*Human Factors in Simulation, Software and Systems Engineering*, Vol. 220, 2026, 89–99
https://doi.org/10.54941/ahfe1007697

# Human Performance Modelling in Virtual Factories: A Simulation-Driven Ergonomics Approach

**Chun Shih Cheng, Chia Chen Kuo, Chien Hsin Yang, and Yu Jie Tsai**

Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung, Taiwan

## ABSTRACT

Traditional manufacturing evaluations often fail to integrate ergonomic risks with task performance. This study proposes a multi-view, vision-based framework to simultaneously capture performance and ergonomic indicators at a precision assembly workstation. Using three synchronized cameras, we recorded 18 operators over 324 trials of picking (T1), fastening (T2), and inspection (T3) tasks. The processing pipeline—incorporating RT-DETRv2, OCSort, and pose estimation—achieved an event-level F1 score of 0.88. Results showed distinct task characteristics: T1 required the highest reach, while T2 involved the greatest shoulder loading. Mixed-effects models confirmed significant task effects on both temporal and postural metrics *($p < 0.01$)*. A subsequent workstation redesign reduced excessive reach by 28%, arm elevation by 19%, and cycle time by 9%, validating the framework's utility for ergonomics-informed design.

**Keywords:** Manual workstation analysis, Human performance measurement, Ergonomics risk assessment, Multi-view vision sensing, Human factors and simulation

## INTRODUCTION

Human Systems Integration (HSI) and Human Factors Engineering (HFE) systematically align human capabilities with system interfaces to optimize life-cycle performance. While Industry 4.0 leverages Digital Twin (DT) technologies to integrate physical and human processes , traditional Industrial Engineering often overlooks ergonomic risks and behavioral factors in manual assembly. This fragmented analysis persists due to three limitations in current vision-based systems: (1) a focus on task classification over ergonomic quantification; (2) single-camera occlusion; and (3) the lack of integrated productivity-ergonomics evaluation. To address this, we propose a multi-view framework for the simultaneous modeling of operational performance and ergonomic risks.

Condensed Research Contributions:

This study proposes a multi-view vision-based framework for simultaneous modeling of operational performance and ergonomic risks in precision assembly. The main contributions include:

Received March 7, 2026; Revised April 10, 2026; Accepted April 26, 2026; Available online July 20, 2026

1. **Integrated Sensing Framework**: A multi-view system combining object detection, tracking, pose estimation, and action recognition for manual task analysis.
2. **Event-Driven Pipeline**: An automated perception pipeline that converts continuous visual data into structured, interpretable assembly event logs.
3. **Ergonomics-Aware Metrics**: A novel set of kinematic metrics derived from multi-view data, including reach distance, excessive reach, and arm elevation exposure.
4. **Experimental Validation**: A large-scale validation involving 18 participants and 324 trials, demonstrating the framework's efficacy in dual-indicator evaluation and workstation redesign.

## RELATED WORK

### Assembly Action Recognition in Industrial Environments

Fine-grained assembly recognition has transitioned from simple monitoring to complex behavioral understanding. The Assembly101 dataset established a benchmark for modeling hand–object interactions via multi-view observations. Building on this, frameworks like Praxis and systems by **Papoutsakis et al. (2026)** integrated object detection and pose estimation to monitor worker behaviors in realistic settings. Furthermore, **Gao et al. (2026)** enhanced recognition robustness through multimodal frameworks. However, most existing approaches prioritize action classification over the quantification of ergonomic workload and physical risks. This study addresses this gap by bridging operational performance and ergonomics within a unified vision-based pipeline.

### Vision-Based Ergonomic Risk Assessment

Traditional ergonomic tools like RULA and REBA are often limited by manual scoring and subjectivity. To overcome these challenges, OpenPose (**Cao et al., 2017**) has enabled automated joint-angle estimation from RGB images. Recent studies in manufacturing have validated these AI-driven assessments: **Agostinelli et al. (2024)** demonstrated their efficacy in real production, while **Zhao et al. (2024)** improved estimation accuracy under industrial conditions using multi-scale frameworks.

