Received 26 July 2023, accepted 16 September 2023, date of publication 22 September 2023,
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318114

RESEARCH ARTICLE

Deep Learning-Based Recognition of Unsafe Acts in Manufacturing Industry

ARSO M. VUKICEVIC[1], MILOS N. PETROVIC[2], (Graduate Student Member, IEEE), NIKOLA M. KNEZEVIC[2], (Graduate Student Member, IEEE), AND KOSTA M. JOVANOVIC[2], (Member, IEEE)

[1]Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia
[2]School of Electrical Engineering, University of Belgrade, 11000 Belgrade, Serbia

Corresponding author: Arso M. Vukicevic (arso_kg@yahoo.com)

This work was supported in part by the Science Fund of the Republic of Serbia, Modular and Versatile Collaborative Intelligent Waste Management Robotic System for Circular Economy–CircuBot, under Grant 6784; and in part by the AI4WorkplaceSafety under Grant 6524219.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Faculty of Medicine, University of Belgrade, under Application No.1322/X-42.

ABSTRACT Despite technological progress and the tendency for automation, the majority of manufacturing workplaces still rely on human labor. Although industrial tasks are frequently composed of simple operator actions, non-ergonomic execution of such repetitive tasks has been reported as the primary cause of musculoskeletal disorders. Considering the sizes of manufacturing halls and large numbers of employees, there is an increasing need for tools that can improve the recognition of unsafe acts. Herein, a deep learning-based procedure for pose safety assessment is proposed and validated using monocular videos captured with a conventional IP camera. The two key composing components of the proposed pipeline are the three-dimensional (3D) pose estimator and mesh classifier. The proposed method was validated experimentally by considering three different methodologically selected industrial tasks: a laborious task that requires all-body effort (pushing and pulling), a task that requires an upper-limb action comprising intensive interaction and motion control (drilling), and a typical collaborative task (polishing with a collaborative robot with variable mechanical impedance). Accuracies of 84.67%, 92%, and 98%, respectively, were achieved. Besides higher accuracy, the proposed method has shown practical advantages over existing alternatives based on analyzing the parameters derived from the human poses. Particularly, we report that the proposed procedure is generic, and it works directly with 3D human body poses, which significantly increases applicability while reducing the complexity and effort needed for data annotation and output interpretation by non-experts.

INDEX TERMS Artificial intelligence, human-centric industry 5.0, industrial engineering, workplace safety.

I. INTRODUCTION

Although automation has brought tremendous progress in many manufacturing sectors (automotive, electronics manufacturing, welding, etc.) over the past decades, the practice has shown that there are still many tasks and workplaces that cannot be adequately or fully automated. As an alternative, there is a trend of supplementing laborious manufacturing workplaces with robots that can collaborate with human operators to reduce the physical load and enhance operator performance [1]. Furthermore, collaborative robot technologies are underexploited in the industry because a reliable and efficient understanding of human intention prediction in collaborative tasks is lacking.

The awareness of human roles in Industry 4.0 is increasing and authorities from both academia and industry have agreed

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

that Industry 5.0 will be highly human-centered [2]. Hence, there is an increasing need to improve the interface between human operators and technology while ensuring the highest standards of workplace safety and the well-being of human operators in an industrial environment.

At the international level, workplace safety is regulated by Occupational Safety and Health (OSH). OSH currently covers a wide range of topics, and the overall goal is to improve safety by reducing the occurrence of workplace hazards, injuries, illnesses, and fatalities. Together with the Hierarchy of Controls, proposed by the Occupational Safety and Health Administration [5], optimizing workplaces for ergonomics is a major preventive measure for injuries in the industry. Ergonomics aims to adapt the workplace to complement human capabilities and minimize discomfort, physical exertion, stress, and risk of injury due to work.

Despite the application of preventive measures, work-related musculoskeletal disorders (WMSD) occur frequently [6]. The timely detection of unsafe postures that lead to WMSD is challenging. In contrast to, for example, the misuse of personal protective equipment or unsafe conditions that can be instantly detected and mitigated [7], [8], WMSD must be considered as the long-term accumulation of negative effects caused by repetitive unsafe acts [9]. Studying and preventing the repetition of unsafe acts is important from various perspectives. In addition to forced retirement and job loss, WMSD have negative long-term consequences such as permanent disabilities and the inability to perform everyday activities (which have negative sociological and economic impacts) [10]. According to World Health Organization reports, approximately 1.71 billion people have musculoskeletal conditions worldwide [11]. Reports indicate that 33% of European workers (42% of male workers and 24% of female workers) have unnatural body postures for >25% of their working time [12]. Approximately 126.6 million Americans - one in two adults have a musculoskeletal disorder, costing an estimated $213 billion in annual treatment, which is 1.4% of the United States' gross domestic product [13].

The critical role in ensuring the implementation of safety and ergonomic recommendations is accomplished by onsite managers (supervisors), who are also responsible for the monitoring of workers. Considering the sizes of manufacturing halls and the number of employees that move across halls, manual supervision of workers is expensive, time-consuming, and ineffective. Thus, there is an increasing need for computerized tools that can ease or automate the process of detecting unsafe acts in industrial environments. Accordingly, the aim of this study was to develop a computer vision procedure for enabling the recognition of unsafe acts in industrial environments by fusing algorithms for pose estimation and shape analysis.

The rest of the paper is organized as follows. Section II gives an overview of related work. The experimental setups and the proposed procedure for unsafe acts recognition are presented in Section III, followed by the results in Section IV. The obtained results are discussed and compared to the previous studies in Section V. Lastly, the paper ends with concluding remarks and future work perspectives in Section VI.

II. BACKGROUND RESEARCH

Recognition of unsafe acts represents a visual task. It is a well-studied topic in the field of computer vision that has rapidly evolved with the advancement of deep learning [14]. In this section, we review studies in which computer-vision techniques or sensor technologies were used to assess workplace safety or recognize unsafe acts.

To analyze and optimize ergonomics, Kim et al. utilized wearables (electromyography) and motion-capture systems [15]. To identify physical risk factors, in addition to a detailed human model and overloading joint torques, the same group of authors introduced compressive forces and a fatigue index [16], [17]. Han and Lee proposed a framework for detecting unsafe actions on construction sites (ladder climbing) using three-dimensional (3D) landmark points (reconstructed from two-dimensional (2D) human skeleton points) and motion templates to detect several climbing actions through pattern recognition [18], [19]. Another approach for safety assessment of ladder climbing tasks was presented that aimed to achieve real-time unsafe-behavior recognition using videos, considering dynamic behavior as a static posture, and using a mathematical model of the human skeleton to identify unsafe behaviors according to value ranges of joint parameters [20], [21]. In previous studies, safety analysis of industrial workers was performed by considering ergonomic parameters obtained from reconstructed 3D body poses from monocular videos [22]. Ding et al. developed a hybrid model that integrated a convolutional neural network (CNN) and long short-term memory (LSTM) for offline detection of safe and unsafe actions conducted by workers from stereo videos [23]. Seo et al. offered a comprehensive review of computer vision-based systems for object detection, object tracking, and action recognition to identify unsafe acts and violations of safety and health rules on construction sites by using 2D and 3D images [24]. To monitor and automatically assess worker activities by integrating red-green-blue (RGB), optical flow, and gray stream CNNs, an improved CNN was proposed [25]. Fang et al. developed an algorithm using a faster region-based CNN for detecting the presence of construction workers and a deep CNN for determining whether they are wearing a safety harness to prevent falls from heights [26]. An individualized system based on CNN classifiers and the Weighted Average of Selected Classifiers (WASC) was presented [27]. It uses human skeletal data for real-time action recognition of assembly workers and correcting mistakes, thus enhancing task performance. Ciccarelli et al. developed SPECTRE (Sensor-independent Parallel dEep ConvoluTional leaRning nEtwork), a learning model that uses CNN to classify human postures in the workplace and assess ergonomic risks in body segments [28].

As an alternative to computer vision, there are various approaches for recognizing human activities using sensor data. Yang et al. proposed a method in which a penalized naive Bayes classifier and radiofrequency identification (RFID) sensor data are used to classify human activities in known classes [29]. Hofmann et al. used LSTM for recognizing wasteful activities in production [30]. The considered activities were walking, sitting, standing, and jogging, and the reported accuracies >98%. Ordóñez et al. proposed the DeepConvLSTM framework (based on convolutional and LSTM recurrent units), which was reported to be suitable for multimodal wearable sensors [31].

