



Article

# Detection of Physical Strain and Fatigue in Industrial Environments Using Visual and Non-Visual Low-Cost Sensors

Konstantinos Papoutsakis <sup>1,\*</sup>, George Papadopoulos <sup>1</sup>, Michail Maniadakis <sup>1</sup>, Thodoris Papadopoulos <sup>1</sup>, Manolis Lourakis <sup>1</sup>, Maria Pateraki <sup>1,2</sup> and Iraklis Varlamis <sup>1,3</sup>

<sup>1</sup> Computational Vision and Robotics Laboratory, Institute of Computer Science, Foundation for Research and Technology—Hellas, 700 13 Heraklion, Greece; gpaps@ics.forth.gr (G.P.); mmaniada@ics.forth.gr (M.M.); thodoris@thodoris.net (T.P.); lourakis@ics.forth.gr (M.L.); pateraki@ics.forth.gr (M.P.); varlamis@hua.gr (I.V.)

<sup>2</sup> Laboratory of Photogrammetry, School of Rural Surveying and Geoinformatics Engineering, National Technical University of Athens, 157 80 Athens, Greece

<sup>3</sup> Department of Informatics and Telematics, School of Digital Technology, Harokopio University of Athens, 177 78 Athens, Greece

\* Correspondence: papouts@ics.forth.gr

**Abstract:** The detection and prevention of workers' body straining postures and other stressing conditions within the work environment, supports establishing occupational safety and promoting well being and sustainability at work. Developed methods towards this aim typically rely on combining highly ergonomic workplaces and expensive monitoring mechanisms including wearable devices. In this work, we demonstrate how the input from low-cost sensors, specifically, passive camera sensors installed in a real manufacturing workplace, and smartwatches used by the workers can provide useful feedback on the workers' conditions and can yield key indicators for the prevention of work-related musculo-skeletal disorders (WMSD) and physical fatigue. To this end, we study the ability to assess the risk for physical strain of workers online during work activities based on the classification of ergonomically sub-optimal working postures using visual information, the correlation and fusion of these estimations with synchronous worker heart rate data, as well as the prediction of near-future heart rate using deep learning-based techniques. Moreover, a new multi-modal dataset of video and heart rate data captured in a real manufacturing workplace during car door assembly activities is introduced. The experimental results show the efficiency of the proposed approach that exceeds 70% of classification rate based on the F1 score measure using a set of over 300 annotated video clips of real line workers during work activities. In addition a time lagging correlation between the estimated ergonomic risks for physical strain and high heart rate was assessed using a larger dataset of synchronous visual and heart rate data sequences. The statistical analysis revealed that imposing increased strain to body parts will results in an increase to the heart rate after 100–120 s. This finding is used to improve the short term forecasting of worker's cardiovascular activity for the next 10 to 30 s by fusing the heart rate data with the estimated ergonomic risks for physical strain and ultimately to train better predictive models for worker fatigue.

**Keywords:** computer vision; sensor fusion; low cost sensors; heart rate; WMSD; fatigue; ergonomic risk; physical strain; working postures; predictive models; occupational health



**Citation:** Papoutsakis, K.; Papadopoulos, G.; Maniadakis, M.; Papadopoulos, T.; Lourakis, M.; Pateraki, M.; Varlamis, I. Detection of Physical Strain and Fatigue in Industrial Environments Using Visual and Non-Visual Low-Cost Sensors. *Technologies* **2022**, *10*, 42. <https://doi.org/10.3390/technologies10020042>

Received: 15 January 2022

Accepted: 9 March 2022

Published: 16 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

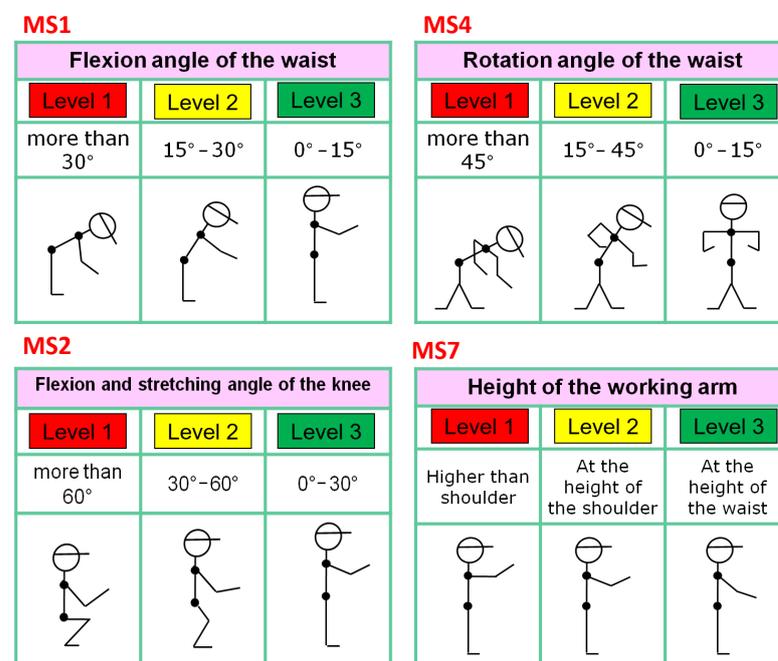
## 1. Introduction

The assessment and prevention of work-related musculo-skeletal disorders (WMSD) and physical fatigue are considered common and critical issues related to occupational safety and well-being in work environments. Especially in the manufacturing industry, labour-intensive assembly works attribute repetitive tasks, often in sustained, sub-optimal working postures [1], also noted as inappropriate or awkward postures, that lead to physical strain, according to several studies on physical ergonomics [2,3], and may also cause abnormal heart rate. Those indicators are known as risk factors for WMSD and

fatigue [3,4]. The case study in this paper addresses the car manufacturing industry and specifically line workers that often work in shifts on a workstation of the assembly line, where the conveyor belt slowly moves at a constant speed. Each worker executes a specific set of car assembly activities (e.g., welding, assembling) that constitute a task cycle for a workstation, lasting for 4 to 5 min and is continuously repeated during the shift.

The current study is part of the sustAGE system (<http://www.sustage.eu>, accessed on 10 January 2022), which is developed to provide a person-centered smart solution to support the employment, safety, and health of ageing workers in occupational contexts. One of the novelties of sustAGE is the adoption and integration of the *Micro-Moments* (MiMos) concept [5,6]. MiMos are used to digitise interactions with the physical environment, repeated patterns, or events occurring in workers' daily living routines and they link with recommended actions targeted directly at the workers themselves or at their supervisors. By issuing recommendations through MiMos, the system capitalizes on the early detection and avoidance of risky and stressful conditions that affect the performance of individual workers or worker groups. Accordingly, recommendations reach the users at the right moment and place, and proposed actions match the users' preferences and current needs.

For being able to issue relevant recommendations and prevent risky situations in an industrial environment, it is important to early identify such events and their underlying conditions. In this direction, we aim to re-actively detect events causing physical strain and fatigue and recommend preventive actions. We rely on visual information acquired by low-cost cameras placed along the production line that support the unobtrusive automatic assessment of awkward ergonomically sub-optimal body postures, as shown in Figure 1, and on heart rate data acquired by smartwatches for monitoring workers' cardiovascular activity. The respective modules that process the two data modalities provide the sustAGE system with information on the detected events and trigger personalised recommendations to the workers aiming to enhance occupational safety in a preventive manner.



**Figure 1.** The proposed vision-based method focuses on the classification of four (4) types (sketches) of ergonomic working postures during assembly activities in videos. The selected types are part of a larger set of working body postures that are widely-used for the assessment of physical ergonomics (i.e., by the MURI risk analysis approach [7,8]). Each type is associated with three variations/deviations of the body configurations of increasing physical discomfort and ergonomic risk for physical strain imposed to specific body joints (image courtesy of Stellantis—Centro Ricerche FIAT (CRF)/SPW Research & Innovation department).

It is important to highlight the need for unobtrusive monitoring mechanisms that allow workers to fully operate in their work environment, without any physical discomfort and distraction from their work duties. It is also necessary to keep the cost of the monitoring infrastructure low, in order to be able to scale up to hundreds of employees that work in a shift in a medium sized factory. Wall or tripod mounted cameras and smartwatches allow the proposed methodology, as part of the sustAGE system to collect valuable data about the workers' body posture during the execution of assembly tasks. The fusion of this data with heart rate measurements from the smartwatches, as shown in this work, is a promising solution for the early detection of fatigue that can be caused by the prolonged execution of a task with an ergonomically sub-optimal posture that imposes increased stress to specific body parts, such as the waist, the legs, or arms.

The main contributions of this work can be summarized in the following:

- An unobtrusive and low cost solution for the detection of physical strain and fatigue during work activities, which is based on the smart fusion of vision-based extracted information (working postures) and non-visual (heart rate) input, regardless of the activity performed. A vision-based approach for the classification of ergonomically sub-optimal working postures that cause increased physical strain is proposed. It relies on the combination of Graph-based Convolutional Networks and the soft-DTW method for pairwise temporal alignment of 3D skeletal data sequences. The proposed approach can achieve real-time/online runtime performance using continuous streams of data acquired by a single camera.
- A predictive model for the early detection of high heart rate incidents, which exploits vision-based extracted information related to the worker physical strain to improve heart rate prediction accuracy.
- A new multi-modal dataset is introduced that comprises synchronized visual information of color and depth image sequences and worker heart rate (HR) data acquired using smartwatches during car assembly activities in an actual manufacturing environment. Annotation data is available for the sequences of assembly actions performed by real line workers and the assessment of posture-based physical ergonomics according to the MURI risk analysis method [7,8].

The combination of the Spatial Temporal Graph-based Neural Network (ST-GCN) model [9] with the soft-Dynamic Time Warping method [10] for the classification of ergonomically sub-optimal working postures during work activities is one of the innovations of the proposed method. ST-GCN has been used in the past for classifying whole video sequences into different categories of human actions using 2D or 3D skeletal data as input. However the existing methods learn an embedding-based representation for the input sequence, and consequently directly classify this information towards an action class. Our proposed method extracts rich information of the observed skeleton-based human motion and actions in 3D space and employs a ST-GCN model to generate sequences of embeddings that allows for the fine-grained representation of the spatio-temporal relationships of body parts throughout the input skeletal sequence. The generated sequence of embeddings is consequently fed to the soft-DTW component, which has the ability to perform non-linear temporal matching of the embeddings to the respective example sequences of known types of working postures. This allows us to assess the temporal evolution and variations between the representations of two comparing sequences of working postures, being also invariant to the duration and the speed of execution sequence. This option better fits to the requirements of the task of classifying body straining working postures.

The proposed vision-based approach performs in real-time given a segmented input video of human actions as input. Moreover, online runtime performance in untrimmed streaming (continuous) data is also feasible as part of a joint temporal localization and classification framework. Such an approach could refer to the typical sliding window design or an elaborate deep neural network model and thus ensures high computational efficiency and online recognition of the type of posture deviations as soon as it happens.

Another advantage of the proposed methodology refers to the processing of multi-modal information that is realized as sequences of physiological measurements acquired from wearable heart-rate sensors and of temporal information regarding the worker physical strain based on vision-based, automatic classification of ergonomic working postures during work activities. The two data modalities are used as input to an LSTM module with the aim to predict the heart rate of the worker for one or more consecutive periods. This combination of inputs improves the predictive performance, compared to merely using the heart rate (HR) measurements.

This study further contributes to the release of new multi-modal dataset of HR, visual data and working body postures-related measurements to assess the ergonomic risk for physical strain and fatigue in automotive manufacturing occupational contexts. It further provides a deep-learning based classification scheme for vision-based postural risk analysis and an analysis on the association of physical strain and HR activity, exploiting the detection of ergonomically sub-optimal postures to improve short- and mid-term heart rate prediction.

In Section 2 that follows, we provide an overview of the related work and datasets. We survey recent methods for the visual estimation of body poses and recognition of human actions in videos, methods for assessing the ergonomic risks, and works that focus on heart rate monitoring. Section 3 provides details on the proposed methodology, while in Section 4 we report the data acquisition and annotation methodology we followed and the obtained results on the prediction of worker heart rate and body strain and fatigue related incidents. In Section 5 the main findings of this work and summarizes our next steps are discussed. This study is an extension of our previous work presented in [11].

## 2. Related Work

Human motions extend from the simplest movement of a limb to complex joint movement of a group of limbs and body, possibly interacting with other entities, i.e., objects, in order to act and accomplish goal-oriented tasks of varying significance and impact. Such motions and postures, especially when they are repetitive and sustained, they can cause an increase in physical stress and lead to injuries and fatigue [12]. The detection of body straining motions or postures can be performed using invasive motion capture systems that require expensive, special body-worn equipment or unobtrusively using visual information from cameras that monitor humans during the execution of daily activities and tasks. The analysis of this information is usually handled as a spatio-temporal mining task on 2D or 3D skeletal body representations into consecutive video frames. Deep Long Short-Term Memory (LSTM) networks [13,14] have been widely applied for the analysis of such data with the main objective to identify user activities [15,16], and they have proven very successful due to their ability to capture sequential features and long-term dependencies in the input image sequences. Convolutional and simple LSTMs have also been used for detecting awkward and stressing postures from raw sensor (wearables) input [17], also relying on the ability of LSTM to capture motion-related long-term dependencies in videos. LSTM models have also recently been used to tackle the problem of car-driver identity recognition [18] using a non-invasive biosensor system that comprises multiple devices placed on the car driver's hand and the car steering. The method proposes a deep learning architecture that comprises blocks that use the Dynamic Time Warping method for non-linear temporal alignment among the input physiological signals of the driver's heart pulse rate and blocks of LSTMs for learning physiological features and finally for estimating the car drive profile class.

