applied sciences

MDPI

Article

Detection of Anomalous Behavior of Manufacturing Workers Using Deep Learning-Based Recognition of Human–Object Interaction

Rita Rijayanti [1], Mintae Hwang [1,*] and Kyohong Jin [2,*]

[1] Department of Information and Communication Engineering, Changwon National University, Changwon 51140, Republic of Korea; rita.rijayanti@gs.cwnu.ac.kr
[2] Department of Electronic Engineering, Changwon National University, Changwon 51140, Republic of Korea
* Correspondence: mthwang@cwnu.ac.kr (M.H.); khjin@changwon.ac.kr (K.J.); Tel.: +82-55-213-3832 (M.H.); +82-55-213-3656 (K.J.)

Abstract: The increasing demand for industrial products has expanded production quantities, leading to negative effects on product quality, worker productivity, and safety during working hours. Therefore, monitoring the conditions in manufacturing environments, particularly human workers, is crucial. Accordingly, this study presents a model that detects workers' anomalous behavior in manufacturing environments. The objective is to determine worker movements, postures, and interactions with surrounding objects based on human–object interactions using a Mask R-CNN, MediaPipe Holistic, a long short-term memory (LSTM), and worker behavior description algorithm. The process begins by recognizing the objects within video frames using a Mask R-CNN. Afterward, worker poses are recognized and classified based on object positions using a deep learning-based approach. Next, we identified the patterns or characteristics that signified normal or anomalous behavior. In this case, anomalous behavior consists of anomalies correlated with human pose recognition (emergencies: worker falls, slips, or becomes ill) and human pose recognition with object positions (tool breakage and machine failure). The findings suggest that the model successfully distinguished anomalous behavior and attained the highest pose recognition accuracy (approximately 96%) for standing, touching, and holding, and the lowest accuracy (approximately 88%) for sitting. In addition, the model achieved an object detection accuracy of approximately 97%.

Keywords: anomalous behavior; human–object interaction; manufacturing worker; MediaPipe Holistic; LSTM; Mask R-CNN; deep learning

Citation: Rijayanti, R.; Hwang, M.; Jin, K. Detection of Anomalous Behavior of Manufacturing Workers Using Deep Learning-Based Recognition of Human–Object Interaction. *Appl. Sci.* **2023**, *13*, 8584. https://doi.org/10.3390/app13158584

Academic Editor: Steven Davy

Received: 23 June 2023
Revised: 18 July 2023
Accepted: 19 July 2023
Published: 26 July 2023

1. Introduction

In Industry 4.0, following the growing public demand for industrial products that significantly contribute to economic perspectives [1,2], manufacturers are striving to produce highly customizable products while maintaining production accuracy and speed [2,3]. In line with this, government regulations have been established with the purpose of regulating the responsibility of company owners to ensure worker safety during working hours [4]. Likewise, monitoring manufacturing environments, particularly human workers, is crucial [5–7]. Implementing these measures enhances quality control, worker productivity, and cost efficiency in addition to workplace safety. Hence, it is necessary to focus on aspects such as human error, machine malfunction, or other possibilities that significantly impact product quality, efficiency, and workplace safety.

As a matter of fact, human error significantly contributes to manufacturing production disturbances and safety problems [1–11]. In addition, humans are considered the most active objects; thus, detecting worker behavior is vital. Anomaly detection, which is a key task of modern automated monitoring systems, highly impacts diverse fields such as health care, sports analysis, industries, and security. Therefore, this study aims to develop

a model that detects anomalous worker behavior based on human–object interactions. The outputs are in the form of descriptive texts that indicate detection results. This supports management teams in mitigating production failure and improving worker performance, as well as workplace safety. Accordingly, an algorithm has been developed for anomaly detection and text generation.

To provide a clear structure and logical flow, this paper is organized into different sections. Section 2 presents a description of related research, and Section 3 contains an explanation of the concept and detection of anomalous behavior. Next, the proposed method is discussed in Section 4, and the experimental results are provided in Section 5. Finally, Section 5 also presents the conclusion and future study directions.

2. Related Works

Computer vision related to the monitoring and detection of anomalous behavior has been studied under different conditions and environments for different purposes using different methods. Some of the studies are addressed below.

The authors of [7] present a system that analyzes human motion from the position and movement of objects within three-dimensional spaces in in-depth images. The system uses a depth sensor and a computer vision algorithm to detect human poses and movements and subsequently executes a machine learning model to recognize specific behaviors. In [12], computer vision monitors worker productivity and safety for the purpose of minimizing manual monitoring in industries using an object recognition algorithm (YOLOv3) trained using images. Likewise, ref. [13] presents a vision-based fault classification approach by utilizing cameras to capture real-time images from robot movements and components to monitor industrial robots. The aim is to ensure worker productivity and product quality. In contrast, ref. [14] focuses on designing and implementing a system that uses a combination of shape- and motion-based features to recognize human actions based on silhouettes. The system consists of preprocessing, silhouette extraction, feature extraction, and classification modules. A support vector machine classifier is implemented to classify human behavior. Furthermore, in [15,16], image analysis is performed on images where objects and body poses remain stable.

Subsequently, ref. [17] proposes a multi-view learning approach to detect anomalous human behavior in complex and dynamic environments using multiple data sources, such as sensor data, video feeds, and contextual information. In addition, the authors of [18] conduct video monitoring using intelligent behavior identification techniques to detect suspicious behavior and alert security personnel. The system utilizes a camera to capture footage that is subsequently processed to identify suspicious worker behaviors based on human body movement analysis. As a result, the proposed system reduces the number of false alarms and enables security personnel to promptly respond to potential threats. Apart from this, ref. [19] focuses on safety monitoring at construction sites to prevent accidents and injuries using an anomaly detection algorithm developed in a random finite set framework. Later, ref. [20] presents an unsupervised anomaly-based approach for pedestrian age classification in surveillance camera footage, and the proposed framework incorporates an adversarial model with skip connections. The results indicate the potential of this approach to classify pedestrian age and detect anomalous age patterns. In addition, in [21], an online and adaptive method is utilized to detect abnormal events in video surveillance using a spatiotemporal ConvNet. The method combines spatial and temporal information and incorporates an adaptive learning approach to dynamically update the model, thus improving the detection accuracy using normal events as training samples. A different study [22] highlights the importance of anomaly detection in the elderly daily behavior domain. The study presents diverse methods and approaches as valuable monitoring tools to detect possible anomalous situations that potentially indicate warning signs for chronic illnesses or initial physical and cognitive decline. The study is specifically designed for caregivers, healthcare professionals, and family members to enhance their overall safety and quality of daily life. Ultimately, the integration of advanced

