Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064250

REVIEW

Monocular 3D Human Pose Estimation for REBA Ergonomics: A Critical Review of Recent Advances

Ahmad Mwfaq Bataineh[1,2,*] **and Ahmad Sufril Azlan Mohamed**[1]

[1]School of Computer Science, Universiti Sains Malaysia, Penang, 11800, Malaysia
[2]School of Computer Science, Prince Sattam Bin Abdulaziz University, Al-Kharj, 16273, Saudi Arabia

*Corresponding Author: Ahmad Mwfaq Bataineh. Email: ahmadbataineh@student.usm.my

Received: 10 February 2025; Accepted: 13 May 2025; Published: 09 June 2025

ABSTRACT: Advancements in deep learning have considerably enhanced techniques for Rapid Entire Body Assessment (REBA) pose estimation by leveraging progress in three-dimensional human modeling. This survey provides an extensive overview of recent advancements, particularly emphasizing monocular image-based methodologies and their incorporation into ergonomic risk assessment frameworks. By reviewing literature from 2016 to 2024, this study offers a current and comprehensive analysis of techniques, existing challenges, and emerging trends in three-dimensional human pose estimation. In contrast to traditional reviews organized by learning paradigms, this survey examines how three-dimensional pose estimation is effectively utilized within musculoskeletal disorder (MSD) assessments, focusing on essential advancements, comparative analyses, and ergonomic implications. We extend existing image-based classification schemes by examining state-of-the-art two-dimensional models that enhance monocular three-dimensional prediction accuracy and analyze skeleton representations by evaluating joint connectivity and spatial configuration, offering insights into how structural variability influences model robustness. A core contribution of this work is the identification of a critical research gap: the limited exploration of estimating REBA scores directly from single RGB images using monocular three-dimensional pose estimation. Most existing studies depend on depth sensors or sequential inputs, limiting applicability in real-time and resource-constrained environments. Our review emphasizes this gap and proposes future research directions to develop accurate, lightweight, and generalizable models suitable for practical deployment. This survey is a valuable resource for researchers and practitioners in computer vision, ergonomics, and related disciplines, offering a structured understanding of current methodologies and guidance for future innovation in three-dimensional human pose estimation for REBA-based ergonomic risk assessment.

KEYWORDS: Human posture estimation; deep neural networks; three-dimensional analysis; benchmark datasets; rapid entire body assessment (REBA)

1 Introduction

Accurate estimation of three-dimensional human pose has significantly advanced ergonomic assessments, particularly within Rapid Entire Body Assessment (REBA) frameworks, by enabling automated, real-time evaluation of workplace health and safety. Conventional ergonomic assessments typically depend on manual observations or wearable sensors to identify musculoskeletal risks, which can be time-consuming and intrusive. By leveraging computer vision, three-dimensional pose estimation provides a non-invasive approach to monitor posture, assess joint positions, and identify risk factors for musculoskeletal disorders. These advancements have the potential to transform workplace safety practices by offering continuous, objective assessments that improve productivity and reduce injuries. REBA assessments specifically require

accurate evaluation of posture angles and limb positions, explicitly considering the three-dimensional spatial orientation (x, y, z) of joints. Traditional two-dimensional pose estimation often fails to capture depth information accurately, resulting in unreliable ergonomic scores. In contrast, three-dimensional human pose estimation directly provides precise depth perception, enabling accurate joint localization in three-dimensional space, thus significantly improving the reliability and objectivity of ergonomic assessments. Despite these benefits, integrating three-dimensional pose estimation into ergonomic assessments poses challenges, such as ensuring high accuracy in diverse environments, addressing depth ambiguities, and adapting models to various skeleton representations.

Beyond ergonomics, estimating three-dimensional human poses from monocular images, particularly for single-person scenarios, remains a central challenge in computer vision. This task involves mapping two-dimensional image coordinates (x, y) to three-dimensional space (x, y, z) and presents unique difficulties due to complex body articulations and the need for models to generalize across diverse conditions. Techniques like deep learning methods, including Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and Transformer architectures have shown remarkable potential by learning spatial relationships between joints and addressing non-linear mappings. However, these methods often require specialized training strategies, such as heatmap regression, attention mechanisms, and kinematic constraints, to refine depth estimation and enhance joint localization.

This article offers a detailed review of recent advancements in monocular-based three-dimensional human pose estimation, focusing on applications in REBA, ergonomic assessments, and the challenges of skeleton format variability. By consolidating state-of-the-art deep learning techniques, approaches for dataset fusion, and strategies to address key challenges, this work aims to guide researchers in advancing three-dimensional pose estimation for ergonomics and beyond. The insights offered in this survey have the potential to accelerate the adoption of three-dimensional pose estimation in real-world ergonomic applications and inspire innovations that make these techniques more robust, scalable, and impactful across diverse fields.

2 Previous Surveys

The purpose of this survey is to build upon and address the gaps identified in previous studies addressing three-dimensional human pose estimation and its practical applications in ergonomic assessment methodologies, such as the Rapid Entire Body Assessment (REBA) approach. Several prior surveys have extensively reviewed methods for transforming two-dimensional pose data into three-dimensional representations, focusing on deep learning models, graph-based approaches, and self-supervised techniques [1–3]. For instance, surveys on three-dimensional human pose estimation have provided comprehensive insights into monocular methods and highlighted challenges like depth ambiguity, occlusions, and dataset variability [4]. However, these works often miss the critical connection between pose estimation and practical applications like ergonomics. Likewise, previous reviews on two-dimensional pose estimation primarily emphasized improvements in accuracy and robustness but lack an in-depth exploration of how two-dimensional techniques contribute to downstream ergonomic analysis [5].

Numerous studies have focused on ergonomic risk prediction using established observational methods such as RULA, REBA, OWAS, PERA, and OCRA in traditional settings or automated systems enhanced by deep learning and vision-based tools. For example, Abobakr et al. [6] proposed a semi-automated RULA system using depth images and a residual CNN to estimate joint angles from human poses. Their proposed system demonstrated an overall prediction accuracy of 89% in real-world scenarios, though it lacked comprehensive ablation studies and comparative model evaluations. Hossain et al. [7] introduced a deep learning framework for predicting REBA scores based on three-dimensional keypoints derived

from the Human3.6M dataset, proving the feasibility of integrating three-dimensional pose estimation with REBA. However, the system was limited to static poses and did not address occlusions or real-time deployment challenges.

Agostinelli et al. [8] presented a comparative analysis of REBA and RULA in industrial environments, showing that REBA was more reliable for lower-body assessment, while RULA excelled in upper-body scoring, though both were affected by partial visibility issues. Beheshti et al. [9] applied the OWAS method in agricultural and industrial settings and found it reliable but less flexible for complex and dynamic tasks, particularly in its neglect of repetitive motion intensity and load variation.

Chander and Cavatorta [10] proposed the PERA method, validated against EAWS in cyclic industrial tasks, but its simplification of force and duration metrics limited its application in high-load scenarios. Erginel and Toptanci [11] used the OCRA index to assess repetitive motion injuries in food industry workers and identified high wrist and back risks, reinforcing the need for automated posture monitoring systems. Wu et al. [12] developed a REBA-based mobile application using Mask R-CNN for real-time ergonomic risk assessment, performing well under controlled conditions but showing limited robustness in noisy or poorly lit environments.

A significant contribution was made by Ghasemi and Mahdavi [13], who proposed a new scoring system for REBA called FBnREBA that incorporates fuzzy sets and Bayesian networks. This approach improved the sensitivity of REBA by allowing for more nuanced scoring of posture risk and better differentiation between similar postures. The system was validated through a case study and demonstrated stronger performance compared to the traditional REBA method.

Stefana et al. [14] systematically reviewed wearable ergonomic monitoring devices such as IMUs, smart insoles, and vibrotactile systems. While accurate, these systems were noted for intrusiveness and setup complexity, emphasizing the need for scalable, non-contact alternatives such as computer vision.

These findings highlight that while traditional methods remain foundational, the integration of deep learning and vision-based pose estimation marks a pivotal shift in ergonomic assessment. However, challenges in validation, scalability, and generalizability remain. This review aims to address these gaps through a focused analysis of monocular three-dimensional HPE systems for REBA evaluation. Specifically, this paper explores the largely unexamined possibility of directly estimating REBA ergonomic risk scores from single RGB images through deep learning-driven monocular three-dimensional human pose estimation techniques. Unlike prior studies that predominantly rely on video sequences, temporal information, or depth sensors, this review targets real-time ergonomic risk analysis in scenarios constrained by limited hardware or computational resources. Our work aims to overcome current limitations by proposing a solution that can generalize across variable skeleton structures, handle partial occlusions, and deliver REBA assessments efficiently from a single static image input.

3 Survey Methodology

The primary goal of this section is to systematically review recent research concentrating on single-stage and two-stage human poses estimation methods, diverse skeletal data formats, and ergonomic assessments utilizing the REBA method with monocular imaging.

Research question:

RQ1: What are the main methodologies and classification schemes employed in three-dimensional human pose estimation from monocular images?

RQ2: What are the key challenges associated with various three-dimensional human pose estimation methodologies?

RQ3: Which datasets are most frequently utilized for performance evaluation in three-dimensional human poses estimation research?

RQ4: What are the current limitations of REBA-based assessments, and what potential improvements could enhance this approach?

4 Taxonomy of the Survey

The primary aim of this review is to provide an extensive and current review of advancements in three-dimensional human pose estimation from monocular images, addressing variations in dataset skeleton formats and applications in REBA ergonomic assessment. In compliance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, research papers published from 2014 through May 2024 were systematically analyzed, ensuring a rigorous selection process and transparent reporting. The search involved in-depth manual exploration of essential publications in human pose estimation, sourcing from well-regarded databases including databases like arXiv, Google Scholar, IEEE Xplore, ScienceDirect, Springer Link, and ACM Digital Library. Relevant studies were identified and compiled in December 2024.

To provide historical context and acknowledge foundational contributions, we also included significant works published up to 2024, specifically selecting those studies that introduced influential methodologies or perspectives to three-dimensional pose estimation. Additionally, manual cross-referencing was performed to identify additional critical studies that may not have appeared in the initial search. Recognizing the integral role of two-dimensional pose estimation as a precursor to three-dimensional tasks, we incorporated a dedicated section on two-dimensional pose estimation, highlighting the most frequently cited and impactful models in the field.

5 Dataset

The availability of extensive datasets with diverse examples is a crucial aspect of advancing computer vision tasks. In the context of three-dimensional human pose estimation, a key challenge is the scarcity of extensive outdoor datasets. While datasets for two-dimensional human pose estimation often include benchmarks collected in diverse indoor and outdoor environments, datasets for three-dimensional human pose estimation are predominantly gathered under controlled laboratory conditions employing motion capture (MoCap) systems. This method of data collection often restricts diversity in backgrounds, camera viewpoints, and illumination conditions. In this section, this section reviews several prominent datasets commonly utilized in evaluating two-dimensional and three-dimensional human pose estimation approaches.

5.1 LSP (2010)

The Leeds Sports Pose (LSP) dataset serves as a benchmark dataset focused on single-person pose estimation. It contains 1000 images annotated with 14 keypoints representing major body joints. The dataset primarily consists of images of individuals engaged in sports activities, which makes it especially useful for assessing the performance of pose estimation techniques under dynamic and varied poses. The evaluation metric used for this dataset is the Percentage of Correct Parts (PCP), an evaluation metric that quantifies the accuracy of predicted limb locations relative to the ground truth [15].

