Generalization of Inverse Kinematics Frameworks Based on Deep Learning to New Motor Tasks and Markersets

Hasnaa Ouadoudi Belabzioui, Charles Pontonnier, Georges Dumont, Pierre Plantard, Franck Multon

Generalization of Inverse Kinematics Frameworks Based on Deep Learning to New Motor Tasks and Markersets

Hasnaa Ouadoudi Belabzioui[a,b], Charles Pontonnier[a], Georges Dumont[a], Pierre Plantard[b] and Franck Multon[a]

[a] *Univ Rennes, Inria, CNRS, IRISA-UMR 6074, 35000, Rennes, France*
[b] *Moovency, Inc, 35051, Cesson-Sévigné, France*

ARTICLE INFO

Keywords:
Musculoskeletal disorder assessment
Motion capture
RGB-based human pose estimation
Opencap
Biomechanics
Generalization of Opencap marker augmenter algorithms for inverse kinematics

ABSTRACT

Several systems propose to monitor the activity of workers in industry, with markerless Human Pose Estimation (HPE) methods based on deep learning. However, these systems simply provide sparse 3D human body keypoints, including noise and missing information. Hence, these sparse and noisy keypoints cannot be directly used to assess the biomechanical constraints associated with professional activity. Indeed, computing these constraints would require more accurate and high frequency motion capture data to compute reliable joint angles, or using inverse kinematics frameworks (such as OpenSim). Deep-learning (DL) based approaches, such as Opencap, have been introduced to estimate additional anatomical markers' positions, to overcome this limitation. However, such DL-based methods rely on training datasets and predefined keypoints and markersets, and their ability to generalize to other tasks or experimental conditions is still unclear. In this paper, we assess the ability of Opencap, pre-trained with bipedal locomotion dataset, to generalize (i.e. estimate reliable 3D positions of additional anatomical markers) to bimanual manipulation and picking tasks, and new markersets. Fine tuning, commonly used in DL to generalize a model to new data, is a promising mean to deal with unseen motions and different experimental conditions, with a few set of new training data. We evaluated the performance of various fine tuning strategies, such as retraining the full model, only the last layers or adding an additional output layer. Our results showed an important decrease of the estimation error when using fine tuning on picking and manipulation tasks, with new markersets, compared to directly applying the pretrained Opencap model. This decrease of error is obtained with a limited training dataset of 140000 poses, which is promising for future use in new measurement conditions and unseen motions, as frequently observed in industry.

Relevance to industry: Accurate Human Pose Estimation on-site is a key challenge to accurately assess musculoskeletal disorders with relevant and reliable biomechanical variables. However, RGB-based HPE used on-site generally provide sparse and noisy postural information, which is not compatible with standard biomechanical frameworks. This paper suggests and evaluates guidelines to overcome this limitation, and to make standard HPE methods be used in biomechanical framework. This open new avenues in estimating biomechanical variables that could improve the estimation of the musculoskeletal disorders risks directly inindustrial context, as it is performed in laboratory conditions. This paper could be viewed as recommendations for companies which develop ergonomic assessment tools usable in industrial context.

1. Introduction

The risk of musculoskeletal disorders is multifactorial, including biomechanical, physiological, psychological, organizational, etc. factors. When dealing with biomechanical factors, most of the assessment methods rely on estimating joint angles. Inverse kinematics (IK) aims at computing these joint angles according to a predefined and scaled skeleton, aligned with some experimental positions of anatomical markers. It is the first step of several assessments, such as filling-in ergonomic assessment grids, or computing mechanical joint constraints (joint forces and torques). Based on precise, low-noise, high-frequency motion data, inverse kinematics is formulated as a global optimization problem at each frame, with the objective of minimizing the distance between experimental markers and kinematic model markers (Lu and O'connor, 1999). Nevertheless, obtaining precise, low-noise and high-frequency motion capture data in real industrial work conditions is impractical due to several constraints: the time required for