Unsafe acts have a wide variety of forms and may depend on the nature of a particular task or environment. Although there are recommendations that cover, for example, minimum safety and health requirements for the workplace (89/654/EEC [32]), manual handling of heavy loads (90/269/EEC [33]), manual handling of pushing and pulling (ISO 11228-2 [34]), and collaborative robots (ISO/TS 15066 [35]), these directives and standards are too broadly defined for application to highly specific workplaces. Thus, their main purpose is to help practitioners optimize workplaces, while safety monitoring for unsafe acts relies on the subjective judgment of onsite supervisors.

However, industrial applications of computerized procedures from the reviewed literature remain limited, as they commonly represent empirical models developed in laboratory conditions (which differ from manufacturing halls). In many studies, Kinect [19], [27] or motion capture system [15] was used to assess body pose and shape, which gives more accurate results in laboratory settings but is not suitable for application in big industrial halls because of its complexity and price. Even though sensor systems for human-activity recognition can successfully classify activities [29], they cannot identify some subtle differences - therefore, this kind of system cannot determine whether someone is conducting a task correctly. An additional limitation of existing procedures is that they frequently rely on 2D pose estimation, which suffers from body-part occlusions, the inability to ensure the optimal viewpoint, etc. Moreover, the proposed body parameters derived from 2D [18] or 3D [19] poses can be difficult for non-experts to interpret in industry practice.

The contribution of this research is a generic two-step procedure for unsafe acts recognition that solves the aforementioned problems. It contributes to the state-of-the-art by presenting an innovative approach that: 1) utilizes videos of arbitrary unsafe acts (captured with conventional devices, e.g., smartphones) and 2) employs realistic 3D silhouettes of a body as inputs/outputs so that analysis and interpretation of pose safety are intuitive for a broad audience. The proposed method is validated on several methodologically selected industrial tasks. For all considered tasks, we assumed that one IP camera is enough to cover the operator's movements so that there is no need for tracking operator through the manufacturing hall. From the end-user viewpoint, the workflow of the proposed solution has three steps: 1) collection of videos that correspond to safe and unsafe execution of the considered task; 2) training of the pose classifier; 3) obtaining the information about the operator's pose safety by inferencing the pipeline on corresponding hardware (that consists of IP camera connected to PC with GPU).

III. MATERIALS AND METHODS

A. EXPERIMENTS AND DATA ACQUISITION

To demonstrate the generality and versatility of the proposed approach, three industrial tasks were selected to cover the entire body and specific body parts, as well as different interaction dynamics: 1) pushing and pulling (P&P) of handcarts, 2) two-handed drilling, and 3) polishing with a collaborative robot. Such diverse tasks were selected to cover the safety challenges of various industries, as workplaces vary from static to dynamic and from collaborative to manual. Collection and processing of data used in this study were done in accordance with the relevant guidelines proposed with the Declaration of Helsinki - after obtaining ethical approval from the Faculty of Medicine, University of Belgrade, Serbia (Approval No. 1322/X-42) and informed consent from all participants. All experiments were designed to replicate experiments found in the literature in an industrial environment and involved five healthy male participants (age: 29.7 ± 4.7 years; body mass: 80.1 ± 10.2 kg; height: 1.77 ± 0.1 m) with no prior experience working in the industry and with no previous health conditions related to WMSD or workplace injuries. Each task was conducted under controlled conditions and repeated multiple times, with each repetition lasting 2-5 min. The experiments were guided Workplace safety and ergonomics experts guided the experiments, ensuring that the acquired video sequences involved both safe and unsafe actions and body postures. The first task - handcart P&P (Fig. 1a) was performed using a dedicated experimental handcart (100 kg), as described in previous work [21]. The second task - standing two-handed drilling (Fig. 1a) was performed in a vertical plane using a power tool. The third task - collaborative polishing (Fig. 1a) was performed with the support of Franka Emika Panda Robot, controlled in the impedance mode with its end-effector stiffness set as 1500 N/m along all translational axes and 100 Nm/rad along all rotational axes, as described in detail in previous work [37]. The experiments were recorded using four DAHUA IPC-HFW2831TP-ZS 8MP WDR IR Bullet IP cameras, with a DAHUA PFS3010-8ET-96 8port Fast Ethernet PoE switch. The host PC had an 1151 Intel Core i3-8100 3.6-GHz 6-MB BOX CPU. From the collected videos, we extracted the 3D poses of the operators using the algorithms described in Section 2.2 (Fig. 1b). By observing the obtained 3D meshes and video streams, safety experts selected 500 safe and 500 unsafe poses for each task (Fig. 1c). These datasets were used to train the mesh classifiers.

B. CLASSIFICATION OF SAFE AND UNSAFE ACTS

The proposed procedure for recognizing unsafe acts is presented in Fig. 1. The key components of the pipeline are 1) database creation and annotation (incorporation of experts'

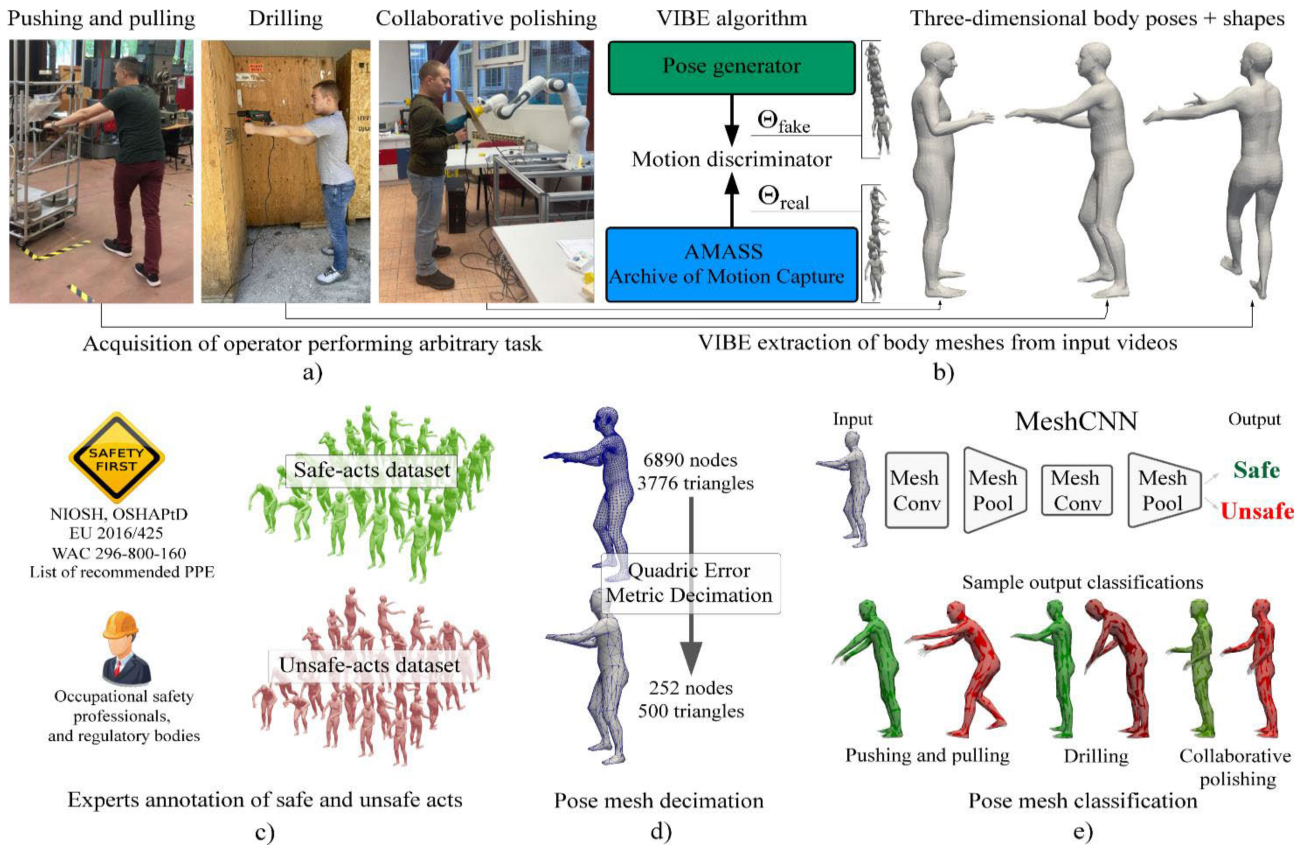


FIGURE 1. Considered industrial tasks and an overview of the proposed procedure.

domain knowledge needed for recognition of unsafe acts in a particular workplace), 2) 3D human pose reconstruction, 3) mesh decimation, and 4) mesh classification. The first two steps are interconnected and rely on a deep-learning module to reconstruct human body shapes from monocular videos.