5.2 LSP-Extended (2011)

The LSP-Extended Dataset builds on the original LSP dataset, providing 10,000 additional single-person images annotated with the same 14 keypoints. It is designed to enhance the training of pose estimation models

by offering an expanded and more varied collection of images. Like its predecessor, it uses the PCP metric for evaluation, ensuring consistency in performance comparisons [16].

5.3 FLIC-Plus (2014)

The FLIC-Plus dataset extends and enhances the original FLIC dataset, featuring 17,000 training images annotated for the same 10 upper-body keypoints. This dataset focuses on improving annotation quality and consistency while retaining the primary purpose of upper-body pose estimation. Evaluation continues to rely on PCP and PCK, providing benchmarks for accuracy in predicting upper-body joint positions.

5.4 MPII (2014)

The MPII Human Pose dataset is an extensive benchmark created to support single-person and multi-person human pose estimation research. It contains 29,000 training images, 3800 validation images, and 1700 test images annotated with 16 body keypoints. It includes diverse human activities captured across various indoor and outdoor scenarios, providing versatility for both training and benchmarking. Evaluation metrics employed are PCPm and PCKh (Percentage of Correct Keypoints normalized by head size), ensuring robust and equitable model performance comparisons [17].

5.5 COCO Dataset

The COCO 2016 dataset serves as a prominent benchmark commonly employed in multi-person two-dimensional pose estimation research. It features 45,000 training images, 22,000 validation images, and 80,000 test images, with annotations for 17 body keypoints. The dataset provides a diverse range of human poses across various activities and environments, enabling models to learn from unconstrained scenarios. The evaluation metric used is Average Precision (AP), which measures the overlap between predicted and ground truth poses. The COCO 2017 is an updated version of the COCO 2016 dataset, expanding the training set to 64,000 images while providing 2700 validation images and 40,000 test images. Like its predecessor, it focuses on multi-person pose estimation with the same number of keypoints in COCO 2016. The dataset's increased size and refined annotations make it a critical resource for improving model generalization. Evaluation is based on AP, consistent with the 2016 version [18].

5.6 HumanEva (2010)

The HumanEva Dataset uses a marker-based motion capture (MoCap) system to collect three-dimensional human pose data in controlled indoor environments. It includes 6 subjects performing 7 actions, totaling 40,000 frames. This dataset supports single-person pose estimation and provides both single-view and multi-view recordings, making it a foundational resource for early three-dimensional pose estimation research [19].

5.7 Human3.6M (2014)

Human3.6M is a widely recognized, comprehensive marker-based motion capture (MoCap) dataset for three-dimensional pose estimation tasks. The dataset includes over 3.6 million frames depicting 11 subjects performing 17 distinct actions captured within indoor environments. Supporting single-person pose estimation, this dataset provides both single-view and multi-view data, serving as a standard benchmark for evaluating three-dimensional human pose estimation methods [20].

5.8 CMU Panoptic (2016)

The CMU Panoptic Dataset employs a marker-less MoCap system in an indoor environment. It includes data from 8 subjects with a total of 1.5 million frames. Supporting single-person pose estimation, the dataset is notable for its extensive multi-view capture setup, enabling high-fidelity three-dimensional pose reconstruction [21].

5.9 MPI-INF-3DHP (2017)

The MPI-INF-3DHP dataset is a markerless motion capture (MoCap) dataset acquired under indoor and outdoor conditions. It contains approximately 1.3 million frames, involving eight subjects performing eight distinct actions. Intended for single-person pose estimation, the dataset includes single-view and multi-view recordings, providing diverse scenarios suitable for training robust pose estimation models [22].

5.10 TotalCapture (2017)

The TotalCapture dataset integrates a marker-based motion capture (MoCap) system with inertial measurement units (IMUs), capturing data exclusively in indoor environments. It includes five subjects performing five distinct actions, totaling 1.9 million frames. Designed specifically for single-person pose estimation, this dataset provides single-view and multi-view recordings to facilitate the evaluation of pose estimation accuracy [23].

5.11 DPW (2018)

The three-dimensional Pose in the Wild (3DPW) Dataset is captured using handheld cameras with IMUs in indoor and outdoor environments. It consists of 7 subjects performing various actions, with 51,000 frames. Supporting single-person pose estimation, it is unique for its focus on real-world, unconstrained scenarios and includes both single-view and multi-view data [24].

5.12 MuPoTS-3D (2018)

The MuPoTS-3D Dataset uses a markerless MoCap system to collect data in indoor and outdoor settings. It includes 8 subjects and 8000 frames, supporting both single-person and multi-person pose estimation. The dataset offers single-view data, making it particularly suitable for evaluating methods designed for outdoor and multi-person scenarios [25].

5.13 SURREAL Dataset (2017)

The SURREAL dataset offers synthetic data for three-dimensional human pose estimation generated using computer graphics techniques. It includes 24 key points based on the SMPL skeleton format, providing detailed annotations of synthetic human poses in various activities and environments. The use of synthetic data allows for the generation of large amounts of annotated data at a lower cost compared to real-world data collection. Models 50 trained on SURREAL can benefit from the diversity and precision of the synthetic annotations. However, the gap between synthetic and real-world data can pose challenges for generalization, as models may need fine-tuning on real-world datasets to perform effectively in practical applications [26].

5.14 PeopleSansPeople Dataset (2022)

The PeopleSansPeople Dataset is a synthetic data generator developed in 2022 using the Unity engine and the Unity Perception package, aimed at advancing human-centric computer vision tasks, including two-dimensional and three-dimensional human pose estimation. This generator provides high-quality,

simulation-ready three-dimensional human assets with diverse ages, ethnicities, clothing, and actions, featuring 28 human models and 39 animations ranging from basic movements to complex interactions. Leveraging domain randomization techniques, it generates datasets with randomized lighting, camera settings, and occluding objects to create diverse and realistic scenarios. PeopleSansPeople produces COCO-compatible annotations [27].

6 Metric Evaluation for 3D and 2D

In three-dimensional and two-dimensional Human Pose Estimation (HPE), accurate evaluation of model performance is essential to assess the quality of predicted poses relative to ground truth. Several standard metrics are commonly used to measure the precision, alignment, and reliability of pose estimation. These include:

6.1 3 DHPE Evaluation Metric

6.1.1 Mean per Joint Position Error (MPJPE)

This metric is frequently employed as an evaluation measure in three-dimensional human pose estimation, quantifying the mean Euclidean distance (in millimeters) between the estimated and ground-truth joint coordinates. It assesses the model's accuracy by averaging errors across all joints and frames.

$$\text{MPJPE} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}|p_i^{(t)} - g_i^{(t)}|_2 \tag{1}$$

where the T is total number of frames in the dataset, N are number of joints per pose, $P_i^{(t)}$ are predicted three-dimensional coordinates of the i-th joint in frame t, and the $g_i^{(t)}$ are Ground truth three-dimensional coordinates of the i-th joint in frame t.

On the Human3.6M dataset, MPJPE evaluation follows standardized protocols. Protocol 1 computes the error directly without any alignment, utilizing subjects 1, 5, 6, 7, and 8 for training and subjects 9 and 11 for testing. Protocol 2, termed Procrustes-aligned MPJPE (P-MPJPE), aligns the estimated poses with ground truth by compensating for translation, rotation, and scaling to emphasize structural differences. Protocol 3 (Normalized MPJPE or N-MPJPE) employs only scale alignment and specifically examines sequences recorded from the frontal camera without subsampling.

6.1.2 Procrustes Aligned MPJPE (PA-MPJPE)

MPJPE assesses the MPJPE calculated after aligning estimated poses to the corresponding ground truth via Procrustes analysis, thereby removing differences in scale, rotation, and translation. This metric specifically quantifies structural pose errors independent of camera alignment or scale factors.

6.1.3 Normalized MPJPE (NMPJPE)

NMPJPE is a variant of MPJPE where predictions are normalized by scale before computing the error. This metric helps in evaluating the structural similarity of poses without the influence of absolute scaling differences.

6.1.4 3D Percentage of Correct Keypoints (3DPCK)

3DPCK measures the percentage of joints where the predicted position is within a certain threshold distance (e.g., 150 mm) from the ground truth. Formally:

$$\text{3DPCK} = \frac{1}{N}\sum_{i=1}^{N} I\left(|J_i - J_i^*|_2 \leq \tau\right) \tag{2}$$

where τ is the distance threshold, and I is the indicator function.

6.1.5 Area Under Curve for Relative Error (AUCrel)

AUCrel evaluates the relationship between the error threshold and the proportion of correctly predicted keypoints, plotting a curve of relative error vs. precision. The area under this curve provides a summary measure of model accuracy across varying thresholds.

6.2 2 DHPE Evaluation Metric

6.2.1 Percentage of Correct Keypoints (PCK)

PCK quantifies keypoint estimation accuracy by determining the proportion of predicted joints that fall within a specified threshold distance from their ground-truth positions. This threshold typically corresponds to a proportion of a reference dimension, such as head size (PCKh) or torso length. For instance, PCK@0.5 denotes the percentage of keypoints detected within 50% of the reference measurement. This evaluation metric is commonly employed to assess the general accuracy of pose estimation models, especially in datasets such as MPII and COCO.

$$\text{PCK} = \frac{1}{N}\sum_{i=1}^{N} I\left(|J_i - J_i^*|_2 \leq \alpha L\right) \tag{3}$$

where L is the reference length, α is the threshold factor, and I is the indicator function.

6.2.2 Object Keypoint Similarity (OKS)

OKS is a COCO-specific metric that evaluates keypoint accuracy based on normalized distances between predicted and ground truth keypoints, weighted by visibility and keypoint type. It incorporates a Gaussian function to account for variability in keypoint size and visibility. OKS serves as the basis for calculating Average Precision (AP) scores in COCO benchmarks, providing a holistic measure of detection and localization performance.

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{|J_i - J_i^*|_2^2}{2s^2\kappa_i^2}\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)} \tag{4}$$

where J_i and J_i^* are the predicted and ground truth keypoints, s is the object scale, κ_i is a per-keypoint constant, and v_i is the visibility flag.

6.2.3 Average Precision (AP)

AP is derived from the OKS metric and measures precision over a range of recall values. It is calculated at multiple OKS thresholds, such as 0.5 0.75, and an average across thresholds from 0.5 to 0.95 (commonly referred to as AP@50, AP@75, and AP@ [0.5:0.95], respectively). This metric evaluates the consistency of keypoint localization across different levels of accuracy, making it a standard for COCO evaluation.

6.2.4 Root Mean Squared Error (RMSE)

RMSE measures the average squared difference between predicted and ground truth keypoint coordinates. It provides a straightforward evaluation of positional accuracy in pixel space. While less commonly used in public benchmarks, RMSE is particularly useful for analyzing model performance in custom datasets or controlled settings.

6.2.5 Normalized Mean Error (NME)

NME is similar to RMSE but normalizes the error by a reference distance, such as the interocular distance for facial keypoints or the torso size for full-body poses. This normalization allows for fair comparisons across datasets with varying scales and resolutions, making it a popular choice for cross-dataset evaluations.

7 3D Human Pose Estimation

Three-dimensional human pose estimation extends the concept of two-dimensional pose estimation, which determines joint positions using coordinates on the x and y axes within a two-dimensional space. In contrast, three-dimensional pose estimation introduces a depth component (z-axis), significantly enhancing the complexity of the estimation task. This third dimension enables crucial applications across various fields, including ergonomic assessments of work-related musculoskeletal disorders (WMSDs), computer vision, and interactive gaming, where accurate spatial depth perception is vital.

