






A vision-based framework for human behavior understanding in industrial assembly lines

Konstantinos Papoutsakis¹, Nikolaos Bakalos^{2,3}, Konstantinos Fragkoulis²,
Athena Zacharia^{2,3}, Georgia Kapetadimitri², and Maria Pateraki^{2,3}

¹ Institute of Computer Science, Foundation for Research and Technology - Hellas,
Greece papouts@ics.forth.gr

² School of Rural Surveying and Geoinformatics Engineering, National Technical
University of Athens, Greece

³ Institute of Communication and Computer Systems, National Technical University
of Athens, Greece
{bakalosnik,kfragoulis,azacharia,mpateraki}@mail.ntua.gr

Abstract. This paper introduces a vision-based framework for capturing and understanding human behavior in industrial assembly lines, focusing on car door manufacturing. The framework leverages advanced computer vision techniques to estimate workers' locations and 3D poses and analyze work postures, actions, and task progress. A key contribution is the introduction of the CarDA dataset, which contains domain-relevant assembly actions captured in a realistic setting to support the analysis of the framework for human pose and action analysis. The dataset comprises time-synchronized multi-camera RGB-D videos, motion capture data recorded in a real car manufacturing environment, and annotations for EAWS-based ergonomic risk scores and assembly activities. Experimental results demonstrate the effectiveness of the proposed approach in classifying worker postures and robust performance in monitoring assembly task progress.

Keywords: Human actions · industrial assembly lines · ergonomic posture evaluation · human pose estimation · video dataset

1 Introduction

In industrial manufacturing, the continuous advancement of automation technologies necessitates a comprehensive understanding of human behavior within assembly lines to enhance productivity and safety. This paper introduces a vision-based framework designed to analyze and interpret human behaviors specifically in the context of industrial assembly lines. By leveraging advanced computer vision techniques, this framework aims to monitor and assess physical ergonomic, and operational aspects of human activities during the assembly process. The framework employs state-of-the-art methods for human pose estimation, ergonomic postural evaluation, and human action monitoring, providing a robust solution to the challenges of human dynamics, severe long-term human body occlusions, and understanding of complex human assembly actions. The integration

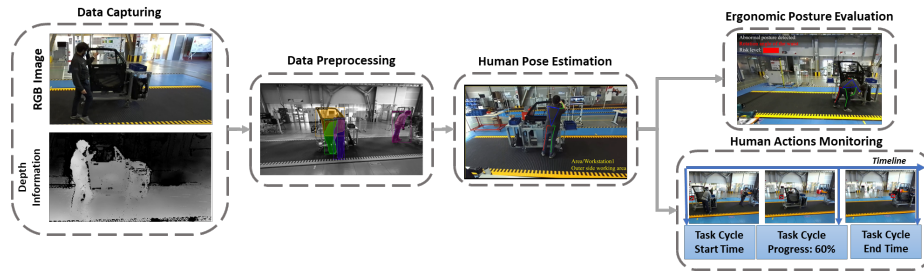


Fig. 1: Workflow of the human behavior understanding framework.

of these methods facilitates the real-time assessment of workers’ postures and actions, enabling the identification of potential ergonomic risks and inefficiencies in the workflow. An important contribution of this research is the development and deployment of the CarDA dataset—a comprehensive multi-modal dataset comprising time-synchronized RGB-D videos and motion capture data recorded in a real-world car manufacturing environment providing in a single dataset ground truth data for 3D human poses, ergonomic risk scores based on the European Assessment Work Sheet (EAWS) posture grid, temporal segmentation, and the classification of assembly actions demonstrated by real workers. Most available datasets published so far deal with daily activities [6, 24, 54] and very few have been published on manufacturing actions, usually adopting an egocentric perspective focusing on finer movements of hands [1, 7, 32, 45, 48, 49]. At the same time, our framework requires camera perspectives overlooking the workstation area where the activity is performed to capture large body movements and human-car door interactions during the assembly task cycle. Furthermore, available datasets are strongly domain-related to specific assembly activities, not transferable to the car door assembly task. This dataset is a critical resource for training and evaluating the proposed vision-based framework, ensuring its applicability and effectiveness in real industrial settings. The dataset will be available freely available online ⁴.

2 Related work

2.1 Human pose estimation

Human pose estimation (HPE) in 2D or 3D, in images or from videos is a challenging yet fundamental Computer Vision task due to depth ambiguity and self-occlusion and the fact that it is an inverse, highly ill-posed problem with multiple feasible solutions. In addition, the rather time-consuming process of collecting high-quality and fine-grained annotations (labeled/ground truth data) for the human body leads to limited access to training data. In particular, annotating 3D data related to human motion requires dedicated, costly motion capture

⁴ <https://zenodo.org/uploads/13370888>

systems, which are not easy to obtain and install to capture human motion and activities in various contexts and environmental conditions. The surveys of [22, 25, 31, 42, 66] reviewed the human motion analysis in many aspects (e.g., detection and tracking, pose estimation, recognition). They described the relation between human pose estimation and other related tasks. More recent surveys mainly focused on subdomains, such as RGB-D-based action recognition [57], 3D human pose estimation [14, 52, 55, 59], model-based human pose estimation [43], body parts-based human pose [44, 69], and monocular-based human pose estimation [41]. Despite the fairly accurate performance of state-of-the-art algorithms for 3D human pose estimation in controlled or semi-controlled settings, coping with complex, realistic scenarios exposes the limits of these algorithms, particularly their effectiveness in handling occlusions. The latter is a key aspect of our approach (sec. 4.1) aiming to derive valid detections in the presence of occlusions by utilizing synchronized camera data.