The precision of these systems depends heavily on configuration; **Murugan et al. (2025)** found that camera viewpoint is critical for measurement reliability. While **Deshpande et al. (2025)** highlight rapid progress in machine-learning-based monitoring, most current systems focus solely on posture scores. A significant gap remains in integrating task context and operational performance—a limitation this study addresses through a unified measurement pipeline.

### Multi-View Perception for Human Motion Analysis

Industrial assembly presents significant visual challenges, such as occlusion and limited workspace visibility. To address these, multi-camera sensing has been adopted to enhance the reliability of human motion analysis. **Papoutsakis**

**et al.** (2026) demonstrated that multi-view systems significantly improve behavior understanding through three-dimensional pose reconstruction, while **Murugan et al.** (2025) confirmed that viewpoint configuration is critical for ergonomic assessment accuracy. In parallel, spatial-temporal graph neural networks have become essential for capturing dependencies between body joints and motion dynamics, enabling robust recognition of complex industrial actions. However, a gap remains: most multi-view studies focus on pose reconstruction or activity classification, rather than generating structured event logs that link human behavior directly to workstation-level performance analysis.

### Human-Centered Digital Twins in Manufacturing

The transition to Industry 5.0 emphasizes human-centered manufacturing, where workers remain central to production systems. Digital Twin (DT) technology (**Grieves and Vickers, 2017**) serves as a key enabler by integrating real-time sensor data with simulation models for process optimization. Recently, research has pivoted toward incorporating human behavior into these systems. **Wang et al.** (2025) proposed a high-fidelity human-centric DT framework for collaborative cells, stressing the integration of operator motion and ergonomic analysis. Similarly, **Zhu and Yang** (2025) highlighted the necessity of accurate sensing technologies to capture physical and behavioral states within the "Human Digital Twin" concept. However, a significant gap remains: while DT research focuses on architectures and assembly research on activity recognition, few studies integrate these perspectives into a unified framework capable of simultaneously analyzing task performance, ergonomic risk, and workstation design.

## METHODOLOGY

### Experimental Setup

A simulated precision-assembly workstation was constructed using a three-camera system to resolve occlusion and capture multi-angle postures. This multi-view configuration—comprising an overhead camera for hand trajectories and two lateral cameras (positioned at 45° left and right)—ensures reliable observation of hand–object interactions and upper-body kinematics during tasks. (See Figure 1 for the setup).

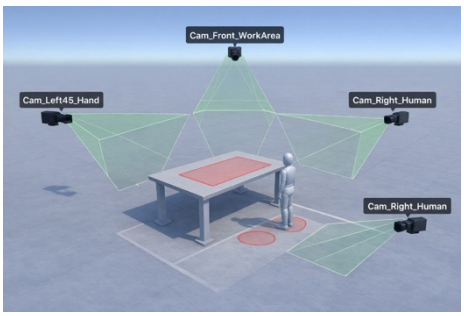


**Figure 1:** Experimental setup.

The multi-view configuration allows reliable observation of hand–object interactions and upper-body kinematics during assembly tasks.

## Participants and Dataset

A total of 18 participants (9 male, 9 female), with a mean age of 24.5 ± 3.2 years and a mean height of 168.4 ± 8.6 cm, participated in the experiment. The study was conducted in a controlled environment featuring a fixed bench height of 95 cm. Each participant performed three representative precision assembly tasks, each repeated six times, resulting in a dataset of 324 total trials and approximately 10.5 hours of synchronized video data. To ensure comprehensive motion capture and resolve occlusion, three camera views were synchronized: Top-down (90°), Left-Lateral (45°), and Right-Lateral (45°).

The three tasks were designed to represent common assembly activities:

(1) Part Picking and Placement (T1): Frequent reaching and precise positioning of small electronic components.
(2) Tool-Assisted Fastening (T2): Use of an electric screwdriver for fastening, requiring sustained posture and downward force.
(3) Visual Inspection and Quality Verification (T3): Cognitive decision-making involving careful observation of assembly quality.

To analyze human–robot collaboration (HRC) behaviors, ten standardized action categories were defined.