Advancements in deep learning have dramatically reshaped this field by effectively modelling nonlinear joint relationships and capturing intricate spatial configurations. Methods for three-dimensional human pose estimation are typically classified into two main groups: single-person and multi-person methods. Single-person approaches, which form the primary focus of this review, concentrate on accurately estimating the pose of one individual at a time. Conversely, multi-person methods address scenarios involving multiple interacting individuals, introducing additional complexity related to joint detection, identification, and assignment.

Furthermore, methods can also be differentiated according to their coordinate systems. In person-centric (root-relative) frameworks, joint positions are expressed relative to a central reference joint—commonly the pelvis—yielding representations that are invariant to the camera location. Alternatively, the camera-centric system positions joints directly within the camera's coordinate reference, accurately representing absolute spatial placement relative to the camera viewpoint.

7.1 Single-Person Approach

In [28–32], Single-person three-dimensional human pose estimation aims to reconstruct the spatial configuration of body joints from a single image or video frame, where the individual is the primary subject. In this setup, the model often operates under the assumption that the person's position, scale, and bounding box are given, which aids in isolating the pose estimation process. However, estimating three-dimensional poses from monocular images presents unique challenges due to the inherent ambiguity of depth information and the difficulty in generalizing to varied environments. This process is typically divided into two main approaches: one-stage and two-stage methods. One-stage approaches directly regress three-dimensional poses from images, offering a streamlined path from image to three-dimensional output, while two-stage methods estimate two-dimensional poses as an intermediate step before mapping them to three-dimensional space, often enhancing reliability by leveraging intermediate joint detection. Additionally, the ill-posed nature of monocular three-dimensional pose estimation, where similar two-dimensional projections can correspond to multiple three-dimensional poses, requires sophisticated learning techniques

to ensure robustness. Deep learning models, trained on large datasets like Human3.6M and HumanEva-I, have proven effective in directly regressing three-dimensional joint positions. Nevertheless, the reliance on controlled dataset environments sometimes limits the generalizability of these models to more diverse, real-world settings. The challenge of single-person three-dimensional pose estimation continues to drive research toward refining algorithms that can overcome depth ambiguities, improve model generalization, and accurately infer three-dimensional human poses from single-camera inputs.

In this section, we aim to provide an overview of the basic concept and recent advances in this field, focusing on diction base and regression base from monocular image to be the core suggestion model for the REBA assessment method.

7.1.1 Single-Stage Pipeline

In this section, we discuss estimating the three-dimensional key points directly without using intermediate two-dimensional key points, which are end-to-end networks; we address this approach in two pipelines, the Regression-based (Fig. 1) and the Detection-based approach (Fig. 2).

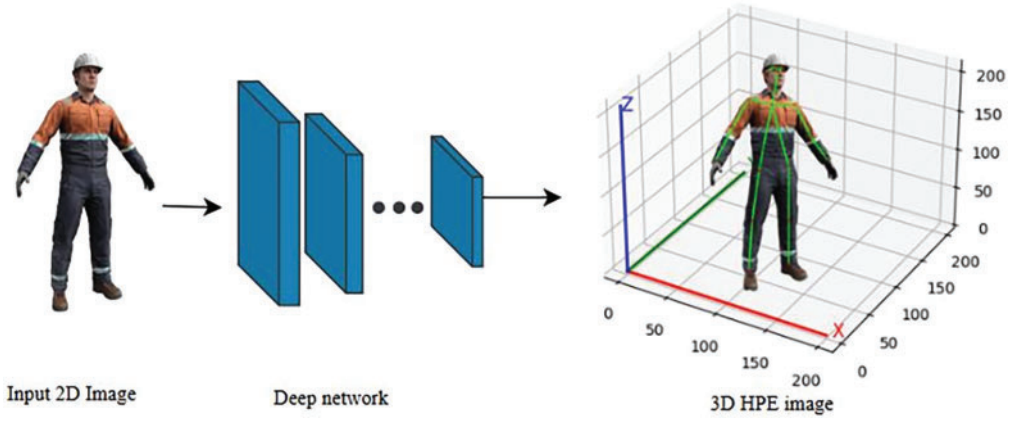


Figure 1: Regression-based Approach for three-dimensional Human Pose Estimation. The figure illustrates a regression-based methodology where a deep neural network is utilized to directly estimate three-dimensional human poses from a single input RGB image

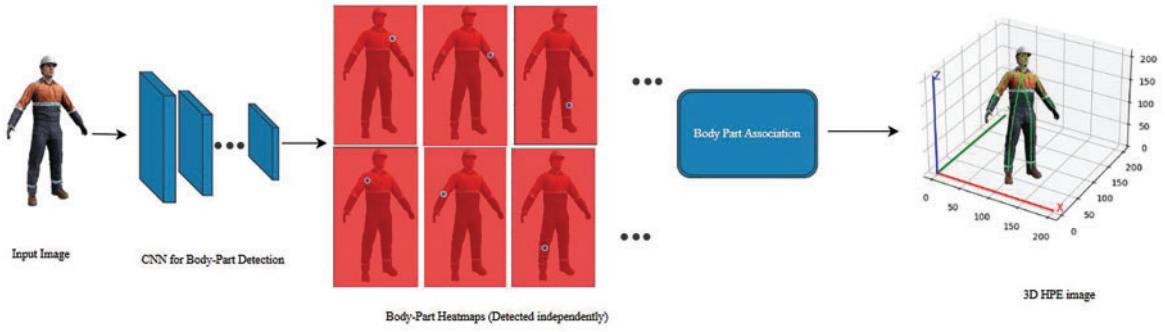


Figure 2: Detection-based Framework for three-dimensional Human Pose Estimation. Input RGB images are processed through a Convolutional Neural Network (CNN) for independent detection of multiple body-part heatmaps. These detected body parts are subsequently integrated using a Body Part Association step to construct an accurate three-dimensional pose representation, enabling detailed ergonomic analysis and reliable joint localization

Detection-Based

In 2014, Tompson et al. [33] presented a novel hybrid architecture that integrates a CNN with a Markov Random Field (MRF) to identify human body joints in complex environments. The heatmap illustrates the probability distribution of each joint's position, with higher intensity values indicating greater confidence in the joint location. In this method, the network learns to identify the likely locations of each key point by generating a series of heatmaps for each joint. CNNs are frequently employed to forecast these heatmaps by analysing image features and generating spatially structured predictions. This capability enables joint localization with remarkable precision. Ref. [34] propose a novel approach for human pose estimation from a single image using a deep learning-based voting scheme, whereas the traditional approach used the graphical model to enforce global pose consistency by relying on relative location statistics of keypoints. Wei et al. [35] present a novel approach to human pose estimation using a sequence of CNNs. This approach, called "Convolutional Pose Machines CPM", focuses on progressive refining pose predictions through multiple stages, where the CPMs employ a multi-stage framework where each stage predicts probability maps for the location of each body part, To address the challenge in pose estimation is capturing dependencies between distant body parts through employing large receptive fields within each stage of the CNNs, which allows for the integration of contextual cues from distant parts. Balut in [36] presents a cascaded CNN architecture for human pose estimation that uses part heatmaps followed by regression. This design enhances pose estimation, particularly in scenarios involving severe occlusion, using the part detection subnetwork, which produces heatmaps for each part using the sigmoid loss function after the regression network then takes these heatmaps as input and predicts confidence maps for the final joint locations then uses an L2 loss function for confidence map regression. For estimating human pose in images, Newell et al. [37] introduce a novel convolutional network architecture for estimating human pose in images. Their approach, known as the Stacked Hourglass Network, effectively captures and consolidates multi-scale spatial relationships associated with human body joints through repeated bottom-up and top-down processing. The core innovation of this work is the hourglass network design. This structure alternates between pooling layers down-sampling and up-sampling, allowing the network to process features across multiple scales and aggregate detailed and contextual information for precise localization of body joints, the outputs of network heatmaps, where each heatmap corresponds to the likelihood of a specific joint's location within the image. The use of heatmaps allows the network to retain spatial precision for each key-point, even in complex poses with occlusions. Pavlakos et al. [38] address the challenge of estimating three-dimensional human poses from a single RGB image by introducing an end-to-end approach that estimates three-dimensional poses directly through Volumetric Representation and Coarse-to-Fine, where they framed three-dimensional pose estimation as a key-point localization problem in a discretized three-dimensional space. Instead of directly predicting joint coordinates, they use a CNN to assign likelihoods to each voxel (three-dimensional pixel), creating a "volumetric heatmap" for each joint. This volumetric approach, which predicts the probability of each joint's location in three-dimensional space, allows for more accurate and stable joint predictions by considering spatial relationships within the volume. To handle the large dimensionality of the three-dimensional space, they employ a coarse-to-fine approach to a low-resolution prediction and progressively increase the resolution of the three-dimensional volume, focusing on refining the depth dimension. Sun et al. [39] address the challenges associated with heatmap-based human pose estimation in deep learning for two-dimensional and three-dimensional human pose prediction where the traditional heatmap methods generate likelihood maps for each joint and locate joints based on the maximum likelihood in these maps. While effective, this approach has limitations, such as being non-differentiable (hindering end-to-end training) and causing quantization errors due to lower-resolution heatmaps. The model unifies heatmap representation and regression. Instead of taking the maximum value from the heatmap, they

compute a weighted integral across all possible joint locations. This "soft-argmax" approach calculates the expected value, or weighted average, across the heatmap, providing continuous joint coordinates while retaining the heatmap's advantages. Integral regression makes the process differentiable, enabling end-to-end learning and avoiding quantization errors. To maintaining high-resolution representations throughout the network rather than relying on conventional low-to-high-resolution processes, Sun et al. [40] introduce the HRNet to maintain high-resolution representations throughout the network that retain high spatial resolution at each stage of processing, which improves key-point localization accuracy. Groos et al. [41] found that traditional models like OpenPose, while popular, require substantial computational resources and can struggle with precision in applications needing high accuracies, such as medical assessments and sports analysis; therefore, they leverage EfficientNet backbones to create a scalable family of models optimized for single-person pose estimation tasks. Tables 1 and 2 summarize detection-based methods for 3D human pose estimation, highlighting their performance across common datasets and the variation in results depending on the dataset and evaluation metrics used.

Table 1: Estimating three-dimensional human position from a Detection-based approach

Article	Methodology	Result
[33]	A hybrid model integrating CNN and MRF, concurrently trained for part detection and spatial consistency reinforcement.	Attained exceptional precision on the FLIC and LSP datasets, demonstrating resilience in the presence of occlusions and intricate poses.
[34]	Deep consensus voting scheme where pixels vote for body keypoints, aggregating data for keypoint prediction.	Competitive outcomes on the MPII and LSP datasets; robust performance in occlusion situations.
[35]	Multi-stage Convolutional Pose Machines (CPMs) refining pose predictions with intermediate supervision.	High accuracy on MPII and LSP datasets, surpassing previous methods and establishing a new benchmark.
[37]	Stacked Hourglass Networks for iterative refinement through downsampling and upsampling processes.	Cutting-edge precision on MPII and FLIC datasets, especially in challenging joint detection.
[36]	Cascaded CNN architecture using part heatmaps followed by regression for refinement.	Robust to occlusions; competitive Outcomes on the MPII and LSP datasets.
[38]	Coarse-to-fine volumetric approach generating three-dimensional voxel-based heatmaps refined iteratively.	Enhanced competitive performance on Human3.6M and HumanEva; increased precision of three-dimensional joint positioning.
[39]	Integral Pose Regression approach unifying heatmap representation and regression with a weighted integral for joint localization.	Cutting-edge outcomes on Human3.6M, MPII, and COCO with enhanced two-dimensional and three-dimensional pose precision.
[40]	HRNet preserves high-resolution representations through parallel high-to-low networks and multi-scale fusion.	Outperformed traditional models on COCO and MPII; effective in video tracking.