When it comes to ergonomic risk analysis, the detection of working postures has to be associated with a potential risk for injury or fatigue, and thus new models and assessment scores have been devised in order to take as input the video or sensor data and provide a risk score as output [19,20]. Recurrent neural networks and LSTM have been widely employed in this case too, this time taking as input carefully devised skeletal features, or sensor patterns that have been extracted from the sensor (or camera) input.

What is still missing from the literature, as explained in the subsections that follow, is the association of physical ergonomics analysis with the accumulated worker fatigue that results in high fatigue incidents, expressed with high heart rates, during the execution of tasks. A good OSH practice in industrial environments is to early detect such hazards and alert workers and their supervisors to preemptively take corrective actions, as explained in the following sections.

In this section, we provide a brief overview of works on visual learning and classification of human actions, specifically of state-of-art methods that rely on skeletal-based representation of the human motion in videos. Moreover, we discuss recent methods proposed to tackle the tasks of automatic vision-based risk analysis of physical ergonomics and the analysis of ergonomics and fatigue based on cardiovascular activity data. Finally, we briefly discuss widely-used and other recently introduced video or multi-modal datasets related to the tasks of action recognition and posture-based ergonomic risk assessment.

### *2.1. Skeleton-Based Action Recognition*

Recently, a significant amount of research has been dedicated to visual understanding of human actions using deep neural network models [21]. In this work, we focus on skeleton-based classification of human actions [22,23] that exploit fine representations of the spatio-temporal configurations of the human body joints as computed by efficient 2D/3D body pose estimation methods [24,25].

Early approaches for skeleton-based action recognition mainly used hand-crafted features to capture the skeletal-based human motion dynamics, such as covariance matrix of the joints' trajectories [26], lie groups [27], decision trees [28] and more. Thus, the spatio-temporal relationships of the human body joints are modelled using effective attention mechanisms [29] and Graph Neural Network-based (GNN) methods [22,30,31], Convolutional Neural Networks (CNN) [32], Recurrent (RNN) [33] or Long-Short Term Memory (LSTM) [14,34] Neural Networks and various versatile and powerful Transformer-style architectures [35,36] that have been recently proposed extending the popular Vision Transformer (ViT) model [37,38]. In the recent work by Plizzari et al. [39], a novel spatio-temporal transformer network model is introduced using spatial and temporal self-attention modules for modelling both the intra-frame interactions between different body parts, and their correlations across time efficiently using the 2D human body joints coordinates.

Motivated by these new powerful methods for fine-grained human behavior monitoring and by emerging applications in Human-Robot Interaction and Collaboration [40–44], researchers in computer vision and robotics [45] have recently joined their efforts to tackle the challenging problem of fine-grained recognition of assembly activities in videos. In this context, the fine-grained recognition problem refers to joint temporal segmentation (action detection) [46] and classification [35] of a sequence of assembly actions that comprise a complex and possibly long assembly activity. A series of methods have been proposed that are able to model both the temporal and spatial structure of assembly procedures in a fine-grained manner in realistic scenarios [47,48] of furniture construction tasks [49], cooking activities [50], toy block building tasks [51] and simple human-robot collaborative assembly tasks [40]. Finally, a recent method [52] focuses on 3D skeletal data to assess sub-skeleton features with trajectory similarity measures and a k-nearest-neighbor-based classifier for the sign language and human action classification problems. Three different speed invariant distance measures are tested for trajectory similarity: the continuous and discrete Fréchet distances and the Dynamic Time Warping (DTW). Despite the fact that no deep learning model is used in this method, the mining techniques developed enable the extraction of efficient sub-skeleton, interpretable representations of human actions showing the information capacity, and the discriminative power of skeletal-based human motion data.

## 2.2. Vision-Based Ergonomic Risk Analysis

Body postures analysis and action recognition in the context of industrial environments is mostly related to the detection of hazardous or body stressing postures, which in the short or long term are related to ergonomic risks for WMSDs. In the paragraphs that follow, we summarize previous work related to the task of vision-based postural ergonomic risk assessment, mainly in the context of work tasks. Parsa et al. [53] introduce a novel approach based on Temporal Convolutional Network (TCN) models for action segmentation in RGB-D videos and subsequently for predicting the REBA ergonomic risk score (Rapid Entire Body Assessment [54]) during object manipulation actions. Nguyen et al. [55] propose a method to extract the working human postures using depth images and to assess the ergonomic safety. Then, in case the ergonomic guidelines based on the EAWS metric are violated, a robotic system re-actively adjusts the height of a workpiece to enable the worker to adapt to an ergonomically safe pose during the working task. The 2D [56] or 3D skeletal body features [57] are extracted using one or two RGB video captured from different viewpoints in order to recognize awkward postures of workers in the context of construction hazard prevention. Shafti et al. [58] focus on a real-time human-robot interaction scenario and extract the 3D skeletal poses of the worker to analyse the safe range of arm motions during welding actions following the RULA posture monitoring method. Mehrizi et al. [59] propose a deep learning approach for markerless 3D pose estimation optimized in the context of object lifting tasks using RGB images from two different viewpoints. Plantard et al. [60] also evaluate the potential WMSDs in a real car manufacturing environment using vision-based extracted 3D skeletal poses of workers to evaluate the RULA ergonomic risk score (Rapid Upper Limb Assessment [61]), while the work proposed by Kim et al. [62] focuses on overloading body joints assessment and user intention recognition to improve worker ergonomics and productivity in a real-time adaptable workstation scenario in manufacturing.

Recently, Parsa et al. [63] proposed a novel approach that is based on Spatio-Temporal Graph Convolutional Networks (ST-GCNs) combined with LSTMs in order to segment and recognize object manipulation actions (lifting, moving boxes etc.) in videos using 3D skeletal features. Finally, the REBA score is estimated for each recognized action. An extension of their work [64] regards a multitask learning paradigm proposed to simultaneously detect actions using an Encoder-Decoder Temporal Convolutional Network and directly predict the REBA score in videos. Finally, another recently proposed method by Konstantinidis et al. [65] introduces a novel multi-stream deep network that acquires 3D skeletal data sequences to compute the REBA score regardless of the activities performed in a video. Each stream is responsible for predicting a partial score that corresponds to a predefined set of body parts prior to their aggregation for the computation of the total REBA score.

## 2.3. Ergonomics and Cardiovascular Activity

Cardiovascular activity is proposed as a key indicator of workload [66,67]. The relationship between cardiovascular activity and ergonomics has been a subject of research for several years [68]. Despite the fact that the majority of existing works focuses on the role of Heart Rate Variability (HRV), the monitoring of Heart Rate (HR) that provides an indirect means to estimate metabolic workload and energy expenditure in work environments [69,70], has also been explored in previous works. The association between HR and ergonomics has been explored in [71] to determine maximum allowed payload lifting. The effects of following instructive ergonomics guidelines during cleaning tasks has been investigated in [72] showing that when ergonomic guidelines are followed, lower cardiovascular load is observed in comparison to non-ergonomics sessions. Similarly, heart rate is linked with cognitive ergonomics [73] with a statistically significant increase observed in tasks with high cognitive demands [74].

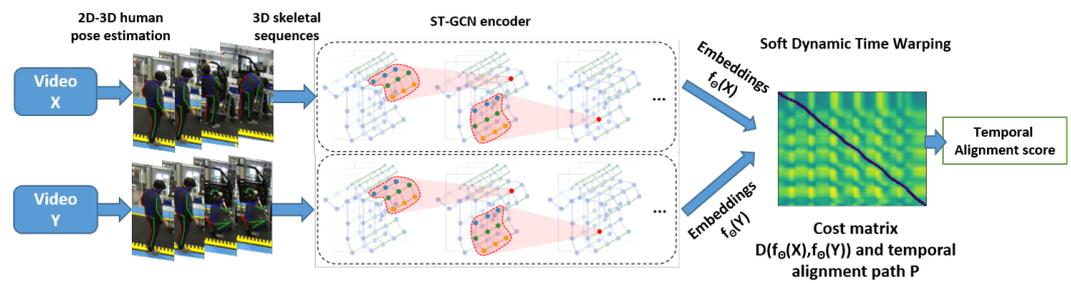
#### 2.4. Datasets

With the advancement of motion capture systems, and the increased interest for crowd sourcing of the video annotation process, the available video and ground truth data have expanded in terms of quantity and acquisition context, i.e., outside the lab environment. A summary of early datasets is provided in [75,76]. However, these datasets are mainly characterised by limited number of image sequences and types of activities that serve specific applications. Among the state of the art benchmarks is the Max Planck Institute for Informatics (MPII) Human Pose Dataset [75] that includes rich annotations, and the joint-annotated Human Motion Database (J-HMDB) [77]. Other recent state-of-the-art datasets for 3D action recognition comprise: the large-scale action recognition NTU RGB+D dataset (120 action classes and 114.480 samples in total) [78], the Kinetics dataset [79], the recently proposed BABEL large scale dataset with language labels describing the actions being performed in mocap sequences and frame-level annotations for fine-grained action analysis [80] and the FineGym dataset providing an insightful hierarchical representation of gymnastic activities for fine-grained action understanding and performance evaluation in sports [81]. Among the few existing datasets related to action recognition or posture-based ergonomic risk assessment in assembly videos, the UW-IOM dataset [53] features a limited number of object manipulation actions involving awkward poses and repetitions, and provides frame-level annotations for scores according to the REBA ergonomic risk index, while the existing TUM Kitchen dataset [82] is also annotated with respect to the REBA scores in the same work. The IKEA furniture-assembly demonstration dataset [40] for human action segmentation and fine-grained recognition provides multifaceted annotation data for a realistic scenario of chair assembly actions in videos captured from different viewpoints.

### 3. Methodology

The main parts of the proposed methodology are described in this section. We rely on synchronous visual and physiological data of workers for the real-time estimation of physical strain and fatigue during work activities in an actual manufacturing environment. It is known that the frequency and the severity of various types of ergonomically sub-optimal postures, also noted as awkward posture deviations, during work activities impose increased physical strain and in the long-term WMSD [3]. Therefore, the detection of these occurrences and the estimation of the ergonomic risks is one of the main objectives of the proposed methodology. With this aim, we firstly focus on the vision-based detection and estimation of the skeleton-based human poses in 3D space using colour image sequences acquired by low-cost camera sensors. The visual information is used to feed state-of-the-art deep-learning based methods that estimate the 3D poses per image in real-time, unobtrusively. The estimated 3D skeletal sequence that represents the observed human motion is subsequently used to feed an enhanced deep-learning based classification approach for estimating the types of sub-optimal working postures performed by the worker and the ergonomic risks for physical strain. The target set of working postures rely on the MURI ergonomic risk analysis approach as shown in Figure 1. A novel aspect of the proposed classification method relies on the combination of a Spatio-Temporal Graph-based Convolutional Network model (ST-GCN) [9] for learning effective spatio-temporal representations of the input 3D skeletal sequences into a new embedding space with the soft Dynamic Time Warping (soft-DTW) method [10] as a classification measure estimated based on the differentiable pairwise temporal alignment between these representations. The pipeline of this approach is presented in Figure 2.

In addition, we examine whether the occurrences of sub-optimal working postures and the estimated ergonomic risk scores can be used as source of information to improve the analysis and the short- and mid-term prediction of the cardiovascular activity of workers, which can be used as an indicator for their fatigue state. To this end, we rely on real-time worker heart rate measurements during assembly activities acquired using smartwatches to feed a Long Short-Term Memory (LSTM) neural network model [14] that has been particularly effective in sequence-to-sequence learning and in time-series forecasting [83].



**Figure 2.** The outline of the proposed approach for vision-based classification of human actions and working postures based on a Spatio-temporal Graph Convolutional Networks [9] and the temporal alignment [10] between two videos of work activities (image of the ST-GCN encoder model was originally presented in [9]).

Finally, we set out to examine the relationship of the detected sub-optimal working postures and the worker cardiovascular activity during work activities in assembly lines, in real-world and demanding industrial environments.

### 3.1. Detecting Worker Physical Strain

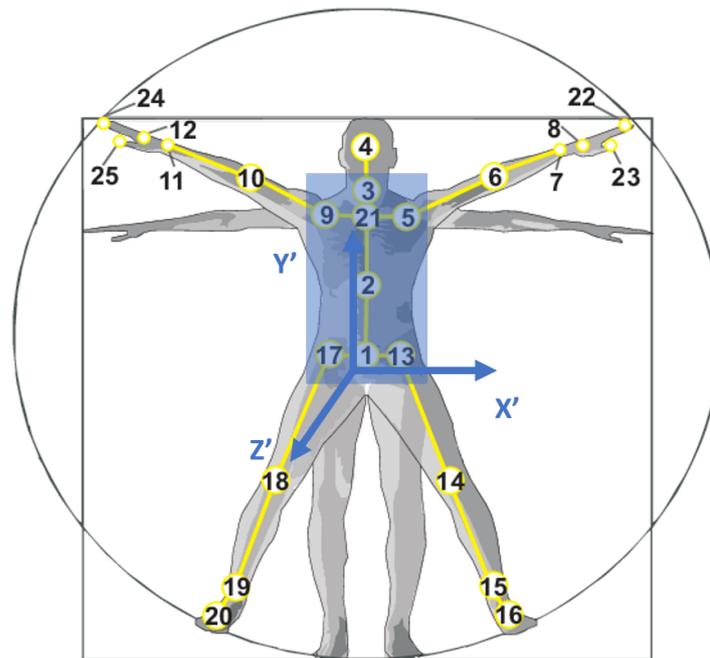
We are interested in four types of ergonomic working postures according to the MURI risk analysis method [7,8] that is commonly used in manufacturing for the evaluation of physical ergonomics of workers, as shown in Figure 1. Each posture type is considered as a time-varying event; thus comprises a sequence of body configurations of a minimum duration. An efficient skeleton-based representation of the human body is extracted per video frame in order to encode the 3D full body configurations during work activities, as shown in Figure 3. The MURI analysis method provides a two-level labelling scheme, where each posture type is associated with three postural variations of increasing level of ergonomic risk for physical body strain/discomfort. Some examples of ergonomically sub-optimal working postures and the classification results are shown in Figure 4.