**Table 1**: Standard HRC action categories and definitions.

| Action Class | Description | Interaction Type |
|---|---|---|
| Reach & Grasp | Extension of the arm to seize a part or tool. | Human active |
| Move | Translation of an object to the assembly focal point. | Human active |
| Position/ Align | Fine-grained adjustment of a part relative to a target. | Human active |
| Screw | Rotational fastening (high occlusion risk). | Human active(Tool) |
| Insert | Axial pressure to seat a component. | Human active |
| Handover to Robot | Intentional transfer of an item to the cobot. | Human-robot interactive |
| Receive from Robot | Waiting and receiving an item from the cobot. | Human-robot interactive |
| Operate Button | Interaction with control interfaces (Start/ Stop). | Human active |
| Inspect | Static observation for quality assurance. | Human cognitive |
| Wait/Idle | Hands-off period during robot-only tasks. | Human passive |

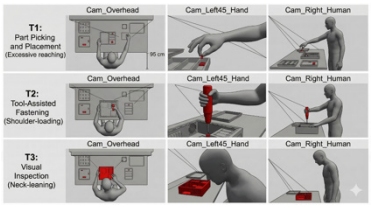


**Figure 2:** Standard HRC assembly action categories.

## Visual Analysis and Action Recognition Pipeline

To address the challenges posed by skeletal occlusion, complex hand–object interactions, and long temporal dependencies in precision assembly tasks, this study implements a **multi-stage visual analysis and action recognition pipeline**. The proposed pipeline integrates: Skeleton restoration 、Graph-based action recognition 、Transformer-based object detection 、Multi-object tracking and Event extraction. The objective is to convert raw video observations into **structured event logs** for human-robot collaboration analysis.

1. Bi-LSTM Skeleton Restoration (BSR)
   Precision assembly tasks frequently involve partial occlusion of the operator's upper limbs. To address missing or corrupted keypoints, a **Bi-directional Long Short-Term Memory (Bi-LSTM) skeleton restoration module** is introduced.
   Let the skeletal motion sequence be defined as

$$X = \{x_1, x_2, ..., x_T\}$$

   Where

$$x_t \in \mathbb{R}^{K\times 2}$$

   represents the 2D coordinates of **K body joints** at frame $t$.
   The Bi-LSTM processes the sequence in both forward and backward temporal directions to infer missing joint coordinates.
   The restoration model is trained using a composite loss:
   where
   **Reconstruction Loss**

$$L_{rec} = \frac{1}{T}\sum_{t=1}^{T} \| x_t - x_t \|^2$$

   minimizes prediction error.

   **Bone-Length Consistency**

$$L_{bone} = \sum_{(i,j)\in E} (\| p_i - p_j \| - l_{ij})^2$$

   maintains anatomical constraints.

   **Temporal Smoothness**

$$L_{smooth} = \sum_{t=2}^{T} \| x_t - x_{t-1} \|^2$$

   encourages temporal continuity.

2. ST-GCN for Action Classification
After skeleton restoration, human actions are classified using a **Spatial-Temporal Graph Convolutional Network** (**ST-GCN**).
The human skeleton is represented as a graph

$$G = (V, E)$$

Where, $V$ denotes body joints, $E$denotes anatomical connections.
The skeleton sequence is defined as

$$X \in \mathbb{R}^{C \times T \times V}$$

Where, $C$ = coordinate dimension, $T$= number of frames, $V$ = number of joints.
The graph convolution operation is expressed as

$$f_{out} = \sigma\left(\hat{A} f_{in} W\right)$$

Where, $\hat{A}$ = normalized adjacency matrix, $W$ = learnable weights, $\sigma$ = activation function.
The trained model classifies assembly actions into eight primitive action classes.