The VIBE architecture is used to solve the 3D pose reconstruction problem in an adversarial manner [38]. It reconstructs 3D actions from videos using previously detected 2D body landmark points (a pre-trained Mask R-CNN keypoint detector is employed) [39]. The pose-generator part $\hat{\Theta}$ is trained to compete with the discriminator part D_M until the discriminator cannot identify the difference between the obtained motion and realistic motion(s) from the AMASS database (~11000 movements of ~300 subjects) [40]. The use of temporal encoder results in reliable pose estimations because the algorithm makes reconstructions by considering a series of neighboring frames, which is important to compensate for glitches and errors caused by the temporary overlapping of particular body parts. The resulting body shape of the generator $\hat{\Theta}$ is forwarded to the discriminator's self-attention layers, which are used to increase the influence of the dominant frames. The final linear layer returns the probability that the predicted manifold $\hat{\Theta}$ belongs to the group of realistic human actions by minimizing the loss function, which is composed of four components that account for the 3D pose error, back-projection error, error of body pose and shape, and likelihood that the obtained motion (sequence of 3D poses) corresponds to motions in the AMASS database. The 3D poses were represented using a Skinned Multi-Person Linear (SMPL) body model, which is a parametric surface representation of a human shape composed of 6890 Quad4 elements. Thus, in addition to 3D body landmarks, we obtained body shapes containing anatomical information, facilitating the data annotation for human experts. Finally, the comparison with the state-of-the-art approach is given where we computed a series of parameters from the reconstructed 3D landmark points that were used as ergonomic indicators [22].

Data annotation was performed with the help of three workplace safety and ergonomics experts, whose task was to classify reconstructed body shapes into "Safe acts" and "Unsafe acts" datasets (Fig. 1c). As safety assessment is a subjective task, the experts reached a consensus so that intra-observer and inter-observer variabilities of the data-annotation process were eliminated. By considering the recorded videos, obtained 3D shapes, and corresponding body-pose parameters, we developed datasets comprising 500 safe acts and 500 unsafe acts for each considered task.

To make the system more efficient, we simplified a triangular human body mesh using Quadric Error Metric Decimation [41]. This algorithm iteratively replaces two vertices with one by applying a pair collapse operator, causing the

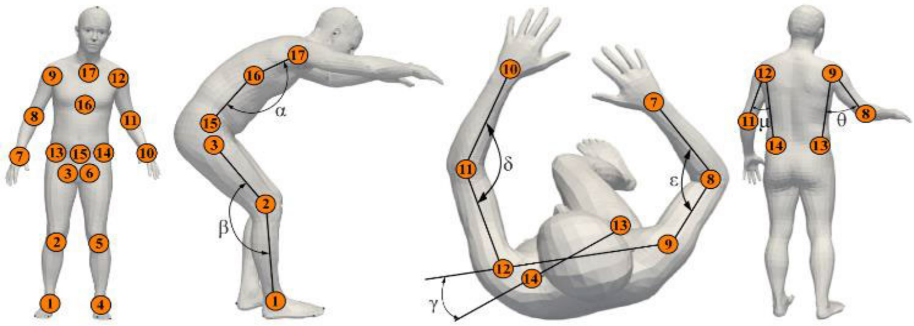


FIGURE 2. SMPL models and considered ergonomic parameters.

neighboring faces to degenerate. The initial output of the VIBE estimation, i.e., the SMPL mesh, had 20666 edges (13776 faces and 6890 vertices), and after applying the decimation, it had approximately 750 edges (Fig. 1d). Simplified meshes were used as inputs for the next stage of the pipeline (pose mesh classifier).

In contrast to previously proposed methods for the classification of unsafe acts using images and videos [26], [42], this study investigated the possibility of using inherent information contained in 3D body meshes. Although CNNs are effective for solving problems represented by regular structures (e.g., images, videos, voxels), extending the paradigm to work with irregular structures (e.g., triangular meshes) is challenging. In this study, the decimated SMPL meshes were classified using the MeshCNN module (Fig. 1e), which was composed of customized convolution and pooling operations designed to work with 3D mesh edges [45]. By accounting for mesh geodesic connections, convolutions processed edges, whereas pooling layers preserved the surface topology through edge collapse. Through successive mesh convolution and pooling, valid mesh connectivities were iteratively generated via learning to discard and collapse redundant features while preserving important ones.

During training, each dataset was randomly split into training (70%), validation (15%), and test (15%) sets. The training was performed using the Adam optimization algorithm [46]. The learning rate of the Adam algorithm was set to 0.0002 and the group norm (g = 20). Data augmentation was performed with 5% edge flips and 15% slide vertices.

IV. RESULTS

All experiments were performed using the Python 3.7.4 programming language, along with PyTorch 1.6.0, Torchvision 0.7.0, and Cuda 10.1 GPU. The computations were performed on a workstation with an AMD Threadripper 3970X (32-core, 3.79 GHz processor) CPU with 128 GB of random-access memory and two Titan RTX (24GB) + NVLink GPUs.

The proposed method was assessed using the following statistical metrics for the binary CNN classification (Table 1): accuracy (percentage of correct classifications, ACC), sensitivity or true positive rate (percentage of true positives classified correctly, TPR), specificity of true negative rate (percentage of true negatives classified correctly, TNR), precision or positive predictive value (percentage of correct positive classifications, PPV), negative predictive value (percentage of correct negative classifications, NPV), and F1 score (harmonic mean of PPV and TPR). For the studied problem, we adopted that unsafe acts belonged to class 1 (positive), and safe acts belonged to class 0 (negative).

TABLE 1. Performance for recognizing safe/unsafe acts in various industry tasks.

	P&P handcarts	Two-handed drilling	Collaborative polishing
ACC	84.67%	92.00%	98.00%
TPR	89.33%	93.33%	98.67%
TNR	80.00%	90.67%	97.33%
PPV	81.71%	90.91%	97.37%
NPV	88.24%	93.15%	98.65%
F1	85.35%	92.11%	98.01%

For comparison with the state-of-the-art approach, from the reconstructed 3D poses, landmark points were used to compute a series of ergonomic parameters. For each task, we computed various parameters, and several of them are presented in Fig. 2: 1) the angle of the spine (15, 16, 17) for P&P handcarts (α); 2) the angle of the right knee (1, 2, 3) (β); 3) the angle of torsion between the shoulders and the hips (13, 14, 12, 9) (γ); 4) the angle of the left elbow (10, 11, 12) (δ); 5) the angle of the right elbow (7, 8, 9) (ε); 6) the angle between the left upper arm and the torso (11, 12, 14) (μ); and 7) the angle between the right upper arm and the torso (13, 9, 8) (θ). Fig. 3 presents the outputs of the proposed procedure - SMPL meshes marked green (safe) and red (unsafe), with the corresponding input data (T is frame number). Fig. 4-6 represents time series of ergonomic parameters extracted from the part of the video sequence for P&P, two-handed drilling, and collaborative polishing, respectively (the x-axis represents the frame number and the y-axis the angle in degrees). In these figures, green and red lines indicate time points of safe and unsafe acts and directly correspond to SMPL meshes presented in Fig. 3.

