discussed in the context of Lean production by authors such as Ojo [1], Goncalves and Salonitis [2]. Efficiently designed workstations are essential to provide both flexibility and mass production in an effective way. Unfortunately, it is common to find industrial workstations built without a purposeful design. The design of the workstation, oriented to both users and tasks requirements, allows organisations to increase their production indicators (less time, space and cost) and quality levels. Proper organisation of workers area, has significant impact for standardisation, procedures and efficiency. Especially when operators must use different tools and equipment or have access to many components. As a result of a wrong design, people will lose production time. From this point of view, the analysis of workers behaviours, movements and other activities, stars to play a key role in production efficiency assessment.

Nowadays, as noted by Li and Lee [3], as well as Gavrila [4], it can be observed that the traditional methods to measure work activities rely upon manual on-site observations which are time-consuming and inefficient to address these limitations, computer vision techniques for worker motion analysis are proposed to automatically identify movements without on-site work interruption [3]. The ability to recognise humans and their activities by vision, is key for a machine to interact intelligently and effortlessly with a human-inhabited environment. Because of many potentially important applications, "looking at people" is currently one of the most active application domains in computer vision [4]. The same approach and usage of vision methods, for employee tracking, can be found in the case of building constructions, described in the article of Roberts et al. [5] and also in the article by Cheng et al. [6]. Analysis of construction resources is performed by manually observing construction operations either in person or through recorded videos. Automating this procedure eliminates these issues and allows project teams to focus on performance improvement [5], especially in large areas or sites [6].

This problem analysis and solving propositions can be found in elaborates and articles by: Bian at al. [7], Kitazawa et al. [8] and Stojadinović et al. [9]. They proposed implementation of computer vision system, supported by artificial intelligence as a tool, to analyse and assess workers movement trajectories. It can be a static images classification and detection, like that proposed by Mizher et al. [10] or more complex movement mapping methods, described in paper of Sawano et al. [11]. A similar approach can be found in [12] where is proposed vision-based method for tracking workers in off-site construction.

According to Bauters et al. [13], the vision system used to monitors workstations must have two main functions: providing basic performance measurements and detecting problems or inefficiencies by recognising abnormal operator behaviour. This means that the most difficult thing for vision systems designers is to create a device and an algorithm, which will be fully automated in data processing, assessment and final result. This system, must be able to deliver high accurate data and allows user to quick review and optimise current production process. In this article, authors presented a system, which used a stereo camera and U-Net convolutional neural network, for tracking designated manufacturing employees. Following Capelin et al. [14], convolutional neural networks have found a lot of applications across a variety of different areas, but one of the most important aspect is their application for image processing.

The introduction of employee monitoring using stereoscopic cameras opens new possibilities for defining and controlling work zones assigned to individual employees or work groups. In the context of production organisation, a key aspect is maintaining a well-structured division of the workspace, which minimises the need for workers to move between different areas. Any unjustified crossing of predefined zone boundaries may indicate improper workstation layout or a mismatch between the production process and actual operational needs.

The use of stereo cameras enables precise reconstruction of worker movement trajectories in 3D space, allowing for the automatic assignment of activities to specific zones and the detection of zone boundary violations.

The use of computer vision systems for tracking workers enables not only the analysis of individual movement paths but also the modelling of interactions between operators and the work environment [15]. Analysing human flow within the production setting can reveal hidden patterns of inefficiency resulting from frequent movements outside of assigned work areas [16]. Identifying such behaviours can serve as a foundation for reorganising workstations and processes in order to minimize unnecessary movements and reduce collisions between workers – factors that directly contribute to shorter production cycle times and improved safety.

The literature highlights that traditional approaches to analysing worker flow – such as spaghetti diagrams or time-consuming manual observations – are often subject to bias and limited in precision [17]. In contrast, the integration of stereoscopic cameras with artificial intelligence algorithms, such as U-Net or re-identification networks, enables the automated and continuous generation of activity maps and zone based heatmaps. As a result, it becomes possible to develop a system that not only records the movement of individuals but also analyses, in real time, the compliance of their location with designated work zones.

Such solutions can support the implementation of Lean Manufacturing principles by reducing waste in the form of unnecessary movements

(MUDA) and enabling the dynamic adaptation of the production space layout. Furthermore, motion analysis in the context of work zones can help identify excessive workload on certain employees or groups, which contributes to improving industrial ergonomics.

Ultimately, monitoring activity within designated work zones can also enhance occupational health and safety efforts, as it enables rapid response in the event of detecting personnel in potentially hazardous areas or outside their assigned work zones [18].

The proposed set of computer vision methods, combined with artificial intelligence tools allowed observing and analysing people behaviour, without any interaction with them, and without any disturbance in their work. Stereoscopy and programmable functionalities, executed in real time, requires proper computational power, which can be available by customized software and hardware architecture [19]. Utilisation of dedicated cameras, supported by artificial intelligence algorithm, allowed building trajectory maps of operator's movements, and presence in specific zones of production areas. In effect, a set of data, which allows identifying most time-consuming moments in process, to rearrange or eliminate them, was obtained.

Stereoscopic cameras, by utilising two lenses placed at a fixed distance from each other, enable depth reconstruction of a scene through the analysis of disparity between the images captured by the left and right optical channels. This process mimics human spatial perception, in which the brain interprets differences between the views from both eyes to estimate the distance of objects from the observer. This technology is widely applied in autonomous navigation systems, where precise spatial mapping is critical for obstacle avoidance and trajectory planning [20].

In industrial environments, the use of stereo cameras enables accurate monitoring of the positions of workers, machinery, or components, thereby supporting the optimisation of manufacturing processes and enhancing workplace safety. Stereo cameras also offer high distance measurement accuracy at relatively low implementation costs compared to alternative technologies such as LiDAR systems. An additional advantage lies in the ability to perform precise calibration using known patterns or feature-based methods, which improves the quality of 3D reconstruction even under varying lighting conditions.

High image resolution and frame synchronisation capabilities make stereo cameras suitable for the dynamic observation of fast-moving objects. Moreover, compatibility with deep learning algorithms – particularly convolutional neural networks – enables automated object detection, segmentation, and tracking in 3D space. This significantly enhances the applicability of the technology within Industry 4.0 contexts [21].

Proposed stereo camera is a device with two separate lenses, which can simulate human binocular vision with ability to create 3D images and visions. Following Imam et al. [22] in recent decades, stereo cameras have become a significant technology in robotics, autonomous vehicles, and augmented reality devices. A wide range of applications and digital interface, which allows introducing digital signal processing operations, causes that stereo cameras are currently typical range sensors, which provide distance measurements, among sensors that can be mounted on mentioned above platforms and systems [23]. Another advantage of stereoscopic cameras is the ease and accuracy of their calibration [24]. For these reasons, stereo cameras are the best devices to use in the manufacturing environment, especially in combination with neural network algorithms.