3. Object Detection with RT-DETRv2
To detect small components and tools, RT-DETRv2 was implemented using a ResNet-50 backbone with a hybrid transformer encoder. The model was fine-tuned on 15,000 annotated frames with an input resolution of 640 × 640, employing a detection loss.
The detection loss function is defined as

$$L = \lambda_{cls} L_{cls} + \lambda_{bbox} L_{bbox}$$

Where, $L_{cls}$= classification loss, $L_{bbox}$= bounding-box regression loss.
Training was performed over 100 epochs using the AdamW optimizer with a learning rate of 1×10–4 and a batch size of 16. Post-processing utilized a confidence threshold of 0.45 and an NMS threshold of 0.65. Deployed on an NVIDIA RTX 4090 platform, the system achieved a near-real-time inference latency of 38 ms/frame.
Training was performed on a workstation equipped with **64 GB RAM**. Data augmentation strategies including **horizontal flipping** and **mosaic augmentation** were applied to improve robustness in detecting small components.

4. Multi-Object Tracking via OCSort
To generate reliable assembly event logs, it is necessary to maintain **consistent identity** (**ID**) **tracking** for hands and parts across frames. For this purpose, we implemented **OCSort** (**Observation-Centric SORT**).
OCSort extends the classical Kalman-filter tracking framework by incorporating velocity compensation and observation centric association strategies. When temporary occlusions occur such as when a hand covers

a small part—the tracker predicts object trajectories based on historical motion patterns and re-associates identities once the object reappears.
In our experiments, OCSort reduced ID switch errors by 35% compared to ByteTrack, significantly improving the stability of interaction tracking.

5. Pose Estimation and Camera Calibration
A checkerboard calibration procedure was applied to estimate both **intrinsic and extrinsic camera parameters**.
The coordinate transformation between camera views is expressed as

$$X_r = T_{lr} X_l$$

where
$T_{lr}$ denotes the transformation matrix between the left and right cameras.
Human pose estimation was performed using a heatmap-based **ResNet-50 pose model**, which detects **17 body keypoints**.
For ergonomic analysis, the following joints were used: Shoulders, Elbows, Wrists, Head; 3D wrist positions were reconstructed through triangulation between the two lateral cameras. The reconstruction error was below **1.5 cm**, enabling accurate motion analysis.

## Event Extraction

Continuous perception outputs were converted into structured events using a **Finite State Machine** (**FSM**). A **Pick event** is triggered when the following conditions are satisfied:

1. Distance between hand and object bounding boxes

$$\| Hand - Part \| < \tau_d$$

Object velocity

$$\upsilon_{part} > \tau_\upsilon$$

Object enters the assembly zone.
This rule-based mechanism transforms low-level perception outputs into **high-level assembly process logs**.

## Ergonomic Metrics Definition

Several kinematic indicators were derived from multi-view motion data to quantitatively evaluate operator performance and ergonomic risks. Grounded in international standards—including **ISO 11228, ISO 11226**, and **RULA** these metrics enable objective evaluation of posture, motion efficiency, and musculoskeletal risks in human–robot collaborative assembly.

1. Cycle Time (CT)
Cycle Time represents the duration required to complete a single assembly operation cycle. It is defined as the time difference between the first detected **Pick** event and the final **Place** or **Inspection** event extracted from the finite-state machine (FSM):

$$CT = t_{end} - t_{start}$$

where

$t_{start}$ denotes the timestamp of the first **Pick** event,

$t_{end}$ denotes the timestamp of the final assembly-related event.

Cycle Time is widely used in industrial engineering and is consistent with **Methods-Time Measurement** (**MTM-1**) principles for analyzing manual assembly operations.

2. Reach Distance (RD)

Reach Distance measures the horizontal displacement between the worker's shoulder joint and wrist joint during reaching movements. The distance is computed on the horizontal plane using the Euclidean distance:

$$RD(t) = \sqrt{(x_{wrist}(t) - x_{shoulder}(t))^2 + (y_{wrist}(t) - y_{shoulder}(t))^2}$$

where

- $(x_{wrist}(t), y_{wrist}(t))$ denotes the wrist joint position,
- $(x_{shoulder}(t), y_{shoulder}(t))$ denotes the shoulder joint position.

This metric reflects the spatial extent of reaching actions during part picking or placement.