technologies and machine learning in eldercare conceivably strengthens the well-being and independence of the elderly population, while facilitating and supporting caregivers in delivering professional duties. Moreover, ref. [23] presents a performance analysis of convolutional neural networks (CNN) using a long short-term memory (LSTM) network to effectively capture and analyze spatiotemporal information from surveillance video data for surveillance systems. The image features are extracted from the image frame sequences using CNN, whereas LSTM uses the gate mechanism to maintain vital information. Afterward, the results are compared with existing detection models, including a mixture of probabilistic principal analysis, motion deep net, social force, and dictionary-based models for performance evaluation. The authors of [24] propose a framework that utilizes deep learning techniques to detect abnormal sea surface temperature (SST) events automatically. The framework consists of two main components: a deep convolutional autoencoder and a classifier. The deep convolutional autoencoder is trained on a large dataset of normal SST patterns. The purpose is to learn the underlying features and patterns to detect abnormal events in sea surface temperature, thus monitoring and managing marine ecosystems and predicting extreme weather events. Subsequently, ref. [25] aims to improve the accuracy and efficiency of SST prediction, as well as reduce computational costs by combining two deep learning models: LSTM networks and CNNs. The hybrid approach consists of two main stages: feature extraction and prediction. During the feature extraction stage, CNN extracts spatial features from satellite SST images. The CNN learns to capture patterns and spatial dependencies in the data; therefore, it allows for the encoding of relevant information. The extracted features are subsequently fed into the LSTM networks in the prediction stage. In this case, LSTM networks are capable of capturing temporal dependencies and long-term patterns suitable for time series prediction tasks. In [26], a valuable overview of deep learning approaches is provided for anomalous event detection in video surveillance that serves as a comprehensive resource for researchers and practitioners inclined toward understanding the advancements and current state-of-the-art techniques in this field.

Equivalently notable, existing studies have not specifically addressed the labeling of anomalous worker behavior using descriptive text. This study, therefore, presents a model that detects anomalous behavior performed by manufacturing workers based on human–object interaction in videos. In practice, the model combines object detection and human pose estimation using Mask R-CNN (R-CNN stands for Region-Based Convolutional Neural Network), MediaPipe Holistic, long short-term memory (LSTM), and worker behavior description algorithm. Mask R-CNNs and LSTMs are part of the most outstanding practices within deep learning. Mask R-CNN is capable of detecting objects, whereas Me-diaPipe works as a pose tracker. In addition, LSTM is utilized to train the model to recognize possible differences in worker behavior within manufacturing environments. Additionally, as an advancement of previous research [15,16], this study focuses on using videos to generate descriptive texts and distinguish normal and anomalous worker behavior.

3. Anomalous Behavior Detection

3.1. Problem Statement

Anomalous behavior is defined as unusual or abnormal activities performed by a human. For example, sleeping in a bedroom is normal, whereas sleeping in a bathroom is unusual; lying on a floor motionless over an extended period is unusual. Likewise, anomalous worker behavior refers to actions or conduct exhibited by workers that deviate from the expected, normal, standard, or desired behavior at a manufacturing site. This behavior compromises a workplace's productivity, safety, and overall functions.

This study addresses anomalous behavior in manufacturing environments in particular. The anomalies are divided into anomalies related to human pose recognition (emergency: worker falls, slippage, or illness) and anomaly related to the recognition of human poses with object positions (tool breakage and machine failure). In this case, an anomaly detection algorithm is developed to detect particular conditions resulting from recognizing unexpected human poses based on body movements. Moreover, a framework

is proposed to detect human poses and interconnected object positions. Normal data consist of a video containing expected worker behavior, and test data contain normal and abnormal behavior. Hence, the proposed method learns behavior patterns from normal data to estimate normal/abnormal responses.

3.2. Body Motion-Based Anomaly Detection

Body motion-based anomaly detection requires thorough human body motion information. The idea is to collect features from the expected body movements provided by the normal data. The model is trained to classify normal or anomalous features and has been tested to verify performance. Correspondingly, the use of MediaPipe Holistic and LSTM to analyze body motion features is proposed. The method is pre-trained to classify purposes in the frames. Frame t, f(t) is the set of all possible feature vectors associated with t. Additionally, the collected b vectors from the normal data are subsequently embedded into a common multidimensional Cartesian space.

4. Proposed Method

Figure 1 illustrates the architecture of the proposed method to provide an overview of the developed model. The model consists of object detection, pose identification, pose classification, and behavior classification description stages. The purpose is to identify worker anomalous behavior in a manufacturing environment. The output is descriptive text generated based on the identified worker pose and interaction with surrounding objects. The stages are detailed as follows.

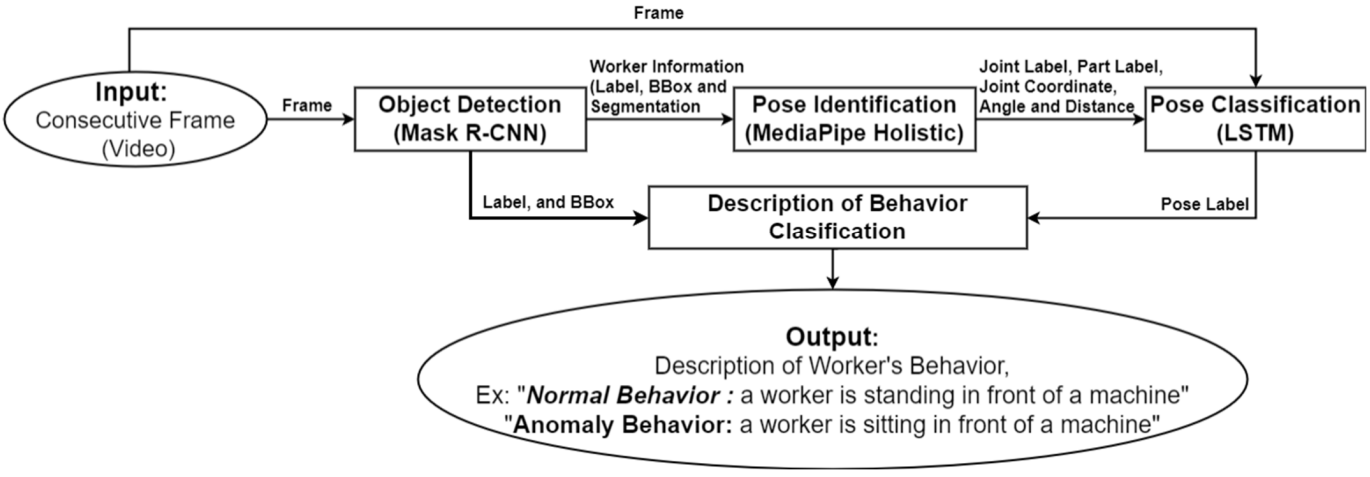


Figure 1. Proposed architecture.