*Corresponding author. ENS Rennes, Campus de KerLann, Avenue Robert Schuman, 35170 Bruz, France
E-mail address: hasnaa.ouadoudi-belabzioui@ens-rennes.fr

installation and subject preparation, the significant space needed for equipment setup, and the extensive processing time involved. Recent advances in computer vision and deep learning offer the possibility to use repeatable posture measurements on site, in industrial context, with a simple RGB camera. For example, Abobakr et al. (Abobakr et al., 2019) leverages deep learning and vision-based techniques to estimate joint angles directly from single depth images. Other authors (Plantard et al., 2017a) showed that correcting Kinect data and adapted inverse dynamics approach, enables to correctly estimate internal joint torques, which provides relevant information for ergonomic assessment in real working environment. Several companies and researchers propose RGB-based Human Pose Estimation (HPE) as a promising alternative for biomechanical analysis of human movement in industry: simply using a smartphone, without calibration, markerless, and almost no constraint for the worker who just has to perform his/her task. Despite the promising advancements, a systematic review in (Egeonu and Jia, 2024) highlighted that RGB-based Human Pose Estimation (HPE), though convenient and minimally intrusive, generally provides sparse 3D keypoints. This sparse noisy information might be not sufficient to compute well admitted ergonomic assessments grids, or compute physical values. Hence, previous studies (Falisse et al., 2023) have reported an inaccuracy of 5 degrees in the joint angle estimation when using HPE compared to those obtained with optoelectronic systems, but these tests are generally performed in laboratory condition. However, HPE generally returns sparse 3D keypoints information, such as 3D joint centers solely. Inverse kinematics based on this sparse data consequently leads to higher error rates, compared to using 3D positions of a large set of anatomical markers as input (Uhlrich et al., 2022).

Opencap (Uhlrich et al., 2023) has recently proposed to overcome this limitation by augmenting the number of anatomical markers based on the sparse joint positions. It consists in an open-source platform for computing both kinematic (i.e., motion) and dynamic (i.e., forces) variables using videos captured from two or more smartphones. The calibrated videos are used by HPE systems to estimate sparse 3D keypoints trajectories. Then, Opencap proposes a marker augmenter (based on deep learning) algorithm that estimates additional anatomical markers positions based on these few available 3D keypoints. The resulting anatomical markers can be used by standard IK algorithm to estimate joint angles, and apply inverse dynamics. Opencap marker augmenter contains two deep learning (DL) models, namely the **Body Model** and the **Arm Model**. The **Body Model** aims at predicting the 3D positions of the lower-limb and torso anatomical markers. The **Arm Model** aims at predicting the 3D positions of the two arms anatomical markers. These models have been trained and tested on a dataset that contains the following motions (Uhlrich et al., 2023): walking, running, squatting, cutting, drop, jumping, and stair ascending and descending. This dataset has also been obtained with a given set of experimental conditions (such as camera intrinsic and extrinsic parameters, 3D keypoints definitions, etc.) and for a given output set of anatomical markers. The ability of these models to generalize to new experimental conditions and set of anatomical markers is not documented yet, up to our knowledge.

In Deep Learning, generalization aims at adapting the model: to understand the patterns and relationships within its training data and apply them to previously unseen examples, from within the same distribution as the training set. A more complex problem consists in extending this generalization to unseen examples from within a different distribution, i.e. a set of examples that have never been used for training and testing. Transfer learning consists in using a model trained on one task as the starting point, as a basis for a model addressing a new task, or on data with different distribution (Zhuang et al., 2020). This is done by transferring the knowledge that the first model has learned about the features of the input and output data to the second model. This is an interesting approach to train a new Opencap marker augmenter model that is able to handle new types of motion and markersets. Hence, fine-tuning or adapting a pre-trained model to a labeled target dataset (Han et al., 2024), represents a prevalent methodology in transfer learning, and is progressively establishing itself as a standard procedure within the computer vision and natural language processing research communities (Shi et al., 2024).