2.2 Vision-based ergonomic posture evaluation

Several screening tools for physical ergonomic analysis are commonly used in industry, based on different evaluation protocols, sets of postures and other related human activities such as grasping and lifting, such as the Rapid Entire Body Assessment (REBA) [30], the European Assembly Worksheet [46] (EAWS), the OCRA checklist [34], the MURI risk analysis [60] and the Ovako Working Posture Analyzing System (OWAS) [18]. Based on these methods, various efficient vision-based approaches have been proposed to tackle the challenging problem of automated posture evaluation during work activities. Parsa et al. [40] exploited spatial and temporal visual features from RGB-D frames using CNNs for predicting the ergonomic risk of object manipulation actions according to the REBA method. Shafti et al. [51] automatically extract 3D skeletal body information using RGB-D frames to continuously analyze the user’s posture and understand the safe range of arm motions during welding actions following the RULA posture monitoring method. Both methods in [50] and [64] address the task of vision-based posture recognition on construction sites using the OWAS approach. Recently, Parsa et al. [38, 39] proposed a novel approach based on Graph-based CNNs and LSTMs to recognize object manipulation actions (lifting, moving boxes, etc.) and predict the REBA scores in videos. Recently, a novel multi-stream deep network has been proposed [20] to compute the REBA score based on 3D skeletal data sequences extracted during work activities. Our proposed framework introduces an efficient deep-learning method using Graph-based Neural Networks for recognizing the ergonomic postures using the 3D skeletal motion of workers during assembly activities in an industrial environment based on the widely used EAWS postural grid.

2.3 Human actions monitoring

Understanding human actions in visual data is linked to advances in areas like human dynamics, domain adaptation, and semantic segmentation. Extensive re-

search in this domain follows the latest deep learning methodologies [16, 33]. Recently, much work has focused on human actions and human-object interactions (HOI) using deep neural networks [10, 21, 23]. These approaches often model coarse geometric and appearance features using spatial regions-of-interest (ROIs) [13] to classify actions in short video clips. More advanced methods represent the temporal and spatial structure of entities, such as 2D/3D skeletal models [5, 44], 2D hand/object masks [3], and 3D poses of hand(s)-object(s) [11, 53].

Spatiotemporal relationships in HOI are modeled using attention mechanisms [26], Graph Neural Networks (GNN) [23], CNNs [19], RNNs, LSTMs [27, 67], and Transformer models [12, 61, 65]. However, many methods treat HOI as non-composite, monolithic activities, limiting their ability to generalize across diverse actions. Recent research considers HOI as composite activities with complex spatiotemporal relationships [28], integrating high-level semantics [2], logic rules [62], and graph-based methods [37]. Novel representations encode human-object relationships over time using spatio-temporal graphs [17], with some focusing on visual reasoning to uncover causal relationships [3]. Our framework extends these methodologies to the structured context of industrial assembly lines, enhancing fine-grained activity recognition and ergonomic assessment.

Dataset	Type of actions	Data	Viewpoint (cameras)	Ergonomics	Videos	Duration (h)
ATTACH [1]	assembly actions	Video RGB-D, IR, 3D poses	exocentric (3)	-	378	17
BRIO-TA [32]	assembly toy set	Video (RGB)	exocentric (1 top-view)	-	75	2.9
IKEA ASM [4]	assembly furniture	Video RGB-D, 3D poses	exocentric (3)	-	381	35.3
IndustReal [48]	maintenance, assembly toy set	Video RGB-D, Gaze	egocentric (1)	-	84	5.8
ENIGMA-51 [45]	repairing electrical board	Video, audio instructions	egocentric (1)	-	51	22
Assembly101 [49]	assembly toy vehicles	Video RGB-D, 3D hand poses	exo-egocentric (8-4)	-	362	167
HA4M [7]	assembly mechanical parts	Video	exocentric (1)	-	217	5.9
sustAGE [35]	car door assembly	Video RGB-D, heart rate	exocentric (2)	MURI	40	0.8
UW-IOM [40]	manipulating boxes-rods	Video RGB-D, 3D poses	exocentric (1)	REBA	20	0.63
AnDy [29]	screwing, manipulating loads	Video RGB-D, motion capture	exocentric (2)	EAWS	390	5
CarDA	car door assembly	Video RGB-D, motion capture	exocentric (2)	EAWS	25	1.48

Table 1: Overview of video datasets focusing on human activity monitoring in industrial or industrial-like environments.

2.4 Datasets

We summarize available video datasets related to the visual understanding of human manufacturing activities in industrial environments, as shown in Table. 1. Only, a few of them provide annotation data related to ergonomic analysis of body postures during work activities. The UW-IOM dataset [40] features limited object manipulation actions involving awkward poses and repetitions and frame-level annotations according to the REBA ergonomic risk index. The sustAGE User Postures & Actions Monitoring dataset [35] includes time-synced video recordings from multiple static RGB-D cameras and heart rate data of real line workers acquired using wearable sensors recorded during manufacturing activities. Annotation data related to assembly actions and ergonomically sub-optimal postures based on the MURI postural grid for physical ergonomics assessment

are provided. Moreover, the AnDy multi-modal dataset [29] captures human motions in industry-like activities (screwing and manipulating loads). Annotation data regards whole-body kinematics of 13 participants recorded with optical motion capture, finger pressure force, videos, and postural events following the EAWS [46]. The proposed CarDA dataset comprises videos (RGB-D) acquired using two stereo cameras, whole-body 3D kinematics acquired using a motion capture system, and annotations related to the assembly actions and the postures using the EAWS postural grid for analysis of physical ergonomics in a realistic industrial environment. The IKEA ASM dataset [58] focuses on realistic chair assembly actions in videos, while the IndustReal [48] and the ENIGMA-51 [45] datasets comprise egocentric videos of maintenance or assembly activities on a construction-toy assembly set and repairing procedures on electrical boards, respectively. The Assembly101 dataset [49] features exocentric and egocentric recordings of people assembling and disassembling 101 “take-apart” toy vehicles. The HA4M multi-modal dataset [7] demonstrates assembly activities for building an Epicyclic Gear Train. The ATTACH dataset [1] demonstrates simulated assembly actions in the context of a human-robot collaborative scenario. Finally, the BRIO-TA [32] dataset simulates operations in factory assembly consisting of scenarios for normal and anomalous work processes.