First, the input feature (video frames) is extracted. In this case, sequential frames are extracted from the video from the top of the convolutional layer to identify the transformation of each movement in each frame. The entire dataset is divided into training and testing data in 75% and 25% ratios. The model classifies objects using an input image size of $224 \times 224 \times 3$ (RGB).

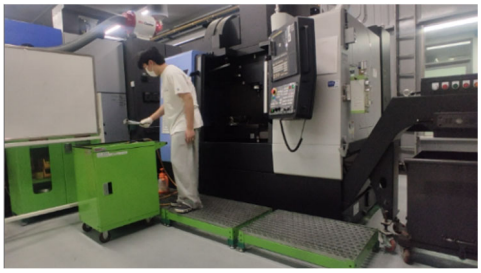
The next step is object detection, where each object in the sequential video frames is detected using Mask R-CNN methodology by producing a fixed size from each feature map. Afterward, a completely connected layer with a similar input size is created to enable convolutional and deconvolution neural networks to detect object classes (labels). Subsequently, it creates bounding boxes (bbox) and segmentation masks. The underlying reason for utilizing R-CNN Mask is its capability to detect each small object and overlapping objects in a more specific and efficient way compared to other models such as Fast R-CNN, Faster R-CNN, Single-Shot MultiBox Detector (YOLO), and Single-Shot MultiBox Detector (SSD). Additionally, Mask R-CNN is suitable for the requirements and focus of the study.

In the next stage, poses are identified and classified to recognize worker body movement. Within the process, the model acquires pose landmarks from the MediaPipe Holistic framework, and subsequently examines the overall body parts to determine the precise locations of human body key points. The purpose is to understand the articulation of an individual's joints and body parts. As a matter of fact, this stage is intended to generate a label for each part of the human body prior to extracting and classifying the workers' poses, such as standing, walking, touching, holding, bending, sitting, and squatting.

Following the forementioned pose classification stage, the model accordingly describes behavior classification (determines possible anomalous behavior) and generates descriptive text based on the identified worker pose and interaction with surrounding objects. In this study, the behavior is identified as normal or anomalous subsequent to adjusting to the real conditions within manufacturing environment. To illustrate this, the sitting pose, in general, is considered a neutral pose; however, in case the sitting pose is performed on an object, a combination of the pose and the object is considered as a behavior and is classified as normal or abnormal based on the location or domain. For a more comprehensive illustration, a student is sitting on a chair in a classroom, for example. This behavior is classified as normal, as sitting in the classroom is normal. However, when the behavior is performed in a manufacturing environment, the behavior is classified as anomalous, as the behavior is forbidden according to the working procedure. For a more detailed example, a worker is sitting in front of a machine located in a manufacturing environment. This is classified as anomalous behavior as it is considered to deviate from the working procedure. In addition, Figure 2 displays anomalous conditions derived from video frames. Figure 2a–c presents normal conditions: a worker is precisely standing in front of a machine (a), a worker is touching a control box in front of a machine (b), and a worker is holding a tool in front of a machine (c). By contrast, Figure 2d–f presents anomalous behavior: a worker is sitting in front of a machine (d), a worker is sitting/slipping in front of a machine (e), and a worker is standing behind a machine (unseen; f).

(**a**)

(**b**)



(**c**)

(**d**)

Figure 2. *Cont.*

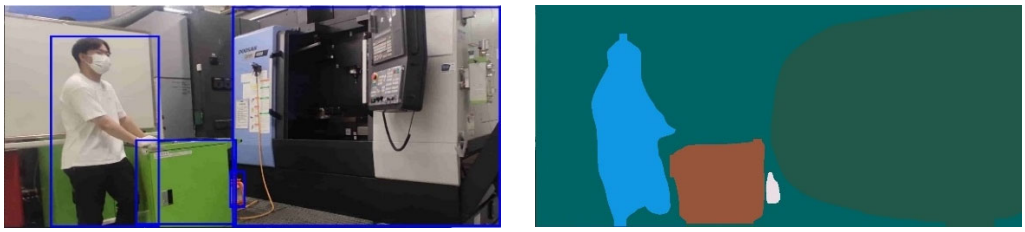
(**e**) (**f**)

Figure 2. Examples of normal and anomalous worker behavior: (**a**) a working is standing, (**b**) a worker is touching the control box, (**c**) a worker is holding a tool, (**d**) a worker is sitting in front of a machine, (**e**) a worker is sitting/slipping, and (**f**) a worker is standing behind a machine (unseen).

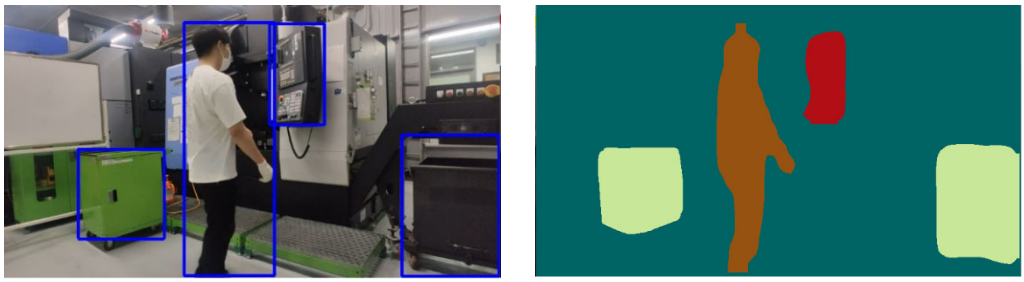
5. Implementation of the Study

5.1. Implementation Result

In practice, a customized dataset was trained by including objects created through a process that was augmented by rotation and occlusion operations. The objects comprised a worker, a machine, a control box, a toolbox, a tool, and a product, which were utilized to perform object detection. Subsequently, the collected data were divided into training data of 75% and testing data of 25%. In this particular context, the model was intended to enhance the previous model [15,16] by adding a dataset and layers. For a more detailed explanation, Figure 3a depicts the model specifically detecting each object in the frames: a worker, a machine, a toolbox, and a tool; Figure 3b depicts a worker, a machine, a control box, and two toolboxes displayed using bbox and segmentation masks. In addition, the results signify high accuracy and fast recognition based on normal conditions.