As a second, software part of the tracking system, the authors proposed the usage of convolutional U-Net type neural network for classifying moving objects and localise them in 3D space. Currently automatic image processing, supported by methods of classification, which are based on neural networks is one of the most prospective directions. Especially these, using U-Net architecture [25–27] The article provided an in-depth analysis of the AI model, based on U-Net neural network, to monitor and optimize employee's movement around production lines and work centres. The results of algorithm included construction of optimal Worker's trajectories and paths. On its basis, it will be possible to rearrange production lines layouts and process steps. It will be useful in the stage of existing process improvement and new technologies design. This approach allows process engineers and technology specialists to increase production efficiency and adjust workstation locations, according to Lean rules implemented into algorithm.

The use of an optimisation system, build from stereo camera device and Artificial Intelligence U-Net algorithm allows improving operations, fully automatically. Analysis of various parameters in long period of time, plus the possibility of work with a large set of data will give in result discover of hidden potential, which cannot be detected using standard Lean Manufacturing tools and methods.

The proposed optimisation system, based on stereo camera technology and the U-Net algorithm, enables fully automated improvement of production processes through continuous monitoring and analysis of spatial data. A key component of this approach is the long-term analysis of parameters, which allows for the identification of employee behaviour patterns and recurring inefficiencies in the spatial arrangement of workstations. The collection and processing of large-scale datasets is essential for obtaining statistically significant insights into process dynamics, thereby minimising the influence of random or incidental events [28].

Large volumes of data also enable the training of more complex machine learning models capable of capturing subtle dependencies and correlations between worker movements and production performance. Real-time data analysis, combined with long-term archiving, allows for the verification of implemented changes and their impact on processes over time. This approach supports not only ongoing optimisation efforts but also facilitates forecasting of potential production bottlenecks.

Systematic monitoring of extensive datasets further fosters the development of adaptive predictive models that can dynamically adjust the layout of the workspace in response to changing production conditions [29]. As a result, it becomes possible to transition from reactive production management to proactive process improvement in line with Lean Manufacturing principles. Ultimately, the implementation of such solutions enhances decision-making accuracy and supports the development of a data-driven culture of continuous improvement.

In this article, the authors present an innovative method that combines stereo imaging with advanced neural network techniques, particularly U-Net, to track the movements of manufacturing employees and optimise production processes. The proposed system not only offers a significant improvement in the accuracy of object identification, surpassing existing methods, but also demonstrates exceptional performance in 3D spatial localisation. By enabling the use of low-resolution images with an accuracy comparable to that of stereo camera systems with wider lens baselines, the approach achieves a high level of precision without the need for high computational resources. The results presented in this study clearly demonstrate the effectiveness and practical applicability of the method, making it a groundbreaking tool for automatic monitoring and optimisation in industrial settings. The innovative integration of stereo vision and machine learning paves the way for future advancements in the field of Lean Manufacturing, offering new opportunities for continuous process improvement and enhanced operational efficiency.

MATERIALS AND METHODS

The issue of tracking an employee's movements was divided into two stages. The first stage involved the automatic identification of a characteristic element of the employee's clothing and its extraction from the image. The second stage was concerned with determining the position of the extracted element.

To enhance the effectiveness of the classification algorithms, it was assumed that the employee would wear a brightly coloured baseball cap. This choice was made deliberately. On the one hand, such an article of clothing stands out in the image, which facilitates identification. On the other hand, it has highly variable geometry, as the presence or absence of the visor depending on the viewing angle significantly alters the shape of the object. The images required for the identification and localisation of objects were captured using a stereo camera. The Waveshare Dual Camera Base module was employed in conjunction with a Raspberry Pi Compute Module 4. This type of setup constitutes an energy-efficient and sufficiently powerful platform for stereoscopic image acquisition under industrial conditions.

The applied stereo camera module includes two IMX219-160 cameras based on the Sony IMX219 sensor, which offers a resolution of 8 MP (3280 × 2464) and a 160° field of view. The cameras are connected via the CSI (Camera Serial Interface).

The selection of this camera module was dictated by its capability to capture synchronous images from the left and right channels, its compact form factor, low power consumption, and the potential for deploying this type of solution in mobile environments, such as robots or drones.

A drawback of the module may be the lack of native support for simultaneous data transfer from both cameras. However, in this case, the sampling time is not a critical parameter. Additionally, the absence of a hardware shutter synchronisation mechanism between the cameras results in slight temporal offsets. Nevertheless, this does not significantly affect the performance of the algorithm.

The cameras and image acquisition process are controlled by the Raspberry Pi CM4, which is equipped with a quad-core ARM Cortex-A72 processor (64-bit, 1.5 GHz) and 4 GB of RAM. Owing to its Ethernet interface (1 Gbps) and Wi-Fi capability (Wi-Fi 802.11 b/g/n/ac), the device can communicate with the central system in two ways.

The camera module and the Raspberry Pi CM are connected to a central control system located on a server, which initiates and coordinates the operation of the cameras. After each image pair is captured, the images are stored locally in the appropriate format and subsequently transmitted to the central server via a REST API interface. Further processing takes place on the server with higher computational capacity, where image matching, depth map generation, and object position estimation are performed. This separation of acquisition from processing enables the use of an energy-efficient mobile device in the field while leveraging the computational resources of the central server to execute complex image analysis algorithms.

In Figure 1a, a colour image with a resolution of 256 × 192 pixels is presented, depicting an employee wearing a cap. For the purpose of analysis, the image resolution was intentionally reduced to better reflect production conditions. This decision is justified by the fact that industrial cameras installed on production floors often deliver images at relatively low resolutions.

Another motivation for the resolution reduction was the decrease in image processing time, which is expected to enable the implementation of the employee trajectory estimation process using relatively limited computational resources.