With the aim to capture the full-body motion of a worker during work activities, we rely on RGB image sequences acquired by conventional, low-cost cameras to estimate rich 3D skeleton-based representation of the human poses using state-of-the-art methods for marker-less 2D and 3D human pose estimation. This feature exempts the requirement of inconvenient and expensive body-worn sensors in the real working environment. The 3D skeletal sequence is then used to train a spatio-temporal Graph Convolutional Network model (ST-GCN) [9] for learning to encode the input representations of the target working postures into a new shared embedding/feature space. Finally, the pairwise temporal alignment cost between the embeddings of an unlabelled 3D skeletal sequence with the embeddings of a 3D skeletal sequence of known class type is computed using the soft Dynamic Time Warping approach (softDTW) [10]. The pairwise cost is used as a similarity measure for the classification of the unlabelled sequence among the target types of working postures. The outline of the proposed approach is shown in Figure 2.

In the following, the main steps of the proposed approach, additional information on the deep learning techniques utilized and the types of the target ergonomic working postures are reported.

#### 3.1.1. Human Pose Estimation

Given a RGB video  $V$  of length  $T$  of a single line worker that performs a single or multiple actions, we use two existing state-of-the-art, deep learning-based methods to estimate the 2D and the 3D skeleton-based body pose per frame/image in real-time. Specifically, we employ the popular OpenPose method [25] to estimate the 2D human body in each frame  $I_t$  in  $V$ .



**Figure 3.** The skeletal body model that was originally presented in [78] to introduce and compile the NTU RGB+D large-scale dataset for 3D human action recognition. The hierarchical skeletal model comprises the following labelled body joints: (1) base of spine, (2) middle of spine, (3) neck, (4) head, (5) left shoulder, (6) left elbow, (7) left wrist, (8) left hand, (9) right shoulder, (10) right elbow, (11) right wrist, (12) right hand, (13) left hip, (14) left knee, (15) left ankle, (16) left foot, (17) right hip, (18) right knee, (19) right ankle, (20) right foot, (21) spine, (22) tip of left hand, (23) left thumb, (24) tip of right hand, (25) right thumb. The 3D user-centric coordinate reference frame (blue axes) is estimated based on the 3D torso frame using the skeletal joints that are included in the shaded blue rectangle area and aligned with the base of spine joint (1).



**Figure 4.** Snapshots of car door assembly activities captured in a real manufacturing environment and experimental results of the estimated 3D human poses (overlaid as colour coded skeletal body model) and the classification of working postures that are associated with the ergonomic risk for increased physical strain (text overlaid). We apply markerless (unobtrusive) vision-based pose estimation to recover the 2D skeletal body poses of a worker using the Openpose [25] method and subsequently lift this information in 3D space using the MocapNet2 [24] model. The sequence of 3D body poses is further analysed using a combination of Spatial Temporal Graph-based Convolutional Network model [9] and soft Dynamic Time Warping [10] to classify into a set of ergonomically sub-optimal working postures.

Openpose [25] is the first open-source real-time approach for multi-person 2D pose detection in images. The core feature of this method refers to an explicit nonparametric representation of the associations among the estimated body keypoints, called Part Affinity Fields (PAFs), which encodes both position and orientation of the candidate human limbs. A Convolutional Neural Network model is used to learn how to estimate the observed

2D body poses that features two branches; the first one predicts a set of 2D confidence maps that indicate the locations of body parts, while the second branch defines a set of 2D vector field of PAFs. In essence, the output of this approach comprises a set of 2D image coordinates  $(x, y)$ , that corresponds to the locations of  $K = 25$  skeletal body joints in the image plane according to the BODY25 pose output format. The coordinate vectors of all joints are concatenated to form a single feature vector per frame, noted as  $P_{2D}(t) = \{j_{2D}(1, t), \dots, j_{2D}(K, t)\}$ , where  $j_{2D}(i, t) = (x(i), y(i))$ ,  $i = [1 \dots K]$  the 2D image coordinates of each joint  $i$  for frame  $t$ .

Since we are interested in exploiting skeleton-based human poses to model and analyse human activities and time-varying postures that involve subtle body configurations and motion patterns, an important step to consider is to lift the estimated 2D information in 3D space. This step will enable the accurate and fine-grained estimation of static body configurations and will allow to disambiguate the spatio-temporal relationships among body parts according to a 3D skeletal body model. Then, the extraction of discriminative spatio-temporal skeletal features will drive the effective classification of the observed sub-optimal working postures. With this aim, the estimated 2D skeletal body pose  $P_{2D}(t)$  is used as input to the MocapNet2 [24] method that is able to predict in real-time the 3D human body pose for the  $I_t$  image frame. The MocapNet2 method encodes the input 2D skeleton (joints) hierarchy into two Normalized Signed Rotation Matrices (NSRMs), one for the upper body and one for the lower body. An NSRM is a translation and scale invariant representation of the 2D human pose that encodes joints in relation to each other by storing their Euclidean distances in the 2D image plane. The two NSRM representations are provided as input to an orientation classifier that comprises an ensemble of Self-normalizing neural network models (SNNs) [84] that uses 8 densely connected hidden layers in order to predict among the front, back, left, right body orientation classes. In the following, another ensemble of SNNs that comprises six hidden layers and is specifically trained for the estimated body orientation is used to estimate the 3D body pose.

It is important to note that the body hierarchy is split into the upper and the lower body parts that are estimated independently. This important design feature allows to tackle cases of extreme body occlusions, that frequently occur especially in real-world scenarios that feature semi- or unconstrained environments. For example, even when the whole lower body is occluded, the method can still efficiently estimate the pose of the upper body parts. Moreover, an Inverse Kinematics mechanism is used for the refinement of the regressed 3D skeletal body pose.

We use the pre-trained model of the the MocapNet2 architecture that encodes its estimated 3D human pose for frame  $I_t$  in  $V$  directly to a Biovision Hierarchy (BVH) character animation file format [85] using 498 body motion fields. These correspond to the degrees of freedom of the depicted human skeletal model (armature) and accommodate the estimated 3D coordinates and angles of the body, the human face, hands and feet. Only 87 of the degrees of freedom of the human skeletal body model are estimated by the method, leaving the rest to their default values in the BVH output file. The output 3D human pose for each frame  $I_t$  of the input video  $V$  is noted as  $P_{BVH}(t)$ .

### 3.1.2. Spatio-Temporal Modelling of the Human Motion

We further process the rich information for the 3D human body pose  $P_{BVH}(t)$  estimated by the MocapNet2 method for each frame  $I_t$  by selecting the 3D degrees of freedom that correspond to 25 main skeletal body joints. These target joints rely on the skeletal body model/configuration used in [78], as shown in the Figure 3. Our aim is to retrieve the 3D information of these skeletal body joints from the  $P_{BVH}(t)$  and transform them to 3D user-centric, rotation/view-invariant 3D coordinates using the concept of the "Torso PCA frame" (TPCAF) introduced in [86,87]. This method considers the human torso as a rigid body, that is a good candidate to define a torso frame, as a robust 3D user-centric coordinate reference system for the skeletal body joints. The torso frame, essentially refers to the orthonormal basis of a 3D plane computed using Principal Component Analysis (PCA) on

a 7-by-3 torso matrix consisting of the coordinates of the most proximate joints to the torso from the  $P_{BVH}(t)$ , as shown in (blue shaded rectangular) Figure 3. Therefore, the three main principal components define the 3D coordinate reference system, while the origin of its axes is conventionally attached to the base spine joint. We finally calculate the 3D coordinates of each body joint with respect to the estimated user-centric coordinate reference system to obtain a robust, rotation-invariant 3D skeletal body representation of the human pose observed in the input frame  $I_t$ . The outcome of this approach is a 75D feature vector  $P_{3D}(t) = \{j_{3D}(1, t), \dots, j_{3D}(K, t)\}$  that includes the 3D coordinates  $j_{3D}(i, t) = [x'_i, y'_i, z'_i]$  for all joints  $i = [1 \dots K]$ , where  $K = 25$ , of the estimated human pose for frame  $I_t$ . We note as  $P_{3D}$  the skeletal sequence of size  $[75 \times T]$  represented by the 3D coordinates of each human joint  $P_{3D}(t)$  for frame  $I_t$  in  $V$ , where  $t = 1 \dots T$ .

We use this information as input to a Spatio-temporal Graph Convolutional Network model (ST-GCNs) [9], as shown in Figure 2 for learning to analyse the spatio-temporal dynamics of body configurations and to extract discriminative representations of the human motion and actions performed throughout the video  $V$ .

The ST-GCN is a neural network model that is constructed as a undirected spatial temporal graph  $G = (V, E)$  based on a skeleton sequence with  $K$  joints and  $T$  frames. The node set  $V = \{u(t, k)\}$  for  $t = [1 \dots T]$ , where  $k = [1 \dots K]$  corresponds to all the body joints in  $P_{3D}$ . In this graph model, the edge set  $E_S$  comprises of intra-body connections between joints in the same frame  $t$  are realized as a subset of graph edges according to the connectivity of the skeletal hierarchy, shown in Figure 3. The second subset of edges  $E_F$  refer to the inter-frame connections of each body joint  $u(t, k)$  to the same type of joint  $u(t + 1, k)$  in consecutive frames. Each graph node  $u(t, k)$  acquires the feature vector  $j_{3D}(k)$  that is the of 3D joint coordinate vector and a confidence value  $C_i(t) \in [0, 1]$  that is also provided by the MocapNet2 method and indicates the visibility of the  $i$ -th joint in frame  $t$ . Essentially, in this model the spatial body configuration, i.e., the locations and the dependencies of the human skeletal joints, is represented as a spatial graph-based CNN for each video frame.

The main strength and novelty of this model is the extension of the standard formulation of 2D convolutions to cases where the input features map resides on a spatial graph (directly connected nodes act as neighbouring image pixels), essentially a static graph-based CNN formulation. The concept of neighbourhood is then extended to the temporal domain to also include temporally connected joints for modelling the spatial temporal dynamics within an input skeleton sequence. Various partition strategies and scoring schemes to define sets of neighbouring nodes are assessed in [9]. The best modelling capacity and classification performance is achieved by the the spatial configuration partitioning, where for each set of neighbouring nodes they are labelled according to their distances to the skeleton gravity centre compared with that of the root node of the set. The ST-GCN model comprises a set of 9 layers of spatial-temporal graph convolution operators that are applied to the input skeletal sequence gradually in order to generate higher-level feature maps of the graph according to the hierarchical representation of the human body. These layers also feature the ResNet mechanism for sharing weights and can be grouped in sets of three with respect to their output channels: 64, 128 and 256 channels, respectively. A global pooling layer is also used on output of the last layer that provides a 256 dimension feature vector for each sequence, which is an encoded feature vector, in other words an embedding of the input sequence to a new embedding space. Then, a SoftMax classifier is used to transform the embedding values of the network to probabilities towards the target classes.

We train a ST-GCN model using a modified SoftMax layer attached to the global pooling layer towards the new set of target classes of interest, in order to optimize for the network weights. The optimized model will essentially be used as an embedding function  $f_\theta(\cdot)$  for the proposed classification approach (Figure 2) to encode an input skeletal sequence  $P_{3D}$  to a sequence  $f_\theta(P_{3D})$  of 256-dimensional embedding vectors. We follow the training protocol and parameter settings described in [9]. The set of target classes will be defined in the following section.

### 3.1.3. Classification of Ergonomic Working Postures

We are interested in classifying a set of working postures in 3D skeletal sequences. The target classes of working postures are derived from the *MURI* risk analysis method [7,8] that directly links the observed body configurations to the ergonomic risk for increased physical strain imposed to specific body joints and parts during work activities. According to the World Class Manufacturing strategy (WCM) [88], the *MURI* risk analysis is a generic and widely-used tool for efficiency evaluation and risk analysis of physical ergonomics in workstations in different production contexts [2] and especially in the automotive industry. Overall, nine types (sketches) of time-varying working postures are assessed by the *MURI* analysis method. In our study, we opted for four of these types that are illustrated in Figure 1. These types of working postures were qualified as they affect different main joints and large parts of the human body for which the physical strain is more critical to assess during work activities, in terms of ergonomics and occupational safety. Moreover, the respective body configurations and specific skeletal features can also be efficiently captured visually and analysed across time, i.e., compared to the rotation angle of the worker's wrists, based on the camera positioning setting that is available in the actual workplace of our use case, that is reported in Section 4.1. The selected types refer to: (a) rotation angle of the waist, (b) flexion, stretching angle of the knees, (c) flexion angle of the waist, (d) height of working arms. Each type is further analysed into three postural variants that are associated with increasing level of ergonomic risk for physical strain/stress imposed to specific body parts/joints. These variants refer to the low ('Level 3'), medium ('Level 2') and high ('Level 1') risk level, that are quantified according to specific criteria linked to the pose-based angles and positions of the body parts. The low risk variants of the postures correspond to a neutral body pose of low or no ergonomic risk for physical strain. An important note is that each working posture is realized as time-varying event; thus, a sequence of body configurations with a duration of at least 4 s. Some of the sub-optimal working postures classified by the proposed method are also shown in Figure 4.