(**a**)



(**b**)

Figure 3. Examples of object detection: (**a**) bbox and segmentation result of a worker, machine, toolbox, and tool; (**b**) bbox and segmentation result of a worker, machine, control box, and two toolboxes.

Moreover, pose classification was performed using LSTM. In the process, a network was built that consisted of three LSTM layers and three dense layers, which therefore, produced an output layer that consisted of seven neurons representing dynamic poses or human behavior: standing, walking, touching, holding, bending, sitting, and squatting. Due to direct connections between the current and previous cells, the earlier-generated information was directly applied to predict target poses; thus, crossing multiple units was unnecessary. The long-range information persisted up to the stage of predicting

the last pose. In addition, enhanced information was employed due to direct use of the previous and last hidden states to predict the current pose. In this model, the previous LSTM cells, including the last cells, were directly connected to the current cells, where the attention mechanism integrated information in different hidden states. For a more precise description, the training and testing processes are presented in Figure 4.

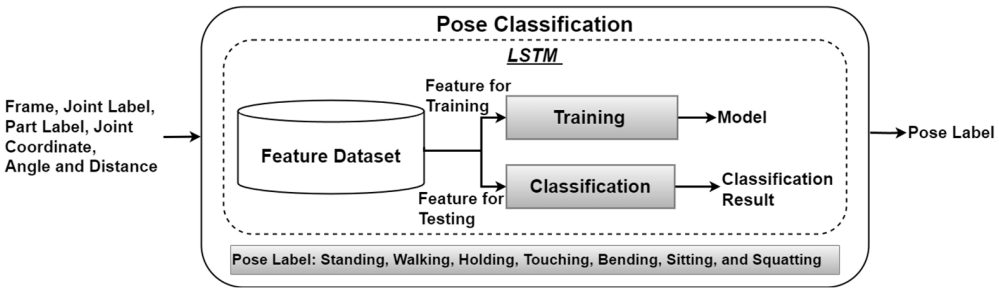


Figure 4. Pose classification.

Table 1 illustrates the hyperparameters for training the model. These parameters are obtained from continuous training of the model that produced the expected results.

Table 1. Hyperparameters of the study.

Hyperparameter	Values
Hidden layer and number of neurons (Three LSTM layers and three dense layers)	LSTM 128, LSTM 64, LSTM 64, Dense 64, Dense 32, Dense 32
Activation function input and hidden layer	Relu
Activation function output	Softmax
Dropout	No
Regularization	No
Optimizer	Adam
Bach size	32
Number of epochs	100

The model was tested on a video sequence to ensure each worker pose was accurately detected, as illustrated in Figure 5.

Turning to another aspect, a worker behavior description algorithm was created to present the worker behavior derived from video images in the form of descriptive texts. In this practice, the logic was based on the worker's position relative to the coordinates of each object and the pose label within each frame. Each behavior was classified as normal or anomalous based on the given object position and pose label. Afterward, descriptive texts were created according to the attained results from human pose estimations and detection on the objects nearest to the worker. In this case, simple description texts were composed of a subject ("("a") + ("worker") + to be + verb + adverb ("in front of") + particle + object").

(**a**) (**b**)

Figure 5. *Cont.*

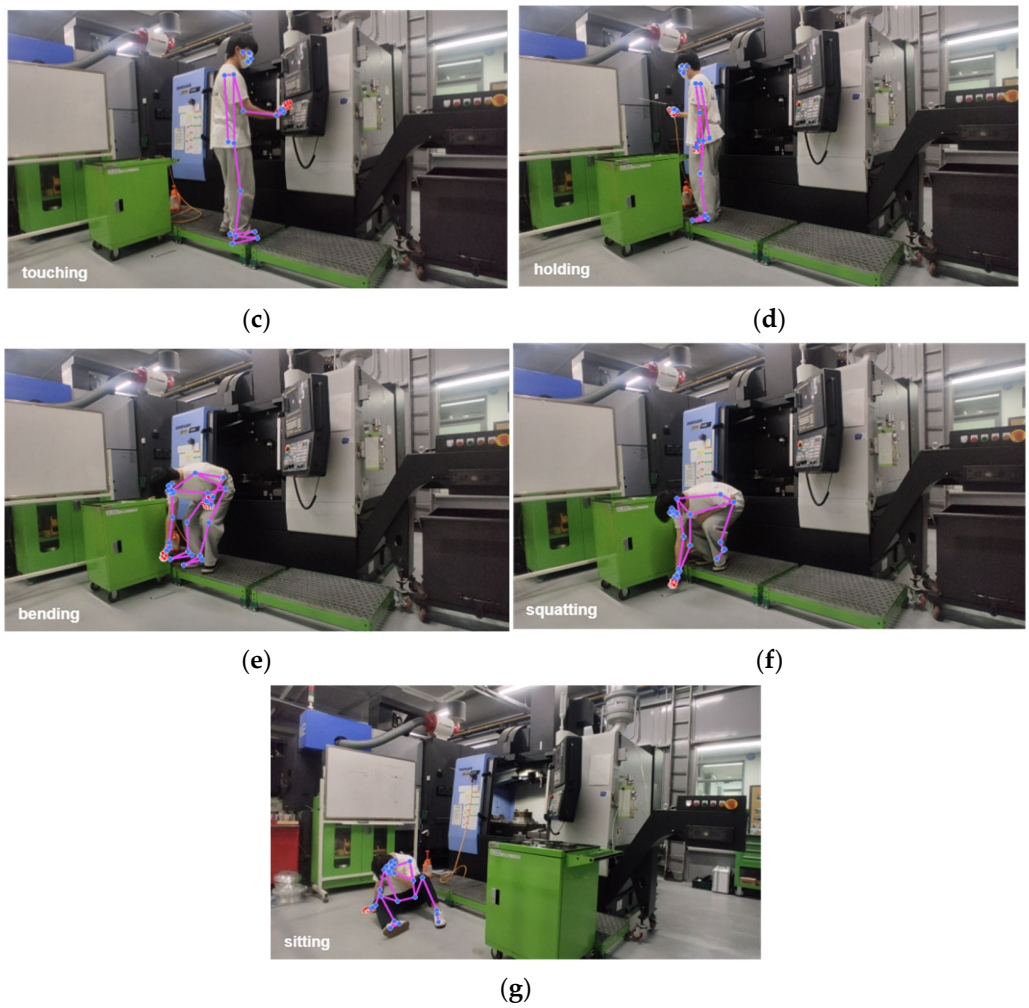


Figure 5. Examples of worker pose classification conducted per frame: (**a**) standing, (**b**) walking, (**c**) holding, (**d**) touching, (**e**) bending, (**f**) squatting, and (**g**) sitting.