3 CarDA - Car door Assembly Activities Dataset

The proposed multi-modal dataset for car door assembly activities noted as CarDA, comprises a set of time-synchronized multi-camera RGB-D videos and motion capture data acquired during car door assembly activities performed by real line workers in an experimental, real-world car manufacturing environment. The following sections provide a detailed description of the deployment environment, and the data acquisition and annotation procedures followed to compile the CarDA dataset. The dataset will be available freely available online ⁵.

3.1 Deployment environment

The use-case scenario concerns a real-world assembly line workplace in an automotive manufacturing industry, as the deployment environment. In this context, line workers simulate the real car door assembly workflow using the prompts, sequences, and tools under very similar ergonomic and environmental conditions as in existing factory shop floors. The assembly line consists of three workstations, overseen by a team leader. A conveyor belt with three virtually separated work areas that correspond to the three workstations, moves at a constant speed, supporting cart-mounted car doors and material storage. A worker is assigned to each workstation. All workers assemble car doors as the belt moves, with each station (WS10, WS20, and WS30). A worker completes a workstation-specific set of assembly actions, noted as a task cycle, lasting about 4 minutes before the

⁵ <https://zenodo.org/uploads/13370888>

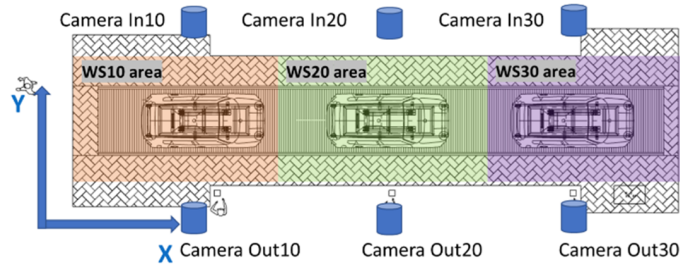


Fig. 2: The shopfloor outline of the real manufacturing environment used to acquire the CarDA dataset. Three workstation areas (color-coded rectangles) are virtually defined on the conveyor belt area (assembly line). A pair of two stereo cameras (InXX, OutXX) (blue cylinders) are installed on both sides of each workstation WSXX.

cart proceeds to the next workstation for further assembly. Upon the successful completion of the task cycle, the cart is left to travel to the virtually defined area of the subsequent workstation where another line worker will continue the assembly process during the new task cycle. Each task cycle lasts approximately 4 minutes and is continuously repeated during the worker’s shift.

A camera system with three pairs of cameras is installed along the production line to monitor all workstations, capturing the assembly activities from both the inner and outer sides. Each RGB-D sensor, mounted on rods at a height of 1.85 meters, is oriented towards the center of a workstation and the conveyor belt for comprehensive visual monitoring.

3.2 Car Door Assembly Activities

The car door assembly dataset contains detailed timing data captured through multiple video recordings of the assembly process, as described above. This temporal data includes the start and end times of various activities, recorded in seconds. Each activity is part of a larger subgoal, which collectively describes the entire car door assembly process. Table 2 presents indicative sub-goals for the first of the three workstations. These subgoals span a period between 22 and 67 seconds each and occur sequentially during a complete assembly task cycle process. For the transformer model described in 4.3 we used these temporal annotations to extract the labels which correspond to the appropriate sub-goal. These labels were used to train the classifier.

3.3 Data Acquisition

Data acquisition involves low-cost, passive RGB-D camera sensors that are installed at stationary locations alongside the car door assembly line and a motion capture system for capturing time-synchronized sequences of images and motion capture data during car door assembly activities performed by real line workers.

No	Sub-Goal	Description
1	Place Cart	Moving the cart at the assembly line to align the anterior wheels.
2	Attach Door Frame	align and secure the door frame to the main body with bolts.
3	Install Window Mechanisms	Fit motorized and manual window components into the door structure.
4	Connect Wiring	Connecting the wiring harness to the door's electrical systems.
5	Test Electrical Components	Verifying the functionality of electrical components (windows, locks).
6	Install Interior Panels	Aligning and securing the inner door panel with screws and clips.
7	Attach Exterior Components	Attach exterior components (handles and trim pieces) to the door.
8	Quality Check	Perform visual inspection for defects, functional tests on moving parts.

Table 2: Indicative sub-goals for the assembly activities in WS10.



Fig. 3: Annotations of the car door and humans area (marked in yellow). Annotations of the car door with (left) and without (middle) the overlapping human area and the extracted 2D segmentation masks of the humans (right).