V. DISCUSSION

The performance metrics of the proposed method are presented in Table 1. As shown, the performance of the

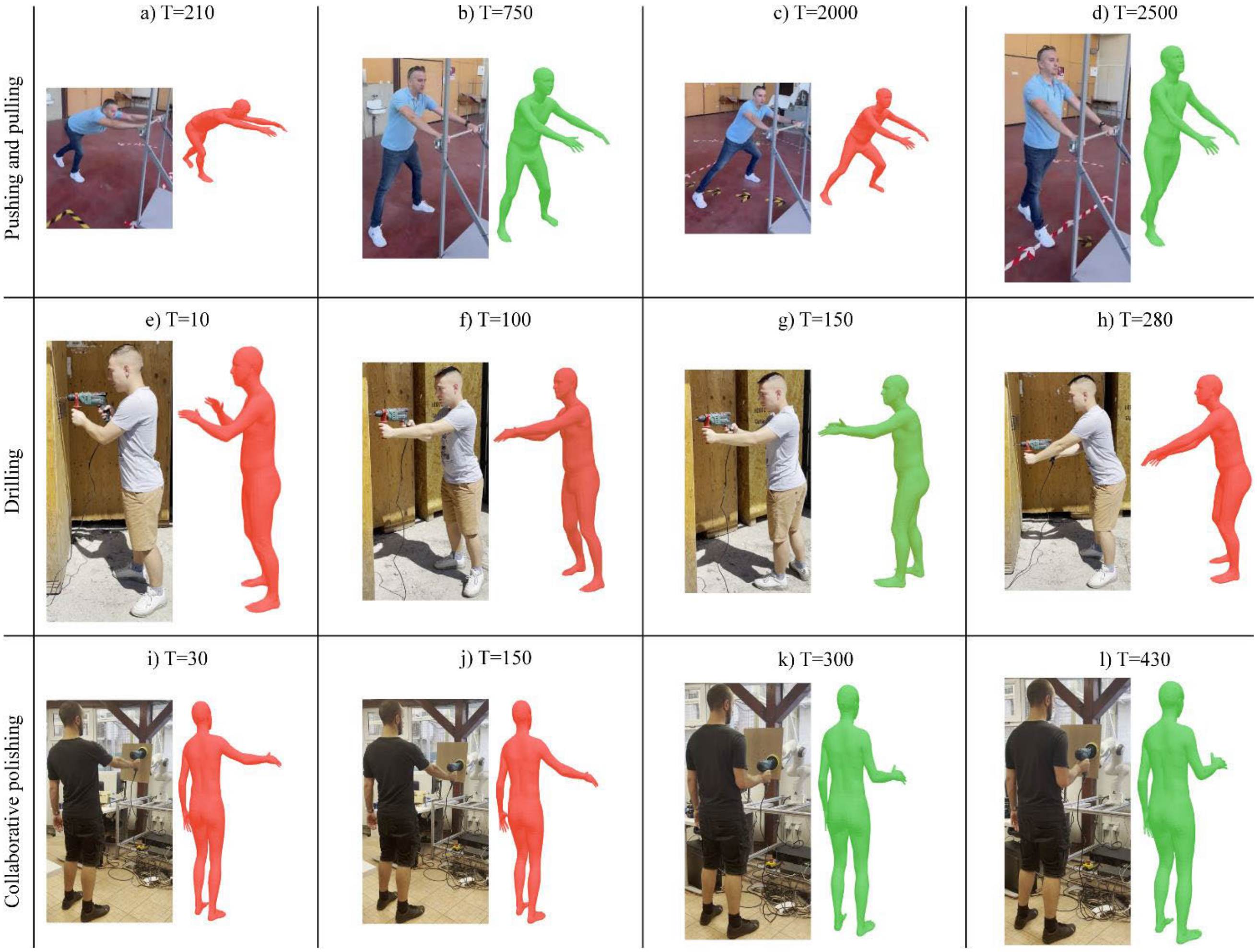


FIGURE 3. Sample inputs and corresponding outputs of the proposed procedure for the three considered industry tasks: (b,d,g,k,l) safe - green; (a,c,e,f,h,i,j) unsafe -red.

procedure was lower when the task involved a larger degree of movement. The most dynamic task (P&P of handcarts) had the lowest ACC, TPR, TNR, PPV, NPV, and F1-score values, whereas the most static task (collaborative polishing) had the highest values. This indicates that the procedure is more suitable in situations where one IP camera is enough to cover the operator's movements and workplace, and where there are a limited amount of movements that one needs to make to accomplish a task. Among the considered metrics, the sensitivity had the highest value, indicating that the proposed procedure is robust for practical applications, as the primary aim is to recognize unsafe acts correctly.

Referring to the high diversity of industrial tasks and the unavailability of appropriate public datasets considered in the literature, a direct comparison of performance metrics between various studies was not doable - although we outperformed them by a considerable margin on the tasks considered in this study. Nevertheless, a comparative overview of literature (which used computer-vision techniques or sensor technologies) for assessing workplace safety or recognition of unsafe acts is presented in Table 2.

In terms of acquisition devices used, only a few relied on the use of conventional monocular cameras [22], [23], [25], [26]. Compared to the previous study focused on analyzing pushing and pulling tasks [22], this study proposes the use of a mesh classification module instead of the traditional analysis of body parameters extracted from 3D human poses.

Fig. 4-6 illustrate the limitations of the existing practice that depends on the assessment of various parameters extracted from 2D or 3D body poses. Following previous studies, we extracted numerous ergonomic parameters that may be used for assessing the ergonomics of different tasks. Interpretation of these ergonomic parameters is shown to be very complex. Even for an expert, it is non-intuitive to make a decision if a particular position is unsafe as several parameters should be considered simultaneously. On the other side, concluding whether an act is safe or unsafe from a 3D human model is considerably more intuitive, as shown in Fig. 3.

TABLE 2. Comparative overview of related studies.

Study	Industry/ Task/ Environment	Method	Aim	Data acquisition	Transfer for studying another task needs domain expertise
[15]	Manufacturing/ Collaborative power tool handling/ Laboratory	Human Whole-Body model with additional sensor measurements	Overloading torque and muscle activity analysis, improving ergonomics	Xsens MVN BIOMECH Suit, Trigno EMG, KistlerForce plate	Yes
[16]	Manufacturing/ Collaborative power tool handling/ Laboratory	Human Whole-Body model with additional sensor measurements	Compressive force analysis, improving ergonomics	Xsens MVN BIOMECH Suit, Trigno EMG, Kistler Force plate	Yes
[17]	Manufacturing/ Collaborative power tool handling/ Laboratory	Calculation of ergonomic indices using Human Whole-Body model	Monitoring kinematic and dynamic quantities of human body	Xsens MVN BIOMECH Suit, Trigno EMG, Kistler Force plate	Yes
[22]	Warehouse/ Pushing and pulling/ Laboratory	Deep learning pose estimation, analysis of motion parameters derived from 3D pose	Analyzing pushing and pulling activities	Monocular cameras	Yes
[18]	N/A/ Ladder climbing/ Industrial environment	3D model reconstructed from 2D skeleton, motion-template pattern recognition	Detection of unsafe actions	Stereo camera system (multiple cameras, 3D camcorder)	Yes
[19]	N/A/ Ladder climbing/ Laboratory	Motion templates pattern recognition	Detection of unsafe actions	Kinect and iPi Soft Motion Capture solution	Yes
[20]	N/A/ Ladder climbing/ Laboratory	Human skeleton mathematical model, joint angle parameters	Unsafe behavior identification	Kinect	Yes
[21]	N/A/ Ladder climbing/ Laboratory	Human skeleton mathematical model, joint angle parameters	Unsafe action identification	Kinect	Yes
[23]	N/A/ Ladder climbing/ Laboratory	CNN and LSTM	Unsafe action detection	Video camera	Yes
[24]	Construction/ Tool and equipment handling/ Construction	Object detection, object tracking, action recognition	Safety and health monitoring	Surveillance, portable and stereo vision cameras, flash LADAR, RGB-D sensor	Yes
[25]	Construction/ Steel banding, transporting and walking/ Construction	CNN	Activity assessment	Monocular cameras	Yes
[26]	Construction/ Construction work/ Construction	Faster R-CNN, CNN	Detection of workers and their harnesses	Monocular cameras	Yes
[27]	Manufacturing/ Assembly task/ Laboratory	CNN and WASC	Action recognition	Kinect	Yes
[29]	N/A	Bayes classifier	Human-activity recognition	RFID sensors	Yes
[30]	N/A	LSTM	Human-activity recognition	Sensor data from smartphone	Yes
[31]	N/A	DeepConvLSTM	Human activity recognition	Triaxial accelerometers, Inertial measurement units	Yes

Body parameter-dependent approaches commonly use human models and consider joint angles as parameters for detecting unsafe acts [20], [21]. Similarly, researchers have recommended safety assessment based on a complex human model obtained from a motion-capture system, and wearable sensor information has been presented [15]. Although