(Continued)

Table 1 (continued)

Article	Methodology	Result
[41]	Scalable architecture based on EfficientNet with MBConv layers and PAFs for efficient keypoint localization.	Higher efficiency and accuracy compared to OpenPose; strong performance on MPII.

Table 2: Performance method estimating 3D/2D human pose on detection-based

Article	Method Name	Dataset	Metric	Result
[33]	Joint CNN + Graphical model	FLIC, Leeds Sports Pose	PCK@0.2	FLIC: 95.0% (elbow), 90.8% (wrist); Leeds: 80.3% (PCK@0.2)
[34]	Deep consensus voting	MPII	PCKh@0.5	88.5% (PCKh@0.5)
[35]	Convolutional Pose Machines (CPM)	MPII, LSP	PCKh@0.5	MPII: 87.95% (PCKh@0.5); LSP: 84.32% (PCK@0.2)
[36]	Convolutional part heatmap regression	MPII	PCKh@0.5	89.7% (PCKh@0.5)
[37]	Stacked hourglass networks	MPII, FLIC, LSP	PCKh@0.5, PCK@0.2	MPII: 90.9% (PCKh@0.5); FLIC: 98.2% (elbow), 96.3% (wrist); LSP: 90.9% (PCK@0.2)
[38]	Coarse-to-fine volumetric prediction	Human3.6M	MPJPE (3D)	51.9 mm (direct 3D prediction)
[39]	Integral human pose regression	Human3.6M	MPJPE (3D)	40.5 mm (with ground truth two-dimensional poses)
[40]	Deep high-resolution representation	COCO, MPII, Human3.6M	AP (COCO), PCKh@0.5 (MPII), MPJPE (Human3.6M)	COCO: 76.3% (AP); MPII: 92.3% (PCKh@0.5); Human3.6M: 34.5 mm (MPJPE)
[41]	EfficientPose	Human3.6M	MPJPE (3D)	–

In summary, Heatmap-based human pose estimation methods offer substantial advantages in terms of spatial accuracy and robustness in joint detection. Their capability to represent joint likelihoods as spatial distributions has driven significant progress in both two-dimensional and three-dimensional posture estimation concerns. However, challenges such as computational demands, depth ambiguity, and handling occluded joints continue to prompt the development of innovative approaches. Techniques like integral

regression and multi-stage refinement architectures represent steps forward in addressing these issues, emphasizing the ongoing need for balance between accuracy, computational efficiency, and adaptability to real-world complexities.

Regression-Based Approach

Single-stage three-dimensional pose estimation models using a regression-based approach directly map two-dimensional image inputs to three-dimensional coordinates, bypassing complex, multi-step processes. This section traces the evolution of these methods, from early foundational models to recent innovations, highlighting their unique methodologies and improvements in robustness, efficiency, and real-world applicability.

Discriminative and Shape-Based Models: The foundational models for single-stage three-dimensional pose estimation often relied on shape-based methods and discriminative techniques. Mori and Malik [42] introduced shape contexts to recover body configurations by matching shapes derived from edge maps in two-dimensional images. While effective in controlled settings, this approach struggled with occlusions and background complexity. Expanding on this idea, Agarwal and Triggs proposed [43] where silhouettes serve as features for pose estimation via relevance vector regression. However, both methods are limited by their reliance on clean edge or silhouette extraction, making them less robust in naturalistic settings. Histogram of Oriented Gradients (HOG) provided another discriminative approach, with Kovashka et al. employing HOG-based classifiers to recognize body parts [30]. Later, Fang et al. applied HOG features in [44] using gradients to infer joint positions. Although computationally efficient, HOG-based methods are sensitive to lighting and edge clarity, limiting their utility in dynamic environments.

Integrating Context and Probabilistic Models: Probabilistic models have been introduced to address the ambiguity in monocular three-dimensional pose estimation. In [45], Jung et al. developed a probabilistic diffusion model that generates multiple hypotheses for joint locations, reducing errors from ambiguous poses. Similarly, Li et al. in [46] leveraged semantic relationships between joints via graph convolutional networks (GCNs), improving the structural accuracy of pose estimations. However, these models are computationally intensive, presenting challenges for real-time applications. In addition, C. Ionescu et al. introduced figure-ground segmentation in [47] to isolate subjects from backgrounds, enhancing pose accuracy. While segmentation improves focus on the subject, it is susceptible to background noise, especially in crowded scenes.

Physics-Based and Kinematic Models: To ensure realistic and anatomically feasible poses, physically-aware and kinematic models introduced physics-based constraints. Ma et al. [48] employed an intuitive physics framework to enforce balance and spatial accuracy. By embedding physical constraints into pose prediction, this model effectively reduces depth errors. Another advancement is seen in Rhodin et al. [49], where joint angles and bone lengths are constrained to maintain plausible body positions. However, such models face limitations in dynamic environments where assumptions about physical constraints may not always apply. Kinematic constraints further ensure realistic limb and joint angles. For instance, Ionescu et al. in [47] applied these constraints to maintain anatomical consistency across various poses. While kinematic constraints enhance pose realism, their rigidity can reduce adaptability to non-standard poses and movements.

Regression-Based CNN and Baseline Models: Convolutional Neural Network (CNN) models marked a turning point for single-stage, regression-based pose estimation by enabling efficient mapping from two-dimensional images to three-dimensional coordinates. Zhou et al. [50] developed one of the first deep CNNs that lifts two-dimensional pose features into three-dimensional space. Zhao et al. later introduced [29] simplified three-dimensional estimation using orthographic projections. These models, while efficient, are

often sensitive to occlusions and variations in body orientation. Xiao and Wei [51] introduced an effective baseline model that uses deconvolutional layers to refine pose predictions directly from two-dimensional features, improving accuracy without complex architectures. Although highly efficient, this baseline does not explicitly handle occlusions or dynamic, real-world settings.

Advanced Integrated Feature-Sharing Models: On the other hand, advanced integrated feature-sharing models are another phase in single-stage models that involve integrated feature-sharing between two-dimensional and three-dimensional tasks. Sun et al. [39] introduced a differentiable integral operation to calculate joint coordinates directly from heatmaps, integrating two-dimensional and three-dimensional learning for more accurate pose estimation. This model, trained on Human3.6M, MPII, and COCO, achieves high precision but can struggle with occlusions. Similarly, Mehta et al. in [22] introduced a transfer learning approach where two-dimensional features from large in-the-wild datasets are shared with the three-dimensional model, enhancing generalization across diverse scenes. Trumble et al.'s introduced weakly supervised model in [28] employs geometric constraints to align two-dimensional and three-dimensional learning for challenging environments, though it remains less effective in low-quality images.

Bone-Based and Context Modeling: Bone-based representations have been introduced to improve anatomical consistency. Wang et al. in [32] developed a model that represents pose based on bone vectors rather than joint coordinates, leading to structurally accurate predictions. Although effective in ensuring anatomical accuracy, this representation can limit adaptability to extreme poses. In recent advancements, Ma et al. proposed a context modeling framework [52] by combining Pictorial Structure Models (PSM) and Graph Neural Networks (GNN). This model leverages contextual information from neighboring joints, improving pose accuracy across complex scenes like Human3.6M and MPI-INF-3DHP. However, its reliance on extensive training data can limit adaptability to new datasets or unlabeled scenarios. Tables 3 and 4 provide a comparative overview of regression-based approaches used in monocular 3D human pose estimation, presenting benchmark results across multiple datasets.

Table 3: Estimating 3D human position from a Regression-based approach

Article	Explanation	Dataset	Metric
[28]	Weakly-supervised model for assessing three-dimensional human posture in outdoor environments	Human3.6M, MPI-INF-3DHP	Mean Per Joint Position Error (MPJPE)
[53]	Structured learning approach with max-margin objectives for three-dimensional pose	Human3.6M, CMU Panoptic	MPJPE, Percentage of Correct Keypoints (PCK)
[29]	Uses orthographic projection to simplify single-image three-dimensional pose estimation	Human3.6M, MPII	MPJPE
[54]	CNN model trained to predict three-dimensional pose based on learned camera viewpoint	Human3.6M	MPJPE, PCK
[54]	Deep CNNs design for three-dimensional pose estimation with two-dimensional picture features	MPII, Human3.6M	MPJPE
[38]	Volumetric representation method for pose prediction from a singular image	Human3.6M	MPJPE, PCK

(Continued)

Table 3 (continued)

Article	Explanation	Dataset	Metric
[55]	A novel approach using human pose as a calibration pattern for camera setup	Custom, MoCap	Root Mean Square Error (RMSE)
[56]	A probabilistic model for multi-view three-dimensional pose estimation without camera calibration	Human3.6M, CMU Panoptic	MPJPE
[57]	Multi-view framework for three-dimensional pose estimation of multiple people simultaneously	Shelf, Campus, CMU Panoptic	MPJPE
[58]	Self-supervised model that adapts based on prediction uncertainty for three-dimensional pose	Human3.6M, MPI-INF-3DHP	MPJPE
[59]	Pyramid network designed for efficient, real-time three-dimensional pose regression	Human3.6M, MPI-INF-3DHP	MPJPE
[49]	A physically aware neural model for robust three-dimensional motion capture from monocular input	Human3.6M, MPI-INF-3DHP	MPJPE, Physio metrics
[48]	Uses physics-based constraints to ensure plausible three-dimensional poses	Human3.6M, Custom MoYo Dataset	MPJPE, Base of Support (BoS) Error
[30]	HOG-based classifier for detecting and recognizing body parts	Leeds Sports Pose (LSP)	Classification accuracy
[44]	HOG-based model for single-image three-dimensional posture estimation	Custom Dataset	Root Mean Square (RMS) error
[43]	RVM-based model for three-dimensional pose from monocular silhouettes	Synthetic Dataset	Angular error
[46]	GCN model leveraging semantic relationships for structured pose regression	Human3.6M	MPJPE
[55]	Uses anatomical and camera knowledge to improve three-dimensional pose estimation	Human3.6M	MPJPE, joint length consistency
[45]	Probabilistic diffusion model with joint-level aggregation for three-dimensional pose	Human3.6M, MPI-INF-3DHP	MPJPE
[42]	Shape context matching for recovering body configurations from a single image	Speed Skating, Custom Dataset	Keypoint accuracy
[60]	Shape context-based model for estimating three-dimensional body pose configurations	Custom Dataset	Joint localization error

(Continued)

Table 3 (continued)

Article	Explanation	Dataset	Metric
[47]	Latent model for three-dimensional pose estimation using figure-ground segmentation	HumanEva-I	MPJPE, Mean Per Joint Angle Error (MPJAE)
[20]	Comprehensive dataset for three-dimensional human sensing with annotated pose data	Human3.6M	MPJPE, Mean Per Joint Localization Error (MPJLE)
[32]	Bone-based pose representation for structural consistency in pose estimation	Human3.6M, MPII	MPJPE, PCK
[51]	Simple, effective baseline model using a deconvolutional network for precise pose tracking	COCO, PoseTrack	Mean Average Precision (mAP), Multi-Object Tracking Accuracy (MOTA)
[39]	Unified regression approach integrating two-dimensional and three-dimensional pose estimation using an integral operation over heatmaps	Human3.6M, MPII, COCO	MPJPE, PCK, mAP
[22]	Improved generalization for in-the-wild three-dimensional pose estimation using transfer learning	Human3.6M, MPI-INF-3DHP, LSP	MPJPE, 3DPCK, AUC
[61]	A weakly supervised model with geometric constraints and re-projection for 3D pose lifting from two-dimensional	Human3.6M, MPII, MPI-INF-3DHP	MPJPE, PCK, AUC