Figure 1 also presents a decomposition of the colour image into six of the most commonly used channels in computer vision methods. The first three channels (b, c, and d) originate from the

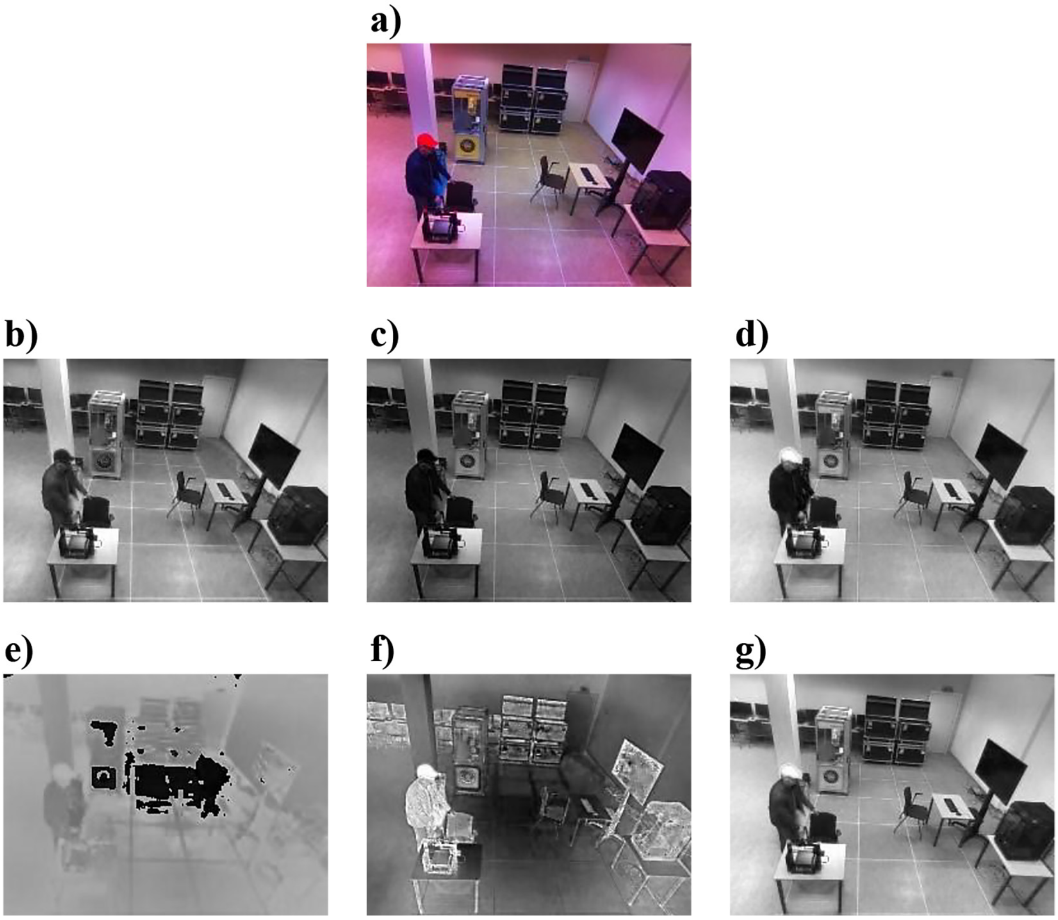

Figure 1. Example of the distribution of a colour 256 × 192 image across 6 most popular channels: (a) colour image, (b) blue, (c) green, (d) red, (e) hue, (f) saturation, (g) value

BGR colour space, while the remaining three (e, f, and g) are from the HSV colour space. These subsets of channels were tested to identify the most efficient method for detecting the target element.

The decomposition clearly demonstrates that, despite the presence of a distinctive clothing element, detecting it based on a single channel could be highly challenging.

A convolutional neural network based on a five-level U-Net architecture [25–27, 30] was employed for the identification of the cap. The network was trained in a supervised manner. The input to the model consisted of images with a resolution of 256 × 192 pixels, with the number of input channels being configurable.

The training process involved optimising the network parameters based on input tensors and their corresponding output tensors. The inputs were images decomposed into selected channels, while the outputs were masks representing the target element within the image. A set of examples input and output tensors is presented in Figure 2. Figure 3 presents the schematic of the network, which accepts three input channels (e.g., the three channels of the HSV or BGR colour space) and produces a single output channel with the same resolution as the input.

The network is composed of five encoding levels and four decoding levels, arranged in a characteristic U-shape specific to this architecture. Each encoding level consists of two convolutional layers, each followed by a nonlinear ReLU activation function. Subsequently, each image dimension is halved using the MaxPooling method, after which the image is forwarded to the next lower level, where the number of channels is doubled. During decoding, encoding channels from the same level are combined with encoding channels from a lower level. Due to dimensional discrepancies, the lower-level encoding channels are scaled to the appropriate dimensions using linear interpolation. Consequently, the first

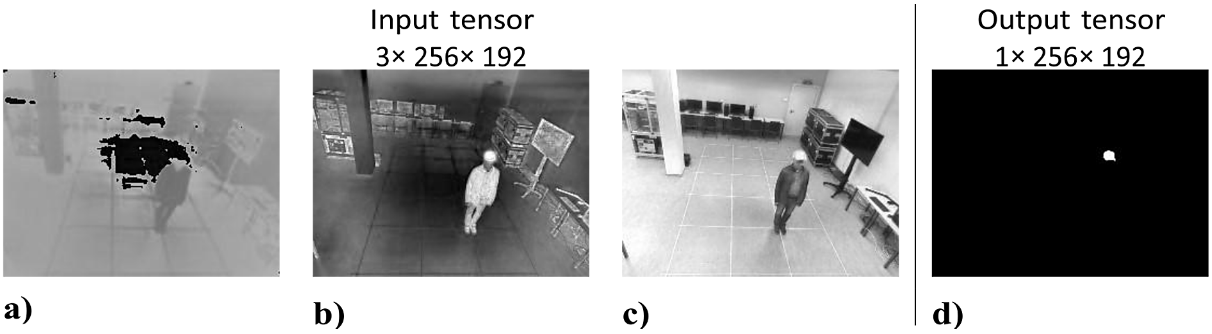

Figure 2. Example of 3 channels input tensor: (a) hue, (b) saturation, (c) value, and 1 channel output tensor: (d) mask

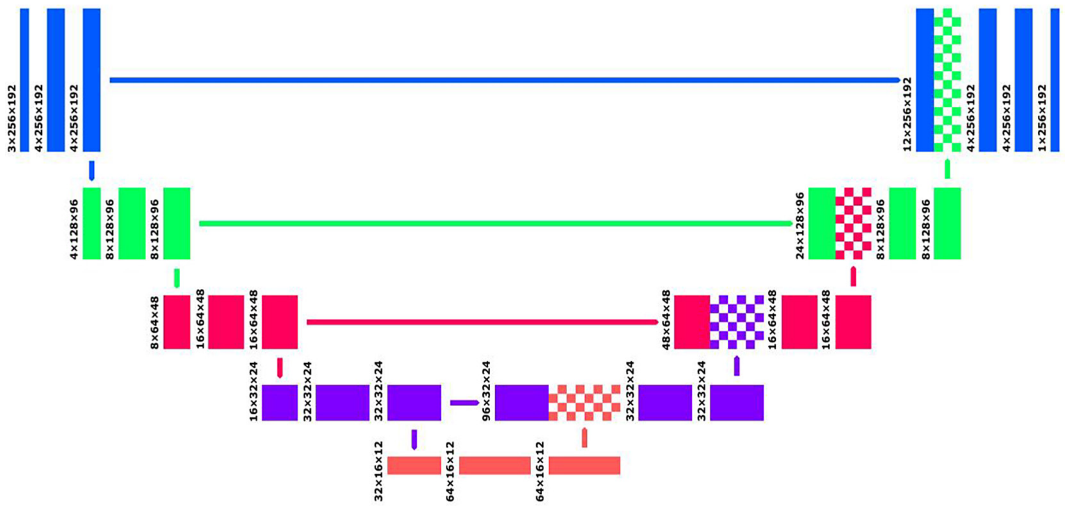

Figure 3. A schematic of a convolutional neural network based on the U-Net architecture designed for object identification in images

convolutional layer of the decoder operates on three times the number of channels compared to the encoding layer at the same level. The decoding step at each level concludes after two convolutional layers, each terminated with a nonlinear ReLU activation function.