More specifically, two stationary StereoLabs ZED2 stereo cameras were installed in each of the three workstations of the car door assembly line. The two stationary, workstation-specific cameras are located at bilateral positions on the two sides of the conveyor belt at the center of the area concerning that specific workstation. The pair of sensors were utilized to acquire stereo color and depth image sequences during car door task cycle executions. Each recording comprises time-synchronized RGB (color) and depth image sequences captured throughout a task cycle execution at 30 frames per second (fps). At the same time, the line worker used a wearable XSens MVN Link suit during work activities to acquire time-synced 3D motion capture data at 60 fps. The motion capture system and the pair of RGB-D cameras are calibrated according to a common (global) coordinate reference system set near the start of the assembly line. A complete task cycle execution has an average duration of 4 minutes and comprises a fixed sequence of car door assembly actions (also noted as assembly subgoals). Two RGB videos, a sequence of depth images, and a BHV file with motion capture data are recorded per task cycle execution. Overall, 25 recordings are available for an equal number of car door task cycles, noted as samples: 8 recordings concern WS10, 10 for WS20, and 7 for WS30. We note that all CarDA samples contain videos, while 12 of the samples correspond to both video and motion capture data (5 for WS10, 4 for WS20, and 3 for WS30).

EAWS evaluation (Video: WS10_HD720)	Temporal segmentation		3. Bent forward	4. Strongly bend forward	5. Elbow at/above shoulder level	6. Hands above head level	Trunk Rotation	Lateral Bending
Subgoal	Start [s]	End [s]	Duration (s)	Duration (s)	Duration (s)	Duration (s)	Duration (s)	Duration (s)
1. Adjust WS	11	52	16					
2. Fixate rearview mirror	52	102	15		19		10	19
3. Fixate tie rod	102	151	21	1				21
4. Fixate plugs	151	180	6					
5. Fixate carrier	180	269	5					10

Table 3: Sample annotation data related to EAWS-based ergonomic postures and work/assembly activities (subgoals) for a complete task cycle in WS10.

3.4 Data Annotation

Ground truth data on human localization and 3D pose estimation: As described above, ground truth data for 3D human motion is acquired using the wearable XSens MVN Link suit. The motion capture system obtains the 3D coordinates and orientations of 24 skeletal body joints throughout a recording in a BVH format, that corresponds to a task cycle execution in our case.

Ground truth data on basic postures of the EAWS ergonomic screening tool: Two experts in manufacturing and ergonomics performed manual annotations related to the EAWS screening tool. Overall, the CarDA dataset contains videos of 25 task cycles for all workstations for which 412 instances concerning the seven target classes of EAWS basic postures are obtained. The average duration of annotated postures is approximately 10 seconds. For each activity subgoal, a single occurrence, if any, and its total duration for any of the posture classes that have been observed is annotated, as shown in Tab. 3. Information for the id-name of each subgoal, and temporal segmentation (start, and end video timestamps, and the duration of each subgoal in seconds) is provided. The total duration of each type of EAWS-based basic posture, if any occurred, during each subgoal is also reported. Postures of duration lower than 4 seconds are not considered valid for postural evaluation.

Ground truth related to temporal segmentation and classification of car door assembly actions (subgoals) during task cycle executions: Manual annotations related to the temporal segmentation and classification of car door assembly actions were performed by personnel working directly on the assembly line for the CarDA dataset, as shown in Tab. 3. It encompasses 25 task cycles (samples), where each comprises on average 5.4 subgoals (minimum of 4 and maximum of 8 subgoals) and has a mean duration of 216 seconds, ranging from 137 to 271 seconds. In total, the set of 25 task cycles comprised 135 different subgoals that were manually for the start and end timestamps (in seconds of video time). Each subgoal has an average duration of 42 seconds, ranging from 22 to 67 seconds.

4 Human behavior understanding framework

In the following sections, the proposed framework for human behavior understanding in the assembly line work environment is presented (Fig. 1). It relies on

visual data that capture stereo RGB (color) and depth images of human workers during work activities on the real shop floor and comprises a set of components and methods for a) detecting the location and the 3D body pose of the human workers in real-time; b) analyzing the spatiotemporal relationships of body joints for monitoring workers' postures in terms of physical ergonomics and c) estimate the work progress during an ongoing assembly task cycle execution. The framework also caters to the estimation of the location of each cart door during assembly activities which will aid in real-time monitoring of worker actions and the task cycle progress estimation.

4.1 Human and car door detection

For the estimation of the human pose, we exploit two state-of-the-art methods to estimate the 2D and subsequently the 3D body articulated (skeleton-based) configuration based on a known tree-based skeletal model. First, the OpenPose [5] deep learning-based method is employed to detect the 2D location of each human in a single image and estimate the 2D body joint coordinates of 25 body joints (Model25). Then, the set of 2D poses is used as input to the MocapNet2 [44] method, an end-to-end approach that relies on ensembles of Deep Neural Networks to regress the 3D coordinates and angles of 15 human body joints per frame as in [36]. The method is further extended to process visual data captured synchronously by a pair of cameras installed on both sides of each workstation of the assembly line (cf. 3.1) to improve the reliability of the recovered 3D human body pose. Given the confidence values of the 2D and 3D estimations for the body joint positions, we opt for the camera view that provides valid detections (more than 50% of body joints provide a confidence score higher than 50%). We utilize the pre-trained Mocapnet2 method without any additional training based on the proposed CarDA dataset.

Beyond the human pose, the car door is an instrumental element in the analysis of human assembly activities. The estimated car door pose (position and orientation) is used in conjunction with the skeletal joints in the human actions monitoring component (sec. 4.3). To estimate the car door pose, the car door area is visually segmented and the obtained fine object mask is subsequently exploited to estimate the car door centroid, orientation, and bounding box both in camera-centric coordinates and global coordinates. For the car door segmentation, the Mask R-CNN deep learning model [15] was exploited as part of the Detectron2⁶ framework. Initially, we used the Mask R-CNN R50 FPN model from the set of pre-trained CNN models from the model zoo, trained on the COCO dataset, to assess the extraction of different class types, including the segmentation of humans (i.e., class *person*). Since the car door is a special class of object not part of available pre-trained models, the car door area had to be annotated in sample images from a series of videos from the assembly line and further finetune the model for the newly introduced car door class together with the 2D segmentation masks of humans (i.e., the class for *person*).