For example, ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) architectures, initially trained on the ImageNet dataset (Deng et al., 2009), are extensively fine-tuned for a multitude of computer vision applications. Concurrently, models such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020), which are pre-trained on extensive corpora, exhibit robust performance across a wide spectrum of NLP tasks. There are multiple approaches to implementing fine-tuning of deep networks in practice. A common method is to optimize all the parameters of the deep network using the target training data, after initializing them with the pre-trained model's parameters. However, when the target dataset is small and the network parameters are numerous, this can lead to overfitting (Yosinski et al., 2014). Alternatively, one may fine-tune only the last few layers of the deep network, while keeping the parameters of the initial layers fixed at their pre-trained values (Mao et al., 2023). This approach is motivated by the limited training data in the target task, and the empirical evidence that initial layers learn low-level features that are transferable across various but similar tasks. However, this approach assumes that the input data have the same nature and distribution,

which may not be the case if a different HPE or set of keypoints is used as inputs of the Opencap marker augmenter. Moreover, determining the optimal number of initial layers to freeze remains a manual and potentially inefficient process, particularly for networks with hundreds or thousands of layers. We also have to figure out that fine tuning generally has to deal with a small dataset containing the new distribution, which may rapidly lead to overfitting.

One of the main objectives of this paper is to evaluate the accuracy of the Opencap marker augmenter (Falisse et al., 2023) when dealing with new types of motion, such as those frequently used in industry: bimanual tasks, including asymmetric handling tasks (denoted Lifting Movement), and handling and picking tasks (denoted Picking Movement). These tests also involve different experimental set-up/conditions and different definitions of the anatomical markers. Each company may have its own markerset, HPE with predefined 3D keypoints, and specific motions. Hence, by performing these evaluations, we aim at dealing with similar constraints than these companies may face to adapt the Opencap system. Hence, for each new task, HPE system or specific markerset, the company should be able to collect a small set of motions (concurrently with the HPE and ground truth values) to retrain the system before exploiting it on several workstations and places.

The second main contribution of this paper is to propose a method to retrain the Opencap marker augmenter to handle such new conditions, with a limited set of examples. To this end, we explored two main fine tuning strategies for the **Body Model** and **Arm Model**. The first strategy consists in retraining all the layers of the DL architecture, assuming that the resulting models could better adapt to the new condition, compared to retraining only part of the network. However, this involves to adapt a huge number of parameters, while the number of examples of the new dataset may be small. Hence, it may lead to overfitting, with difficulties to generalize to new data in the future. The second strategy consists in tuning only the last output layers (to deal with the different output markerset), while freezing the remaining of the network. It leads to a smaller number of parameters to adapt, which may be more appropriate for the available small dataset of new examples.

2. Materials and methods

In this section, we introduce the experimental data and methods used to evaluate and train the Opencap marker augmenter models. We also describe the fine tuning processes used to adapt the models to new upper-limb industrial motions.

2.1. Overview

In this study, we evaluated two fine tuning training strategies to adapt the Opencap marker augmenter models to new motions, input and output data. These models aim at estimating a dense set of anatomical markers based on sparse 3D video keypoints computed by HPE methods. Our proposed experimental pipeline consists of two phases. In the initial phase, we fine tuned Opencap's marker augmenter models (the **Body Model** and **Arm Model**) using two different strategies. In the subsequent phase, we applied geometric calibration and inverse kinematics based on the resulting anatomical markers to compute joint angles, as illustrated in Figure 1.

In this section, we first recall the Opencap marker augmenter models (see Subsection 2.2). Next, we describe the fine tuning process of these models (see Subsection 2.3). To evaluate the two fine tuning strategies, we collected a dataset of upper-limb motions (see subsection 2.4). We then applied geometric calibration and inverse kinematics to estimate joint angles (see subsection 2.5). Finally, we evaluated the resulting anatomical markers and joint angles against ground truth values (see subsection 2.6).

2.2. Opencap marker augmenter models

The Opencap marker augmenter models (Uhlrich et al., 2023) aim at computing dense anatomical markers position according to sparse 3D video keypoints provided by HPE methods. The 3D video keypoints delivered by the HPE model, and the output anatomical markers are detailed in Supplementary material section 7.1.

As described above, the Opencap marker augmenter is based on two models associated with various body parts. The **Body model** architecture comprised four Long Short-Term Memory (LSTM) layers, each with 128 units, followed by an output layer, as illustrated in figure 2. It aims at predicting the 3D positions of 35 body anatomical markers thanks to 15 3D positions of lower-limb and torso 3D video keypoints, along with subjects-specific parameters such as height and weight.