Table 4: Performance method estimating three-dimensional/two-dimensional human pose on regression-based approach

Article	Method Name	Dataset	Metric	Result
[47]	Latent structured models	HumanEva	MPJPE	HumanEva: 60.9 mm (MPJPE)
[62]	Maximum-margin structured learning	HumanEva	MPJPE	HumanEva: 52.8 mm (MPJPE)
[54]	Camera viewpoint CNN	Human3.6M	MPJPE	Human3.6M: 54.3 mm (MPJPE)
[50]	Lifting from the deep	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 62.0 mm (MPJPE), 47.7 mm (PA-MPJPE); MPI-INF-3DHP: 88.4 mm (MPJPE)

(Continued)

Table 4 (continued)

Article	Method Name	Dataset	Metric	Result
[46]	Semantic graph convolutional networks	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 50.1 mm (MPJPE), 37.9 mm (PA-MPJPE); MPI-INF-3DHP: 74.2 mm (MPJPE)
[51]	Simple baselines	COCO, MPII	AP (COCO), PCKh@0.5 (MPII)	COCO: 73.7% (AP); MPII: 90.5% (PCKh@0.5)
[63]	Weakly supervised three-dimensional pose estimation	Human3.6M	MPJPE	Human3.6M: 58.4 mm (MPJPE)
[52]	Context modeling in three-dimensional pose estimation	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 45.2 mm (MPJPE), 33.1 mm (PA-MPJPE); MPI-INF-3DHP: 65.4 mm (MPJPE)
[58]	Uncertainty-aware adaptation	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 43.5 mm (MPJPE), 31.8 mm (PA-MPJPE); MPI-INF-3DHP: 63.2 mm (MPJPE)
[45]	Diffusion-based three-dimensional pose estimation	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 46.2 mm (MPJPE), 34.1 mm (PA-MPJPE); MPI-INF-3DHP: 67.3 mm (MPJPE)
[48]	Intuitive physics-based three-dimensional pose estimation	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 48.7 mm (MPJPE), 35.6 mm (PA-MPJPE); MPI-INF-3DHP: 69.8 mm (MPJPE)
[56]	Probabilistic triangulation	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 44.8 mm (MPJPE), 32.9 mm (PA-MPJPE); MPI-INF-3DHP: 64.7 mm (MPJPE)

(Continued)

Table 4 (continued)

Article	Method Name	Dataset	Metric	Result
[59]	SSP-Net	Human3.6M, MPI-INF-3DHP	MPJPE, PA-MPJPE	Human3.6M: 42.7 mm (MPJPE), 31.2 mm (PA-MPJPE); MPI-INF-3DHP: 62.5 mm (MPJPE)

7.1.2 Two-Stage Pipeline

Two-stage pipelines have become a predominant methodology for monocular three-dimensional human pose estimation. This paradigm divides the process into two main stages: (1) the estimation of two-dimensional human poses and (2) lifting these two-dimensional poses to three-dimensional space. Stage one, we explored previously using the regression approach and heatmap approach where the second stage benefits from leveraging the accuracy of state-of-the-art two-dimensional pose detectors, enabling a more refined three-dimensional reconstruction process. Fig. 3 presents a structured two-stage deep learning approach for estimating three-dimensional human poses from single-view two-dimensional images. In the first stage (two-dimensional Keypoint Extraction), an input image undergoes feature extraction using Convolutional Neural Networks (CNNs) to identify and represent the human joints as two-dimensional keypoints. Subsequently, these extracted two-dimensional keypoints serve as input for the second stage (three-dimensional Keypoint Estimation), which employs an Autoencoder neural network composed of an encoder-decoder structure. The encoder transforms the two-dimensional keypoints into a compressed latent space representation, capturing essential spatial relationships, while the decoder reconstructs this representation into estimated three-dimensional keypoints. The final output provides a precise three-dimensional human pose estimation, facilitating accurate, ergonomic assessments and detailed pose analysis in diverse practical scenarios.

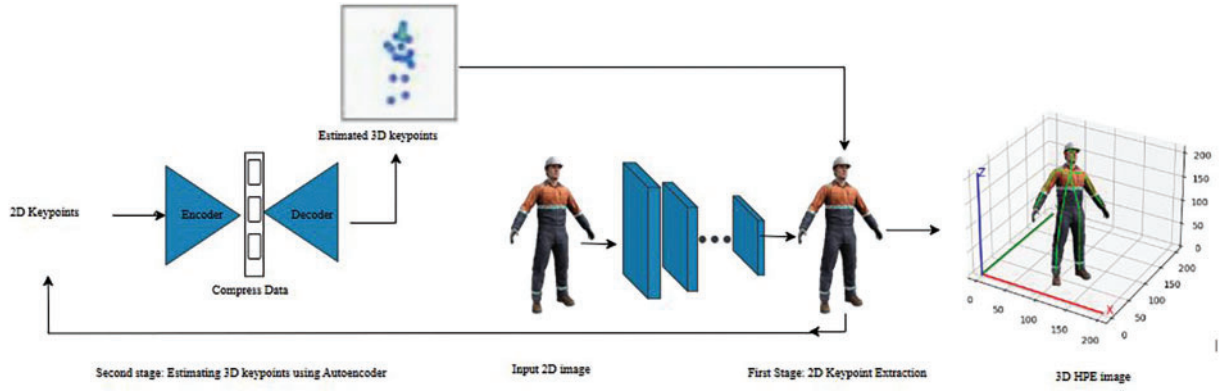


Figure 3: The first stage extracts two-dimensional human keypoints from a single RGB image using a convolutional neural network (CNN). These 2D keypoints are then input into a neural network-based encoder-decoder architecture in the second stage to regress the corresponding three-dimensional joint positions

Classification problem/discriminative: In contrast to modern techniques employing deep neural networks to discern the relationship between two-dimensional and three-dimensional poses, previous

studies concentrated on translating two-dimensional poses into three-dimensional by determining the best suitable three-dimensional configuration that aligns with the two-dimensional observations. These methods often utilized a comprehensive dictionary of three-dimensional poses derived from large three-dimensional pose datasets through techniques such as Principal Component Analysis (PCA) or dictionary learning methods. For instance, early approaches like those presented in [64] leveraged sparse representations of three-dimensional poses with an over-complete dictionary. Used projected matching pursuit algorithms to recover three-dimensional poses from only two-dimensional projections and designed loss functions to optimize dictionary coefficients and joint speed [57]. Later works, such as those referenced in [56] introduced probabilistic triangulation techniques for robust multi-view three-dimensional pose estimation focusing on handling errors due to camera calibration inconsistencies by employing convex optimization techniques and expectation maximization algorithms to jointly estimate sparse representation coefficients. Some methods, such as [65] combined two independent training datasets to improve three-dimensional pose generalization. Used a dual-source matching algorithm to minimize projection errors under constraints from estimated two-dimensional poses this founded strategies like matching the depth of two-dimensional poses to their three-dimensional counterparts using k-nearest neighbor algorithms.

In contrast, the approach detailed in [66] introduced a consensus-based optimization algorithm for uncalibrated multi-view data. Used a single monocular training process to fuse multi-view predictions for robust three-dimensional pose estimation which coordinates minimized projection errors under constraints that ensured proximity to retrieved poses, further refining pose reconstruction accuracy. In classification task where images were assigned to predefined pose classes. For example, Ref. [67] used classification over pose classes for three-dimensional reconstruction, ensuring valid predictions. However, these approaches were limited by the granularity of pose classes, as increasing the number of classes improved precision but also complicated discrimination tasks. These models laid foundational techniques for three-dimensional pose estimation but were gradually complemented and, in some cases, replaced by neural network-based methods that directly optimize correlations between two-dimensional and three-dimensional key points for improved accuracy and flexibility.

2D-to-3D lifting techniques: The evolution of two-dimensional-to-three-dimensional lifting techniques has progressed significantly, transitioning from traditional dictionary-based methods to sophisticated deep learning approaches. Each innovation in this field addresses specific challenges such as depth ambiguity, occlusions, computational efficiency, and the accuracy of intermediate representations. Below, we delve into the advancements, focusing on deep learning models, the kinematic model, self-supervised and weakly supervised methods, Graph Convolutional Networks (GCNs), and transformer networks for three-dimensional reconstruction.

Moreno-Noguer [68] introduced a framework for three-dimensional human pose estimation from a single image, utilizing Euclidean Distance Matrices (EDMs) to encode structural relationships between joints. The approach first detects two-dimensional joint positions using CNN-based detectors and then regresses three-dimensional poses through EDMs. Martinez et al. in [69] introduced fully connected residual networks, which regress three-dimensional joint locations from two-dimensional key points through a stacked hourglass network. This method highlighted the potential of neural networks in lifting two-dimensional to three-dimensional poses by leveraging robust intermediate two-dimensional pose detectors. To enhance this approach, Tekin et al. in their model include the confidence map stream, image stream, and a fusion stream, utilizing two-dimensional heatmaps instead of raw two-dimensional key-points. Heatmaps encode spatial distributions and improve robustness by capturing additional spatial context. Despite achieving state-of-the-art results at the time, it suffered from reconstruction ambiguities, especially when input two-dimensional poses were noisy or occluded. To enhance this approach, their model includes

the confidence map stream, image stream, and a fusion stream, utilizing two-dimensional heatmaps instead of raw two-dimensional key-points. Heatmaps encode spatial distributions and improve robustness by capturing additional spatial context. Wang et al. in [70] addressed depth ambiguity and proposed a ranking-based network that incorporates depth matrices into the lifting process. This model was more effective than standard regression techniques. Zhou et al. [71] introduce a new method that leverages Part-Centric Heatmap Triplets (HEMlets), which encode both two-dimensional joint positions and relative depth relationships of skeletal joints to enhance pose accuracy in complex scenes with a 20% improvement over existing methods. Wu et al. [72] introduced a Locally Connected Mixture Density Network (LCMDN) to improve feature extraction by leveraging structural relationships between joints.

Graph Convolutional Networks (GCNs): Have been pivotal in refining two-dimensional-to-three-dimensional lifting by representing the human skeleton as a graph, with joints as nodes and bones as edges. This framework allows models to exploit the structural relationships within the human body, improving pose estimation robustness under diverse conditions. Choi et al. [73] introduced the Pose2Mesh model to directly recover three-dimensional human pose and mesh from two-dimensional poses using PoseNet and MeshNet. Zhao et al. in [46] introduced on Semantic-GCN (SemGCN) regression base; the model processes the input as a structured graph where nodes represent joints and edges capture their spatial relationships. Training employs a combination of joint and bone constraints, ensuring physical plausibility. Zhao et al. in [74] integrates graph convolutions with transformers. This hybrid approach captures both explicit joint relationships and hidden dependencies, enhancing pose estimation for occluded or ambiguous configurations. Another model by Zhou et al. proposed a Modulated GCN in [75] identifies two critical limitations in existing GCNs for three-dimensional Human Pose Estimation with weight sharing and Suboptimal Graph Structures for that introduced weight modulation and affinity modulation to dynamically adapt the graph structure, enabling more accurate feature transformations across diverse pose configurations. Choi et al. [76] introduced the DiffuPose model to leverage a Denoising Diffusion Probabilistic Model (DDPM) for generating multiple plausible three-dimensional human pose hypotheses from a single two-dimensional pose input by Multi-Hypothesis and Combining a reconstruction Loss Function for denoising and an auxiliary two-dimensional supervision loss to ensure consistency with two-dimensional input.