Identification of the correct element within the image constitutes the initial step in the localisation process. Thus, features are extracted from the binary mask obtained through the U-Net network to determine the spatial position of the object in three-dimensional space. Assuming that the system simultaneously captures two images using the left and right cameras, two binary masks – the left and the right – are consequently obtained. For each mask, the centroid of the object is calculated using the following equations:

$$u_L = \left(\sum_{x=0}^{255}\sum_{y=0}^{191} x \cdot m_L(x,y)\right) \Big/ \left(\sum_{x=0}^{255}\sum_{y=0}^{191} m_L(x,y)\right) \quad (1)$$

$$v_L = \left(\sum_{x=0}^{255}\sum_{y=0}^{191} y \cdot m_L(x,y)\right) \Big/ \left(\sum_{x=0}^{255}\sum_{y=0}^{191} m_L(x,y)\right) \quad (2)$$

$$u_R = \left(\sum_{x=0}^{255}\sum_{y=0}^{191} x \cdot m_R(x,y)\right) \Big/ \left(\sum_{x=0}^{255}\sum_{y=0}^{191} m_R(x,y)\right) \quad (3)$$

$$v_R = \left(\sum_{x=0}^{255}\sum_{y=0}^{191} y \cdot m_R(x,y)\right) \Big/ \left(\sum_{x=0}^{255}\sum_{y=0}^{191} m_R(x,y)\right) \quad (4)$$

where: x and y denote the width and height coordinates of each pixel, respectively, while m_L and m_R indicate the pixel colour (0 – black, 1 – white). The four numerical values thus determined serve as the starting point for reconstructing the spatial coordinates (X, Y, Z).

The reconstruction process was implemented using two approaches. The first is an analytical method based on the known projection matrices P_L, $P_R \in \mathbb{R}^{3\times4}$ for the left and right cameras, respectively. The second approach involves utilising a neural network trained on an appropriately selected dataset. Both methods were implemented using a set of vectors $S = (u_{L,i}, v_{L,i}, u_{R,i}, v_{R,i}, X_i, Y_i, Z_i)$, which relates the position of the marker in the images to its actual spatial location in 3D space.

The camera projection matrices P_L, P_R were determined using the Singular Value Decomposition (SVD) method [31] applied to a selected subset of S. Generally, if P_* denotes one of the projection matrices (P_L or P_R), the following relationship holds between the coordinates of the vectors in set S:

$$\begin{pmatrix} u_{*,i} \\ v_{*,i} \\ 1 \end{pmatrix} = \begin{pmatrix} p_{*,1,1} & p_{*,1,2} & p_{*,1,3} & p_{*,1,4} \\ p_{*,2,1} & p_{*,2,2} & p_{*,2,3} & p_{*,2,4} \\ p_{*,3,1} & p_{*,3,2} & p_{*,3,3} & p_{*,3,4} \end{pmatrix} \cdot \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} \quad (5)$$

If |S| > 6, these equations lead to an overdetermined system of equations, where the unknowns are the coefficients of the projection matrix. This system can be solved, among other methods, by decomposing the system's matrix into a product of rotation, scaling, and rotation matrices using the Singular Value Decomposition (SVD) technique. Given the coefficients of both projection matrices P_L and P_R, it becomes possible to determine the real-world coordinates (X, Y, Z) based on the coordinates of the object (u_L, v_L) and (u_R, v_R) localised in the images taken by two independent cameras. These coordinates are computed as $X = t_1/t_4$, and $Z = t_3/t_4$, where t_1, t_2, t_3, t_4 are solutions to the following system of equations:

$$\begin{pmatrix} u_L P_{L,3} - P_{L,1} \\ v_L P_{L,3} - P_{L,2} \\ u_R P_{R,3} - P_{R,1} \\ v_R P_{R,3} - P_{R,2} \end{pmatrix} \cdot \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (6)$$

where: $P_{L,i}$ and $P_{R,i}$ are the i-th rows of the projection matrices P_L and P_R, respectively.

The second approach involved utilising the set S as a training dataset for a neural network designed to reconstruct the coordinates (X, Y, Z) based on the centroid coordinates of the marker (u_L, v_L, u_R, v_R) identified in the images from the left and right cameras. For this task, an MLP (Multilayer Perceptron) [32, 33, 34, 35] network with an input dimension of 4 and an output dimension of 3 was selected. Each hidden layer was separated

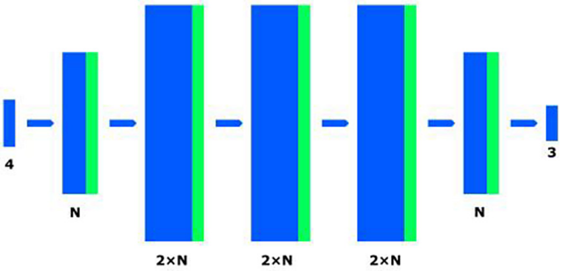

Figure 4. Diagram of a network based on MLP architecture designed for determining the position of objects in 3D space

by a nonlinear ReLU activation layer. Several models of such a network were developed for this study, differing in the number of hidden layers and the number of neurons per layer. An example scheme of such a network with 5 hidden layers is presented in Figure 4.

RESULTS

The training process for cap recognition was performed using 346 images captured with a worker in standing and sitting positions, as well as without any worker present. All images were scaled to a resolution of 256 × 192 pixels. To avoid the issues related to model overfitting, it was decided that images simultaneously captured by both cameras would form inseparable pairs. In practice, this implied that both images from such a pair were assigned either entirely to the training set or entirely to the validation set. Additionally, each image was vertically flipped, horizontally flipped, and rotated by 180°. Consequently, the training procedure was conducted using 1384 images along with their corresponding masks. The entire dataset was partitioned into training and validation sets according to a 3:1 ratio.