Additionally, adverbs were used to express the worker's position relative to other objects based on the camera's viewing angle. This process was conducted by encoding the object into a feature vector among the objects in the frame. Thereafter, the language model adopted vectors to generate descriptive texts, as illustrated in Figure 6—("*" is indicated as the beginning of the sentence structure). The proposed algorithm that describes worker behavior is presented in Algorithm 1, while the detection and labeling results are in Figure 7 (normal and anomalous conditions).

5.2. Performance Evaluation

In evaluating the performance, assessing accuracy was conducted on the proposed model's object detection and human pose identification. Table 2 presents two experimental results for the object detection accuracy, precision, and recall (with 1000 and 12,000 datasets). When implementing 1000 datasets in real time, the model produced an approximate average accuracy of 96%. On the other hand, when implementing 12,000 data, the average accuracy was approximately 97%. The results suggest that employing both datasets provides insignificantly different results. Additionally, Figure 8 depicts a confusion matrix of six objects: worker, machine, control box, toolbox, tool, and product.

Algorithm 1: Worker behavior description

Input: Frame (f(t)), Object_label, Bbox, Pose_label
Output: Descriptive text of worker behavior (normal/anomalous behavior)

Generate Human–Object {
    f(t), object_label, bbox, pose_label Associate each pose_label, object_lable, and Bbox per frame f(t) (distance threshold)
    Perform action recognition: object_label[n], pose_label[n], distance[i][j], sentance_list
    return human_object_interaction }
Generate descriptive_text {
    F(t), human_object_interaction, list_behavior
    Compare the interaction (human_object_interaction, list_behavior) per frame (f[(t − 1), (t), (t + 1)])
    Analyze the context and flag_clasification (normal/anomalous).
      if flag_clasification = True:
    Descriptive_text1 ← "Anomalous Behavior"
      else:
        Descriptive_text1 ← "Normal Behavior"
    return Text (Descriptive text of worker behavior, normal/anomalous })

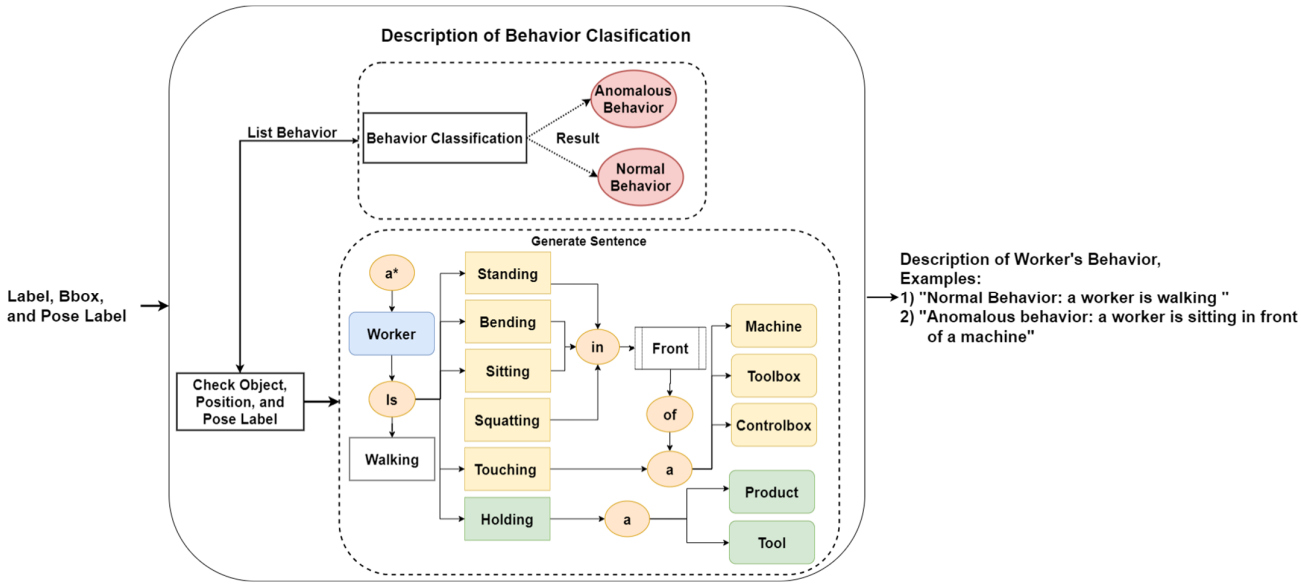


Figure 6. Behavior classification and recognition toward the relation between workers and objects in creating descriptive texts.

Table 3 details the model's performance on human poses in terms of accuracy, precision, and recall from 21 to 31 videos. In this case, the video contains a total of 39,005 frames and 37 anomalous events: sitting, bending in front of the machine, and when the worker is not seen in the frame. For training and testing, when implementing 21 videos, the model achieved satisfying accuracy rates of 94% and 93%, respectively. Moreover, when implementing 31 videos, the model achieved a similarly satisfying accuracy of approximately 95% and 94%, respectively; in addition, no overfitting was found. Furthermore, the model successfully performed the highest accuracy for standing, touching, and holding poses (approximately 95% for 21 videos and 96% for 31 videos). Out of the overall poses, when implementing 21 and 31 videos, the sitting pose performed the lowest accuracy of, approximately 88% and 90%, respectively. For a clearer description, the confusion matrix of the seven poses is as follows: standing, walking, touching, holding, bending, sitting, and squatting (illustrated in Figure 9).