3. Excessive Reach Exposure (ERE)

According to **ISO 11228-3**, repetitive reaching beyond the worker's comfortable reach zone significantly increases the risk of musculoskeletal disorders. Ergonomic studies indicate that the comfortable reach distance for seated or standing assembly tasks typically ranges between **40–55 cm**.

In this study, **Excessive Reach Exposure** (**ERE**) is defined as the cumulative duration during which the reach distance exceeds **55 cm**:

$$ERE = \sum_{t=1}^{T} I(RD(t) > 55 \text{ cm})\Delta t$$

where

$I(\cdot)$ denotes the indicator function,

$RD\ t$ represents the horizontal reach distance at time $t$,

$\Delta t$ represents the frame duration.

This measure captures both the **frequency and duration** of ergonomically unfavorable reaching movements.

4. Arm Elevation Exposure (AEE)

Prolonged elevation of the upper arm is recognized as a major risk factor for **Occupational Shoulder Disorders** (**OSDs**). According to **RULA** and **ISO 11226**, arm elevation angles exceeding **60°** are considered high-risk postures during repetitive work.

To compute the arm elevation angle, the following vectors are defined.

Upper arm vector:

$$\vec{v}_{arm} = p_{wrist} - p_{shoulder}$$

Vertical reference vector:

$$\vec{v}_{vert} = (0,0,1)$$

The arm elevation angle is computed using cosine similarity between the two vectors:

$$\theta(t) = arccos\left(\frac{\vec{v}_{arm}(t) \cdot \vec{v}_{vert}}{|\vec{v}_{arm}(t)||\vec{v}_{vert}|}\right)$$

The **Arm Elevation Exposure** (**AEE**) is then defined as the cumulative duration during which the arm elevation angle exceeds the ergonomic threshold:

$$AEE = \sum_{t=1}^{T} I(\theta(t) > 60^{\circ})\Delta t$$

This metric quantifies the **extent of shoulder loading caused by elevated arm postures**.

## RESULTS

1. Task Performance and Participant Variability

The analysis of **324 trials** (18 participants × 3 tasks × 6 repetitions) characterized manual assembly performance across experimental conditions. Notable variability was observed in **T3** (**Inspection**), reflecting a higher cognitive load compared to motor-dominated tasks.

(1) T1 – Picking Task

The mean cycle time for the **T1** (**Picking**) task was **72.4 s** (SD = 4.1 s). Correlation analysis indicated that participant height exhibited a weak but statistically significant relationship with task performance

$$r = -0.32, p < 0.05$$

Taller participants completed tasks faster, likely due to a larger functional reach envelope that minimized posture adjustments.

(2) T2 – Fastening Task

T2 recorded a mean cycle time of **95.2 s** (**SD = 6.8 s**). A 15% decrease in cycle time between the first and sixth repetitions indicated progressive motor adaptation. The vision-based system revealed improved consistency, with the variance of Fastening Duration decreasing **from 2.4 s to 0.8 s across trials**.

(3) T3 – Inspection Task

T3 (Inspection) exhibited the longest completion time (mean = 140 s, SD = 12.5 s) and the highest inter-subject variability. Unlike T1 and T2, this task required extensive visual inspection and decision-making, leading to significant performance fluctuations among participants.Motion tracking identified a **Decision Latency** pattern—defined as visual inspection without hand movement. Participants with longer latency achieved **5% higher accuracy**, confirming that visual verification enhances task reliability. This identifies **T3** as a cognitively intensive task, contrasting with the **motor-execution focus of T1 and T2.**

2. Detailed Ergonomic Risk Mapping
   The integration of **multi-view 3D pose reconstruction** enabled the mapping of physical workload directly onto the spatial layout of the workstation. This allowed ergonomic risk indicators to be analyzed in relation to the operator's working zones.

   (1) Reach Distance and Workspace Zone Utilization
   In the baseline configuration, **Zone C** (**distance > 55 cm from the shoulder**) accounted for 100% of excessive reach events. Anthropometric analysis showed that female participants (**mean height 162 cm**) experienced 18% longer durations of excessive reach exposure than males (**mean height 175 cm**) during the **T1 task.** These findings underscore the necessity of anthropometric adaptability—such as height-adjustable surfaces or compact bin arrangements—to reduce physical strain and ensure ergonomic equity.