⁶ <https://ai.meta.com/tools/detectron2/>

The different experiments for finetuning the model either exploiting only the *car door* or both the *person* and *car door* classes and validating the results on image data from available video sequences are presented in sec. 5.1.

Once the 2D segmentation mask of the car door is detected, we utilize the depth information and the camera extrinsic parameters to estimate the car door pose. By transforming the mask’s extremities to the camera and global coordinate system, we deduce the door’s orientation and centroid. Furthermore, the 2D human masks extracted with Mask R-CNN are also used to extract the skeletal data in each dilated human mask area and select the optimal 3D human pose based on the 3D distance to the cart location. This significantly improves the results in the case of multiple humans present in the scene as in these cases the assumption to select the 2D pose with the highest confidence score as the optimal candidate and active worker for 3D pose estimation would be less robust.

4.2 Analysis of ergonomic work postures in videos

We focus on seven types of dynamic body postures based on the basic postures of the EAWS screening tool [29, 46, 47]. The EAWS tool is commonly employed in manufacturing to assess workers’ physical workload and ergonomics. Each basic posture type is treated as a time-varying event, consisting of a sequence of body configurations with a minimum duration of 4 seconds. The EAWS tool uses a scoring/labeling scheme based on the occurrences and total duration of each posture type per minute throughout a work activity. Some of the basic postures concern similar types of body configurations with increased levels of body strain. The selected posture types are the following (by keeping the original numbering/naming): 1. Standing & walking in alteration, 3. Bend forward, 4. Strongly Bend forward, 5. Upright with elbows at or above shoulders, 6. Upright with elbows above the head, and finally the "Trunk Rotation" and the "Lateral Trunk Bending" types from the Asymmetric postures. Static postures of duration at least 4 seconds are identified as valid, according to the EAWS instructions.

We formulate the evaluation of ergonomic work postures during manufacturing activities as a vision-based multi-class action detection and classification problem. To this end, given a video demonstrating a subgoal during a car door task cycle for a specific workstation, we extract the sequence of 3D body poses using the MocapNet2 method for each frame. We adapt and re-trained the previously proposed approach in [35] for the detection and classification of EAWS-based postures throughout a video. The 3D skeletal sequences are used to re-train a spatiotemporal Graph Convolutional Network model (ST-GCN) [63], which learns to encode the input representations of the target working postures into a shared embedding/feature space. Finally, we compute the pairwise temporal alignment cost between the embeddings of an unlabeled 3D skeletal sequence and those of a 3D skeletal sequence of a known class type using the soft Dynamic Time Warping approach (softDTW) [8]. This pairwise cost serves as a similarity measure for classifying the unlabeled sequence among the target types of working postures. We employ a sliding window scheme of 8 seconds, with an overlap

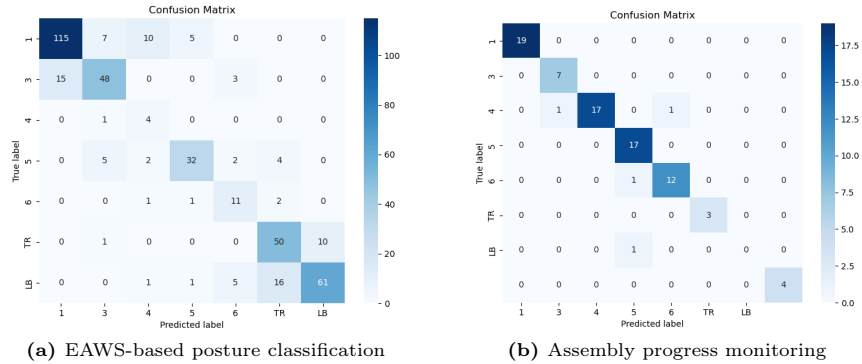


Fig. 4: (a) Experimental results are illustrated as a confusion matrix for (a) the vision-based classification of EAWS-based ergonomic postures (TR: trunk rotation, LB: lateral bend) as shown in Sec. 4.2, (b) the progress monitoring of car door assembly activities during task cycles using the proposed Car Door Assembly dataset.

of 4 seconds, to detect and classify the occurrences of any ergonomic posture classes using our classifier.

4.3 Human actions monitoring

To monitor and interpret actions in the car door assembly line, a transformer model was chosen for its ability to capture long-range dependencies in sequence-to-sequence tasks [68]. This model processes input data, including the car door’s location, the human operator’s position, and skeletal joint measurements, all mapped in world coordinates to maintain consistency [9, 56]. The car door’s position is derived from the center of mass of its segmentation mask, identified using the Mask R-CNN model (cf. Section 4.1). Human skeletal data is transformed into a body coordinate system, transposing movements relative to the base spine joint, which is set as the origin (0,0). All other joints are transposed relative to this base joint, focusing on relative movements rather than absolute positions. This transformation normalizes the data and helps the model learn movement patterns effectively. .

The model handles 10 non-overlapping sets of measurements, embedding spatial and kinematic data. Using transformer encoder layers with multi-head self-attention, the model condenses the output into a vector, which is classified through a fully connected and softmax layer. The model is trained using cross-entropy loss and the Adam optimizer, with a 70-10-20 train-validation-test split, to infer the subgoal of the assembly task based on these measurements.