The **Arm model** architecture is composed of five stacked Long Short-Term Memory (LSTM) layers, each comprising 128 units, followed by an output layer (as illustrated in figure 2). It aims at predicting the 3D positions

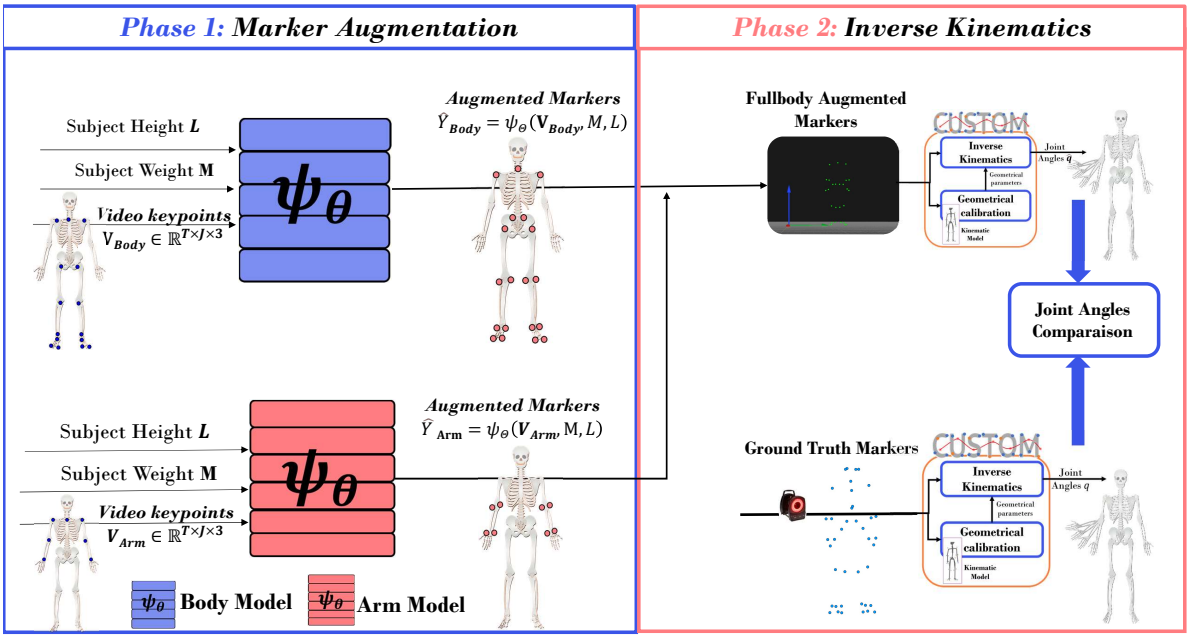

Figure 1: The Proposed Pipeline: Fine tuning the Opencap marker augmenter **Body Model** and **Arm Model** Models to better estimate anatomical markers based on sparse 3D video keypoints, followed by using Custom software to calibrate a biomechanical model and apply inverse kinematics to compute the related joint angles.

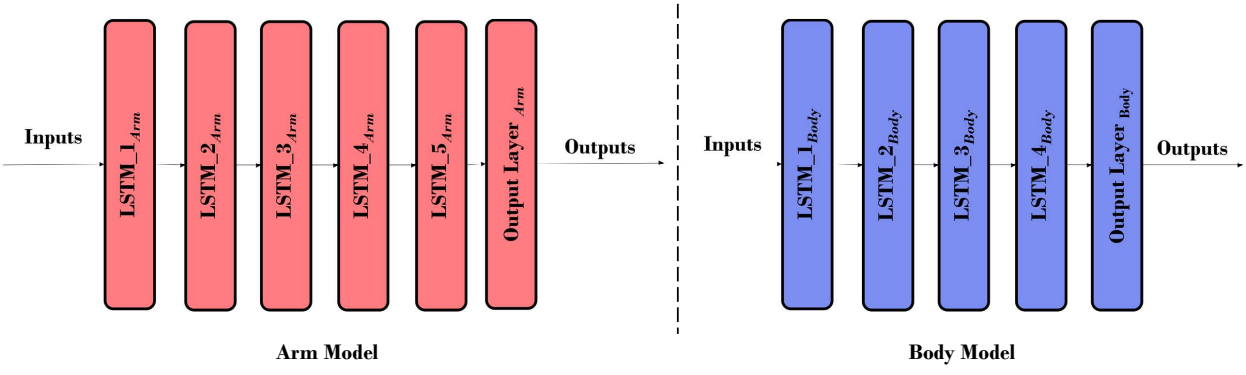

Figure 2: Detailed architecture of Opencap learning models (**Body Model** and **Arm Model**)

of 8 arm anatomical markers using 7 3D positions of arm and torso video keypoints, along with subject height and weight.