Based on this analysis, a set of 9 (nine) target classes of working postures is defined that comprise the high-risk ('Level 1') and the medium-risk ('Level 2') variants for each of the four main types of working postures and a single class of low-risk for all four types, considered as an optimal or neutral working body posture. Specifically, we note the set of target classes  $C = \{C_1, \dots, C_L\}$ ,  $L = 9$ , that correspond to the labels: flexion-waist-L1 and -L2, rotation-waist-L1 and -L2, flexion-stretch-knee-L1 and -L2, height-arm-L1 and -L2 and neutral-L3. We train the ST-GCN model using a training set of labelled 3D skeletal sequences  $\{P_{3D}\}$  against the set of the nine target classes  $C$  to learn the embedding function  $f_\theta$ , as described in Section 3.1.2.

Rather than using the outcome of the SoftMax layer of the trained ST-GCN model to directly classify an input skeletal sequence, noted as  $X_{3D}$ , we formulate a classification scheme that relies on pairwise comparisons between the sequence of embeddings of  $X_{3D}$  towards a support set of sequences comprising one or more labelled skeletal sequences  $\{Y_{3D}(i)\}$  that are considered as representative training examples for each of the target classes  $\{C_i\}$ . The estimated measure scores can be used to assign the unlabelled sequence to the most similar class using a simple 1-nearest-neighbour (1-NN) classification scheme, which can easily be extended to generic  $K$ -NN. This would allow to tackle the problem of high intra-class variability of the different types of time-varying working postures and will enable the fine-grained representation of their temporal evolution. Additional problems that we need take into account regards the speed of execution of the postural events, the varying sampling conditions, the fact that the working postures of interest are not synchronized temporally within the captured videos/skeletal sequences and their duration may vary significantly, while other postural events that might be performed before or after the target working postures are also captured as part of the captured sequence, especially during work activities captured in a real-world scenario. We argue that this design approach better fits to the requirements of the task of classifying body straining working postures.

With this aim, we define a classification measure based on the temporal alignment between skeletal sequences for measuring their discrepancies that also seems an promising approach to effectively encode the above invariances. Specifically, we use the soft Dynamic Time Warping method (softDTW) [10], an extension of the the classical Dynamic Time Warping (DTW) approach [89], to estimate the non-linear temporal alignment cost between two skeletal sequences as our classification measure, as shown in the outline of the proposed approach in Figure 2. A brief description of both approaches is provided in the following.

Given two multivariate data sequences of varying length  $X = (x_1, \dots, x_l) \in \mathbb{R}^{n \times l}$  and  $Y = (y_1, \dots, y_m) \in \mathbb{R}^{n \times m}$ , the classical Dynamic Time Warping (DTW) approach [89] uses as input a cost matrix  $D(X, Y) = [d(x_i, y_j)]_{ij} \in \mathbb{R}^{l \times m}$ , where  $d(x, y)$  is the Euclidean distance between any pair of timestamped  $p$ -dimensional feature vectors  $x_i$  and  $y_j$ . We also define  $\Pi$ , the set of all continuous and monotonic paths that realizes any temporal alignment between  $X$  and  $Y$ , connecting the upper-left to the lower-right of the matrix  $D$ . Finally, let  $\pi \in \Pi$  be one of all those alignments. The inner product  $\langle \pi, D(X, Y) \rangle$  yields the alignment score associated with  $\pi$ . DTW uses dynamic programming to estimate the minimum-cost temporal alignment between  $X$  and  $Y$  sequences, that is their discrepancy. On the basis of the above notation, this is:  $DTW(X, Y) = \min_{\pi \in \Pi} D(X, Y)$ .

Soft-DTW [10] builds upon the original DTW measure and considers a generalized soft minimum operator applied to the distribution of all costs spanned by all possible alignments between two data sequences of variable size. It also provides a differentiable loss function that can be computed with quadratic time/space complexity. Given the following generalized minimum operator, subject to a smoothing parameter  $\gamma \geq 0$ ,

$$\min \gamma(\pi_1, \dots, \pi_k) = \begin{cases} \min_{i \leq k} \pi_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^k e^{\pi_i/\gamma} & \gamma > 0, \end{cases} \quad (1)$$

the soft-DTW score is defined as:

$$sdtw_{\gamma}(X, Y) = \min^{\gamma} \{ \langle \pi, D(X, Y) \rangle, \pi \in \Pi \}. \quad (2)$$

The original DTW score is obtained by setting  $\gamma = 0$ .

In order to further formulate the proposed classification scheme, we select a video  $V_{Y,i}$  of length  $M$  annotated for each of the target classes  $C_i$  of working postures, that captures representative examples of postural performance towards this class. Each of these sequences will be compare with an input, unlabelled video  $V_X$  of length  $N$  in terms of pairwise temporal alignment. We estimate the 3D skeletal sequences  $P_{3D}(X), P_{3D}(Y_i)$  that are subsequently transformed to sequences of embeddings using the ST-GCN encoder, noted as  $f_{\theta}(P_{3D}(X)), f_{\theta}(P_{3D}(Y_i))$ , respectively. The pairwise Euclidean distance-based cost matrix  $D(i, j) = |f_{\theta}(P_{3D}(X)) - f_{\theta}(P_{3D}(Y_i))|^2$ , where  $D \in \mathbb{R}^{N \times M}$  is computed and used as input to the softDTW approach in order to estimate the soft-minimum alignment cost  $sdtw^{\gamma}(P_{3D}(X), P_{3D}(Y_i))$  between  $P_{3D}(X), P_{3D}(Y_i)$ . We set  $\gamma = 0.1$  for our experiments. Finally, we normalize the alignment cost using the  $|L_{X,Y_i}|$  length of the alignment path  $L$  in order to set the measure invariant to variable lengths of the input sequences:

$$sdtw_{\gamma, norm}(P_{3D}(X), P_{3D}(Y_i)) = \frac{sdtw_{\gamma}(P_{3D}(X), P_{3D}(Y_i))}{|L_{X,Y_i}|}. \quad (3)$$

In the following, the distance-based measure scores between  $X$  and  $Y_i$  for each target class  $C_i$  is computed in order to estimate the probabilities  $p(V_X, C_i)$  using the SoftMax function. Finally, the unlabelled input video  $V_X$  is assigned to the class

$$C_i = \arg \max_{C_i \in C} [1 - p(V_X, C_i)]. \quad (4)$$

Figure 2 illustrates the outline of the proposed approach. At the inference stage, the proposed approach is able to perform real-time classification of ergonomic working

postures using a coarsely segmented 3D skeletal data sequence of duration up to 30 s with a sampling rate of 25–30 fps, as input. Online performance of the method (real-time with low latency upon the occurrence of a temporal events) is also feasible for processing untrimmed streaming (continuous) visual data, if our approach is combined with a typical sliding window design [90] or a deep learning-based technique for online, joint action detection (temporal localization and classification) [91,92].

### 3.2. Worker Heart Rate Forecasting

As also mentioned above, the present work is part of a broader complex system that monitors assembly workers in order to estimate their physiological state and trigger personalised recommendations that will help to enhance their occupational safety. In such real-world applications it is useful to make predictions about the evolution of the employee's heart rate, in order to proactively foresee adverse, high-risk events, and be able to take timely corrective actions. It is therefore critical to examine whether the detection of ergonomically sub-optimal postures can be used as information-rich inputs that improve short- and mid-term heart rate predictions.

This can be formulated as a regression problem for forecasting heart rate (dependent variable) using Recurrent Neural Networks (RNNs) by considering that future heart rate values are affected by the current heart rate and physical strain due to visually detected ergonomically sub-optimal postures performed by the worker (independent variables). RNNs are frequently used in the literature to address regression problems with input/output looping [93]. However, most RNNs suffer from the problem of vanishing/exploding gradients, which hampers learning of long data sequences and makes them impractical for applications assuming real-time operation without human supervision. The Long Short-Term Memory (LSTM) neural networks [13,14] have showed significant less sensitivity on the vanishing/exploding gradients problem, because of their internal mechanism for balancing between updating and forgetting. Accordingly LSTMs have been particularly effective in learning long-term dependencies and dealing with pattern recognition problems in timeseries where the order of input is a key factor for signal evolution.

As there several variations of the LSTM that fit to different types of problems, the current work has considered the use of three different LSTM variants to identify the one that suits more to the heart rate prediction task. In particular we have examined a single layer classic LSTM, stacks of classic LSTM, and bidirectional LSTM. The first approach assumes the simplest architecture, which was unable to adequately cope with HR regression. On the other side the third, bidirectional approach, has introduced unreasonable complexity into the network which demands much more computational resources for training, without actually exploiting the power of bidirectionality since HR prediction is in fact a unidirectional problem. The second stacked LSTM approach has provided a balanced, powerful enough approach to implement HR prediction based on historical data. The stacked LSTM can effectively exploit the two different types of input data considered in the present study, namely (i) past heart rate and (ii) past physical strain indicators estimated based on the detection of ergonomically sub-optimal working postures, to implement models that forecast the worker's heart rate several seconds ahead. Accordingly, we consider two different LSTM architectures, one that uses only past heart rate data, and another that uses both past heart rate and past physical strain indicators. If future heart rate and the current postural performance are correlated, then it is expected that the second network will have better performance in forecasting the upcoming heart rate of users at work.

### 3.3. Associating Worker Heart Rate with Physical Strain

In order to examine the association between the physical strain and the heart rate that is an indicator of fatigue, it is necessary to monitor workers' activities on the assembly production line for a long period. This monitoring will provide two separate synchronized data streams for each worker, as a pair of time series, for (a) estimated types (risk-level scores) of the sub-optimal working postures performed by the worker during assembly

activities and (b) the worker heart rate, and consequently the correlation between the two time-series will denote an association between the two. We obtain the required pairs of time-series for the postural performance and the heart rate data of two different workers over 8 long periods of assembly activities, as described in detail in Section 4.1.

The next step is to examine the correlation between the different types of working postures and the resulting heart rate. This correlation can be measured directly between any two time series using any correlation coefficient such as the Pearson product-moment correlation coefficient (PPMCC) or Spearman's rank correlation coefficient [94], in order to examine whether an ergonomically sub-optimal working posture has a direct effect (i.e., increase) on the heart rate and a normal and non-straining body posture helps in reducing heart rate back to its normal values. However, since we need to measure the correlation between time series, any cross-correlation metric that takes into account the lag that may exist between the two time series (i.e., body strain and heart rate) is more appropriate. In addition to this, we must also consider the fact that the body strain time series in the assembly line is non-stationary [95], since there is a periodicity in the tasks performed in the line, and consequently in the expected strain in the different body parts. The time-lagged cross correlation between any body-strain time series and the heart rate time series can be measured using any correlation coefficient, such as PPMCC. Consequently it will reveal whether a continuous, sustained period of physical strain results in an increased level of worker heart rate and the opposite.

#### 4. Data Acquisition and Experimental Evaluation

In this section we describe the requirements, the acquisition and annotation process of a new dataset for detecting physical strain of line workers during assembly activities. The dataset comprises synchronized videos and sequences of worker physiological (heart rate) data acquired during work activities in a realistic manufacturing environment. In the following, the experimental evaluation of the proposed methodology using the new dataset is analysed in three parts. Firstly, the quantitative analysis of the proposed vision-based approach for detecting physical strain of workers in videos is presented in comparison with two baseline classification techniques. In addition, we assess the performance of the proposed LSTM-based approach for forecasting worker heart rate data during work activities. Finally, we investigate the correlation of workers' cardiovascular activity with the ergonomically sub-optimal postures detected during work activities and their role in predicting the increased ergonomic risk for physical strain imposed on the worker's body.

##### 4.1. Data Acquisition

To facilitate the implementation and the evaluation of the proposed methodology, we collected synchronous visual and physiological data for 2 workers during car-door assembly activities for a simulated production line in a realistic manufacturing workplace, that is available at the CRF-SPW Research & Innovation department of the Stellantis group in Melfi, Italy. Overall, data were captured at four random days in a single month, during the morning or the afternoon work shift for different workstations of the production line. The data comprise in total 8 work sessions, each with 5 consecutive task cycles, performed by an individual worker that was assigned to a specific workstation for the session. Each task cycle comprises a workstation-dependent sequence of up to 30 assembly actions and has a duration of approximately 4 min, while each work session has a duration of between 17–23 min.

The data collected, formed a new multimodal dataset with 40 task cycle executions of time-synchronized visual (RGB and depth) data and physiological data. A subset of 12 task cycles was selected and annotated by experienced professionals in manufacturing and ergonomics with respect to the MURI-based ergonomic working postures [7,8] (Figure 1). Additional information on the available annotations is provided in the following.

We follow a low cost, unobtrusive (non-invasive) sensing approach for the acquisition of visual data and cardiovascular activity of workers that allows them to perform ordinary

assembly activities in the real working environment without the need for the installation of special expensive equipment and wearable suits/reflectors (i.e., a motion capture (mocap) system or medical devices). Thus, the proposed solution is potentially applicable across the whole production line.

#### 4.1.1. Visual Data and Annotations

We have acquired visual sensory data of the workers using a set of four low-cost stereo cameras (*StereoLabs<sup>TM</sup>* ZED sensors) installed in the real manufacturing workplace. Each stationary camera is placed at a height of 1.8 m and captures time-synchronized visual data that comprise a stereo RGB image sequence and a depth image sequence of 1080p resolution at 30 frames per second. A pair of cameras is placed at each side of and along the production line to simultaneously capture the human assembly activities from 4 different viewpoints, thus to monitor both the inner-door and outer-door working areas of the workstations. Moreover, the sequences of visual data acquired by the cameras are time-synchronized using a common reference clock. Specifically, visual data for each of the 40 task cycles was captured from a single viewpoint located at the one side or from two viewpoints located at both sides of the observed workstation, therefore monitoring the activities from both the inner and the outer side working areas.