Metrics/Joints	Neck	Trunk	Knees	Shoulders	Elbows	Hands	Hips	Toes	Mean
MJPE	98.4	115.4	156.4	102.7	157.5	171.8	127.8	175.6	138.2
PCK@150	81.3	82.9	56.4	81.1	58.6	49.3	75.6	65.2	68.8

Table 4: Experimental results for 3D pose estimation using the MocapNet2 method [43]. MJPE (Mean Per Joint Position Error). PCK@150 (Percentage of Correct Key-points metric between 3D poses using the commonly used threshold of 150mm). The average scores for pairs of symmetric body joints are reported, e.g. hands.

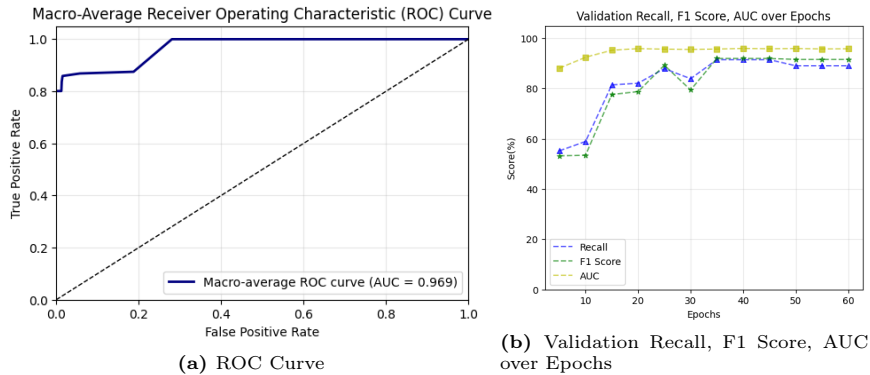


Fig. 5: Performance evaluation of the human action monitoring (a) Macro-Average Receiver Operating Characteristic curve, (b) Recall, F1, AUC scores (validation set).

5 Experiments

5.1 Car Door Detection

We have used 257 frames from [35] for annotation for the *car door* class and two sets of annotation data were derived. The first set comprised masks of the car door with overlapping human areas in the delineated car door area, while in the second set, mask parts attributed to humans were excluded from the estimated car door area/mask (Figure 3). Initially, we trained the model individually with the two sets of car door classes and without the 2D segmentation masks for humans. Overall, the results with the first set exhibited sufficient accuracy, with 85.07% and 80.86% average precision for the bounding boxes and the segmentation masks, respectively. With the second set, the average precision on segmentation masks was lower (66.98%) due to the intricate shape of the car door annotations. Overall, the car door segmentation masks were correctly estimated in cases where humans were not present. We have further fine-tuned the model using annotation data of both sets containing the 2D mask of the car door and those of the humans. The results did not exhibit significant differences in the way the model dealt with the prediction of the two classes, as both the car doors and humans were correctly detected with a confidence score above 95%,

for both classes. The average precision with the COCOEvaluator is 88.20% on bounding boxes and 79.63% on segmentation masks.

5.2 Vision-based Human Detection

We evaluate the performance of the MocapNet2 [43, 44] method for 3D human body pose estimation using ground truth motion capture data and visual information from the CarDA dataset. A subset of 18 samples of the proposed dataset is used for this evaluation.

Each sample comprises time-synchronized video acquired by one camera and ground truth 3D motion capture data acquired using the Xsens MVN Link motion capture system during assembly activities. For each image sequence, the 3D human poses are computed (cf. Section 4.1). To compare the estimated 3D pose to the ground truth data, we consider the 3D body center (Torso) position as the origin of the coordinate system for both sequences of 3D skeleton-based motion data. The pre-trained MocapNet2 model was deployed, with no fine-tuning on the ground truth motion capture data. Moreover, a generic 3D human body model is used to infer the 3D human pose per frame.

Firstly, the Mean Per Joint Position Error (MPJPE) is used to assess the average Euclidean distance between the 2D or 3D positions of predicted joints of the estimated skeleton-based human pose and the ground truth joints in a given dataset. Moreover, the Percentage of Correct Keypoints (PCK) metric is used. The PCK considers correctly estimated detected joints as the ones that have a distance below a certain threshold compared to the true location. We use a fixed threshold of 150 mm that is commonly used in pose estimation benchmarks. Table. 4 provides the experimental results, where the average MPJPE over all joints and all samples of the CarDA subset is below 140 mm and 68.8% of the 3D joints are considered successfully detected throughout all data sequences based on the PCK@150 metric.

5.3 Vision-based analysis of Ergonomic Work Postures

The experimental evaluation of vision-based physical ergonomics is posed as an action classification task given the 25 samples of the CarDA dataset and available EAWS-based annotations. We estimate the performance of the proposed approach based on the accuracy and precision scores, as shown in Tab. 5 and Tab. 4a. Overall, the proposed approach can identify the correct class of postures performed by line workers with 76.7% accuracy and 68.27% mAP scores. Given the challenging conditions captured by the CarDA dataset in a real environment of simulated manufacturing activities, the obtained results indicate a promising potential for the proposed approach in real-world applications. These findings demonstrate the effectiveness of vision-based physical ergonomics in accurately classifying worker postures, even under complex and dynamic conditions, and severe and long-term body occlusions.

Postures & evaluation metrics	1	3	4	5	6	Trunk rotation	Lateral Bending	Mean
Accuracy	0.8394	0.7273	0.8000	0.7273	0.7333	0.8197	0.7262	0.7670
Precision	0.8846	0.7742	0.2222	0.8205	0.5238	0.6944	0.8592	0.6827

Table 5: Experimental results for the vision-based classification of EAWS-based ergonomic work postures (as described in Sec. 4.2) using the Car Door Assembly dataset. The mean accuracy and mean average precision (mAP) are reported (last column).