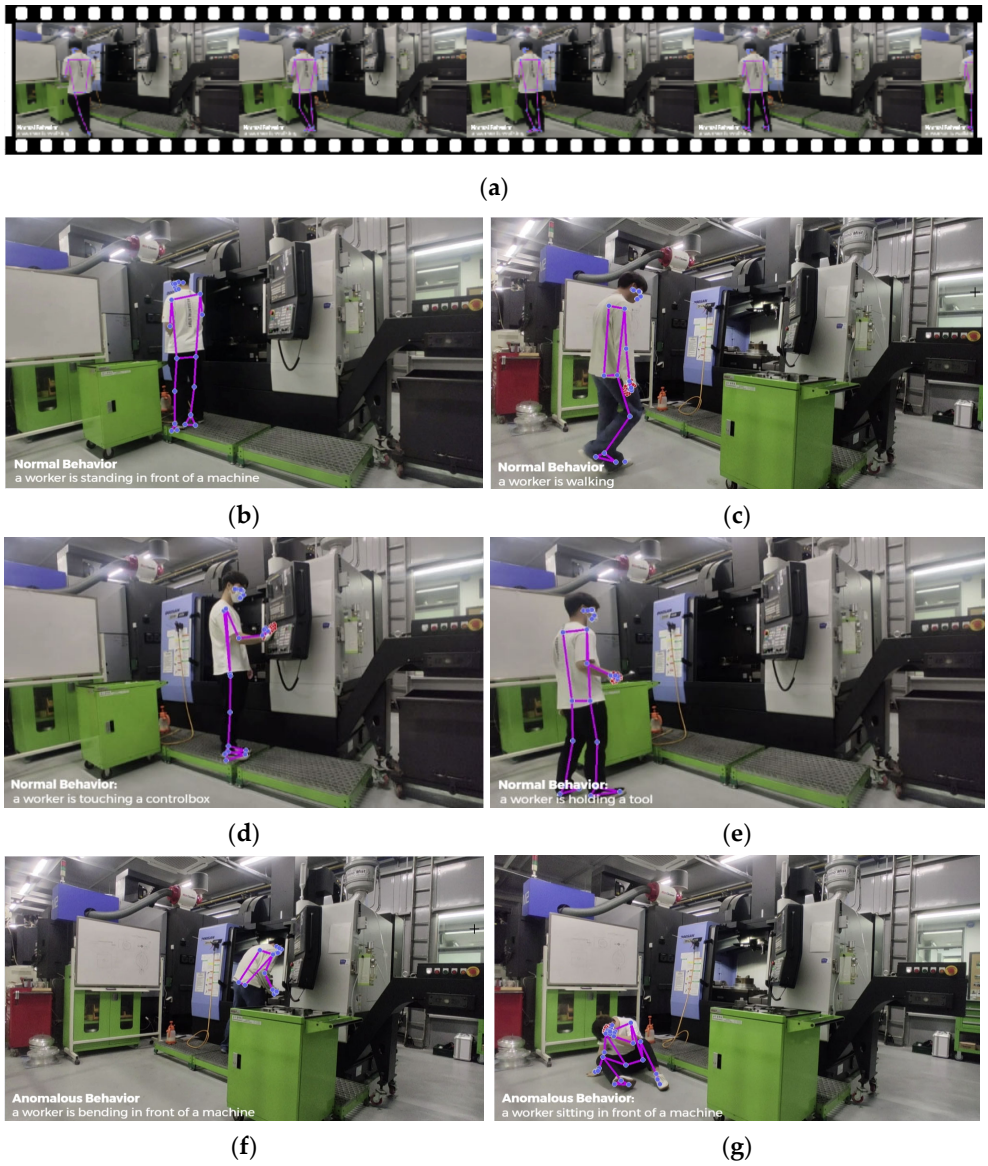
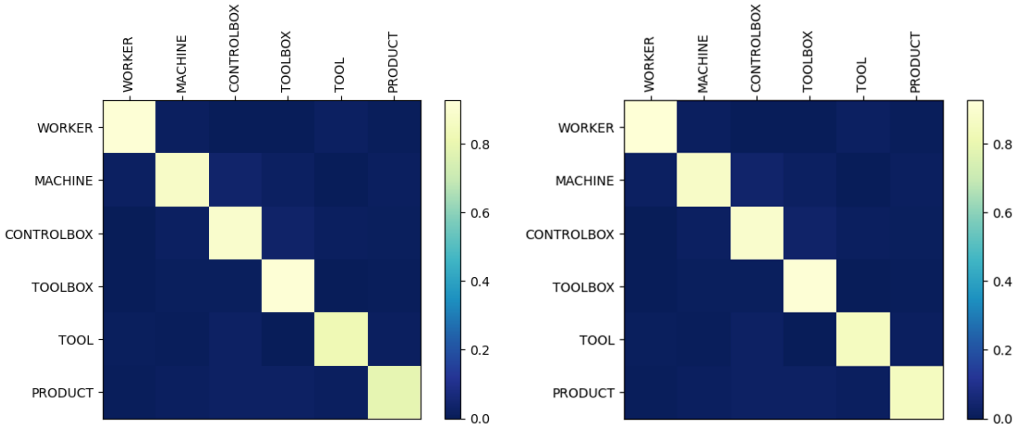


Figure 7. Examples of worker behavior and description (normal/anomalous) results within frames. Normal behavior: (**a**) the result in the form of a video frame, (**b**) a worker is standing in front of a machine, (**c**) a worker is walking, (**d**) a worker is touching the control box, and (**e**) a worker is holding a tool. Anomalous behavior: (**f**) a worker is bending in front of a machine and (**g**) a worker is sitting in front of a machine.

Table 2. Evaluation of object detection performance.

Detection	Accuracy (1000 Dataset)	Accuracy (12,000 Dataset)	Precision (1000 Dataset)	Precision (12,000 Dataset)	Recall (1000 Dataset)	Recall (12,000 Dataset)
Worker	0.98	0.99	0.9830	0.9912	0.9734	0.9817
Machine	0.96	0.97	0.9719	0.9732	0.9617	0.9751
Control box	0.95	0.96	0.9464	0.9621	0.9470	0.9587
Toolbox	0.98	0.98	0.9851	0.9878	0.9722	0.9772
Tool	0.94	0.95	0.9472	0.9553	0.9419	0.9454
Product	0.93	0.94	0.9353	0.9425	0.9221	0.9533

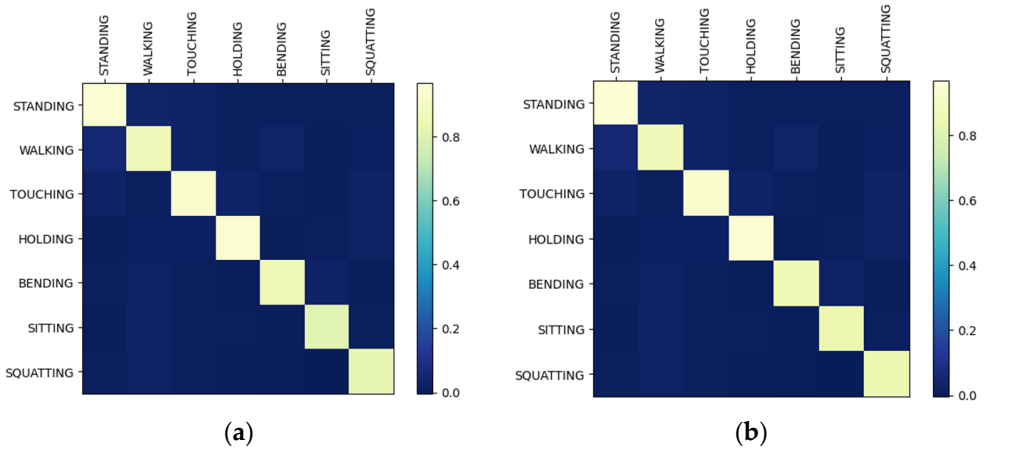


(**a**) (**b**)

Figure 8. Confusion matrix for object detection: (**a**) the result of using 1000 dataset, and (**b**) 12,000 dataset.