The kinematic model: Leverages skeletal joint connectivity, fixed bone-length ratios, and joint rotation properties to ensure plausible pose estimation. For that, Zhou et al. [77] used their model from single-stage embedded kinematic constraints as layers in their network to enforce anatomical consistency and better geometric accuracy. Xu et al. in [78] systematically integrate kinematics-based constraints into a deep learning framework, addressing noise in two-dimensional detections and improving the reliability and accuracy of monocular three-dimensional pose estimation through enforcing perspective projection constraints and introducing temporal CNN to improve the temporal consistency of two-dimensional inputs.

To tackle the scarcity of labeled three-dimensional data, self-supervised and weakly supervised approaches have emerged; Zhou et al. in [28] their weakly supervised transfer learning framework, utilized two-dimensional annotations from in-the-wild images as weak labels to infer three-dimensional poses. This method included a three-dimensional pose estimation module connected to the two-dimensional pose estimation network, enabling joint optimization. Their weakly supervised transfer learning framework utilized two-dimensional annotations from in-the-wild images as weak labels to infer three-dimensional poses. This method included a three-dimensional pose estimation module connected to the two-dimensional pose estimation network, enabling joint optimization. Habibie et al. [79] tailored a projection loss to refine three-dimensional human poses without requiring explicit three-dimensional annotations, relying on two-dimensional-three-dimensional reprojection consistency. Chen et al. [80] introduced a geometric self-consistency loss, leveraging lift-reproject-lift properties to ensure that re-lifted poses remain consistent

with the original three-dimensional structure. Luvizon et al. [66] focus on the challenges of absolute three-dimensional human pose estimation in camera coordinates. The proposed method introduces two key contributions: View Frustum Space Pose Estimation, which predicts three-dimensional poses by separating absolute and relative depth estimation in the view frustum and converts predictions into camera coordinates using inverse projection, and Consensus-Based Multi-View Optimization, which aligns three-dimensional pose predictions from uncalibrated views by optimizing camera parameters. By combining multi-view predictions, the method improves accuracy and resolves occlusions, leveraging both two-dimensional and three-dimensional annotations for effective training. Wandt et al. in [81] proposed single-stage model that the proposed approach introduces key innovations for unsupervised three-dimensional pose estimation. It predicts the camera elevation angle for each two-dimensional pose, learning the distribution of camera orientations rather than relying on predefined priors, ensuring upright and realistic three-dimensional pose alignment. Normalizing flows are used to model the distribution of plausible two-dimensional poses and estimate the likelihood of reconstructed three-dimensional poses, offering a probabilistic measure of pose validity and improving accuracy. To stabilize training, two-dimensional poses are projected to a lower-dimensional subspace using Principal Component Analysis (PCA), reducing noise and redundancy and ensuring the effective training of the lifting network. The end-to-end pipeline integrates depth prediction for three-dimensional pose reconstruction, with loss functions combining pose likelihood, bone length consistency, and reprojection errors for optimization. CameraPose model introduces a novel weakly-supervised framework for monocular three-dimensional human pose estimation [63], addressing key challenges such as reliance on expensive three-dimensional annotations, poor generalization to in-the-wild poses, and the propagation of errors from noisy two-dimensional keypoints. By incorporating a camera parameter branch, the method effectively aligns two-dimensional keypoints with the three-dimensional camera space, enabling the use of unpaired two-dimensional datasets for training. A refinement network improves the accuracy of detected two-dimensional keypoints using confidence-guided loss, while a GAN-based pose generator and discriminator enhance pose diversity through realistic augmentation.

7.2 Limitations and Comparisons

The evolution of three-dimensional human pose estimation methods showcases diverse approaches, each with unique advantages and trade-offs. Two-stage pipelines, which separate two-dimensional keypoint detection and three-dimensional lifting, benefit from modularity and the accuracy of state-of-the-art two-dimensional pose detectors but suffer from error propagation between stages. Dictionary-based methods, though interpretable and computationally efficient for small-scale problems, struggle with generalizing to complex, in-the-wild poses due to their reliance on pre-learned three-dimensional dictionaries. Deep learning-based lifting techniques, such as those leveraging Euclidean Distance Matrices (EDMs) and heatmap triplets, provide scalability and robustness to noisy inputs but demand large labelled datasets and high computational resources. Graph Convolutional Networks (GCNs) capitalize on the skeletal structure's inherent graph properties to model joint dependencies, making them robust to occlusions, though they can be computationally intensive and rely heavily on optimal graph designs. Kinematic models ensure anatomical consistency and realistic poses through constraints like fixed bone-length ratios, but they introduce complexity and require prior anatomical knowledge.

Meanwhile, self-supervised and weakly supervised methods, such as ElePose and CameraPose, address the scarcity of three-dimensional annotations and generalize well to in-the-wild scenarios by leveraging two-dimensional datasets and reprojection losses. Yet, they often underperform compared to fully supervised techniques in controlled environments. Lastly, probabilistic and multi-hypothesis models like DiffuPose

effectively handle depth ambiguities and occlusions by generating plausible pose distributions but are computationally expensive and require advanced optimization strategies. These methods collectively highlight the trade-offs between accuracy, efficiency, generalization, and computational cost, enabling tailored choices for different application requirements.

8 REBA (Rapid Entire Body Assessment)

The field of ergonomic risk assessment has evolved from manual methods to automated, AI-driven systems leveraging computer vision and deep learning. Early work by Abdollahzade and Mohammadi [82] relied on statistical analysis and manual observations to predict hospital REBA scores, establishing a baseline for risk quantification. Chen and Qiu [83] introduced tensor decomposition and SVM classifiers to identify awkward postures in construction, though without validation on a standardized dataset.

Subsequent advances addressed methodological gaps: Erginel and Toptanci [11] applied fuzzy logic to handle uncertainty in posture assessments, while Anghel et al. [84] combined biomechanical load analysis with RULA to optimize automotive assembly lines. Seo and Lee [85] achieved 89% accuracy in supervised vision-based posture classification, demonstrating the potential of automated systems.

By 2020, machine learning dominated the field. Ryu et al. [86] used unsupervised clustering to categorize ergonomic risks, and Ghasemi and Mahdavi [13] enhanced REBA's sensitivity by integrating fuzzy sets and Bayesian networks (FBnREBA). Vision-based tools emerged as practical solutions: Wu et al. [12] developed a Mask R-CNN app (~90% accuracy on COCO images), and Jeong and Kook [87] created CREBAS, a MediaPipe-based REBA system matching manual evaluations. Yan et al. [88] achieved 89% accuracy in two-dimensional posture recognition using OWAS.

Post-2020, deep learning refined automation. Tellaeche et al. [89] automated REBA with variational deep networks (UW-IOM/TUM datasets), while Lin et al. [90] integrated OpenPose with REBA/RULA/OWAS for real-time analysis. Massiris Fernández et al. [91] automated RULA scoring (Cohen's kappa > 0.6), and Kim et al. [6] showed OpenPose outperformed Kinect in joint angle computation.

Recent studies focused on scalability and precision. Namwongsa et al. [92] linked smartphone use to musculoskeletal risks via RULA. Li et al. [93] proposed a deep learning RULA model (Human3.6M dataset, 96.7% accuracy), and Kumar Nayak and Kim [94] achieved high consistency (ICC > 0.75) in automated RULA. Abobakr et al. [95] used RGB-D data (89% accuracy, 3.19° MAE), and Hossain et al. [7] predicted REBA scores from three-dimensional keypoints (89.07% accuracy). Finally, Ionescu et al. [47] advanced monocular pose estimation with latent structured models, paving the way for future ergonomic tools. Table 5 presents a summary of ergonomic risk assessment studies that utilize computer vision techniques, highlighting their methodologies, environments, and evaluation settings.

Table 5: Summary of ergonomic risk assessment studies using computer vision

Article	CV Tool	Environment	Methodology	Dataset	2D/3D
[82]	–	Hospital	Manual statistical analysis of REBA scores	Hospital worker observations	–
[83]	–	Construction site	Tensor decomposition + SVM for posture classification	Unspecified construction site data	3D

(Continued)

Table 5 (continued)

Article	CV Tool	Environment	Methodology	Dataset	2D/3D
[11]	–	Laboratory	Fuzzy logic for posture uncertainty handling	Synthetic/manual posture assessments	2D
[84]	CATIA	Automotive assembly	Biomechanical load analysis + RULA	CAD simulations (CATIA)	3D
[85]	Supervised vision	Construction site	Vision-based posture classification (89% accuracy)	Construction site videos	3D
[86]	–	Laboratory	Machine learning clustering for risk categorization	Unspecified lab data	3D
[13]	–	Industrial	FBnREBA: Fuzzy sets + Bayesian networks for REBA sensitivity	Case study (gas cooker assembly line)	2D
[12]	Mask R-CNN	Laboratory	REBA scoring via Mask R-CNN (~90% accuracy)	MS COCO 2017 dataset	2D
[87]	MediaPipe	Laboratory	CREBAS: Automated REBA using MediaPipe	MediaPipe pose estimates	2D
[88]	Two-dimensional vision	Laboratory	OWAS-based posture recognition (89% accuracy)	Unspecified two-dimensional pose data	3D
[89]	Variational DNN	Laboratory	Deep networks for automated REBA (UW-IOM/TUM datasets)	UW-IOM, TUM Kitchen datasets	3D
[90]	OpenPose	Laboratory	OpenPose + REBA/RULA/OWAS for real-time analysis	OpenPose outputs	3D
[91]	Computer vision	Laboratory	Automated RULA scoring (Cohen's kappa > 0.6)	Unspecified lab recordings	2D
[6]	OpenPose	Laboratory	OpenPose-based joint angle computation (outperformed Kinect)	OpenPose outputs	3D
[92]	–	Laboratory	RULA for smartphone ergonomics	User smartphone usage data	–
[93]	Deep learning	Laboratory	Deep learning RULA (Human3.6M, 96.7% accuracy, MAE 3.33%)	Human3.6M dataset	3D
[94]	Vision-based	Laboratory	Fully automated RULA (ICC > 0.75)	MS COCO 2017 dataset	2D
[95]	RGB-D + CNN	Laboratory	RGB-D ergonomic assessment (89% accuracy, 3.19° MAE)	RGB-D sensor data	3D

(Continued)

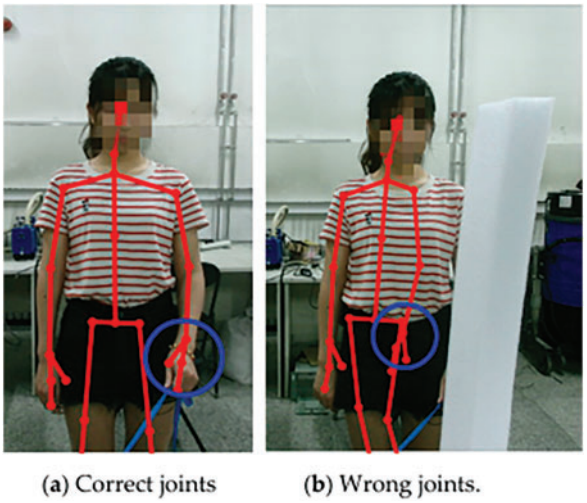
Table 5 (continued)

Article	CV Tool	Environment	Methodology	Dataset	2D/3D
[7]	three-dimensional keypoints	Laboratory	REBA prediction from three-dimensional keypoints (89.07% accuracy)	Human3.6M dataset	3D

Limitations and Comparisons

A critical examination of recent advancements in automated ergonomic risk assessment reveals significant gaps concerning the integration of 3D Human Pose Estimation (3D HPE) within REBA frameworks, particularly regarding dataset diversity and methodological robustness. Although numerous studies—such as those by Anghel et al. [84], Wu et al. [12], and Hossain et al. [7], have achieved high accuracy using standardized or synthetic datasets like Human3.6M, COCO, or CATIA simulations, their findings remain constrained by limited real-world applicability. The absence of cross-dataset validation and the narrow representation of occupational postures hinder the generalizability of these models across dynamic workplace environments.