A total of 15 object identification models were subjected to effectiveness analysis. The loss function for each model was defined using Binary Cross Entropy with Logits Loss, which is expressed by the following formula

$$\mathcal{L}_1 = -\frac{1}{N}\sum_{i=1}^{N}\left[a_i \cdot log\left(\sigma(b_i)\right) + \right.$$

$$\left. + (1 - a_i) \cdot \log\left(1 - \sigma(b_i)\right)\right] \qquad (7)$$

where: N denotes the number of model outputs, a_i is the i-th expected output value, b_i is the corresponding predicted output value, and $\sigma(b_i)$ is the Sigmoid activation function applied to b_i: $\sigma(x) = (1 + e^{-x})^{-1}$.

The analysed models differed in their selection of input channels and the number of channels in the first level of the model. These two parameters uniquely identified each model, since every subsequent level contained twice as many channels as the previous one. Simultaneously, each level reduced the image resolution by halving both spatial dimensions.

Accordingly, the models were labelled using identifiers indicating the chosen input channels: B (Blue), G (Green), R (Red), H (Hue), S (Saturation), and V (Value), along with the number of channels in the first layer. For example, the label BGR-04 refers to a model fed with three BGR channels and using 4 channels at the first level. In contrast, the label RS-08 denotes a model using the Red and Saturation channels with 8 channels at the first level.

The training results for the selected 15 models are presented in Table 1. It should be emphasized that all models demonstrated very high effectiveness. The training time largely depended on the specific set of model parameters randomly initialised at the beginning of the training process. The behaviour of the ADAM optimisation algorithm used for training was highly influenced by these initial parameter values. Nevertheless, the applied method exhibited consistent and predictable convergence, and after a sufficient number of epochs, the model parameters stabilised at values yielding satisfactory results.

The obtained results clearly indicate that even the smallest models can achieve high performance, provided that their parameters are appropriately selected. Figure 5 presents exemplary images with overlaid masks generated by the BGRHSV-08 model. The highly accurate results in cap identification enabled progression to the next stage of the study, which involved the evaluation of 3D spatial localisation methods. In this phase, two approaches operating on the same sets of binary masks were compared. The first method relied on determining the projection matrices for both cameras, followed by the reconstruction of 3D coordinates through analysis of the cap's coordinates in the images captured by the left and right cameras. The second method involved employing a neural network to predict spatial coordinates.

To compare both approaches, the same loss function was applied—namely, the Mean Squared Error (MSE), defined by the following formula:

$$\mathcal{L}_2 = -\frac{1}{N}\sum_{i=1}^{N}\left[(a_{i,1} - b_{i,1})^2 + \right.$$

$$\left. + (a_{i,2} - b_{i,2})^2 + (a_{i,3} - b_{i,3})^2\right] \qquad (8)$$

where: N is the number of model outputs, $(a_{i,1}, a_{i,2}, a_{i,3}$ denotes the i-th expected 3D spatial location, and $(b_{i,1}, b_{i,2}, b_{i,3})$ represents the i-th predicted spatial location (with individual components corresponding to the X, Y and Z axes, respectively).

Table 1. Summary of training times and memory requirements for 15 U-Net neural network model variants for image recognition

Model	RAM [MB]	Epoch training time [s]	Epochs to train end	Final loss [%]
HSV-04	0.469	19.918	46	0.0611
HSV-08	1.873	33.181	11	0.0580
HSV-16	7.486	68.316	46	0.0571
RS-04	0.469	20.511	50	0.0728
RS-08	1.873	33.415	37	0.0549
RS-16	7.486	67.459	50	0.0519
RHSV-04	0.469	21.018	47	0.0597
RHSV-08	1.873	34.136	41	0.0546
RHSV-16	7.486	70.035	38	0.0557
BGR-04	0.469	19.541	20	0.0541
BGR-08	1.873	33.765	14	0.0722
BGR-16	7.486	68.962	49	0.0504
BGRHSV-04	0.469	20.634	41	0.0574
BGRHSV-08	1.874	33.770	43	0.0526
BGRHSV-16	7.488	70.716	45	0.0540

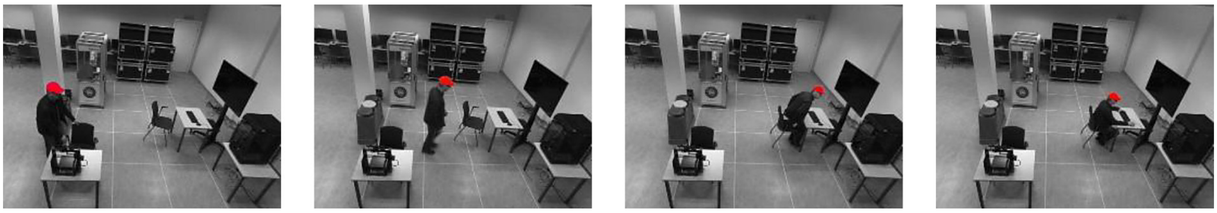

Figure 5. Example of cap detection on test images using the BGRHSV-08 model

To illustrate the impact of image resolution on localisation accuracy, experiments were conducted using masks of four different dimensions: 256 × 192, 512 × 384, 1024 × 768, and 2048 × 1536 pixels. A total of 308 masks were analysed, each depicting the cap positioned at a specific location within the observed space.

Table 2 presents the results of 3D real-world coordinate reconstruction using the projection matrices of the left and right cameras. The obtained results clearly indicate that reconstruction accuracy strongly depends on image resolution. For a resolution of 256 × 192, the mean squared error reaches 1.434 m^2, which corresponds to a root-mean-square error in distance estimation of approximately 1.197 m. This is relatively high, especially considering that the analysed workspace was a rectangle measuring 3 × 5 m. This error decreases significantly with higher resolutions. For instance, at a resolution of 2048 × 1536, the error drops to 0.424 m. Such precision enables meaningful conclusions regarding the worker's movement within a given workstation. However, this increase in accuracy comes at the cost of significantly larger image/mask sizes, which in turn leads to higher computational complexity. The time required to compute object parameters at a resolution of 2048 × 1536 is approximately 16 times longer than at 256 × 192.