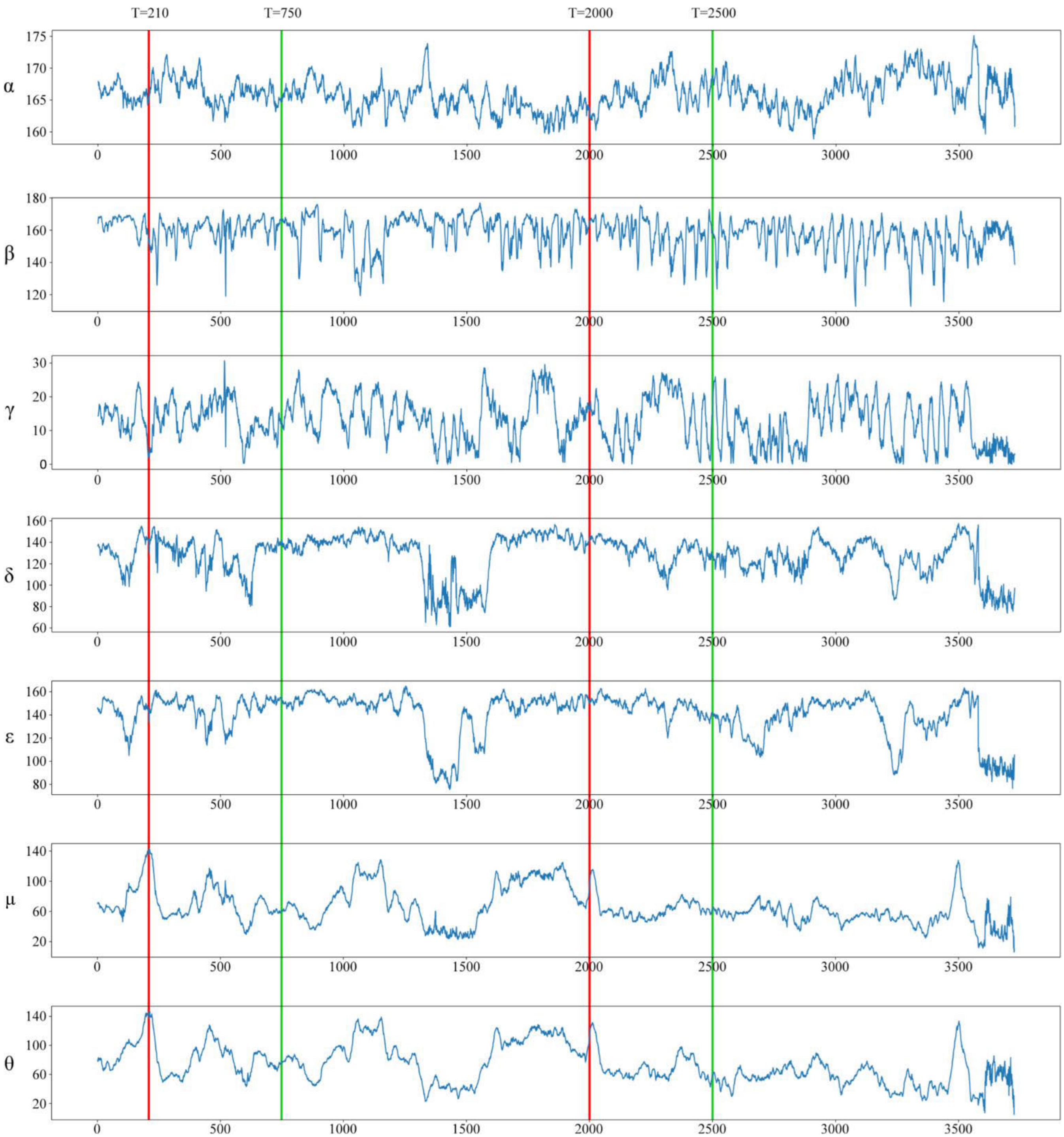


FIGURE 4. Ergonomic parameters for the P&P task and corresponding acts – safe (green) and unsafe (red) from sample outputs presented in Fig 3.

motion-capture systems (compared with closed-circuit television cameras) provide more accurate results under laboratory conditions, their price and complexity of use limit their applicability in conventional industrial environments. In general, previous studies on vision-based recognition of unsafe actions have been restricted to specific and predefined actions [18], [19]. While there are reviews on the use of computer vision for safety and health monitoring on construction sites [24], differences in task execution and their implications for the derived ergonomic parameters were not considered

A. PROGRESS BEYOND STATE-OF-THE-ART

To the best of our knowledge, this was the first study in which deep-learning techniques were used to assess

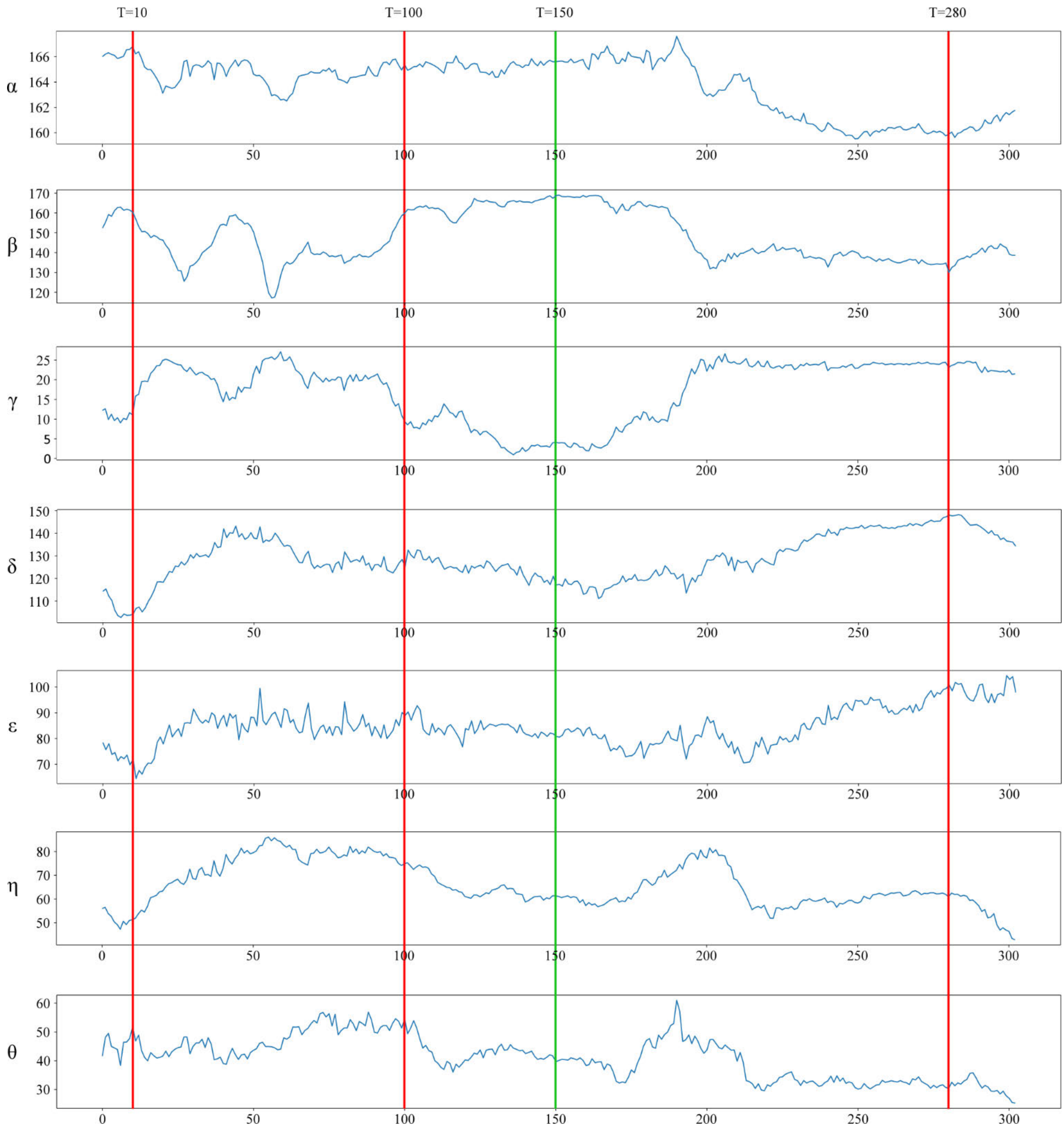


FIGURE 5. Ergonomic parameters for the drilling task and corresponding acts – safe (green) and unsafe (red) from sample outputs presented in Fig. 3.

workplace safety via 3D body mesh analysis. Previously, CNNs were used to integrate RGB, optical flow, and gray stream data to automatically assess worker activities on construction sites (walking, transporting, and steel bending), with an average accuracy of 85% [25]. Similarly, a combination of a CNN and LSTM was proposed for classifying safe and unsafe actions in ladder climbing according to image features, and an accuracy of ∼92% was achieved [23].