Annotation data for the selected 12 task cycles include the following information, as also shown in the example annotation table in Figure 5: (a) the temporal boundaries (start and end timestamps) and the semantic label for each assembly action (one action per row) performed by the worker during the task cycle, (b) the instances of the target types of ergonomic working postures of interest (noted in columns) for each assembly action (row), and (c) the overall ergonomic risk score for the task cycle execution estimated according to the MURI risk analysis method [7,96], as shown in Figure 1 and described in Section 3.1. Annotations toward the three risk levels for each type of working posture correspond to high, medium, low risk as semantic labels, to red, yellow and green as color-coded labels, and to integers 1, 2, 3 as numerical scores, respectively, that correspond to the labels shown in Figure 1. As a reminder, we note that each working posture is realized as time-varying event, represented as a sequence of body poses of minimum duration 4 s.

MURI ANALYSIS						FLEXION ANGLE OF THE WAIST		ROTATION ANGLE OF THE WAIST		HEIGHT OF THE WORKING ARM		FLEXION AND STRETCHING ANGLE OF THE KNEE							
TASK DESCRIPTION: User 124_WS30_May 24 morning						3	2	1	3	2	1	3	2	1					
Assembly front door						3	2	1	3	2	1	3	2	1					
Level 1 = red (high risk) 3 points Level 2 = yellow (medium risk) 2 points Level 3 = green (low risk) 1 point		Execution time in seconds (s)				> 30°	15° + 30°	0° + 15°	> 45°	15° + 45°	0° + 15°	> SHOULDER	= SHOULDER	= WAIST	> 60°	30° + 60°	0° + 30°		
N	operazione	ACTIVITY		start	end	ts	te												
1	10	Take screwdriver on line side and 4 screws from pouch		7:47	8:00														
2		Take one front left speaker from carriage and insert it correctly into retaining slots on carrier		8:00	8:08			2											
3		Place one screw at a time on the screwdriver tip and tighten four screws on the loudspeaker as shown in the sketch.		8:08	8:36			3			2								
4		Leave the screwdriver on the line side		8:36	8:45					1		1							
5	20	Take the door panel from cart and place on top of door frame.		8:45	8:50					2									
6		With both hands hook the superior part of the door panel on the door frame (inferior part of the window)		8:50	9:15					1			2				2		

**Figure 5.** A sample of annotation data for the posture-based ergonomic risk analysis (MURI analysis method [8,96]) of car-door assembly actions performed during a task cycle execution. Annotations were provided by experts in automotive manufacturing and ergonomics based on video observations. For each assembly action (rows), the ergonomic risk level for physical strain is noted towards each working posture type (columns) (image courtesy of Stellantis—Centro Ricerche FIAT (CRF)/SPW Research & Innovation department).

The set of 12 annotated task cycles comprise 310 assembly actions, each of average duration of 13 s, while the annotated instances for each type of working postures are reported in Table 1. Those instances were used for training the proposed vision-based classification approach using the combination of ST-GCN (<https://github.com/yysijie/st-gcn>, accessed on 10 January 2022)) and softDTW (<https://github.com/mblondel/soft-dtw>, accessed on 10 January 2022) methods. Finally, we apply the proposed classification approach to infer the occurrences of working postures for each worker in the videos of the assembly actions from the 28 unlabelled task cycles. The occurrences of working postures for all the 40 task cycles are synchronized with the acquired heart rate data for assessing the correlation between the two and to help in the prediction of near-future worker heart rate data, as described in Sections 4.3 and 4.4.

**Table 1.** The annotated instances of four types of ergonomically sub-optimal working postures, as shown in Figure 1, performed by 2 line workers during 12 task cycle executions of car door assembly activities. The videos of the assembly activities were analysed and annotated by experts in automotive manufacturing and ergonomics, as part of the multi-modal dataset of visual and heart-rate data of the workers presented in our study.

Posture Type	Flexion Angle of Waist			Rotation Angle of Waist			Height of Working Arm			Flexion/Stretch Angle of Knees		
	1	2	3	1	2	3	1	2	3	1	2	3
Risk level	1	2	3	1	2	3	1	2	3	1	2	3
Total	9	31	266	0	52	254	18	36	247	5	7	298

#### 4.1.2. Cardiovascular Activity Data

At the same time, each assembly worker was wearing a Garmin Vivoactive 3 smartwatch that provides measurements on their cardiovascular activity. In particular, a Heart Rate and Heart Rate Variability measurement can be obtained every second using the smartwatch. However, the Heart Rate Variability data are very unstable and sensitive to smartwatch misplacement resulting into long sequences of null values. Therefore, they are omitted from the present study which focuses on exploiting the much more stable heart rate data. The two data streams are synchronized to facilitate contrasting and fuzzing the two modalities as described below. The heart rate data sequences captured are synchronized with the visual data for all the 8 work sessions, that is the 40 task cycle executions that were recorded in the real workplace.

#### 4.2. Worker Posture Classification

We use the set of 12 annotated task cycles for the quantitative analysis of vision-based classification of ergonomic working postures. We temporally segment each assembly action of each task cycle using the annotation data in order to create a set  $\{V\}$  of 305 short videos for training and testing the proposed approach. Then, our goal is to classify each video against the set of target classes of ergonomic postures defined in Section 3.1. To measure the classification performance of the classification task, we employ the metrics of Precision, Recall and F1 score. Those metrics are commonly used in the fields of statistics, data science and information retrieval to evaluate the performance of classification models, by comparing the estimation obtained by such a model with annotation (ground truth) data. All metrics provide values in the range  $[0, 1]$ . The F1 score metric is the harmonic mean of Precision and Recall, where 1 indicates perfect precision and recall. We generate additional training samples using data from other video datasets that demonstrate similar human motion patterns in order to augment the training process of our approach. These regard a set of 600 skeletal sequences that were manually selected, segmented and annotated from videos of the TUM Kitchen Activity (<https://ias.in.tum.de/dokuwiki/software/kitchen-activity-data>, accessed on 10 January 2022), the Berkeley MHAD [97], and the NTU-RGB+D datasets [78].

Two additional classification approaches, namely rule-based and SVMs-based classification, were also developed to assess the performance of the 3D skeletal features extracted

from videos and to compare with the proposed method for detecting the physical strain during assembly activities. More details are reported in the following paragraphs. We use the following notations for the types of working postures to ease the description and the evaluation of those methods. We define the types of ergonomic working postures, shown in Figure 1, as  $[P_A, P_B, P_C, P_D]$ , that correspond to (a) the flexion of the waist, (b) the rotation angle of the waist, (c) the height of the working arms and (d) the flexion/stretching angle of the knees, respectively. The 3 postural variations of each posture type  $P_X$  are labelled based on the set  $L_X = [L_{X,1}, L_{X,2}, L_{X,3}]$ , that correspond to the associated high, medium, and low ergonomic risk level for physical discomfort and strain.

#### 4.2.1. Rule-Based Classification

A heuristic rule-based classification method [98] is developed to identify the postural variations  $L_X$  for each type of postures  $P_X$  in a 3D skeletal sequence  $S$ . The feature vector in  $S$  comprises the 3D body joints coordinates estimated per frame for a video of assembly activities, as described in Sections 3.1.1 and 3.1.2. To this end, we design a classification rule for each  $P_X$ , which is noted as  $R_X$  and consists of 3 branches. The conditions of each rule rely on a single pose-based attribute that is the feature value  $f_X$  as indicated by the specifications of  $P_X$ , as shown in Figure 1; e.g., the  $S$  is classified as  $L_1$  of posture type  $P_A$ , if the forward inclination of the skeletal body representation is more than 30 degrees, etc.

Then, for each  $P_X$ , the input sequence  $S$  is encoded to a 1D sequence of  $f_X$  values with the aim to apply the  $R_X$  and classify each value against the set of labels  $L_X = L_{X1}, L_{X2}, L_{X3}$ . Finally, all sub-sequences of contiguous  $L_X$  labels, where each has a duration of at least 4 s, are extracted and sorted according to the label priority  $L_{X1}, L_{X2}, L_{X3}$ , that is high, medium, low risk, respectively. The  $S$  is assigned the label of the first sub-sequence in this list. In essence, given a data sequence, if at least one sub-sequence is detected and classified as  $L_{X1}$ , along with one or more sub-sequences labelled as  $L_{X2}$  or  $L_{X3}$ , the sequence is labelled as  $L_{X1}$ , indicating postural performance of high ergonomic risk for physical strain. The same applies for the priority between the  $L_{X2}$  and  $L_{X3}$  labels, that indicate working postures of medium and low ergonomic risk for physical strain, respectively.

#### 4.2.2. Multi-Class SVM-Based Classification

We also formulate a supervised learning approach for the classification of working postures based on multi-class Support Vector Machine (SVM) models [99,100]. In this case, we treat the 4 top-level types of ergonomic working postures  $[P_A, P_B, P_C, P_D]$ , shown in Figure 1, in a different manner. We opt to train a multi-class SVM-based model  $M_X$  for each posture type  $P_X$ . The SVM models are trained independently to discriminate between the mutually-exclusive risk labels  $L_X$  of each  $P_X$ . We use the 3D skeletal sequences of the annotated videos of assembly actions, as described in Section 3.1, to build a vocabulary of  $B$  codebooks using the popular Bag-of-Features [101]. We use  $B = 200$  for the number of codebooks. Essentially, a new compact representation  $H_i \in R^{1 \times B}$  is generated that encodes each input the 3D skeleton-based sequence  $V_{skel,i}$ . In the following, we form the set of training samples  $T_X = [H_i, L_X]$  for the model  $M_X$ . A one-versus-rest training scheme was selected to train each model by using a linear Support Vector Classification (SVC) kernel model [99], a hinge loss function, the  $L_2$  measure as a penalty parameter and a 4-fold cross validation training setting for the set of training pairs  $T_X$ . Furthermore, we use a balanced class weighting scheme to adjust the weights inversely proportional to class frequencies in the input data and overcome the potential unbalanced frequencies of occurrences among the three  $L_X$  target classes in the training data.

#### 4.2.3. Quantitative Evaluation

The performance of the postural classification approaches is evaluated using the set of 305 videos of assembly actions that are temporally segmented from the 12 annotated task cycles, as mentioned in Section 4.1.1.

An overview of the performance for each classification approach is provided in Table 2. The average scores of the evaluated Precision, Recall and F1-score metrics are reported based on the classification results of each method for all types of ergonomic working postures. A comparison between the methods is also presented in Figure 6 based on the mean F1-scores achieved for each type of ergonomic working postures. Finally, Table 3 provides an analytic report of the F1-score values achieved by each classifier towards the  $L_X$  classes for all the types  $P_X$  of ergonomic sub-optimal working postures that are considered in our study.

**Table 2.** Experimental results obtained for the detection of physical strain during assembly activities using the videos of the 12 annotated task cycle executions. The mean value of each classification metric is computed for every type of ergonomic working postures.

Classification Method	Precision	Recall	F1-Score
Rule-based classifier	0.527	0.583	0.516
multi-class SVMs	0.603	0.860	0.680
ST-GCNs [9] + softDTW [10] (proposed)	0.653	0.822	0.710

**Table 3.** Experimental results on the detection of physical strain in videos of car door assembly activities performed by line workers in a real car manufacturing environment. F1-scores obtained by three classification approaches is presented for each of the four types of ergonomic working postures (see Figure 1 that are associated with physical strain, according to the MURI risk analysis method [7]).

Types of Working Postures Ergonomic Risk Level/Methods	Flexion Angle of the Waist L1/L2/L3	Rotation Angle of the Waist L1/L2/L3	Height of the Working Arm L1/L2/L3	Flexion/Stretching Angle of the Knee L1/L2/L3	Mean F1-Score
Rule-based classifier	0.34/0.56/0.77	-/0.24/0.70	0.30/0.60/0.70	0.28/0.42/0.75	0.516
multi-class SVMs	0.72/0.70/0.87	-/0.50/0.77	0.70/0.68/0.82	0.50/0.30/0.90	0.680
ST-GCN + softDTW	0.74/0.80/0.90	-/0.63/0.90	0.80/0.61/0.89	0.25/0.38/0.87	0.710



**Figure 6.** The average F1-score scores of each classification method are calculated and presented separately for each type of the ergonomic working postures (Figure 1).

The efficiency of the proposed deep-learning based classifier is evident based on the experimental evaluation conducted using the subset of annotated visual data and its superior performance compared with the two baseline methods.

Overall, the proposed framework can achieve real-time or on-line performance based on continuous streams of visual data acquired by a single camera in a real industrial setting. All methods of the proposed pipeline (data acquisition, 2D and 3D human pose estimation and posture classification) can be processed using a mid- to high-end PC equipped with a

CUDA-enabled GPU (in our case NVidia GeForce 1080ti or better; Compute Capability 6.1 or higher).

As we have already noted, each type of working posture is realized as a time-varying event that has a duration of at least 4 s. The proposed classification method can be combined with any action localization/detection approach for the online processing of continuous streams of visual data for the estimation of sub-optimal working postures. Therefore, in a real industrial setting the proposed approach is able to provide an online performance with a latency of some seconds upon the completion of any observed postural event based on continuous streams of visual data. Finally, we note that a first prototype of the proposed framework has already been tested in an unattended operation mode in a real car manufacturing environment with real line workers (Stellantis—Centro Ricerche FIAT (CRF)/SPW Research & Innovation department in Melfi, Italy) in the context of the sustAGE project for a period of 3 weeks (approximately 5 h per day). Specifically, the respective software module was able to classify the observed ergonomic working postures based on a continuous stream of visual data acquired from a single camera during work activities and to achieve online runtime performance (with a latency of 5–10 s) and satisfactory results. Collective results obtained from the analysis of worker's postural performance after the end of a task cycle per workstation were reported (approximately after 4–5 min).