The application of 3D HPE to REBA introduces unique challenges that further complicate its practical deployment. Occlusions and partial visibility, frequently caused by tools, machinery, or body self-occlusion, often lead to incomplete or inaccurate joint detection, directly impacting the calculation of REBA-relevant angles such as trunk flexion or leg positioning. Similarly, complex postures involving twisting in synthetic datasets, leading to posture ambiguity and misclassification of risk scores (see Fig. 4). Moreover, the widespread reliance on monocular camera views in laboratory studies restricts the ability of 3D reconstruction methods to capture full-body kinematics necessary for accurate ergonomic assessments in real-world settings.



(a) Correct joints (b) Wrong joints.

Figure 4: Challenges in 3D HPE for REBA. In subfigure **(a)**, labelled "Correct joints", the subject is fully visible with no obstructions. In subfigure **(b)**, labelled "Wrong joints", a large object partially occludes the subject's right arm [96]

Limitations in current 3D HPE methodologies are also evident in their scalability and lack of multi-perspective training. While methods like variational deep networks [89] and OpenPose-based pipelines [90] report high accuracy, their validation has been limited to single datasets under controlled conditions. This hinders their adaptability to diverse and complex workplace tasks that require robust, multi-angle, and occlusion-resilient pose estimation.

Moving forward, improving REBA automation will require the integration of multi-source datasets combining real-world occupational motion data with lab-captured ground truth annotations. Developing 3D HPE models that are specifically tailored to handle occlusion, asymmetry, and variable viewing conditions will be critical for accurate ergonomic risk classification. These efforts will bridge current gaps and enable robust deployment of 3D pose-based ergonomic assessment systems in practical occupational settings.

9 Conclusion

This review has examined recent developments in monocular 3D human pose estimation and their integration into REBA-based ergonomic risk assessment. While advancements in deep learning—particularly CNNs, GCNs, and transformer-based models—have significantly improved the accuracy and applicability of 3D pose estimation from monocular images, substantial challenges remain. These include limitations in generalizability across diverse workplace scenarios, the scarcity of annotated real-world datasets, and difficulties in handling occlusion and skeleton variability.

We highlighted that most existing approaches rely on controlled datasets and often overlook the complexity of real-world ergonomic applications. The review also emphasized a major research gap: the limited capability of current models to estimate REBA scores directly from single RGB images without depth sensors or sequential inputs.

Future research should prioritize the development of lightweight, generalizable models capable of functioning effectively in unconstrained environments. Incorporating synthetic datasets, domain adaptation techniques, and multi-view learning could further enhance model robustness. By addressing these challenges, researchers can advance the deployment of automated ergonomic risk assessment tools that are scalable, accurate, and practical for use in occupational health and safety.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design: Ahmad Mwfaq Bataineh. Data collection: Ahmad Mwfaq Bataineh, Ahmad Sufril Azlan Mohamed. Analysis and interpretation of results: Ahmad Mwfaq Bataineh, Ahmad Sufril Azlan Mohamed. Draft manuscript preparation: Ahmad Mwfaq Bataineh. All authors reviewed the results and approved the final version of manuscript.

Availability of Data and Materials: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Gamra MB, Akhloufi MA. A review of deep learning techniques for 2D and 3D human pose estimation. Image Vis Comput. 2021;114:104282. doi:10.1016/j.imavis.2021.104282.
2. Nguyen TD, Kresovic M. A survey of top-down approaches for human pose estimation. arXiv:2202.02656. 2022.
3. Perez-Sala X, Escalera S, Angulo C, Gonzàlez J. A survey on model based approaches for 2D and 3D visual human pose recovery. Sensors. 2014;14(3):4189–210. doi:10.3390/s140304189.
4. Chen Y, Tian Y, He M. Monocular human pose estimation: a survey of deep learning-based methods. Comput Vis Image Underst. 2020;192:1–23. doi:10.1016/j.cviu.2019.102897.
5. Dang Q, Yin J, Wang B, Zheng W. Deep learning based 2D human pose estimation: a survey. Tsinghua Sci Technol. 2019;24(6):663–76. doi:10.26599/tst.2018.9010100.
6. Abobakr A, Nahavandi D, Iskander J, Hossny M, Nahavandi S, Smets M. A kinect-based workplace postural analysis system using deep residual networks. In: Proceedings of the 2017 IEEE International Systems Engineering Symposium (ISSE); 2017 Oct 11–13; Vienna, Austria. p. 1–6.
7. Hossain MS, Azam S, Karim A, Montaha S, Quadir R, De Boer F, et al. Ergonomic risk prediction for awkward postures from 3D keypoints using deep learning. IEEE Access. 2023;11:114497–508. doi:10.1109/access.2023.3324659.
8. Agostinelli T, Generosi A, Ceccacci S, Mengoni M. Validation of computer vision-based ergonomic risk assessment tools for real manufacturing environments. Sci Rep. 2024;14(1):1–19. doi:10.1038/s41598-024-79373-4.
9. Beheshti M, Firoozi Chahak A, Alinaghi Langari A, Poursadeghiyan M. Risk assessment of musculoskeletal disorders by OVAKO Working posture Analysis System OWAS and evaluate the effect of ergonomic training on posture of farmers. J Occup Health Epidemiol. 2015;4(3):131–8. doi:10.18869/acadpub.johe.4.3.130.
10. Chander DS, Cavatorta MP. An observational method for Postural Ergonomic Risk Assessment (PERA). Int J Ind Ergon. 2017;57(1):32–41. doi:10.1016/j.ergon.2016.11.007.
11. Erginel N, Toptanci S. Intuitionistic fuzzy REBA method and its application in a manufacturing company. Adv Intell Syst Comput. 2019;792:27–35. doi:10.1007/978-3-319-94000-7_3.
12. Wu S, Chen Z, Zhao X, Yao M, Wang Z, Kuang S. Design of an ergonomic App for entire rapid body assessment based on Mask RCNN. In: Proceedings of the Third International Conference on Mechanical, Electric and Industrial Engineering; 2020 May 23–25; Kunming, China.
13. Ghasemi F, Mahdavi N. A new scoring system for the Rapid Entire Body Assessment (REBA) based on fuzzy sets and Bayesian networks. Int J Ind Ergon. 2020;80:103058. doi:10.1016/j.ergon.2020.103058.
14. Stefana E, Marciano F, Rossi D, Cocca P, Tomasoni G, Lopomo F, et al. Wearable devices for ergonomics: a systematic literature review. Sensors. 2021;21(3):777. doi:10.3390/s21030777.
15. Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference; 2010 Aug 31–Sep 3; Aberystwyth, UK.
16. Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation. In: Proceedings of the Computer Vision and Pattern Recognition; 2011 Jun 20–25; Colorado Springs, CO, USA. p. 1465–72.
17. Andriluka M, Pishchulin L. 2D human pose estimation: new benchmark and state of the art analysis [Internet]. [cited 2025 Feb 10]. Available from: http://openaccess.thecvf.com/content_cvpr_2014/html/Andriluka_2D_Human_Pose_2014_CVPR_paper.html.
18. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of the Computer Vision-ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland. Cham, Switzerland: Springer; 2014. p. 740–55.
19. Sigal L, Balan AO, Black MJ. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int J Comput Vis. 2010;87(1–2):4–27. doi:10.1007/s11263-009-0273-6.
20. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell. 2014;36(7):1325–39. doi:10.1109/tpami.2013.248.
21. Joo H, Simon T, Li X, Liu H, Tan L, Gui L, et al. Panoptic studio: a massively multiview system for social interaction capture. IEEE Trans Pattern Anal Mach Intell. 2019;41(1):190–204. doi:10.1109/tpami.2017.2782743.

22. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, et al. Monocular 3D human pose estimation in the wild using improved CNN supervision. In: Proceedings of the 2017 International Conference on 3D Vision (3DV); 2017 Oct 10–12; Qingdao, China. p. 506–16.
23. Trumble M, Gilbert A, Malleson C, Hilton A, Collomosse J. Total capture: 3D human pose estimation fusing video and inertial sensors. In: Proceedings of the 28th British Machine Vision Conference; 2017 Sep 4–7; London, UK. p. 1–13.
24. von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G. Recovering accurate 3D human pose in the wild using IMUs and a moving camera [Internet]. [cited 2025 Feb 10]. Available from: http://virtualhumans.mpi-inf.mpg.de/3DPW.
25. Mehta D, Sotnychenko O. Single-shot multi-person 3D pose estimation from monocular RGB [Internet]. [cited 2025 Feb 10]. Available from: https://ieeexplore.ieee.org/abstract/document/8490962/.
26. Varol G, Romero J, Martin X, Mahmood N, Black MJ, Laptev I, et al. Learning from Synthetic Humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA; 2017. p. 109–17.
27. Ebadi SE, Jhang YC, Zook A, Dhakad S, Crespi A, Parisi P, et al. PeopleSansPeople: a synthetic data generator for human-centric computer vision [Internet]. [cited 2025 Feb 10]. Available from: https://arxiv.org/abs/2112.09290v2.
28. Zhou X, Huang Q, Sun X, Xue X, Wei Y. Towards 3D human pose estimation in the wild: a weakly-supervised approach [Internet]. [cited 2025 Feb 10]. Available from: https://github.com/.
29. Zhang Y, You S, Gevers T. Orthographic projection linear regression for single image 3D human pose estimation. In: Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR); 2021 Jan 10–15; Milan, Italy. p. 8109–16.
30. Newlin Shebiah R, Aruna Sangari A. Classification of human body parts using histogram of oriented gradients. In: Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS); 2019 Mar 15–16; Coimbatore, India. p. 958–61.
31. Zhang L, Chen S, Zou B. Estimation of 3D human pose using prior knowledge. J Electron Imaging. 2021;30(4):1–5. doi:10.1117/1.jei.30.4.040502.
32. Sun X, Shang J, Liang S, Wei Y. Compositional human pose regression. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2621–30.
33. Tompson J, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human pose estimation [Internet]; 2014. [cited 2025 Feb 10]. Available from: https://arxiv.org/abs/1406.2984v2.
34. Lifshitz I, Fetaya E, Ullman S. Human pose estimation using deep consensus voting. In: Lecture Notes in Computer Science. In: Proceedings of the Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands. Berlin/Heidelberg, Germany: Springer; 2016. p. 246–60.
35. Wei SE, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 4724–32.
36. Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression. In: Proceedings of the Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands. Cham, Switzerland: Springer; 2016. p. 717–32.
37. Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision; 2016 Oct 11–14; Amsterdam, The Netherlands. Cham, Switzerland: Springer; 2016. p. 483–99.
38. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–Jul 26; Honolulu, HI, USA. p. 1263–72.
39. Sun X, Xiao B, Wei F, Liang S, Wei Y. Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. Berlin/Heidelberg, Germany: Springer; 2018. p. 536–53.