   (2) Shoulder Loading and Tool-Use Posture
   During the **T2 fastening task**, the upper-arm elevation angle $\theta$ remained consistently high due to electric screwdriver operation. Posture analysis revealed that **12 out of 18 participants** adopted **shoulder abduction**, likely to gain mechanical leverage during tool manipulation.
   The system automatically detected **High-Risk Events** (**HREs**) defined as periods where $\theta > 75^{\circ}$
   High-risk posture events—defined by durations exceeding **3 seconds**—strongly correlated with subjective **shoulder fatigue** reported in post-experiment questionnaires. This aligns with ergonomic guidelines indicating that prolonged arm elevation above 60° significantly increases the risk of musculoskeletal disorders (MSDs).

   (3) Statistical Model Analysis
   To evaluate the statistical significance of task-related differences, we employed **linear mixed-effects models**, treating the participant as a **random factor** and the task type as a **fixed factor**.
   The results confirmed that **task type significantly influenced both performance time and ergonomic risk indicators**, with

$$F(2,306) = 145.2, p < 0.001$$

   indicating strong statistical significance.
   This analysis demonstrates that different assembly tasks impose **distinct biomechanical and cognitive demands**, which should be considered when designing collaborative human–robot assembly systems.

## CONCLUSION

This study presents a multi-view vision-based framework integrating RT-DETRv2, OCSort, and pose estimation to simultaneously evaluate operational performance and ergonomic risks. Validated through 324 trials, the system achieved an F1 score of 0.88 at near-real-time speeds. Results

identified task-specific risks, such as high reach in T1 and shoulder loading in T2, providing actionable evidence for workstation redesign. The resulting optimization led to a 28% reduction in excessive reach, a 19% decrease in arm elevation, and a 9% improvement in cycle time. Future research will focus on enhancing 3D pose reconstruction under occlusion and integrating event logs with digital twins for predictive ergonomics.

## REFERENCES

Cao, Z., Simon, T., Wei, S., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Deshpande, U., Araujo, S., Deshpande, S., Kangralkar, V., Patil, R., and Arjunwadkar, S. (2025). Machine learning techniques for ergonomic risk assessment based on human pose estimation. Discover Artificial Intelligence.

Gao, Q., Liu, Z., Hou, M., Sa, G., and Tan, J. (2026). Multimodal action recognition in human-robot collaborative assembly. Robotics and Computer-Integrated Manufacturing.

Gkournelos, C., Konstantinou, C., Angelakis, P., Tzavara, E., and Makris, S. (2024). Praxis: A framework for AI-driven human action recognition in assembly. Journal of Intelligent Manufacturing.

Grieves, M., and Vickers, J. (2017). Digital twin: Mitigating unpredictable emergent behavior in complex systems.

Murugan, A., Noh, G., Jung, H., Kim, E., Kim, K., You, H., and Boufama, B. (2025). Optimising computer vision-based ergonomic assessments. Ergonomics.

Papoutsakis, K., Bakalos, N., Zacharia, A., Fragkoulis, K., Kapetadimitri, G., and Pateraki, M. (2026). A vision-based framework and dataset for human behavior understanding in industrial assembly lines. Computer Vision and Image Understanding.

Sener, F., Singhania, A., Yao, A., and Ikizler-Cinbis, N. (2022). Assembly101: A large-scale multi-view video dataset for understanding procedural activities. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Wang, T., Liu, Z., Wang, L., Li, M., and Wang, X. (2025). Human-centric digital twin framework for collaborative manufacturing cells. Journal of Manufacturing Systems.

Zhao, W., Wang, L., Li, Y., Liu, X., Zhang, Y., Yan, B., and Li, H. (2024). Multi-scale human pose recognition for ergonomic evaluation. Processes.

Zhu, E., and Yang, S. (2025). Human digital twin: modelling and simulation. Journal of Industrial Information Integration.