Considering that many industrial systems provide image resolutions around 640 × 480, such high-resolution approaches offer limited practical feasibility for accurate localisation. It is also worth noting that the computed projection matrices for both cameras were found to be very similar. This is due to the fact that the applied stereo camera model uses lenses spaced only 60 mm apart. As the analysis shows, this distance is insufficient for accurate localisation of objects situated several meters away from the cameras. The most effective camera setup in such scenarios would involve placing the cameras on two mutually perpendicular walls of the workspace. This configuration would significantly improve the

Table 2. Projection matrices and localisation error for different image resolutions

Resolution	Projection matrix	$\mathcal{L}_2$[m^2]	$\sqrt{\mathcal{L}_2}$ [m]
256×192	$P_L = 10^{-2}\begin{pmatrix} 28.38 & 12.73 & -8.15 & 20.61 \\ 3.06 & -2.56 & -31.48 & 86.79 \\ 0.017 & 0.123 & -0.092 & 0.468 \end{pmatrix}$ $P_R = 10^{-2}\begin{pmatrix} 29.55 & 12.47 & -7.62 & 15.64 \\ 2.98 & -3.10 & -31.93 & 87.35 \\ 0.017 & 0.128 & -0.091 & 0.474 \end{pmatrix}$	1.434	1.197
512×384	$P_L = 10^{-2}\begin{pmatrix} 28.46 & 12.81 & -8.10 & 20.64 \\ 3.06 & -2.48 & -31.37 & 86.79 \\ 0.008 & 0.062 & -0.046 & 0.233 \end{pmatrix}$ $P_R = 10^{-2}\begin{pmatrix} 29.37 & 12.44 & -7.81 & 16.08 \\ 2.95 & -3.01 & -31.96 & 87.31 \\ 0.009 & 0.064 & -0.047 & 0.237 \end{pmatrix}$	0.454	0.674
1024×768	$P_L = 10^{-2}\begin{pmatrix} 28.34 & 12.78 & -8.07 & 20.76 \\ 3.03 & -2.46 & -31.36 & 86.82 \\ 0.004 & 0.031 & -0.023 & 0.117 \end{pmatrix}$ $P_R = 10^{-2}\begin{pmatrix} 29.35 & 12.46 & -7.724 & 16.03 \\ 2.94 & -2.99 & -31.92 & 87.34 \\ 0.004 & 0.032 & -0.023 & 0.118 \end{pmatrix}$	0.212	0.460
2048×1536	$P_L = 10^{-2}\begin{pmatrix} 28.35 & 12.79 & -8.06 & 20.76 \\ 3.04 & -2.45 & -31.34 & 86.82 \\ 0.002 & 0.015 & -0.011 & 0.058 \end{pmatrix}$ $P_R = 10^{-2}\begin{pmatrix} 29.33 & 12.48 & -7.76 & 16.14 \\ 2.94 & -2.97 & -31.89 & 87.33 \\ 0.002 & 0.016 & -0.012 & 0.059 \end{pmatrix}$	0.179	0.424

accuracy of coordinate reconstruction, particularly along the *X* and *Y* axes.

Another method used for reconstructing real-world 3D coordinates was the MLP neural network, the architecture of which is illustrated in Figure 4. The available set of 308 masks was divided into training and validation subsets in a 3:1 ratio. The tested models were named according to the value of the MLP network parameter *N* and the dimensions of the masks from which training data were extracted. A name in the format MLP-N-WxH denotes an MLP model with architecture based on the specified value *N*, and source masks of width *W* pixels and height *H* pixels.

In total, 16 configurations were evaluated. The results of the training process are presented in Table 3. Analysis of the results indicates that, for the proposed architecture, the resolution of the masks had virtually no impact on the final performance. Notably, all trained models achieved lower errors than any of the configurations using projection-matrix-based methods.

The model MLP-32-2048x1536 achieved the lowest error, suggesting that the average localisation error did not exceed 30 cm. However, it should be noted that the computational cost for MLP models is significantly higher compared to the projection matrix method. Even the smallest MLP model required over 8 million multiplications, whereas the projection matrix method reduces to solving a system of four linear equations with four unknowns.

DISCUSSION

In terms of object identification, the proposed approach based on the u-net architecture has been shown to be highly effective. In comparison, Cheng et al. [6] present an analysis of the performance of two architectures: ResNet (Residual Network) [36] and VGG (Visual Geometry Group) [37]. These are used for the identification of the workers equipped with a helmet, a vest, as well as both a helmet and a vest. The reported identification accuracy is highest for workers wearing both a helmet and a vest, reaching nearly 97.5% for the ResNet architecture and nearly 96.5% for the VGG architecture. However, this is clearly lower than the method considered in the present article, where the accuracy exceeded 99.9% for all examined variants. On the other hand, Bian et al. [7] used the YOLO algorithm for object identification, which allowed for the identification of the following objects in a 3D printing process: extruder (99.96%), print bed (99.90%), operator’s finger (92.50%),

Table 3. Summary of training times and memory requirements for 16 MLP neural network model variants for 3D coordinates determination

Model	RAM [MB]	Epoch training time [s]	Epochs to train end	$\mathcal{L}_2$ [m^2]	$\sqrt{\mathcal{L}_2}$[m]
MLP-32-256x192	0.049	0.025	173	0.126	0.355
MLP-64-256x192	0.191	0.028	186	0.112	0.334
MLP-128-256x192	0.757	0.035	161	0.129	0.359
MLP-256-256x192	3.015	0.053	188	0.110	0.331
MLP-32-512x384	0.049	0.027	196	0.123	0.351
MLP-64-512x384	0.191	0.029	145	0.137	0.370
MLP-128-512x384	0.757	0.035	149	0.110	0.332
MLP-256-512x384	3.015	0.054	177	0.122	0.349
MLP-32-1024x768	0.049	0.027	132	0.126	0.355
MLP-64-1024x768	0.191	0.028	193	0.173	0.416
MLP-128-1024x768	0.757	0.035	189	0.135	0.367
MLP-256-1024x768	3.015	0.055	177	0.128	0.358
MLP-32-2048x1536	0.049	0.027	185	0.084	0.290
MLP-64-2048x1536	0.191	0.029	191	0.150	0.387
MLP-128-2048x1536	0.757	0.035	171	0.140	0.374
MLP-256-2048x1536	3.015	0.055	186	0.168	0.409

and operator silhouette (96.50%). It should be noted that only for the extruder and the print bed were the results on par with those presented in this article. In contrast, Han et al. [12] utilised the Mask R-CNN architecture [38] for object identification. Although an accuracy of 96.4% is achieved, this result pertains to tracking and identifying multiple objects simultaneously.

Reconstructing the position of the cap in spatial coordinates has been found to be associated with a relatively large error. However, when comparing the results obtained for the projection matrix with those reported by Imam et al. [22], no better results should be expected. In the cited publication, the authors reported errors for objectives at a distance of 250 mm ranging from 0.200 m to 1.000 m for objects placed in the same distance range as the cap. It should be emphasised that this error pertains only to distance estimation from the camera. In practice, this means that the actual determination of spatial coordinates may involve an even larger error. Given that in the present article cameras were used with objectives only 60 mm apart, the error resulting from the use of the projection matrix should be considered expected. Therefore, it is important to note that the estimation of the position using a neural network significantly improves the accuracy of the obtained results. Seo et al. [23] conducted a study on the effectiveness of stereo vision combined with the Monocular Depth method [39] for the cameras with lens separations of 50 mm and 120 mm. In their approach, the average relative error for distances not greater than 10 m was always found to exceed 9%. In the context of the present article, this corresponds to absolute values ranging from 0.3 to 0.8 m. It is thus evident that the results obtained by the authors fall within the lower bounds of the error reported in other publications.