#### 4.3. Worker Heart Rate Forecasting

The recommendation system we are implementing assumes reliable predictions of workers' heart rate, which are used as a prevention mechanism for adverse, high-risk events. As mentioned above, we are interested to develop a module that predicts future heart rate of assembly workers, by using recent past measurements of heart rate and posture deviation observations. This data is used to train an LSTM neural network that aims to implement forecasting.

To prepare the dataset for LSTM training, we exploit the data collected following the procedure summarized above. The formatting of input/output pairs is implemented as follows. At any given moment, a 25 s length window of past data is used to create the input sequence used as an input vector. This data may include heart rate and posture deviations, depending on the LSTM configuration used. In the first case it results into an input vector of 25 (heart rate) values, while in the second it results into an input vector of 50 (heart rate + posture deviation) values. The posture deviation values referred to as total score which is the summary of annotated value at a specific time-frame. The corresponding value targeted by the output of the LSTM is a scalar produced by taking the measured heart rate either 10, or 20, or 30 s ahead the current time (the same applies for all LSTM configurations). The window of input data and the targeted value are moved in a step by step manner, one second each time, to produce the whole sequence of shifted pairs to create the input/output dataset. Following the above, we got 1256 input/output pairs which are used to train the LSTMs and assess their performance. In particular, 80% of the data are used for training and the remaining 20% are used for performance validation. In order to configure the parameters of the LSTM we use a validation dataset that includes the 25% of the training data.

To implement the heart rate forecasting module, we use bidirectional LSTMs of four stacked layers, each one consisting of 10 memory cells with hyperbolic tangent activation function. Additionally, a dropout layer of 0.1 is used between all connected layers to prevent over-fitting on training data and at the end of the network a dense layer is added that is deeply connected with its preceding layer which means the neurons of the layer are connected to every neuron of its preceding layer. The LSTM was trained for 10,000 epochs with Adam optimizer with learning rate 0.001 to optimize 3.891 trainable parameters targeting the minimization of the absolute error between the targeted and the predicted heart rate values. The above are the same for all input/output combinations examined.

The results of training the LSTMs on heart rate forecasting using either only past heart rate data (HR only), or both past heart rate and posture deviation data (HR + PD) are

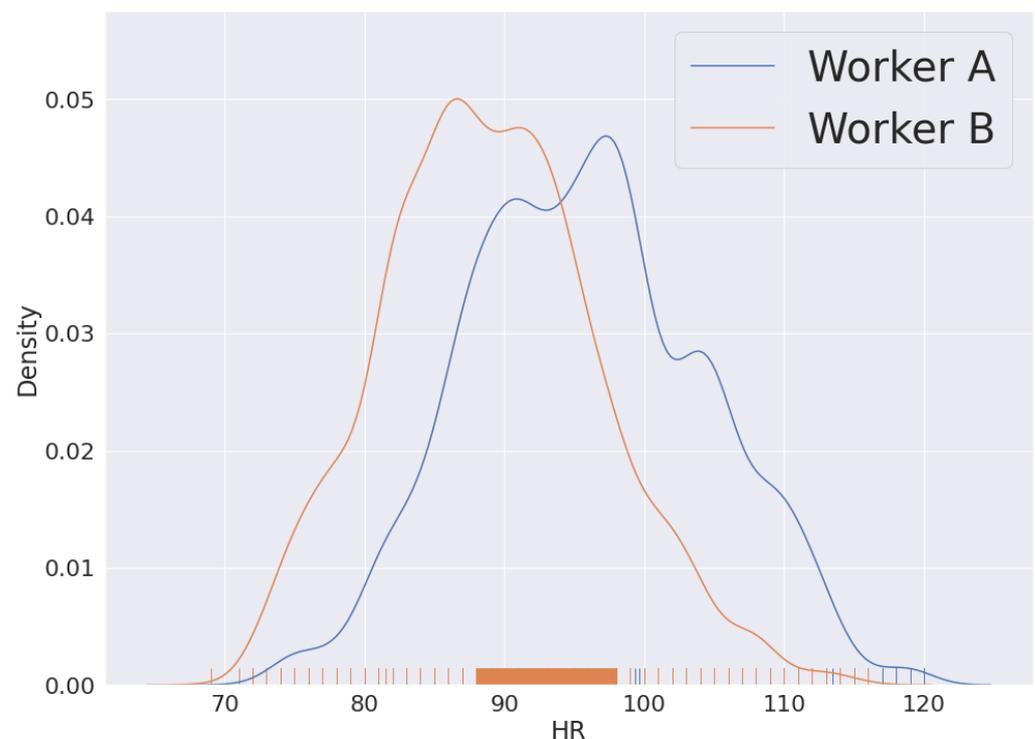
summarized in Table 4. The table reports the mean absolute error values for the training and the validation dataset, both averaged over the last 10 training epochs for a group of five training sessions on each dataset. According to these results, feeding the forecasting module with posture deviation data improves the LSTM training in all three cases, having a more beneficial effect on short term forecasting. The assessment of the forecasting module on the validation dataset provides information about its ability to generalize. Clearly both the HR-only and the HR + PD solutions achieve medium quality results with the HR-only solution being slightly better. We believe this is due to the small size of the datasets examined. It is expected that by increasing the available data, the forecasting error on the validation dataset will move closer to the forecasting error on the training dataset, which in our case seems to be in favour of the HR + PD solution.

**Table 4.** The results of LSTM training on heart rate forecasting for the next 10, 20, or 30 s.

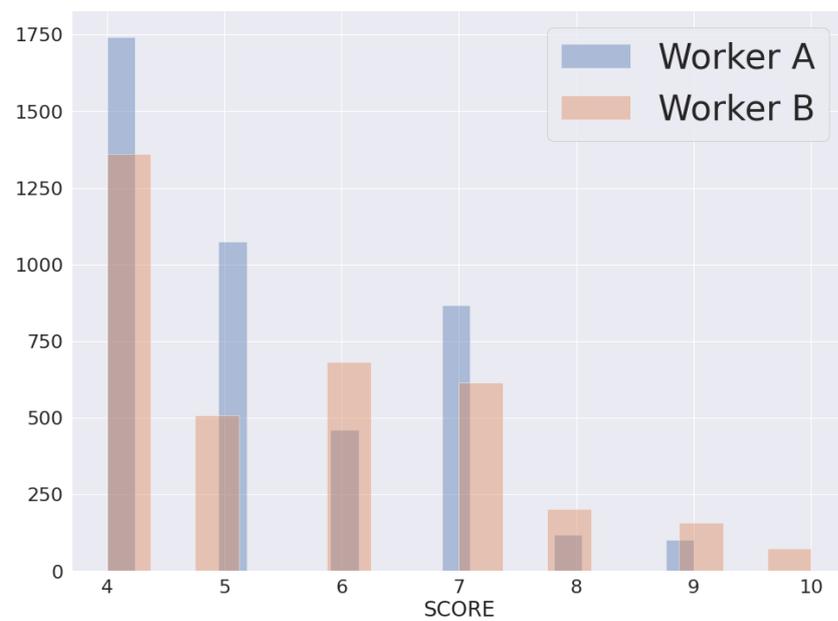
Input	Prediction 10 s		Prediction 20 s		Prediction 30 s	
	Training	Validation	Training	Validation	Training	Validation
HR only	2.02	4.86	2.32	7.93	3.86	7.38
HR + PD	1.02	5.03	1.93	8.34	1.96	6.22

#### 4.4. Integration Aspects

Since we compared the time series for two workers it is important to check their profiles first. As shown in Figure 7 the two workers have some differences in the frequency distribution of their HR values in all datasets. In general worker B has a higher HR than worker A. What is also interesting is that worker A, in general, stresses less than worker B, at least by avoiding either stressing a specific body part too much or multiple body parts at the same time, as shown in Figure 8 that depicts the distribution of the total stress score for the two workers in all files.



**Figure 7.** Workers' heart rate values distribution.



**Figure 8.** Workers' total body stress values distribution.

When we examine the correlation between the different body straining working postures and the resulting heart rate we see that there is no clear pattern between any of the postures (and the cumulative body stress) at any time and the heart rate, as shown in Table 5 that summarizes the correlation in all the 8 work sessions of the proposed dataset. This can be easily explained, since taking a specific occurrence of any working posture does not instantly triggers heart rate to increase, thus it is more important to examine if there is a time lagged correlation between HR and strenuous postures and identify this lag.

**Table 5.** Pearson correlation between the different body stressing postures and the corresponding heart rate.

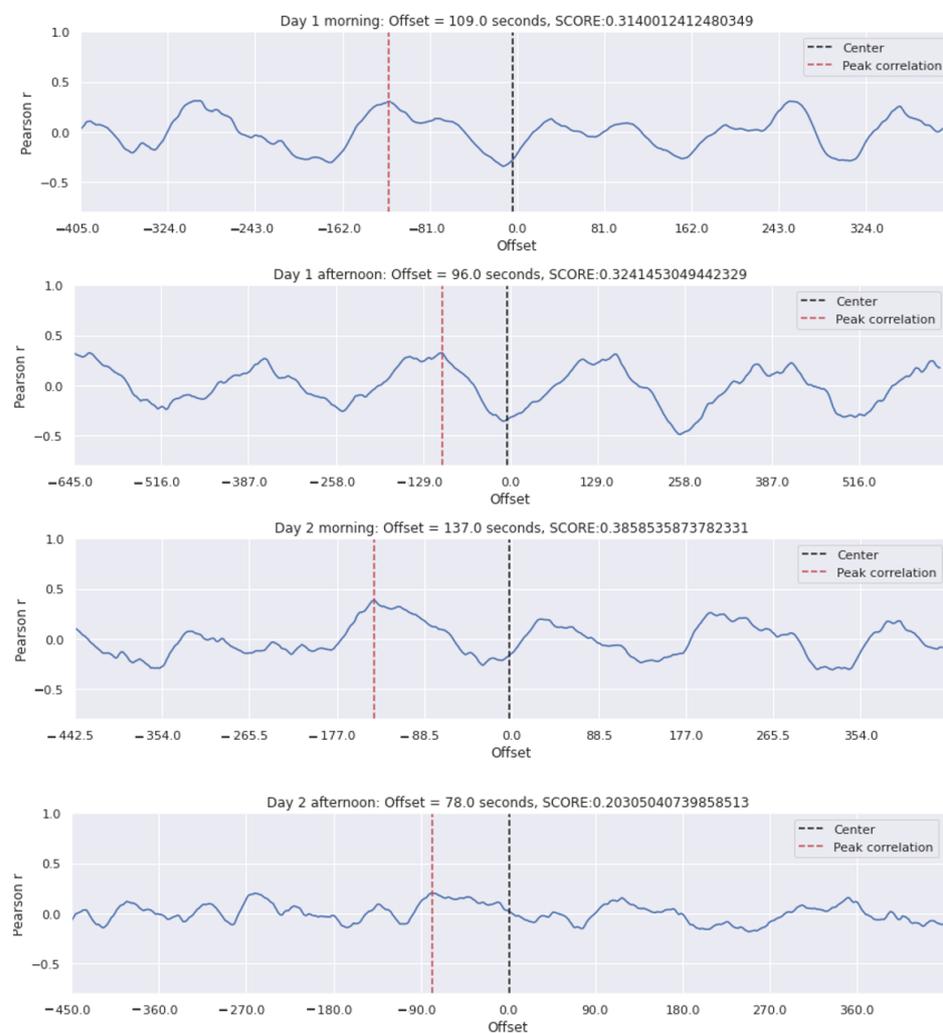
	Flexion Angle of the Waist	Rotation Angle of the Waist	Height of the Working Arm	Flexion and Stretching Angle of the Knee	Stress Score
Worker 1	0.11	0.21	0.01	0.11	0.23
	−0.18	−0.06	−0.44	0.08	−0.34
	−0.29	−0.14	0.17	−0.04	−0.17
	0.01	0.01	0.02	0.00	0.02
Worker 2	−0.17	−0.34	−0.15	0.10	−0.29
	−0.05	−0.10	0.01	−0.03	−0.07
	−0.03	−0.01	0.11	0.06	0.03
	−0.08	−0.07	−0.03	0.03	−0.08

In order to examine the occurrence of time-lagging correlation, we consider the heart rate time series of each work session as is and we gradually shift the respective body posture time series by a certain amount of seconds each time. Then we measure the Pearson correlation between the HR and the shifted body stress time series and produce a plot which is similar to the one depicted in Figure 9.

The blue line depicts the correlation between HR and the total (summed) stress level in all body parts if we shift the body posture earlier in time (negative offset) or later in time (positive offset). The black dashed vertical line maps the correlation at zero offset, which as shown in the previous table is usually close to 0. The red dashed vertical line depicts

the maximum correlation found between the two time series and identifies, in this case, a negative offset. This means that there is a maximum time lagged correlation of 0.278 between the total body strain score and heart rate, which occurs with a delay of 106 s. This means that the worker is undergoing a sub-optimal working posture imposing increased strain to body parts and almost 1.5 min later this results in an increase to the HR. Similarly, avoiding sub-optimal working postures for a while results in a gradual decrease in heart rate back to its normal levels.