2.3. Fine tuning the Opencap marker augmenter models

In this subsection, we describe how the Opencap marker augmenter is fine tuned to adapt to new anatomical landmark and to a new dataset composed of unseen motion. The same datasets, asymmetric handling tasks (denoted as "Lifting Movement") and handling and picking tasks (denoted as "Picking Movement"), were used to train and test both fine-tuning strategies. We also tested the direct use of the pretrained Opencap augmenter models, denoted ***Inference*** in the remaining of the paper.

For all the strategies, the objective of the fine tuning process is to learn the mapping function:

$$\psi_\Theta(\boldsymbol{V}, \boldsymbol{M}, \boldsymbol{L}) = \hat{Y} \tag{1}$$

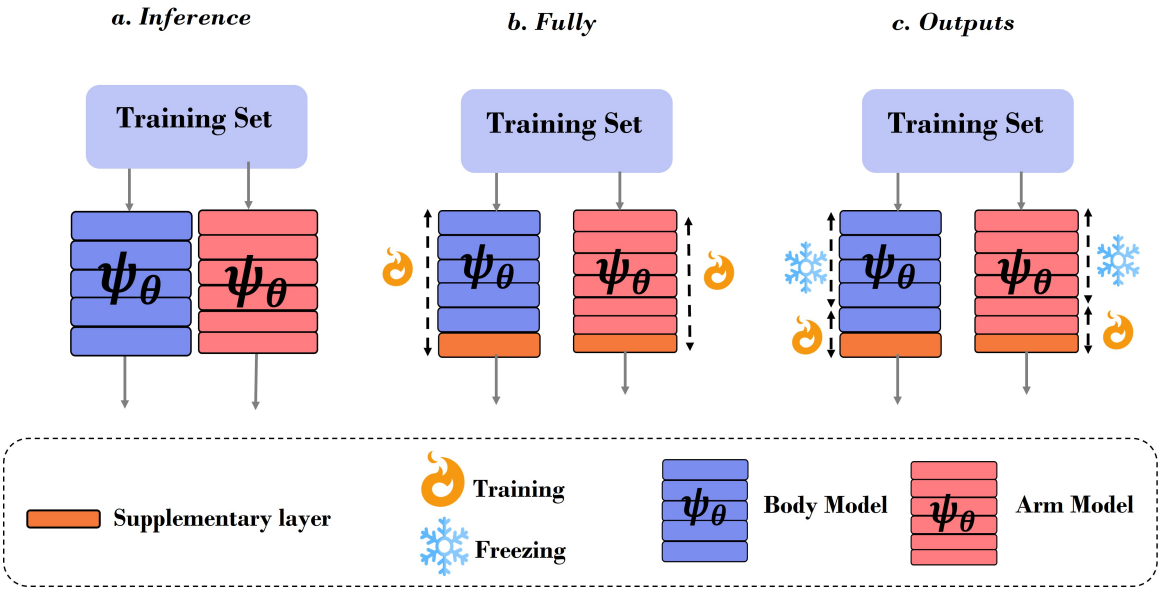

Figure 3: Inference and the two Fine-tuning strategies. a. Inference consists in directly applying the pretrained marker augmenter models on the new dataset. b. ***Fully*** consists in retraining all the original network + an additional output layer with a subset of the new dataset. c. ***Outputs*** consists in adding an output layer and retraining only the two resulting output layers (the remaining layers are frozen) with a subset of the new dataset.

where $\boldsymbol{V}$ stands for the input features (the sparse 3D keypoints obtained with the HPE at time t), $\boldsymbol{M}$ and $\boldsymbol{L}$ are subject's weight and height respectively. The output of this function is the position of the additional anatomical markers $\hat{Y}$ at a given time step.