5.4 Vision-Based Progress Monitoring of Assembly Activities

The performance of the transformer model for human actions monitoring in the car door assembly line demonstrates strong results (Figure 5). The first graph shows the validation recall, F1 score, and AUC over 60 epochs, indicating significant improvements early in the training process, with the metrics stabilizing at high values, around 90-100%. The macro-average ROC curve reflects an AUC of 0.969, signifying excellent discriminative capability. The confusion matrix further supports these findings, showcasing precise classification across different subgoals, with most true labels accurately predicted, and minimal misclassifications, particularly for subgoals with fewer samples. These results confirm the model’s robustness and effectiveness in real-time monitoring applications.

6 Conclusion

In this paper, we have introduced a comprehensive vision-based framework for real-time monitoring of human behavior in industrial settings, specifically focusing on the manufacturing sector and the car door assembly process. We introduced the CarDA dataset, a new dataset that comprises time-synchronized multi-camera video with 3D motion capture data of trained line workers during assembly activities in a real-world manufacturing environment. A rich set of annotations, related to the assembly activities, motion capture data, and information on EAWS-based physical ergonomics is provided to support the multifaceted human behavior. The utilization of the MocapNet2 method allowed for robust 3D pose estimations in real-time, which are crucial for accurate ergonomic assessments. Moreover, a transformer-based approach is presented to monitor the work progress during assembly activities using visual information on the human and car door motion in the work environment. Yet another deep learning-based approach for the classification of the EAWS work postures using sequences of 3d skeletal poses is presented. Despite the challenging conditions presented in the CarDA dataset, our approach achieved commendable performance, highlighting its robustness and reliability. Future work will focus on enhancing the model’s performance by incorporating additional data and refining the pose estimation techniques. Furthermore, we aim to explore the applicability of our framework to other industrial processes and ergonomic evaluation protocols, thereby broadening its impact on occupational safety and health.

Acknowledgements

This work was funded by the European Union’s Horizon 2020 programme FELICE (GA No 101017151) European Union’s Horizon Europe programme SOPRANO (GA No 101120990). Special thanks to Stellantis—Centro Ricerche FIAT (CRF)/ SPW Research & Innovation department in Melfi, Italy for supporting the dataset collection.

References

1. Aganian, D., Stephan, B., Eisenbach, M., Stretz, C., Gross, H.M.: Attach dataset: Annotated two-handed assembly actions for human action understanding. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11367–11373 (2023). <https://doi.org/10.1109/ICRA48891.2023.10160633>
2. Bacharidis, K., Argyros, A.: Extracting action hierarchies from action labels and their use in deep action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 339–346. IEEE (2021)
3. Baradel, F., Neverova, N., Wolf, C., Mille, J., Mori, G.: Object level visual reasoning in videos. In: ECCV (2018)
4. Ben-Shabat, Y., Yu, X., Saleh, F., Campbell, D., Rodriguez-Opazo, C., Li, H., Gould, S.: The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 847–859 (January 2021)
5. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence* **43**(1), 172–186 (2019)
6. 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. pp. 6299–6308 (2017)
7. Cicirelli, G., Marani, R., Romeo, L., Domínguez, M.G., Heras, J., Perri, A.G., D’Orazio, T.: The ha4m dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. *Scientific Data* **9**(1), 745 (2022)
8. Cuturi, M., Blondel, M.: Soft-dtw: A differentiable loss function for time-series. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 894–903. ICML’17, JMLR.org (2017)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6201–6210 (2019). <https://doi.org/10.1109/ICCV.2019.00630>
11. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2018)
12. 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. pp. 244–253 (2019)
13. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8359–8367 (2018)
14. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13041–13051 (2023)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
16. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition. *Image Vision Comput.* **60**(C), 4–21 (04 2017). <https://doi.org/10.1016/j.imavis.2017.01.010>, <https://doi.org/10.1016/j.imavis.2017.01.010>

17. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10236–10247 (2020)
18. Karhu, O., Kansil, P., Kuorinka, I.: Correcting working postures in industry: A practical method for analysis. *Applied ergonomics* **8**(4), 199–201 (1977)
19. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). pp. 1623–1631. IEEE (2017)
20. Konstantinidis, D., Dimitropoulos, K., Daras, P.: Towards real-time generalized ergonomic risk assessment for the prevention of musculoskeletal disorders. In: 14th ACM International Conference on Pervasive Technologies Related to Assistive Environments Conference (PETRA) (June–July 2021)
21. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015)
22. Lei, Q., Du, J.X., Zhang, H., Ye, S., Chen, D.: A survey of vision-based human action evaluation methods. *Sensors (Basel, Switzerland)* **19** (2019)
23. 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 Transactions on Pattern Analysis and Machine Intelligence* (2021)
24. 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 Transactions on Pattern Analysis and Machine Intelligence* (2019). <https://doi.org/10.1109/TPAMI.2019.2916873>
25. Liu, W., Bao, Q., Sun, Y., Mei, T.: Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective. *ACM Computing Surveys* **55**(4), 1–41 (2022)
26. 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. pp. 6790–6800 (2018)
27. Mahasseni, B., Todorovic, S.: Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3054–3062 (2016). <https://doi.org/10.1109/CVPR.2016.333>
28. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-else: Compositional action recognition with spatial-temporal interaction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2020)
29. Maurice, P., Malaisé, A., Amiot, C., Paris, N., Richard, G.J., Rochel, O., Ivaldi, S.: Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *The International Journal of Robotics Research* **38**(14), 1529–1537 (2019)
30. McAtamney, L., Hignett, S.: Rapid entire body assessment. In: Handbook of Human Factors and Ergonomics Methods, pp. 97–108. CRC Press (2004)
31. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* **104**(2), 90–126 (2006). <https://doi.org/https://doi.org/10.1016/j.cviu.2006.08.002>, <https://www.sciencedirect.com/science/article/pii/S1077314206001263>, special Issue on Modeling People: Vision-based understanding of a person’s shape, appearance, movement and behaviour