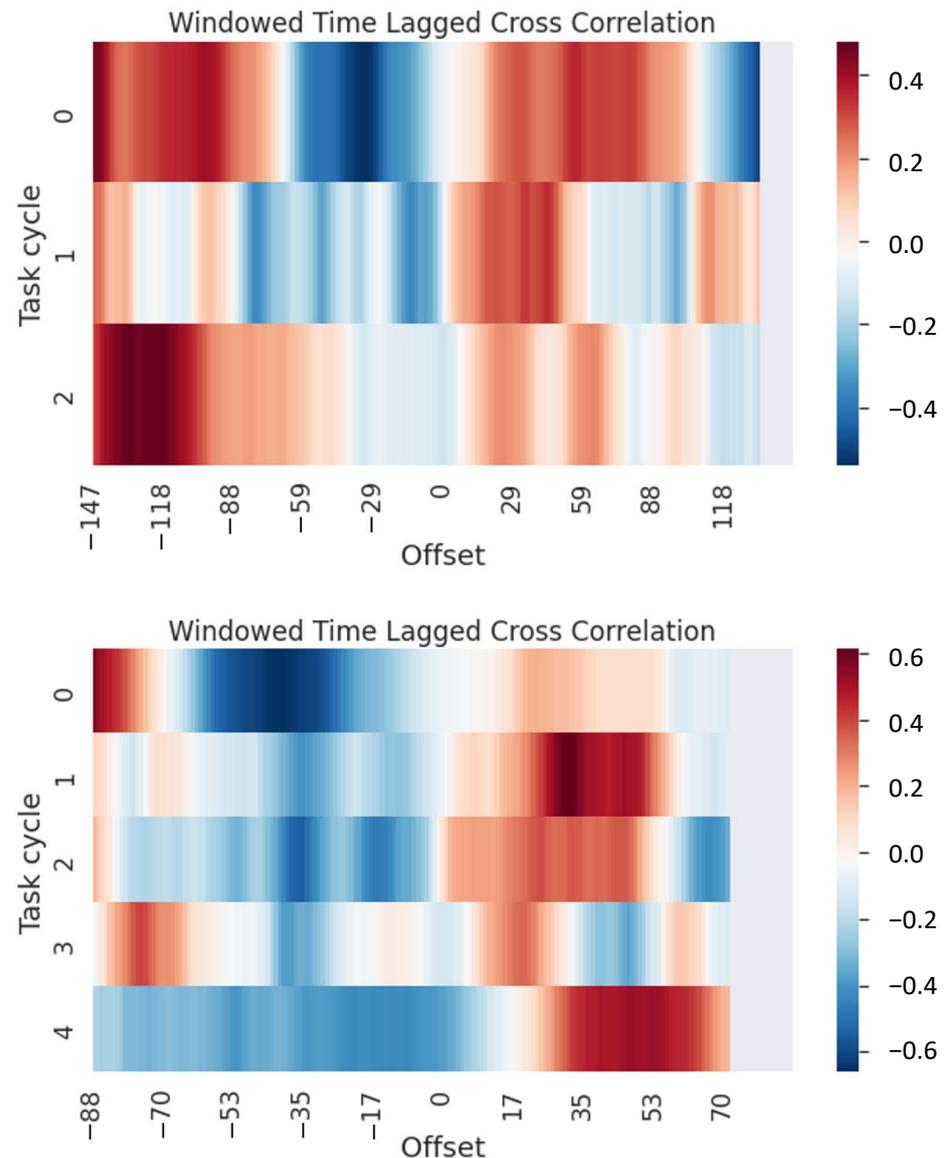
As shown in Figure 9 the pattern repeats in almost all work session analysed for both workers, which suggest that there is a time lagged correlation between HR and physical strain during assembly activities. The statistical significance of the reported correlation values has a  $p$ -value that is always smaller than 0.01, which indicates that the correlation does not occur randomly. What is even more interesting is the repeating pattern in most time lagged correlation plots. This repeating pattern matches the repeating nature of the tasks performed by the workers, which results in repeatedly making the same body stressing postures and consequently have an effect on their heart rate.



**Figure 9.** The time lagged correlation plots for Worker A. The dotted red line denotes the lag that gives the maximum Pearson correlation for the two series. A negative offset denotes that the heart rate time series follows the body stress one.

In order to further analyse the effect of this pattern on the correlation between HR and sub-optimal body postures, we provide an additional visualization that splits each time series to subsets of equal length (windows) and then computes the time lagged correlation

for each window separately. Since each dataset contains work cycles of 4 min duration, we split the dataset in windows that have a size of 240 s and apply time lagged correlation, using Pearson correlation again. The result is as shown Figure 10. The plot shows that continuous sub-optimal postures during the work cycle add a burden to the worker and result in an increase to the heart rate especially with a delay of approximately 2 min. The  $p$ -value for this correlation is higher but still less than 0.05 in all cases.



**Figure 10.** Windowed time lagged cross correlation of the body stressing working postures in each work cycle and the respective heart rate. The plots show a periodicity in the high correlation values during the cycles (rows).

## 5. Discussion and Conclusions

This paper studies methods of detecting the physical strain imposed by ergonomically sub-optimal body postures performed by line workers and examines their relationship to the worker cardiovascular activity during work activities in a real-world and demanding industrial environment. A key characteristic of the present work regards the use of low-cost technology for the acquisition of synchronous visual and cardiovascular data sequences of workers, based on stationary camera sensors and wearable sensors (smartwatches), respectively. The visual information stored in videos or acquired as untrimmed stream-

ing data would enable the unobtrusive and real-time detection and tracking of worker's 3D skeleton-based body configuration, using state-of-the-art vision-based deep learning methods. Therefore, no significant changes in the workplace or the use of markers on the worker's body is required, which would impede their movement and would reduce their physical comfort and productivity. According to the results presented above, the proposed vision-based classification approach acquires a coarsely segmented 3D skeletal data sequence and is able to efficiently assess the ergonomic suitability of worker's postural performance and the risk for physical strain during work activities. With this aim, the observed worker's postures are analysed according to a set of ergonomic working postures, of known ergonomic risk scores, suggested by the well established MURI analysis tool that is widely used for the analysis of physical ergonomics in different occupational contexts. It performs in real-time using a coarsely segmented 3D skeletal data sequence as input, of length up to 30 s and acquisition rate at 30 fps. Moreover it is able to perform in an online setting, as part of a detection and classification scheme using untrimmed streaming data as input.

In addition, we examine the correlation of workers' cardiovascular activity with the occurrences of ergonomically sub-optimal postures detected during work activities, that indicate the risk for physical strain imposed to worker's body. Although the connection between the two sources of information is not visually obvious, the current work shows that there is a latent correlation or relationship between the two sources of information. According to the assessment presented above, an estimation of increased risk for physical strain of a worker in a specific amount of time seem to affect the evolution of heart rate. This assumption is supported by the fact that by using the estimated risk for physical strain based the detected working postures as input to a heart rate prediction module the accuracy of predictions improves. Interestingly, further investigation has revealed a time lagged correlation between HR and physical strain for more than a minute.

Overall, the present work presents an early analysis of a dataset that combines the visual analysis of sub-optimal working postures and heart rate of line workers that perform repetitive tasks. More experiments would be required to further validate its findings. However, the initial results suggest a correlation between ergonomically sub-optimal postural performance by the worker and an increase in heart rate. In addition, the results show that the analysis of worker postures can help in the prediction of future heart rate values, and preemptively notify workers and their supervisors in the workplace about near-future high heart rate incidents.

Moreover, an additional goal for future work refers to the development of a novel end-to-end deep neural network model for multi-modal representation learning and classification of human actions and sub-optimal working postures. Such a model will rely on the fusion of visual and HR data and exploit our findings on the correlation of working postures and worker HR measurements. Finally, we plan to acquire new data and annotations for more work sessions in the real-world manufacturing environment and augment our multi-modal dataset for the detection of physical strain and fatigue during assembly activities.

**Author Contributions:** Conceptualization, K.P., I.V., M.P. and M.M.; methodology, K.P., G.P., M.M., M.P., I.V. and G.P.; software, K.P., G.P., M.M., I.V. and T.P.; validation, M.M., I.V. and M.L.; data curation, K.P., T.P. and G.P.; writing—original draft preparation, K.P., I.V., M.M. and M.P.; writing—review and editing, K.P., M.P., M.L., I.V. and M.M.; visualization, K.P., G.P., M.M. and I.V.; supervision, M.P. and I.V.; project administration, M.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has received funding from the European Union's Horizon 2020 research and innovation programs under grant agreement No.826506 (sustAGE) and grant agreement No. 101017151 (FELICE).

**Institutional Review Board Statement:** The study uses visual data and heart rate measurements of humans acquired during work activities in a pilot environment. The study was conducted according to the principles and procedures specified in: (a) General Data Protection Regulation (GDPR) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC, (b) Commission proposal for a regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data, COM(2012) 11. It will be considered by sustAGE when applicable, (c) ePrivacy Directive 2002/58/EC, as revised by Directive 2009/136/EC, in relation of the store or access to information stored in devices of users located in the European Economic Area, (d) Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine, 4.IV.1997, (e) National legal and ethical requirements as defined by the national data protection law and authorities of the Countries where data collected by sustAGE are controlled, in particular those that transpose the Directives/Regulations mentioned in the previous points. The study was approved by the Research Ethics Committee (REC) of the Foundation for Research and Technology (FORTH) (Protocol code 23/21 November 2018 and date of approval 24 September 2019). Written informed consent has been obtained from the participant(s) for the publication of results in anonymous form and in scientific publications aiming at informing the public and /or the scientific community.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The new multimodal dataset introduced in this study and used for the experimental evaluation of the proposed methodology will be available at ZENODO.

**Acknowledgments:** The authors thank Consortium partner Stellantis—Centro Ricerche FIAT (CRF)/SPW Research & Innovation department in Melfi, Italy, for their valuable feedback in the implementation and evaluation of this work.

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

## References

1. Vieira, E.R.; Kumar, S. Working postures: A literature review. *J. Occup. Rehabil.* **2004**, *14*, 143–159. [CrossRef] [PubMed]
2. Brito, M.F.; Ramos, A.L.; Carneiro, P.; Gonçalves, M.A. Ergonomic analysis in lean manufacturing and industry 4.0—A systematic review. In *Lean Engineering for Global Development*; Springer: Cham, Switzerland, 2019; pp. 95–127.
3. Bao, S.; Howard, N.; Lin, J.H. Are work-related musculoskeletal disorders claims related to risk factors in workplaces of the manufacturing industry? *Ann. Work Expo. Health* **2020**, *64*, 152–164. [CrossRef] [PubMed]
4. Pateraki, M.; Fysarakis, K.; Sakkalis, V.; Spanoudakis, G.; Varlamis, I.; Maniadaakis, M.; Lourakis, M.; Ioannidis, S.; Cummins, N.; Schuller, B.; et al. Biosensors and Internet of Things in smart healthcare applications: Challenges and opportunities. In *Wearable and Implantable Medical Devices*; Elsevier: Amsterdam, The Netherlands, 2020.
5. Ramaswamy, S. How Micro-Moments Are Changing the Rules. 2015. Available online: <https://www.thinkwithgoogle.com/marketing-resources/micro-moments/how-micromoments-are-changing-rules/> (accessed on 16 October 2010).
6. Athanassiou, G.; Pateraki, M.; Varlamis, I. Micro-moment-based Interventions for a Personalized Support of Healthy and Sustainable Ageing at Work: Development and Application of a Context-sensitive Recommendation Framework. In Proceedings of the 13th International Joint Conference on Computational Intelligence—SmartWork, Online, 25–27 October 2021; pp. 409–419.
7. Womack, J.P.; Jones, D.T. Lean thinking—Banish waste and create wealth in your corporation. *J. Oper. Res. Soc.* **1997**, *48*, 1148. [CrossRef]
8. Ciccarelli, M.; Papetti, A.; Cappelletti, F.; Bruzini, A.; Germani, M. Combining World Class Manufacturing system and Industry 4.0 technologies to design ergonomic manufacturing equipment. *Int. J. Interact. Des. Manuf. (IJIDeM)* **2022**, *16*, 263–279. [CrossRef]
9. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
10. Cuturi, M.; Blondel, M. Soft-DTW: A Differentiable Loss Function for Time-Series. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 894–903.
11. Papoutsakis, K.; Thodoris, P.; Maniadaakis, M.; Lourakis, M.; Pateraki, M.; Varlamis, I. Detection of physical strain and fatigue in industrial environments using visual and non-visual sensors. In Proceedings of the 14th Pervasive Technologies Related to Assistive Environments Conference (PETRA 2021), Corfu, Greece, 29 June–2 July 2021; pp. 270–271.
12. Mueller, M.J.; Maluf, K.S. Tissue adaptation to physical stress: A proposed “Physical Stress Theory” to guide physical therapist practice, education, and research. *Phys. Ther.* **2002**, *82*, 383–403. [CrossRef]
13. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
14. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

15. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
16. Liu, J.; Wang, G.; Hu, P.; Duan, L.Y.; Kot, A.C. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017; pp. 1647–1656.
17. Zhao, J.; Obonyo, E. Convolutional long short-term memory model for recognizing construction workers' postures from wearable inertial measurement units. *Adv. Eng. Inform.* **2020**, *46*, 101177. [[CrossRef](#)]
18. Rundo, F. Deep LSTM with Dynamic Time Warping Processing Framework: A Novel Advanced Algorithm with Biosensor System for an Efficient Car-Driver Recognition. *Electronics* **2020**, *9*, 616. [[CrossRef](#)]
19. Kuschan, J.; Krüger, J. Fatigue recognition in overhead assembly based on a soft robotic exosuit for worker assistance. *CIRP Ann.* **2021**, *70*, 9–12. [[CrossRef](#)]
20. Alam, M.A.U. Activity-Aware Deep Cognitive Fatigue Assessment using Wearables. In *Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Virtual, 1–5 November 2021; pp. 7433–7436.
21. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
22. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [[CrossRef](#)]
23. Lei, Q.; Du, J.X.; Zhang, H.; Ye, S.; Chen, D. A Survey of Vision-Based Human Action Evaluation Methods. *Sensors* **2019**, *19*, 4129. [[CrossRef](#)] [[PubMed](#)]
24. Qammaz, A.; Argyros, A.A. Occlusion-tolerant and personalized 3D human pose estimation in RGB images. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR 2020)*, Milan, Italy, 10–15 January 2021.
25. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [[CrossRef](#)]
26. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, 3–9 August 2013.
27. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23 June 2014; pp. 588–595.
28. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. Real time action recognition using histograms of depth gradients and random decision forests. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, USA, 24–26 March 2014; pp. 626–633.
29. Ma, C.Y.; Kadav, A.; Melvin, I.; Kira, Z.; AlRegib, G.; Graf, H.P. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6790–6800.
30. Heidari, N.; Iosifidis, A. Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 10–15 January 2021; pp. 7907–7914.
31. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Virtual, 14–19 June 2020.
32. Kim, T.S.; Reiter, A. Interpretable 3D human action analysis with temporal convolutional networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 21–26 July 2017; pp. 1623–1631.
33. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22–29 October 2017; pp. 2117–2126.
34. Zhang, S.; Liu, X.; Xiao, J. On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks. In *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 24–31 March 2017; pp. 148–157.
35. Zhang, C.; Gupta, A.; Zisserman, A. Temporal Query Networks for Fine-grained Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 20–25 June 2021; pp. 4486–4496.
36. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 15–20 June 2019; pp. 244–253.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.