Table 3. Evaluation of human pose recognition performance.

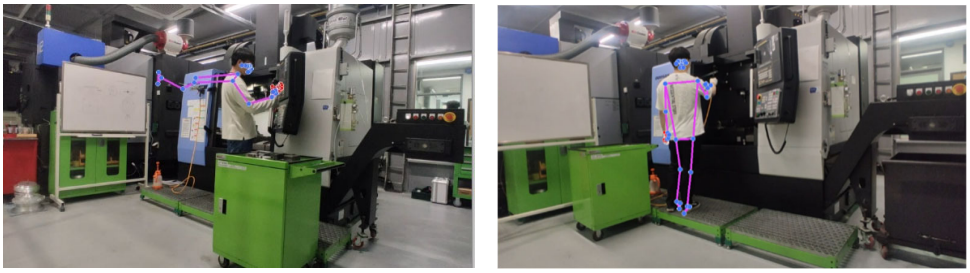
Pose Label	Accuracy (21 Videos)	Accuracy (31 Videos)	Precision (21 Videos)	Precision (31 Videos)	Recall (21 Videos)	Recall (31 Videos)
Standing	0.95	0.96	0.9530	0.9646	0.9534	0.9644
Walking	0.94	0.95	0.9319	0.9534	0.9257	0.9512
Touching	0.95	0.96	0.9564	0.9581	0.9582	0.9634
Holding	0.95	0.96	0.9551	0.9614	0.9522	0.9548
Bending	0.94	0.94	0.9472	0.9495	0.9417	0.9432
Sitting	0.88	0.90	0.8751	0.8984	0.8619	0.9021
Squatting	0.89	0.90	0.8851	0.9058	0.8717	0.9049



(**a**) (**b**)

Figure 9. Confusion matrix for worker pose recognition: (**a**) result of 21 videos, and (**b**) result of 31 videos.

Continuing to the next point, the model was set to 100,000 iterations to avoid overfitting, and the accuracy was approximately 93% for 21 videos and 94% for 31 videos. However, analysis results presented the possibility of incorrectly classifying a frame despite successfully classifying the anomaly, which indicated low false negatives. For example, the landmarks generated by the model did not align with the landmarks of the human body (Figure 10a). The hand landmarks were not correctly identified as covered by the body landmarks (Figure 10b), thus, leading to incorrect classification. However, despite the misalignment, the model successfully labeled the anomalous behavior.



(**a**) (**b**)

Figure 10. Examples of output snapshots with incorrect classification: (**a**) incorrect alignment of positions of overall body landmarks and (**b**) failure to identify hand landmarks as being covered by body landmarks.

6. Conclusions

To resolve problems in manufacturing industries, this study provides a model that detects anomalous behavior in manufacturing environments. The focus is to classify anomalous worker behavior to subsequently label the behavior by utilizing descriptive texts. To provide the expected results, relevant scenarios have been conducted that encompass recognizing the behavior from object detection and human pose based on a threshold. Afterward, classifying the anomalies related to human pose recognition (emergency: worker falls, slippage, or illness) and anomalies related to the recognition of human poses with object positions (tool breakage and machine failure). Next, the worker behavior within the manufacturing environment was classified as normal or anomalous behavior using a worker behavior description algorithm. In this case, the proposed model successfully detected anomalous behavior within the tested videos consisting of anomalous worker behavior and unexpected worker interaction with surrounding objects, while effectively preserving the target with an acceptable object detection accuracy of approximately 97% and the highest pose recognition accuracy of approximately 96% (for standing, touching, and holding); the lowest accuracy was approximately 88% (for the sitting pose). The results did not exhibit significant differences, and the gap was no more than one percent. In fact, some poses were incorrectly detected. For a more detailed example, in one case, the model could not properly align the positions of the overall body landmarks, thus leading to landmark misalignment. In a different case, the hand landmarks were not identified as covered by the body landmarks.

Therefore, a betterment of the current work is required in future studies by expanding requirement scenarios on abnormal real-time localization in terms of target tracking from multiple-target anomalous behavior to detect anomalies. In this particular context, the studies are expected to enhance the scope from individual workers to interactions between multiple workers and objects. The goal is to provide a more comprehensive understanding of the manufacturing environment by adding larger and more diverse datasets to improve the performance. In addition, implementing and deploying the model in real-time monitoring systems are required to create practical and useful models in real-world manufacturing environments. These practices involve optimizing the model's computational efficiency, ensuring low latency, and developing user-friendly interfaces for workers and supervisors to interpret and respond to anomalous alerts effectively. Moreover, advanced Mask R-CNN versions are feasibly pre-trained to enhance object detection.