40. Sun K, Xiao B, Liu D. IEEE JWP of the, 2019 undefined. Deep high-resolution representation learning for human pose estimation [Internet]. [cited 2025 Feb 10]. Available from: http://openaccess.thecvf.com/content_CVPR_2019/html/Sun_Deep_HighResolution_Representation_Learning_for_Human_Pose_Estimation_CVPR_2019_paper.html.
41. Groos D, Ramampiaro H, Ihlen EA. EfficientPose: scalable single-person pose estimation. Appl Intell. 2021;51(4):2518–33. doi:10.1007/s10489-020-01918-7.
42. Mori G, Malik J. Estimating human body configurations using shape context matching. In: Proceedings of the Computer Vision—ECCV 2002: 7th European Conference on Computer Vision; 2002 May 28–31; Copenhagen, Denmark. Berlin/Heidelberg, Germany: Springer; 2002. p. 666–80.
43. Agarwal A, Triggs B. 3D human pose from silhouettes by relevance vector regression. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2004 Jun 02–27; Washington, DC, USA.
44. Onishi K, Takiguchi T, Ariki Y. 3D human posture estimation using the HOG features from monocular image. In: Proceedings of the 2008 19th International Conference on Pattern Recognition; 2008 Dec 8–11; Tampa, FL, USA. p. 1–4.
45. Shan W, Liu Z, Zhang X, Wang Z, Han K, Wang S, et al. Diffusion-Based 3D human pose estimation with multi-hypothesis aggregation. In: Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 14715–25.
46. Zhao L, Peng X, Tian Y, Kapadia M. Metaxas DiN. Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–17; Long Beach, CA, USA. Washington, DC, USA: IEEE Computer Society. p. 3420–30.
47. Ionescu C, Li F, Sminchisescu C. Latent structured models for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain. p. 2220–7.
48. Tripathi S, Müller L, Huang CHP, Taheri O, Black MJ, Tzionas D. 3D human pose estimation via intuitive physics. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 4713–25.
49. Shimada S, Golyanik V, Xu W, Pérez P, Theobalt C. Neural monocular 3D human motion capture with physical awareness. ACM Trans Graph. 2021;40(4):1–15. doi:10.1145/3450626.3459825.
50. Tome D, Russell C, Agapito L. Lifting from the deep: convolutional 3D pose estimation from a single image. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. Piscataway, NY, USA: IEEE; 2017. p. 5689–98.
51. Xiao B, Wei Y. Simple baselines for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 1–16.
52. Ma X, Su J, Wang C, Ci H, Wang Y. Context modeling in 3D human pose estimation: a unified perspective. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 6234–43.
53. Li S, Zhang W, Chan AB. Maximum-margin structured learning with deep networks for 3D human pose estimation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 2848–56.
54. Ghezelghieh MF, Kasturi R, Sarkar S. Learning camera viewpoint using CNN to improve 3D body pose estimation. In: Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV); 2016 Oct 25–28; Stanford, CA, USA. Piscataway, NY, USA: IEEE; 2016. p. 685–93.
55. Takahashi K, Mikami D, Isogawa M, Kimata H. Human pose as calibration pattern: 3D human pose estimation with multiple unsynchronized and uncalibrated cameras. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2018 Jun 18–22; Salt Lake, UT, USA. p. 1856–63.
56. Jiang B, Hu L, Xia S. Probabilistic triangulation for uncalibrated multi-view 3D human pose estimation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 01–06; Paris, France. p. 14804–14.

57. Wang T, Zhang J, Cai Y, Yan S, Feng J. Direct multi-view multi-person 3D pose estimation. In: advances in neural information processing systems. Neural Inf Process Syst Found. 2021;34:13153–64.
58. Kundu JN, Seth S, Ym P, Jampani V, Chakraborty A, Babu RV. Uncertainty-aware adaptation for self-supervised 3D human pose estimation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Piscataway, NY, USA: IEEE; 2022. p. 20416–27.
59. Carbonera Luvizon D, Tabia H, Picard D. SSP-Net: scalable sequential pyramid networks for real-time 3D human pose regression. Pattern Recognit. 2023;142(1):109714. doi:10.1016/j.patcog.2023.109714.
60. Mori G, Malik J. Recovering 3D human body configurations using shape contexts. IEEE Trans Pattern Anal Mach Intell. 2006;28(7):1052–62. doi:10.1109/tpami.2006.149.
61. Biswas S, Sinha S, Gupta K, Bhowmick B. Lifting 2D human pose to 3D: a weakly supervised approach. In: Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN); 2019 Jul 14–19; Budapest, Hungary. p. 1–9.
62. Li S, Chan AB. 3D human pose estimation from monocular images with deep convolutional neural network. Asian Conf Comput Vis. 2014;9004:332–47. doi:10.1007/978-3-319-16808-1_23.
63. Yang CY, Luo J, Xia L, Sun Y, Qiao N, Zhang K, et al. CameraPose: weakly-supervised monocular 3D human pose estimation by leveraging in-the-wild 2D annotations. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 2923–32.
64. Zhou X, Zhu M, Leonardos S, Daniilidis K. Sparse representation for 3D shape estimation: a convex relaxation approach. IEEE Trans Pattern Anal Mach Intell. 2017;39(8):1648–61. doi:10.1109/tpami.2016.2605097.
65. Yasin H, Iqbal U, Kruger B, Weber A, Gall J. A dual-source approach for 3D pose estimation from a single image. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 4948–56.
66. Luvizon DC, Picard D, Tabia H. Consensus-based optimization for 3D human pose estimation in camera coordinates. Int J Comput Vis. 2022;130(3):869–82. doi:10.1007/s11263-021-01570-9.
67. Du Y, Wong Y, Liu Y, Han F, Gui Y, Wang Z, et al. Marker-less 3D human motion capture with monocular image sequence and height-maps. In: Lecture Notes in Computer Science. In: Proceedings of the Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands. Berlin/Heidelberg, Germany: Springer; 2016. p. 20–36.
68. Moreno-Noguer F. 3D human pose estimation from a single image via distance matrix regression. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1561–70.
69. Martinez J, Hossain R, Romero J, Little JJ. A simple yet effective baseline for 3D human pose estimation. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–19; Venice, Italy. p. 2659–68.
70. Wang M, Chen X, Liu W, Qian C, Lin L, Ma L. Drpose3D: depth ranking in 3D human pose estimation. arXiv:1805.08973. 2018.
71. Zhou K, Han X, Jiang N, Jia K, Lu J. HEMlets pose: learning part-centric heatmap triplets for accurate 3D human pose estimation. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 02; Seoul, Republic of Korea. p. 2344–53.
72. Wu Y, Ma S, Zhang D, Huang W, Chen Y. An improved mixture density network for 3D human pose estimation with ordinal ranking. Sensors. 2022;22(13):1–13. doi:10.3390/s22134987.
73. Choi H, Moon G, Lee KM. Pose2Mesh: graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In: Proceedings of the Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. Cham, Switzerland: Springer; 2020. p. 769–87.
74. Zhao W, Tian Y, Ye Q, Jiao J, Wang W. GraFormer: graph convolution transformer for 3D pose estimation [Internet]. [cited 2025 Feb 10]. Available from: http://arxiv.org/abs/2109.08364.
75. Zou Z, Tang W. Modulated graph convolutional network for 3D human pose estimation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 11457–67.

76. Choi J, Shim D, Kim HJ. DiffuPose: monocular 3D human pose estimation via denoising diffusion probabilistic model. In: Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2023 Oct 1–5; Detroit, MI, USA. p. 3773–80.
77. Zhou X, Sun X, Zhang W, Liang S, Wei Y. Deep kinematic pose regression. Lecture Notes in Computer Science. In: Proceedings of the Computer Vision-ECCV, 2016 Workshop; 2016 Oct 8–10 and 15–16; Amsterdam, The Netherlands. Cham, Switzerland: Springer; 2016. p. 186–201.
78. Xu J, Yu Z, Ni B, Yang J, Yang X, Zhang W. Deep kinematics analysis for monocular 3D human pose estimation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 896–905.
79. Habibie I, Xu W, Mehta D, Pons-Moll G, Theobalt C. In the wild human pose estimation using explicit 2D features and intermediate 3D representations [Internet]. [cited 2025 Feb 10]. Available from: http://openaccess.thecvf.com/content_CVPR_2019/html/Habibie_In_the_Wild_Human_Pose_Estimation_Using_Explicit_2D_Features_CVPR_2019_paper.html.
80. Chen CH, Tyagi A, Agrawal A, Drover D, Mv R, Stojanov S, et al. Unsupervised 3D pose estimation with geometric self-supervision [Internet]. [cited 2025 Feb 10]. Available from: http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Unsupervised_3D_Pose_Estimation_With_Geometric_Self-Supervision_CVPR_2019_paper.html.
81. Wandt B, Little JJ, Rhodin H. ElePose: unsupervised 3D human pose estimation by predicting camera elevation and learning normalizing flows on 2D poses. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 6625–35.
82. Abdollahzade F, Mohammadi F. Working posture and its predictors in hospital operating room nurses [Internet]. [cited 2025 Feb 10]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4847110/.
83. Chen J, Qiu J. Construction worker's awkward posture recognition through supervised motion tensor decomposition. nurses [Internet]. [cited 2025 Feb 10]. Available from: https://www.sciencedirect.com/science/article/pii/S0926580517300651.
84. Anghel DCC, Niţu ELLI, Rizea ADD, Gavriluţa AA, Gavriluţa AA, Belu N, et al. Ergonomics study on an assembly line used in the automotive industry. In: MATEC Web of Conferences; 2019; Les Ulis, France: EDP Sciences; 2019. 12001 p.
85. Seo JO, Lee SH. Automated postural ergonomic risk assessment using vision-based posture classification. Autom Constr. 2021;128(1):103725. doi:10.1016/j.autcon.2021.103725.
86. Ryu JH, McFarland T, Haas CT, Abdel-Rahman E. Automatic clustering of proper working postures for phases of movement. Autom Constr. 2022;138(3):104223. doi:10.1016/j.autcon.2022.104223.
87. Jeong SO, Kook J. CREBAS: computer-based REBA evaluation system for wood manufacturers using MediaPipe. Appl Sci. 2023;13(2):938. doi:10.3390/app13020938.
88. Yan X, Li H, Wang C, Seo JO, 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(2):152–63. doi:10.1016/j.aei.2017.11.001.
89. Tellaeche A, Chatzis T, Konstantinidis D, Dimitropoulos K. Automatic ergonomic risk assessment using a variational deep network architecture. Sensors. 2022;22(16):6015. doi:10.3390/s22166051.
90. Kim W, Sung J, Saakes D, Huang C, Xiong S. Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). Int J Ind Ergon. 2021;84(2):103164. doi:10.1016/j.ergon.2021.103164.
91. MassirisFernández M, Fernández JÁ, Bajo JM, Delrieux CA. Ergonomic risk assessment based on computer vision and machine learning. Comput Ind Eng. 2020;149(3):106816. doi:10.1016/j.cie.2020.106816.
92. Namwongsa S, Puntumetakul R, Neubert MS, Chaiklieng S, Boucaut R. Ergonomic risk assessment of smartphone users using the Rapid Upper Limb Assessment (RULA) tool. PLoS One. 2018;13(8):e0203394. doi:10.1371/journal.pone.0203394.
93. Li L, Xu X, Fitts EP. A deep learning-based RULA method for working posture assessment. Proc Hum Factors Ergon Soc Annu Meet. 2019;63(1):1090–4. doi:10.1177/1071181319631174.

94. Kumar Nayak G, Kim E. Development of a fully automated RULA assessment system based on computer vision. Int J Ind Ergon. 2021;86(12):103218. doi:10.1016/j.ergon.2021.103218.
95. Abobakr A, Nahavandi D, Hossny M, Iskander J, Attia M, Nahavandi S, et al. RGB-D ergonomic assessment system of adopted working postures. Appl Ergon. 2019;80(99):75–88. doi:10.1016/j.apergo.2019.05.004.
96. Niu J, Wang X, Wang D, Ran L. A novel method of human joint prediction in an occlusion scene by using low-cost motion capture technique. Sensors. 2020;20(4):1119. doi:10.3390/s20041119.