In practice, conclusions drawn by visually observing images of complex 3D shapes (including human bodies) are known to be biased. This is the major reason why previously proposed procedures considered workers from a profile viewpoint, where 2D key points of the human body could be reconstructed and tracked more precisely. This technical limitation of 2D pose estimation algorithms increases the complexity of determining

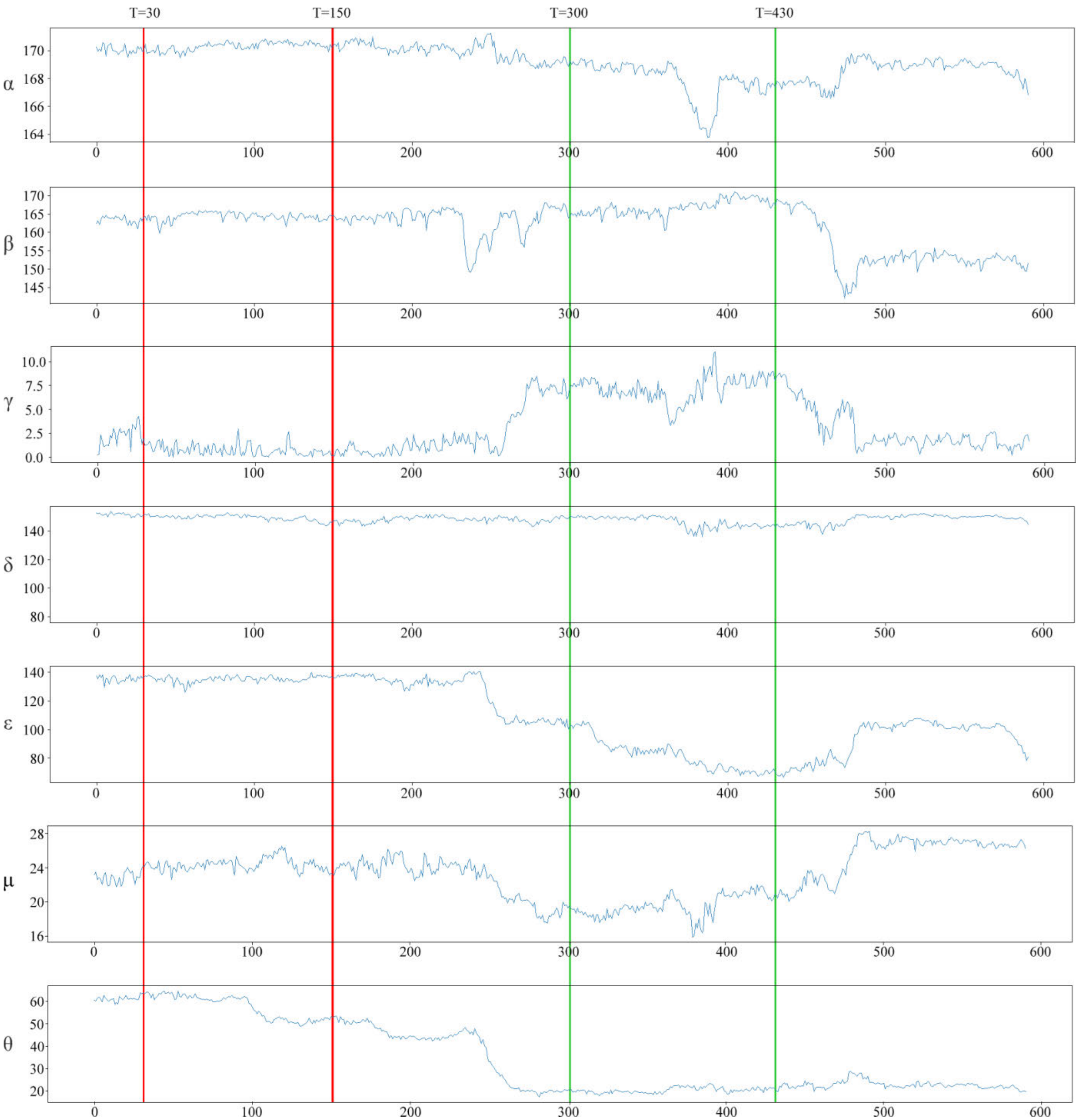


FIGURE 6. Ergonomic parameters for the collaborative polishing task and corresponding acts – safe (green) and unsafe (red) from sample outputs presented in Fig 3.

the optimal camera viewpoint and avoiding body-part occlusions.

The obtained results confirmed the hypothesis of this study that the direct consideration of 3D body poses can resolve the aforementioned limitations and accelerate the development of general-purpose solutions for assessing workplace safety. Although CNNs have proven to be effective for solving problems represented by regular structures (e.g., images), extending the paradigm to irregular structures (e.g., 3D meshes) is a challenging topic. In comparison with the more straightforward analysis [47] and classification [48] of point clouds, polygonal meshes, although more complex, provide a more efficient and hierarchical representation of 3D body shapes. In particular, they explicitly capture the body shape surface and topology, accounting for the volumetric parameters and body constitution, which are individual parameters that can be used for personalizing the safety assessment. Moreover, since input features of MeshCNN are similarity-invariant, applying rotation, translation, and isotropic scaling does not affect input meshes, which is a significant advantage compared to conventional deep CNN (one mesh contains information equal to several input images from different angles).

B. IMPLICATIONS AND APPLICABILITY OF THE FINDINGS

With regard to practical application and adopting a procedure developed for safety assessment of a task-specific method for assessing other tasks, the major challenges that all computer vision-based procedures face are intra-observer and inter-observer variability of data labeling and risk assessment. In a study focused on this problem, it is reported that experienced safety engineers could only achieve 70% accuracy in labeling safety-rule violations of complex scenes [49]. The proposed procedure overcomes this limitation because it is based on direct 3D body shape analysis, which ensures intuitive annotation of data and analysis and interpretation of results. Although it is straightforward to calculate and extract numerous ergonomic parameters from 2D or 3D body shapes, simultaneous tracking and interpretation of the corresponding time series are difficult, even for ergonomics experts (Fig. 4-6). More importantly, parameter selection and analysis are highly task-dependent (parameters relevant for one task may be irrelevant for another), which explains the limited industrial applications of such procedures.

Although international organizations have proposed industrial and ergonomic standards to facilitate the analysis and optimization of workplace ergonomics, the inability of safety experts to track and detect unsafe acts in real-time remains the primary cause of workplace injuries. Moreover, owing to the large number and variability of workplaces in the industry, many cases are not precisely covered by standards, leaving onsite supervisors to judge unsafe acts. Accordingly, there is a practical need for a computerized tool that allows practitioners to intuitively collect, visually define, and classify safe and unsafe acts for specific tasks.

From a methodological perspective, the proposed method is the first to allow the direct derivation of a model from the collection of safe/unsafe videos. In most industries, videos can be acquired from existing surveillance systems, although the use of a conventional smartphone camera is also suitable. Thus, the analysis of three separate industrial tasks indicated that the proposed procedure is generic. In contrast to previously reported methods that depend on parameters derived from pose estimation [20], [21], our deep-learning pipeline implicitly considers ergonomic parameters from shapes.

The proposed pipeline offers a significant advantage by diminishing the reliance on safety/ergonomics experts when it comes to creating or adjusting computerized procedures for the automatic detection of unsafe acts. This advantage is a key characteristic of the pipeline, as it allows practitioners to employ the proposed method in real-world scenarios without relying heavily on the expertise of safety or ergonomics professionals. Additionally, the utilization of 3D representations within the pipeline enhances its versatility, making it suitable for a wide range of applications in various industries. The potential for expanding the use of these 3D representations to other fields and industries further highlights the flexibility and scalability of the proposed method. As a result, practitioners can benefit from the proposed pipeline by effectively addressing safety concerns and recognizing unsafe acts in diverse real-life situations, having the potential to extend its application beyond its current scope.