Author Contributions: Conceptualization, R.R., M.H. and K.J.; methodology, R.R., M.H. and K.J.; software, R.R.; validation, R.R., M.H. and K.J.; formal analysis, R.R., M.H. and K.J.; investigation, R.R., M.H. and K.J.; resources, K.J.; data curation, R.R.; writing—original draft preparation, R.R., M.H. and K.J.; writing—review and editing, R.R. and K.J.; visualization, R.R.; supervision, K.J.; project administration, K.J.; funding acquisition, K.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2016-0-00318), supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Forkan, A.R.M.; Montori, F.; Georgakopoulos, D.; Jayaraman, P.P.; Yavari, A.; Morshed, A. An Industrial IoT Solution for Evaluating Workers' Performance via Activity Recognition. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 1393–1403.
2. Daud, M.M.; Saad, H.M.; Ijab, M.T. Conceptual Design of Human Detection via Deep Learning for Industrial Safety Enforcement in Manufacturing Site. In Proceedings of the 2021 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS 2021), Shah Alam, Malaysia, 26–26 June 2021; pp. 369–373.
3. Han, Y.; Yu, D.; Han, F.; Liu, Y.; Zhao, Q. Industrial APP Design for Data Dynamic Monitoring of Strip Production Process. In Proceedings of the 2019 Chinese Control and Decision Conference (CCDC), Nanchang, China, 3–5 June 2019; pp. 2865–2869.
4. Savitha, C.; Ramesh, D. Motion Detection in Video Surveillance: A Systematic Survey. In Proceedings of the IEEE 2nd international conference on inventive systems and control (ICISC), Coimbatore, India, 19–20 January 2018.
5. Statue–Serious Accident Punishment Act. Available online: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%A4%91%EB%8C%80%EC%9E%AC%ED%95%B4%EC%B2%98%EB%B2%8C%EB%93%B1%EC%97%90%EA%B4%80%ED%95%9C%EB%B2%95%EB%A5%A0 (accessed on 11 May 2023).
6. Kang, S.; Kim, M.; Kim, K. Safety Monitoring for Human Robot Collaborative Workspaces. In Proceedings of the 2019 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 15–18 October 2019.
7. Wang, S.; Yu, S.; Wang, H.; Wu, D.; Zhou, W.; Luo, H. Research and Design of Human Behavior Recognition Method in Industrial Production Based on Depth Image. In Proceedings of the 2022 4th International Conference on Industrial Artificial Intelligence (IAI), Shenyang, China, 24–27 August 2022; pp. 1–6.
8. Bokrantz, J.; Skoogh, A.; Ylipää, T.; Stahre, J. Handling of Production Disturbances in the Manufacturing Industry. *J. Manuf. Technol. Manag.* **2016**, *27*, 1054–1075, ISSN: 1741-038X. [CrossRef]
9. Yang, L.; Su, Q.; Shen, L. A Novel of Analyzing Quality Defect Due to Human Error in Engine Assembly Line. In Proceedings of the 2012 International conference on International Management, Innovation Management and Industrial Engineering, Sanya, China, 20–21 October 2012.
10. Addanke, S.; Krishna, M.V.; Pradeep, K.V.; Jayant, K.P.; Ilyassova, K.; Sainath, K.L. Machine Learning on the Role of Eliminating Human Error on the Manufacturing Industry. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022.
11. Kaiming, H.; Georgi, G.; Piotr, D.; Ross, G. Mask r-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
12. Shetye, S.; Shetty, S.; Shinde, S.; Madhu, C.; Mathur, A. Computer Vision for Industrial Safety and Productivity. In Proceedings of the 2023 International Conference on Communication System, Computing and IT Applications (CSCITA), Mumbai, India, 31 March 2023; pp. 117–120.
13. Fan, S.; Zhang, L.; Wang, J.; Wang, Y.F.; Zhang, Q.S.; Zhao, H.; Fault, V.-B. Classification for Monitoring Industrial Robot. In Proceedings of the 2018 37th Chinese Control Conference, CCC, Wuhan, China, 25–27 July 2018.
14. Wenjuan, G.; Xuena, Z.; Jordy, G.; Andrews, S.; Thierry, B.; Changhe, T.; El-Hadi, Z. Human Pose Estimation from Monocular Images: A Comprehensive Survey. *Sensors* **2016**, *16*, 1966.
15. Rijayanti, R.; Hwang, M.; Jin, K.H. Extraction of Worker Behavior at the Manufacturing Site Using Mask R-CNN and Dense-Net. In Proceedings of the KIICE Spring Conference, Busan, Republic of Korea, 26–28 May 2022.
16. Rijayanti, R.; Hwang, M.; Jin, K.H. Worker behavior Identification from Manufacturing Site Images using Mask R-CNN and MediaPipe. *Korean Inst. Inf. Commun. Eng.* **2023**, *27*, 281–290.
17. Deep, S.; Tian, Y.; Lu, J.; Zhou, Y.; Zheng, X. Leveraging Multi-view Learning for Human Anomaly Detection in Industrial Internet of Things. In Proceedings of the Physical and Social Computing, Rhodes, Greece, 2–6 November 2020; pp. 533–537.
18. Lindemann, B.; Jazdi, N.; Weyrich, M. Anomaly Detection and Prediction in Discrete Manufacturing Based on Cooperative LSTM Networks. In Proceedings of the 2020 IEEE 16th International Conference on Automation Science and Engineering (Case), Hong Kong, China, 20–21 August 2020; pp. 1003–1010.
19. Kamoona, A.M.; Gostar, A.K.; Tennakoon, R.; Bab-Hadiashar, A.; Accadia, D.; Thorpe, J.; Hoseinnezhad, R. Random Finite Set-Based Anomaly Detection for Safety Monitoring in Construction Sites. *IEEE Access* **2019**, *7*, 105710–105720. [CrossRef]

20. Baydargil, H.B.; Park, J.; Ince, I.F. Unsupervised Anomaly Approach to Pedestrian Age Classification from Surveillance Cameras Using an Adversarial Model with Skip-Connections. *Appl. Sci.* **2021**, *11*, 9904. [CrossRef]
21. Bouindour, S.; Snoussi, H.; Hittawe, M.M.; Tazi, N.; Wang, T. An On-Line and Adaptive Method for Detecting Abnormal Events in Videos Using Spatio-Temporal ConvNet. *Appl. Sci.* **2019**, *9*, 757. [CrossRef]
22. Parvin, P.; Paternò, F.; Chessa, S. Anomaly Detection in the Elderly Daily Behavior. In Proceedings of the 2018 14th International Conference on Intelligent Environments (IE), Rome, Italy, 5–28 June 2018; pp. 103–106.
23. Esan, D.O.; Owolawi, P.A.; Tu, C. Detection of Anomalous Behavioral Patterns in University Environment Using CNN-LSTM. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–8.
24. Hittawe, M.M.; Afzal, S.; Jamil, T.; Snoussi, H.; Hoteit, I.; Knio, O. Abnormal events detection using deep neural networks: Application to extreme sea surface temperature detection in the Red Sea. *J. Electron. Imaging* **2019**, *28*, 0210121–0210128. [CrossRef]
25. Hittawe, M.M.; Langodan, S.; Beya, O.; Hoteit, I.; Knio, O. Efficient SST prediction in the Red Sea using hybrid deep learning-based approach. In Proceedings of the 2022 IEEE 20th International Conference on Industrial Informatics (INDIN), Perth, Australia, 25–28 July 2022; pp. 107–117.
26. Jebur, S.A.; Hussein, K.A.; Hoomod, H.K.; Alzubaidi, L.; Santamaría, J. Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. *Electronics* **2023**, *12*, 29. [CrossRef]