In the context of position tracking, it is also worth referring to the results obtained by Bauters et al. [13]. They employed four cameras positioned at the corners of the workspace and one camera with a fisheye lens at the centre of the workspace. This arrangement of cameras allows for relatively precise mapping of the object position in space. The space is filled with a virtual cubic grid with an edge length of 0.020 m, which the authors suggest reflects the precision of object localisation. Unfortunately, the authors did not compare the obtained results with the actual position, as the main focus of the article was on classifying the trajectories along which the worker moves. The method for worker identification itself is based on visual background subtraction techniques [40] and is not very robust to the objects appearing around the worker. However, in future research, it would be valuable to consider the camera placements that will significantly diversify the perspectives of the captured images in the employed method, which should also contribute to improving the precision of object localization.

It should be noted that the presented results are based on tests carried out in a fully controlled environment. The next very important step will be the adaptation of the developed methods to real-world conditions, which are characterised by much greater variability. In particular, the applied methods will have to be adjusted to variable lighting conditions and the presence of a larger number of workers.

The essence of the presented method is the tracking of a worker's movement at a selected workstation. Therefore, identification based on a characteristic feature of clothing is considered very convenient. However, it should be pointed out that the presented approach will not be applicable in the cases where the characteristic element is worn by many workers (e.g., a part of the company uniform).

For this reason, the presented approach is not adequate for continuous tracking of movements. The use of the proposed algorithms should be limited to strictly defined time frames for the purpose of analysing the performance of a specific workstation.

CONCLUSIONS

The method proposed in this article represents a significant advancement in both object identification and spatial localisation within manufacturing environments. The results demonstrate that the proposed U-Net based approach for object identification is notably more effective compared to other methods tested in similar contexts. The identification accuracy achieved by the U-Net architecture consistently exceeded 99.9% across all examined variants, significantly outperforming alternative models, such as those based on ResNet and VGG, which reported maximum accuracies of approximately 97.5% and 96.5%, respectively. This highlights the exceptional performance of the proposed system, making it a highly reliable tool for automatic monitoring of worker movements.

In terms of localisation, the results show that the proposed method for 3D spatial localization, using stereo images combined with neural networks, allows for accurate positioning even with low-resolution images. Despite the relatively low image resolution of 256 × 192 pixels, the localization error of the proposed method was comparable to the results of other techniques tested on stereo cameras with narrow lens baselines. This is particularly impressive given that the achieved localisation error using the projection matrix approach was considerably higher under similar conditions. The use of neural networks in this context significantly improved the accuracy, allowing for reliable 3D localisation even when computational resources were limited.

The approach demonstrated in this study offers significant advantages in industrial applications, particularly where computational efficiency is crucial, and high-resolution imaging is not feasible. The system's ability to work effectively with lower-resolution images while maintaining high accuracy in object identification and localisation makes it an ideal solution for real-time monitoring and optimisation of production processes. Moreover, the ability to adapt to various camera configurations and environments opens up additional possibilities for its integration into existing industrial systems.

Future improvements could involve testing the system with cameras having a wider lens baseline to further enhance localisation precision, as well as incorporating additional algorithms for collision detection, activity prediction, and automatic classification of worker tasks. These enhancements would expand the functionality of the system, making it even more versatile and capable of supporting a wider range of industrial applications.

In conclusion, the proposed method presents a groundbreaking solution for automatic worker identification and trajectory tracking, with remarkable accuracy even under challenging conditions. The combination of stereo vision and neural networks offers an innovative approach to optimising production processes and provides valuable insights for future advancements in industrial automation and smart manufacturing systems.

REFERENCES

1. Ojo S.O., Bailey D.P., Chater A.M., Hewson D.J. The impact of active workstations on workplace productivity and performance: a systematic review. International journal of environmental research and public health. 2018; 15(3): 417. https://doi.org/10.3390/ijerph15030417
2. Goncalves M.T. and Salonitis K. Lean assessment tool for workstation design of assembly lines. Procedia Cirp. 2017; 60: 386–391. https://doi.org/10.1016/j.procir.2017.02.002.