32. Moriwaki, K., Nakano, G., Inoshita, T.: The brio-ta dataset: Understanding anomalous assembly process in manufacturing. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 1991–1995 (2022). <https://doi.org/10.1109/ICIP46576.2022.9897369>
33. Morshed, M.G., Sultana, T., Alam, A., Lee, Y.K.: Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors* **23**(4) (2023). <https://doi.org/10.3390/s23042182>, <https://www.mdpi.com/1424-8220/23/4/2182>
34. Occhipinti, E.: Ocr: a concise index for the assessment of exposure to repetitive movements of the upper limbs. *Ergonomics* **41**(9), 1290–1311 (1998)
35. 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* **10**(2) (2022). <https://doi.org/10.3390/technologies10020042>, <https://www.mdpi.com/2227-7080/10/2/42>
36. 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* **10**(2) (2022). <https://doi.org/10.3390/technologies10020042>, <https://www.mdpi.com/2227-7080/10/2/42>
37. Papoutsakis, K.E., Argyros, A.A.: Unsupervised and explainable assessment of video similarity. In: BMVC. p. 151 (2019)
38. 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). pp. 2352–2362 (January 2021)
39. Parsa, B., narayanan, A.L., Dariush, B.: Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (March 2020)
40. 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 Robotics and Automation Letters* **4**(4), 3153–3160 (2019). <https://doi.org/10.1109/LRA.2019.2925305>
41. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 459–468 (2018)
42. Poppe, R.: A survey on vision-based human action recognition. *Image and Vision Computing* **28**(6), 976–990 (2010). <https://doi.org/https://doi.org/10.1016/j.imavis.2009.11.014>, <https://www.sciencedirect.com/science/article/pii/S0262885609002704>
43. Qammaz, A., Argyros, A.: A unified approach for occlusion tolerant 3d facial pose capture and gaze estimation using mocapnets. In: IEEE/CVF International Conference on Computer Vision Workshops (AMFG 2023 - ICCVW 2023). pp. 3178–3188. IEEE, Paris, France (October 2023)
44. Qammaz, A., Argyros, A.A.: Occlusion-tolerant and personalized 3D human pose estimation in rgb images. In: IEEE International Conference on Pattern Recognition (ICPR 2020) (January 2021), http://users.ics.forth.gr/argyros/res_mocapnet_II.html

45. Ragusa, F., Leonardi, R., Mazzamuto, M., Bonanno, C., Scavo, R., Furnari, A., Farinella, G.M.: Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4549–4559 (2024)
46. Schaub, K., Caragnano, G., Britzke, B., Bruder, R.: The european assembly worksheet. *Theoretical Issues in Ergonomics Science* **14**(6), 616–639 (2013)
47. Schaub, K.G., Mühlstedt, J., Illmann, B., Bauer, S., Fritzsche, L., Wagner, T., Bullinger-Hoffmann, A.C., Bruder, R.: Ergonomic assessment of automotive assembly tasks with digital human modelling and the ‘ergonomics assessment worksheet’(eaws). *International Journal of Human Factors Modelling and Simulation* **3**(3-4), 398–426 (2012)
48. Schoonbeek, T.J., Houben, T., Onvlee, H., Van der Sommen, F., et al.: Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4365–4374 (2024)
49. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhanian, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
50. Seo, J., Lee, S.: Automated postural ergonomic risk assessment using vision-based posture classification. *Automation in Construction* **128**, 103725 (2021). <https://doi.org/https://doi.org/10.1016/j.autcon.2021.103725>, <https://www.sciencedirect.com/science/article/pii/S092658052100176X>
51. Shafti, A., Ataka, A., Lazpita, B.U., Shiva, A., Wurdemann, H., Althoefer, K.: Real-time robot-assisted ergonomics*. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 1975–1981 (2019). <https://doi.org/10.1109/ICRA.2019.8793739>
52. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: *Computer Vision and Pattern Recognition (CVPR)* (2024)
53. Tekin, B., Bogo, F., Pollefeys, M.: H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2019)
54. Tenorth, M., Bandouch, J., Beetz, M.: The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In: *IEEE International Workshop THEMIS, ICCV2009* (2009)
55. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
57. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: RGB-D-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding* **171**, 118–139 (2018)
58. 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. pp. 639–648 (2020)
59. Wang, Y., Daniilidis, K.: Refit: Recurrent fitting network for 3d human recovery. In: *International Conference on Computer Vision (ICCV)* (2023)

60. Womack, J.: From lean tools to lean management. Lean Enterprise Institute **21** (2006)
61. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13577–13587 (2022). <https://doi.org/10.1109/CVPR52688.2022.01322>
62. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
63. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)
64. 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. *Advanced Engineering Informatics* **34**, 152–163 (2017). <https://doi.org/https://doi.org/10.1016/j.aei.2017.11.001>
65. 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. pp. 4486–4496 (2021)
66. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X., Chen, D.S.: A comprehensive survey of vision-based human action recognition methods. *Sensors* **19**(5) (2019). <https://doi.org/10.3390/s19051005>, <https://www.mdpi.com/1424-8220/19/5/1005>
67. 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. pp. 2117–2126 (2017)
68. Zhang, Y., Shen, H., Liu, Y., et al.: Efficient long-range transformers: You need to attend more, but not necessarily at every layer. arXiv preprint arXiv:2310.12442 (2023)
69. Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M.: Deep learning-based human pose estimation: A survey. *ACM Computing Surveys* **56**(1), 1–37 (2023)