C. LIMITATIONS OF THE PRESENT STUDY AND FUTURE WORK

Finally, we report the major drawbacks of the proposed study, which are directions for future research on this topic. Since two modules in the pipeline are independent deep learning architectures that perform distinct tasks, data exchange and redundant computations slow down the execution time. Furthermore, considering the recent progress of 3D human pose estimation from monocular videos, there is a possibility for improving the speed of the pose estimation module by considering alternatives to the VIBE algorithm (which currently represents the bottleneck for real-time applications). Moreover, there is a possibility of applying point-cloud classification instead of mesh classification, which was used in this study as the VIBE outputs were in structured mesh form. Another possible improvement that needs to be made is the incorporation of temporal information into the decision-making, which could be done by using the LSTM modules (enabling one to consider safe/unsafe acts as action classification instead of posture classification). In this sense, it is important to compare such approaches with deep learning models that perform action classification directly from videos (e.g. 3D ResNet).

According to the insights gained in this study, it is more likely to expect that application of such solutions will be first made for static workplaces – while to be able to cover dynamic workplaces the pipeline should also consider operator tracking, which is not the subject of this study and arises other issues, such as workers' privacy and identification in manufactures. To enable further progress towards mentioned challenges, the major prerequisite is a development of an appropriate public data set, which could enable objective evaluation of developed procedures as well as their improvement by the scientific community.

VI. CONCLUSION

In this study, we proposed a deep learning-based approach for pose safety assessment from monocular videos. The experimental results showed that the proposed method has better performance while providing practical advantages over existing procedures based on 2D pose estimation.

Specifically, the proposed procedure offers a more intuitive and comprehensive approach to a safety assessment by direct classification of 3D body shapes. It allows for more accurate annotation of data, analysis, and interpretation of results.

Furthermore, the procedure reduces the reliance on safety and ergonomics experts, making it more accessible and adaptable for real-world scenarios. The versatility and scalability of the proposed method make it suitable for various industries and tasks, addressing safety concerns and recognizing unsafe acts in diverse real-life situations.

We report the pose estimation stage as a bottleneck of this algorithm. When a partial occlusion happens (left or right arm or leg), especially if it lasts more than a few frames, body parts that are missing from the video cannot always be correctly evaluated. The imprecisely estimated SMPL model is propagated through the whole algorithm and can lead to misclassification (which influences the accuracy).

Further improvements in procedures for 3D pose reconstruction from monocular videos are expected to play a key role in improving the limitations of the proposed approach. As an alternative to the classification module, the use of more conventional point-cloud classification and analysis can also be considered. Accordingly, our future research will be directed toward 1) the evaluation of more diverse industrial tasks, 2) enabling the recognition of unsafe acts of multiple people simultaneously, 3) improving and testing new 3D pose and shape estimation and classification algorithms, and 4) modifying the architecture to develop more robust decision support systems while enabling real-time data processing [36]. To reach these goals, the development of appropriate public datasets will be crucial in enabling progress within the scientific community focused on the application of artificial intelligence for improving workplace safety.

TABLE 3. Abbreviation list.

Abbreviation	Meaning
WMSD	Work-Related Musculoskeletal Disorders
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
P&P	Pushing and Pulling
SMPL	Skinned Multi-Person Linear Model

REFERENCES

[1] R. R. Galin and R. V. Meshcheryakov, "Human-Robot interaction efficiency and human-robot collaboration," in *Robotics: Industry 4.0 Issues & New Intelligent Control Paradigms*, A. G. Kravets, Ed. Cham, Switzerland: Springer, 2020, pp. 55–63.

[2] M. Zarte, A. Pechmann, and I. L. Nunes, "Principles for human-centered system design in industry 4.0—A systematic literature review," in *Advances in Human Factors and Systems Interaction* (Advances in Intelligent Systems and Computing), vol. 1207, I. Nunes, Ed. Cham, Switzerland: Springer, 2020, pp. 140–147, doi: 10.1007/978-3-030-51369-6_19.

[3] H. N. Ngoc, G. Lasa, and I. Iriarte, "Human-centred design in industry 4.0: Case study review and opportunities for future research," *J. Intell. Manuf.*, vol. 33, no. 1, pp. 35–76, Jan. 2022, doi: 10.1007/s10845-021-01796-x.

[4] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and industry 5.0—Inception, conception and perception," *J. Manuf. Syst.*, vol. 61, pp. 530–535, Oct. 2021.

[5] *Recommended Practices for Safety and Health Programs*, Occupational Safety and Health Administration, Washington, DC, USA, 2016.

[6] I. Cowi, *Evaluation of the Practical Implementation of the EU Occupational Safety and Health (OSH) Directives in EU Member States-Synthesis Report*. Brussels, Belgium: DG Employment, Social Affairs and Inclusion, European Commission, 2015.

[7] A. M. Vukicevic, M. Djapan, V. Isailovic, D. Milasinovic, M. Savkovic, and P. Milosevic, "Generic compliance of industrial PPE by using deep learning techniques," *Saf. Sci.*, vol. 148, Apr. 2022, Art. no. 105646, doi: 10.1016/j.ssci.2021.105646.

[8] V. Isailovic, M. Djapan, M. Savkovic, M. Jovicic, M. Milovanovic, M. Minovic, P. Milosevic, and A. Vukicevic, "Compliance of head-mounted personal protective equipment by using YOLOv5 object detector," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET)*, Dec. 2021, pp. 1–5, doi: 10.1109/ICECET52533.2021.9698662.

[9] S. P. Anderson and J. Oakman, "Allied health professionals and work-related musculoskeletal disorders: A systematic review," *Saf. Health at Work*, vol. 7, no. 4, pp. 259–267, Dec. 2016, doi: 10.1016/j.shaw.2016.04.001.

[10] E. EU-OSHA. (2019). *Third European Survey of Enterprises on New and Emerging*. [Online]. Available: https//visualisation.osha.Eur.eu/esener

[11] A. Cieza, K. Causey, K. Kamenov, S. W. Hanson, S. Chatterji, and T. Vos, "Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: A systematic analysis for the Global Burden of Disease Study 2019," *Lancet*, vol. 396, no. 10267, pp. 2006–2017, 2020, doi: 10.1016/S0140-6736(20)32340-0.

[12] E. Occhipinti and D. Colombini, "A toolkit for the analysis of biomechanical overload and prevention of WMSDs: Criteria, procedures and tool selection in a step-by-step approach," *Int. J. Ind. Ergonom.*, vol. 52, pp. 18–28, Mar. 2016, doi: 10.1016/j.ergon.2015.08.001.

[13] *One in two Americans Have a Musculoskeletal Condition: New Report Outlines the Prevalence, Scope, Cost and Projected Growth of Musculoskeletal Disorders in the U.S.*, vol. 1, ScienceDaily. ScienceDaily, American Academy of Orthopaedic Surgeons, Rosemont, IL, USA, 2016.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[15] W. Kim, L. Peternel, M. Lorenzini, J. Babič, and A. Ajoudani, "A human-robot collaboration framework for improving ergonomics during dexterous operation of power tools," *Robot. Comput.-Integr. Manuf.*, vol. 68, Apr. 2021, Art. no. 102084, doi: 10.1016/j.rcim.2020.102084.

[16] L. Fortini, M. Lorenzini, W. Kim, E. De Momi, and A. Ajoudani, "A real-time tool for human ergonomics assessment based on joint compressive forces," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 1164–1170, doi: 10.1109/RO-MAN47096.2020.9223565.

[17] M. Lorenzini, W. Kim, and A. Ajoudani, "An online multi-index approach to human ergonomics assessment in the workplace," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 5, pp. 812–823, Oct. 2022, doi: 10.1109/THMS.2021.3133807.

[18] S. Han and S. Lee, "A vision-based motion capture and recognition framework for behavior-based safety management," *Autom. Construct.*, vol. 35, pp. 131–141, Nov. 2013, doi: 10.1016/j.autcon.2013.05.001.

[19] S. Han, S. Lee, and F. Peña-Mora, "Vision-based detection of unsafe actions of a construction worker: Case study of ladder climbing," *J. Comput. Civil Eng.*, vol. 27, no. 6, pp. 635–644, Nov. 2013.

[20] H. Guo, Y. Yu, Q. Ding, and M. Skitmore, "Image-and-skeleton-based parameterized approach to real-time identification of construction Workers' unsafe behaviors," *J. Construct. Eng. Manage.*, vol. 144, no. 6, Jun. 2018, Art. no. 4018042.