3. Li C. and Lee S. Computer vision techniques for worker motion analysis to reduce musculoskeletal disorders in construction. American Society of Civil Engineers. 2011; 380–387. https://doi.org/10.1061/41182(416)47
4. Gavrila D.M. The visual analysis of human movement: A survey. Computer Vision and Image Understanding. 1999; 73(1): 82–98. https://doi.org/10.1006/cviu.1998.0716
5. Roberts D., Calderon W.T., Tang S., Golparvar-Fard M. Vision-based construction worker activity analysis informed by body posture. Journal of Computing in Civil Engineering. 2020; 34(4): 04020017. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000898
6. Cheng J.P., Wong P.K.-Y., Luo H., Wang M., Leung P. Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification. Automation in Construction. 2022; 139: 104312. https://doi.org/10.1016/j.autcon.2022.104312
7. Bian S., Li C., Fu Y., Ren Y., Wu T., Li G.-P., Li B. Machine learning-based real-time monitoring system for smart connected worker to improve energy efficiency. Journal of Manufacturing Systems. 2021; 61: 66–76. https://doi.org/10.1145/2675743.27718
8. Kitazawa M., Takahashi S., Takahashi T.B., Yoshikawa A., Terano T. Combining workers' behaviour data and real time simulator for a cellular manufacturing system. In 2016 World Automation Congress (WAC). IEEE. 2016; p. 1–6. https://doi.org/10.1109/WAC.2016.7583024
9. Stojadinović A., Stojanović N., Stojanović L. Dynamic monitoring for improving worker safety at the workplace: use case from a manufacturing shop floor. In Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems. Association for Computing Machinery. 2015; 205–216. https://doi.org/10.1145/2675743.2771881
10. Mizher H.S., Taher M.A., Falih B.S., Gierz Ł.A., Warguła Ł., Wieczorek B. Detecting and classifying media images of athletes using convolutional neural networks – case study: Individual sports images. Advances in Science and Technology Research Journal. 2025; 19(6): 152–166. https://doi.org/10.12913/22998624/199798
11. Sawano Y., Miura J., Shirai Y. Man chasing robot by an environment recognition using stereo vision. In Arai E., Arai T., Takano M., editors, Human Friendly Mechatronics, 241–246. Elsevier Science, Amsterdam, 2001. https://doi.org/10.1016/B978-044450649-8/50041-3
12. Xiao B., Xiao H., Wang J., Chen Y. Vision-based method for tracking workers by integrating deep learning instance segmentation in off-site construction. Automation in Construction. 2022; 136: 104148. https://doi.org/10.1016/j.autcon.2022.104148
13. Bauters K., Cottyn J., Claeys D., Slembrouck M., Veelaert P., van Landeghem H. Automated work cycle classification and performance measurement for manual work stations. Robotics and ComputerIntegrated Manufacturing. 2018; 51:139–157. https://doi.org/10.1016/j.rcim.2017.12.001
14. Capelin M., Martinez G.A.S., Xing Y., Siqueira A.F., Qian W.-L. Analysis of wire rolling processes using convolutional neural networks. Advances in Science and Technology Research Journal. 2024; 18(2): 103–114. https://doi.org/10.12913/22998624/183699
15. Han S., Lee S., Pen˜a-Mora F. Vision-based motion detection and safety monitoring for construction workers. Automation in Construction. 2013; 35: 131–141. https://doi.org/10.1061/9780784412329.104
16. Li N., Becerik-Gerber B., Krishnamachari B., Soibelman L. A bim centred indoor localization algorithm to support building fire emergency response operations. Automation in Construction. 2012; 42: 78– 89. http://dx.doi.org/10.1016/j.autcon.2014.02.019
17. Sacks R., Koskela L., Dave B., Owen R. Interaction of lean and building information modelling in construction. Journal of Construction Engineering and Management. 2010; 136(9): 968–980. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000203
18. Buniya M.K., Othman I., Sunindijo R.Y., Kineber A.F., Mussi E., Ahmad H. Barriers to safety program implementation in the construction industry. Ain Shams Engineering Journal. 2021; 12(1): 65–72. https://doi.org/10.1016/j.asej.2020.08.002
19. Heinzle S., Greisen P., Gallup D., Chen C., Saner D., Smolic A., Burg A., Matusik W., Gross M. Computational stereo camera system with programmable control loop. ACM Transactions on Graphics (TOG). 2011; 30(4): 1–10. https://doi.org/10.1145/2010324.1964989
20. Geiger A., Lenz P., Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012; 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074
21. Chen X., Ma H., Wan J., Li B., Xia T. Multi-view 3d object detection network for autonomous driving. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; 6526–6534. https://doi.org/10.1109/CVPR.2017.691
22. Imam F., Setyanto A., Kusnawi. Optimising baseline distance in stereo cameras: an experimental approach to enhance object distance accuracy. International Journal of Research and Review. 2025; 12(5): 94– 107. https://doi.org/10.52403/ijrr.20250512
23. Seo B.-S., Park B., Choi H. Sensing range extension for short-baseline stereo camera using monocular depth estimation. Sensors. 2022; 22(12): 4605. https://doi.org/10.3390/s22124605

24. Qin X., Yang J., Liang W., Pei M., Jia Y. Stereo camera calibration with an embedded calibration device and scene features. In 2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012 - Conference Digest. 2012; 2306–2310, https://doi.org/10.1109/ROBIO.2012.6491313

25. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. In Navab N., Hornegger J., Wells W., Frangi A., editors, Medical image computing and computer-assisted intervention – MICCAI 2015: 18th international conference, Munich, Germany, October 5-9. Springer. 2015; III(18): 234–241. https://doi.org/10.1007/9783-319-24574-4 28

26. Li A., Li X., Ma X. Residual dual u-shape networks with improved skip connections for cloud detection. IEEE Geoscience and Remote Sensing Letters. 2024; 21: 1–5. https://doi.org/10.1109/LGRS.2023.3337860

27. Williams C., Falck F., Deligiannidis G., Holmes C.C., Doucet A., Syed S. A unified framework for u-net design and analysis. Advances in Neural Information Processing Systems. 2023; 36: 27745–27782

28. Wuest T., Weimer D., Irgens C., Thoben K.-D. Machine learning in manufacturing: advantages, challenges, and applications. Production & Manufacturing Research. 2016; 4(1): 23–45. https://doi.org/10.1080/21693277.2016.1192517

29. Lee J., Davari H., Singh J., Pandhare V. Industrial artificial intelligence for industry 4.0-based manufacturing systems. Manufacturing Letters. 2018; 18: 20–23. https://doi.org/10.1016/j.mfglet.2018.09.002

30. Chmielowiec A. Utilization of neural networks and u-net architecture for cladding area detection for collaborative robots. In Machado J., Trojanowska J., Soares F., Rea P., Butdee S., Gramescu B., editors, Innovations in Mechatronics Engineering IV, 253–264, Cham, 2025. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-94223-5

31. Akritas A.G. and Malaschonok G.I. Applications of singular-value decomposition (svd). Mathematics and Computers in Simulation. 2004; 67(1): 15–31. https://doi.org/10.1016/j.matcom.2004.05.005

32. Chmielowiec A., Żurawski P., Sikorska-Czupryna S., Klich L., Organiściak P. Application of neural networks for defect detection in rotationally symmetric components. Advances in Science and Technology Research Journal. 2024; 18(8): 403–415. https://doi.org/10.12913/22998624/194891

33. Vladov S., Yakovliev R., Bulakh M., Vysotska V. Neural Network Approximation of Helicopter Turboshaft Engine Parameters for Improved Efficiency. Energies. 2024; 17(9): 2233. https://doi.org/10.3390/en17092233

34. Vladov S., Bulakh M., Baranovskyi D., Kisiliuk E., Vysotska V., Romanov M., Czyżewski J. Application of the integral energy criterion and neural network model for helicopter turboshaft engines' vibration characteristics analysis. Energies. 2024; 17(22): 5776. https://doi.org/10.3390/en17225776

35. Vladov S., Bulakh M., Baranovskyi D., Sokurenko V., Muzychuk O., Vysotska, V. Helicopter turboshaft engines combustion chamber monitoring neural network method. Measurement. 2025; 242:116267. https://doi.org/10.1016/j.measurement.2024.116267

36. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; 770–778. https://doi.org/10.1109/CVPR.2016.90

37. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. https://arxiv.org/abs/1409.1556

38. K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In 2017 IEEE International Conference on Computer Vision (ICCV), 2017; 2980–2988. https://doi.org/10.1109/ICCV.2017.322

39. D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems, 27, 2014.

40. M. Slembrouck, J.N. Castaneda, G. Allebosch, D. Van Cauwelaert, P. Veelaert, and W. Philips. High performance multi-camera tracking using shapes-from-silhouettes and occlusion removal. In Proceedings of the 9th international conference on distributed smart cameras, 2015; 44–49. https://doi.org/10.1145/2789116.2789127