39. Plizzari, C.; Cannici, M.; Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.* **2021**, *208–209*, 103219. [[CrossRef](#)]
40. Wang, Y.; Ajaykumar, G.; Huang, C.M. See what i see: Enabling user-centric robotic assistance using first-person demonstrations. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, UK, 23–26 March 2020; pp. 639–648.
41. Cramer, M.; Cramer, J.; Kellens, K.; Demeester, E. Towards robust intention estimation based on object affordance enabling natural human-robot collaboration in assembly tasks. *Procedia CIRP* **2018**, *78*, 255–260. [[CrossRef](#)]
42. Colim, A.; Faria, C.; Cunha, J.; Oliveira, J.; Sousa, N.; Rocha, L.A. Physical Ergonomic Improvement and Safe Design of an Assembly Workstation through Collaborative Robotics. *Safety* **2021**, *7*, 14. [[CrossRef](#)]
43. Kim, W.; Lee, J.; Peternel, L.; Tsagarakis, N.; Ajoudani, A. Anticipatory Robot Assistance for the Prevention of Human Static Joint Overloading in Human–Robot Collaboration. *IEEE Robot. Autom. Lett.* **2018**, *3*, 68–75. [[CrossRef](#)]
44. Kim, W.; Lee, J.; Tsagarakis, N.; Ajoudani, A. A real-time and reduced-complexity approach to the detection and monitoring of static joint overloading in humans. In Proceedings of the 2017 International Conference on Rehabilitation Robotics (ICORR), London, UK, 17–20 July 2017; pp. 828–834.
45. Fukuda, K.; Ramirez-Alpizar, I.G.; Yamanobe, N.; Petit, D.; Nagata, K.; Harada, K. Recognition of assembly tasks based on the actions associated to the manipulated objects. In Proceedings of the 2019 IEEE/SICE International Symposium on System Integration (SII), Paris, France, 14–16 January 2019; pp. 193–198.
46. Kahatapitiya, K.; Ryoo, M.S. Coarse-Fine Networks for Temporal Activity Detection in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8385–8394.
47. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 6201–6210.
48. Roitberg, A.; Somani, N.; Perzylo, A.; Rickert, M.; Knoll, A. Multimodal Human Activity Recognition for Industrial Manufacturing Processes in Robotic Workcells. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 259–266.
49. Jones, J.D.; Cortesa, C.; Shelton, A.; Landau, B.; Khudanpur, S.; Hager, G.D. Fine-Grained Activity Recognition for Assembly Videos. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3728–3735. [[CrossRef](#)]
50. Yang, Y.; Aloimonos, Y.; Fermüller, C.; Aksoy, E.E. Learning the Semantics of Manipulation Action. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; Long Papers; Association for Computational Linguistics: Beijing, China, 2015; Volume 1, pp. 676–686.
51. Jones, J.; Hager, G.D.; Khudanpur, S. Toward computer vision systems that understand real-world assembly processes. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 426–434.
52. Gudmundsson, J.; Seybold, M.P.; Pfeifer, J. Exploring Sub-skeleton Trajectories for Interpretable Recognition of Sign Language. In Proceedings of the 27th International Conference on Database Systems for Advanced Applications (DASFAA-2022), Hyderabad, India, 11–14 April 2022.
53. Parsa, B.; Samani, E.U.; Hendrix, R.; Devine, C.; Singh, S.M.; Devasia, S.; Banerjee, A.G. Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3153–3160. [[CrossRef](#)]
54. McAtamney, L.; Hignett, S. Rapid entire body assessment. In *Handbook of Human Factors and Ergonomics Methods*; CRC Press: Boca Raton, FL, USA, 2004; pp. 97–108.
55. Nguyen, T.D.; Kleinsorge, M.; Krüger, J. ErgoAssist: An assistance system to maintain ergonomic guidelines at workplaces. In Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA), Barcelona, Spain, 16–19 September 2014; pp. 1–4.
56. Yan, X.; Li, H.; Wang, C.; Seo, J.; Zhang, H.; Wang, H. Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. *Adv. Eng. Inform.* **2017**, *34*, 152–163. [[CrossRef](#)]
57. Li, C.; Lee, S. Computer Vision Techniques for Worker Motion Analysis to Reduce Musculoskeletal Disorders in Construction. In *Computing in Civil Engineering*; American Society of Civil Engineers: Miami, FL, USA, 2011; pp. 380–387.
58. Shafti, A.; Ataka, A.; Lazpita, B.U.; Shiva, A.; Wurdemann, H.; Althoefer, K. Real-time Robot-assisted Ergonomics. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 1975–1981.
59. Mehri, R.; Peng, X.; Tang, Z.; Xu, X.; Metaxas, D.N.; Li, K. Toward Marker-Free 3D Pose Estimation in Lifting: A Deep Multi-View Solution. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 485–491.
60. Plantard, P.; Shum, H.; Pierres, A.S.; Multon, F. Validation of an ergonomic assessment method using Kinect data in real workplace conditions. *Appl. Ergon.* **2016**, *65*, 562–569. [[CrossRef](#)]
61. Mcatamney, L.; Corlett, E.N. RULA: A survey method for the investigation of work-related upper limb disorders. *Appl. Ergon.* **1993**, *24*, 91–99. [[CrossRef](#)]

62. Kim, W.; Lorenzini, M.; Balatti, P.; Nguyen, P.D.; Pattacini, U.; Tikhanoff, V.; Peternel, L.; Fantacci, C.; Natale, L.; Metta, G.; et al. Adaptable Workstations for Human-Robot Collaboration: A Reconfigurable Framework for Improving Worker Ergonomics and Productivity. *IEEE Robot. Autom. Mag.* **2019**, *26*, 14–26. [[CrossRef](#)]
63. Parsa, B.; Narayanan, A.L.; Dariush, B. Spatio-Temporal Pyramid Graph Convolutions for Human Action Recognition and Postural Assessment. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020.
64. Parsa, B.; Banerjee, A.G. A Multi-Task Learning Approach for Human Activity Segmentation and Ergonomics Risk Assessment. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 2352–2362.
65. Konstantinidis, D.; Dimitropoulos, K.; Daras, P. Towards Real-time Generalized Ergonomic Risk Assessment for the Prevention of Musculoskeletal Disorders. In Proceedings of the 14th ACM International Conference on Pervasive Technologies Related to Assistive Environments Conference (PETRA), Virtual, 29 June–1 July 2021.
66. Wu, H.C.; Wang, M.J.J. Relationship between maximum acceptable work time and physical workload. *Ergonomics* **2002**, *45*, 280–289. [[CrossRef](#)]
67. Velásquez, J.; Briceno, L.; Ortiz, L.; Solarte, S.; Agredo, R. Maximum Acceptable Work Time for the Upper Limbs Task and Lower Limbs Task. *Procedia Manuf.* **2015**, *3*, 4584–4590. [[CrossRef](#)]
68. Burger, G.C.E. Heart Rate and the Concept of Circulatory Load. *Ergonomics* **1969**, *12*, 857–864. [[CrossRef](#)]
69. Kamalakannan, B.; Groves, W.; Freivalds, A. Predictive Models for Estimating Metabolic Workload based on Heart Rate and Physical Characteristics. *J. SH&E Res.* **2007**, *4*, 1.
70. Sgarbossa, F.; Calzavara, M.; Persona, A.; Visentin, V. A device to monitor fatigue level in order-picking. *Ind. Manag. Data Syst.* **2018**, *118*, 714–727.
71. Widodo, L.; Daywin, F.; Nadya, M. Ergonomic risk and work load analysis on material handling of PT. XYZ. In Proceedings of the IOP Conference Series: Materials Science and Engineering, 11th ISIEM (International Seminar on Industrial Engineering & Management), Technology and Innovation Challenges Towards Industry 4.0 Era, Makasar, South Sulawesi, Indonesia, 27–29 November 2018.
72. Samani, A.; Holtermann, A.; Søgaard, K.; Holtermann, A.; Madeleine, P. Following ergonomics guidelines decreases physical and cardiovascular workload during cleaning tasks. *Ergonomics* **2012**, *55*, 295–307. [[CrossRef](#)] [[PubMed](#)]
73. Ye, T.; Pan, X. Fatigue, Cognitive Performance, and Subjective Recovery Time Estimation in High-Intensity Work. *IIE Trans. Occup. Ergon. Hum. Factors* **2016**, *4*, 141–150. [[CrossRef](#)]
74. Hidalgo-Muñoz, A.R.; Mouratille, D.; Matton, N.; Causse, M.; Rouillard, Y.; El-Yagoubi, R. Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation. *Int. J. Psychophysiol.* **2018**, *128*, 62–69. [[CrossRef](#)]
75. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
76. Gong, W.; Zhang, X.; González, J.; Sobral, A.; Bouwmans, T.; Tu, C.; Zahzah, E. Human pose estimation from monocular images: A comprehensive survey. *Sensors* **2016**, *16*, 1966. [[CrossRef](#)]
77. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3192–3199.
78. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [[CrossRef](#)]
79. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
80. Punnakkal, A.R.; Chandrasekaran, A.; Athanasiou, N.; Quiros-Ramirez, A.; Black, M.J. BABEL: Bodies, Action and Behavior with English Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 722–731.
81. Shao, D.; Zhao, Y.; Dai, B.; Lin, D. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
82. Tenorth, M.; Bandouch, J.; Beetz, M. The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 1089–1096.
83. Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent neural networks for time series forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [[CrossRef](#)]
84. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-Normalizing Neural Networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 972–981.
85. Meredith, M.; Maddock, S. *Motion Capture File Formats Explained*; Department of Computer Science, University of Sheffield: Sheffield, UK, 2001; Volume 211, pp. 241–244.

86. Raptis, M.; Kirovski, D.; Hoppe, H. Real-Time Classification of Dance Gestures from Skeleton Animation. In Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Vancouver, BC, Canada, 5–7 August 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 147–156.
87. Theodorakopoulos, I.; Kastaniotis, D.; Economou, G.; Fotopoulos, S. Pose-based human action recognition via sparse representation in dissimilarity space. *J. Vis. Commun. Image Represent.* **2014**, *25*, 12–23. [\[CrossRef\]](#)
88. Schonberger, R.J. *World Class Manufacturing*; Simon and Schuster: New York, NY, USA, 2008.
89. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE ICASSP* **1978**, *26*, 43–49. [\[CrossRef\]](#)
90. Shou, Z.; Pan, J.; Chan, J.; Miyazawa, K.; Mansour, H.; Vetro, A.; Nieto, X.G.; Chang, S.F. Online action detection in untrimmed, streaming videos-modeling and evaluation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Volume 1, p. 5.
91. Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; Lin, D. Temporal action detection with structured segment networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2914–2923.
92. Baptista Ríos, M.; López-Sastre, R.J.; Acevedo-Rodríguez, F.J.; Martín-Martín, P.; Maldonado-Bascón, S. Unsupervised Action Proposals Using Support Vector Classifiers for Online Video Processing. *Sensors* **2020**, *20*, 2953. [\[CrossRef\]](#)
93. Sutskever, I. *Training Recurrent Neural Networks*; University of Toronto: Toronto, ON, Canada, 2013.
94. Chatfield, C. *The Analysis of Time Series: An Introduction*; Chapman and Hall/CRC: London, UK, 2003.
95. Shen, C. Analysis of detrended time-lagged cross-correlation between two nonstationary time series. *Phys. Lett. A* **2015**, *379*, 680–687. [\[CrossRef\]](#)
96. Womack, J. *From Lean Tools to Lean Management*; Lean Enterprise Institute: Brookline, MA, USA, 2006; Volume 21.
97. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; pp. 53–60.
98. Pisharady, P.K.; Saerbeck, M. Recent methods and databases in vision-based hand gesture recognition: A review. *Comput. Vis. Image Underst.* **2015**, *141*, 152–165. [\[CrossRef\]](#)
99. Cutler, R.; Davis, L.S. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 781–796. [\[CrossRef\]](#)
100. Kosmopoulos, D.; Papoutsakis, K.; Argyros, A. A Framework for Online Segmentation and Classification of Modeled Actions Performed in the Context of Unmodeled Ones. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 2578–2590. [\[CrossRef\]](#)
101. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-Time Representation of People Based on 3D Skeletal Data. *Comput. Vis. Image Underst.* **2017**, *158*, 85–105. [\[CrossRef\]](#)