[21] Y. Yu, H. Guo, Q. Ding, H. Li, and M. Skitmore, "An experimental study of real-time identification of construction workers' unsafe behaviors," *Autom. Construct.*, vol. 82, pp. 193–206, Oct. 2017, doi: 10.1016/j.autcon.2017.05.002.

[22] A. M. Vukicevic, I. Macuzic, N. Mijailovic, A. Peulic, and M. Radovic, "Assessment of the handcart pushing and pulling safety by using deep learning 3D pose estimation and IoT force sensors," *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115371, doi: 10.1016/j.eswa.2021.115371.

[23] L. Ding, W. Fang, H. Luo, P. E. D. Love, B. Zhong, and X. Ouyang, "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory," *Autom. Construct.*, vol. 86, pp. 118–124, Feb. 2018, doi: 10.1016/j.autcon.2017.11.002.

[24] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Adv. Eng. Informat.*, vol. 29, no. 2, pp. 239–251, Apr. 2015, doi: 10.1016/j.aei.2015.02.001.

[25] H. Luo, C. Xiong, W. Fang, P. E. D. Love, B. Zhang, and X. Ouyang, "Convolutional neural networks: Computer vision-based workforce activity assessment in construction," *Autom. Construct.*, vol. 94, pp. 282–289, Oct. 2018, doi: 10.1016/j.autcon.2018.06.007.

[26] W. Fang, L. Ding, H. Luo, and P. E. D. Love, "Falls from heights: A computer vision-based approach for safety harness detection," *Autom. Construct.*, vol. 91, pp. 53–61, Jul. 2018, doi: 10.1016/j.autcon.2018.02.018.

[27] M. Al-Amin, R. Qin, M. Moniruzzaman, Z. Yin, W. Tao, and M. C. Leu, "An individualized system of skeletal data-based CNN classifiers for action recognition in manufacturing assembly," *J. Intell. Manuf.*, vol. 34, no. 2, pp. 633–649, Feb. 2023, doi: 10.1007/s10845-021-01815-x.

[28] M. Ciccarelli et al., "SPECTRE: A deep learning network for posture recognition in manufacturing," *J. Intell. Manuf.*, vol. 34, pp. 3469–3481, 2023, doi: 10.1007/s10845-022-02014-y.

[29] J. Yang, J. Lee, and J. Choi, "Activity recognition based on RFID object usage for smart mobile devices," *J. Comput. Sci. Technol.*, vol. 26, no. 2, pp. 239–246, Mar. 2011, doi: 10.1007/s11390-011-9430-9.

[30] C. Hofmann, C. Patschkowski, B. Haefner, and G. Lanza, "Machine learning based activity recognition to identify wasteful activities in production," *Proc. Manuf.*, vol. 45, pp. 171–176, Jan. 2020, doi: 10.1016/j.promfg.2020.04.090.

[31] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016, doi: 10.3390/s16010115.

[32] (1989). *Council Directive 89/654/EEC of 30 November 1989 Concerning the Minimum Safety and Health Requirements for the Workplace*. Accessed: Jan. 4, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31989L0654

[33] (1990). *Council Directive 90/269/EEC of 29 May 1990 on the Minimum Health and Safety Requirements for the Manual Handling of Loads Where There is a Risk Particularly of Back Injury to Workers*. Accessed: Jan. 4, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31990L0269

[34] (2007). *ISO 11228-2:2007 Ergonomics Manual handling—Part 2: Pushing and Pulling*. Accessed: Jan. 4, 2023. [Online]. Available: https://www.iso.org/standard/26521.html

[35] (2016). *ISO/TS 15066:2016 Robots and Robotic Devices Collaborative Robots*. Accessed: Jan. 4, 2023. [Online]. Available: https://www.iso.org/standard/62996.html

[36] A. M. Vukicevic, M. Djapan, M. Stefanovic, and I. Macuzic, "SafE-tag mobile: A novel Javascript framework for real-time management of unsafe conditions and unsafe acts in SMEs," *Saf. Sci.*, vol. 120, pp. 507–516, Dec. 2019, doi: 10.1016/j.ssci.2019.07.024.

[37] M. Petrovic, A. M. Vukicevic, B. Lukic, and K. Jovanovic, "Assessment of the human-robot collaborative polishing task by using EMG sensors and 3D pose estimation," in *Advances in Service and Industrial Robotics* (Mechanisms and Machine Science), vol. 120, A. Müller and M. Brandstötter, Eds. Cham, Switzerland: Springer, 2022, pp. 564–570, doi: 10.1007/978-3-031-04870-8_66.

[38] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5252–5262.

[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[40] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Oct. 2015.

[41] M. Garland and P. S. Heckbert, "Surface simplification using quadric error metrics," in *Proc. 24th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1997, pp. 209–216.

[42] P. D. Hung and N. T. Su, "Unsafe construction behavior classification using deep convolutional neural network," *Pattern Recognit. Image Anal.*, vol. 31, no. 2, pp. 271–284, Apr. 2021, doi: 10.1134/S1054661821020073.

[43] N. Khan, M. R. Saleem, D. Lee, M.-W. Park, and C. Park, "Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks," *Comput. Ind.*, vol. 129, Aug. 2021, Art. no. 103448, doi: 10.1016/j.compind.2021.103448.

[44] W. Fang, L. Ding, P. E. D. Love, H. Luo, H. Li, F. Peña-Mora, B. Zhong, and C. Zhou, "Computer vision applications in construction safety assurance," *Autom. Construct.*, vol. 110, Feb. 2020, Art. no. 103013, doi: 10.1016/j.autcon.2019.103013.

[45] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: A network with an edge," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Jul. 2019, doi: 10.1145/3306346.3322959.

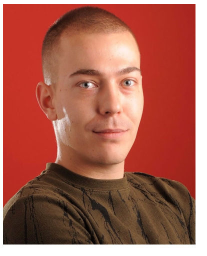
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] T. Xu, D. An, Y. Jia, and Y. Yue, "A review: Point cloud-based 3D human joints estimation," *Sensors*, vol. 21, no. 5, p. 1684, Mar. 2021.

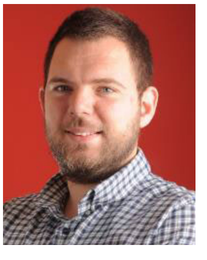
[48] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1588–1597.

[49] P.-C. Liao, J. Ding, and X. Wang, "Enhancing cognitive control for improvement of inspection performance: A study of construction safety," in *Proc. Int. Conf. Eng. Psychol. Cogn. Ergonom.*, 2016, pp. 311–321.

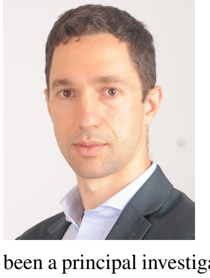
ARSO M. VUKICEVIC received the Ph.D. degree from the University of Kragujevac, Serbia. He is currently a Senior Research Associate with the Faculty of Engineering, University of Kragujevac. His current research interests include computer vision, machine learning, software engineering, and visual computing with applications in industrial and biomedical engineering.



MILOS N. PETROVIC (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the School of Electrical Engineering, University of Belgrade, Serbia, in 2018 and 2019, respectively, where he is currently pursuing the Ph.D. degree. He is a Research Assistant with the School of Electrical Engineering, University of Belgrade. His research interests include computer vision, machine learning, ergonomics, and collaborative robotics.



NIKOLA M. KNEZEVIC (Graduate Student Member, IEEE) received the bachelor's and master's degrees from the Faculty of Electrical Engineering, University of Belgrade, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree. He is a Teaching Assistant with the Faculty of Electrical Engineering, University of Belgrade. He is engaged in the control of robotic systems with compliant joints, machine learning, and bio-inspired robotics.



KOSTA M. JOVANOVIC (Member, IEEE) received the Ph.D. degree in robotics from the University of Belgrade, in 2016. He is currently an Associate Professor with the School of Electrical Engineering (ETF), University of Belgrade, and the Head of the ETF Robotics Laboratory (www.robot.etf.rs). He gained professional experience with the German Aerospace Center (DLR), Germany; the Technical University of Munich, Germany; and SMS Siemag, Germany. He has been a principal investigator (PI) on behalf of ETF in several H2020 projects in robotics and the project ForNextCobot funded by the Science Fund of the Republic of Serbia under the call of Programme for excellent projects of young researchers. His research interests include modeling and control of collaborative robots, mechanical impedance estimation, and physical human–robot collaboration in medical and industrial applications.