2.3.1. Fully strategy

A first strategy consists in retraining all the network, including the input, the intermediate LSTM, and the output layers. Hence, we updated the parameters of all layers in the network based on gradients computed from the new dataset (Fu et al., 2023). We assumed that fine-tuning all layers of the pre-trained model will allow it to better learn features related to the new tasks/motions at all the layers of the network. As the set of output markers is slightly different from the one initially used in Opencap, we need either to adapt and retrain the output layer, or to add a new output layer. As the number and definition of the output markers may differ, we preferred the latter solution: we added an output layer corresponding to the new set of anatomical markers, as illustrated in figure 2. The already existing layers were initialized with the pretrained values of Opencap marker augmenter model to start from a pretrained initial point. As no pretrained value is available for the new output layer, it was initialized using a normal distribution, with a mean of 0 and a standard deviation of 0.022 ($\mathcal{N}(0, 0.02^2)$). During the fine tuning phase, a weight decay of 0.01 was applied to the parameters of the last added layers, excluding biases, in accordance with the methodology outlined by (Barone et al., 2017; Li and Zhang, 2021), with the aim to avoid overfitting.

2.3.2. Outputs strategy

The second strategy, denoted ***Outputs*** strategy consists in retraining only the last output layer, while freezing the remaining ones. We hypothesized that features in the first layers were strongly linked to the input data processing, which is supposed to be similar in both the new dataset and those used to train the original Opencap marker augmenter models. However, this hypothesis is not completely true, as the 3D video keypoints may also differ a bit depending on the HPE that is used. Similarly, the last layers are supposed to be linked to the output data, namely the output estimated anatomical markers (Bordes et al., 2022), which are supposed to be different from the ones used to train the initial Opencap marker augmenter. Hence, we propose to freeze all layers except the last one (output layer). As for the ***Fully*** strategy, we also added a supplementary output layer to handle the new markerset. This method presupposes that the pretrained model has acquired valuable hierarchical features transferable to the new task. By preserving these features

and solely adjusting the output layers (both the original output and new inserted layers), the model could swiftly adjust to the new task while mitigating the risk of overfitting, especially when working with a limited dataset.

Let $\Theta_{\text{past\_output}}$ represent the parameters of the past output layers (respectively, $\Theta_{\text{past\_output}_{\text{Body}}}$ and $\Theta_{\text{past\_output}_{\text{Arm}}}$), and let $\Theta_{\text{new\_output}}$ represent the parameters of the new output layers. Here, X and Y represent the input and output data for this stage, respectively. The objective function J quantifies the performance of this stage's model.

$$J(\Theta_{\text{past\_output}}, \Theta_{\text{new\_output}}, X, Y) = J_{\text{task}}(\Theta_{\text{past\_output}}, \Theta_{\text{new\_output}}, X, Y) + \lambda R(\Theta_{\text{new\_output}})$$

Where J_{task} denotes the original task loss, which in this case is the mean squared error. The regularization term R is introduced to prevent overfitting; in this implementation, L^2 regularization is applied with a regularization parameter of 0.01. The hyperparameter λ controls the regularization strength, determining the trade-off between fitting the training data and minimizing the complexity of the model.

2.4. Datasets

As the Opencap augmenter models were originally mainly trained on lower-limbs motions, such as locomotion, we collected motion capture data associated with upper-limb motions, as mostly used in industry. Hence, we used data collected in two different experiments: asymmetric handling tasks, and handling and picking tasks. Not only the motion are different, but also the markersets, which is an interesting property for testing the fine tuning strategies.

The denoted "Lifting dataset" consists in asymmetric handling tasks (Muller et al., 2019a). It involves thirteen male participants who had to move a load between three areas, leading to cycles of three displacements: from area 1 to area 2, area 2 to area 3, and area 3 back to area 1. Each participant completed two cycles with a standard load of 6.9 kg and two cycles with an additional 3 kg load. The experimental setup included 47 motion capture markers on standardized anatomical markers, following the recommendations of the International Society of Biomechanics (Wu et al., 1995). Motion capture data was recorded at 200 Hz using a 16-camera Vicon motion capture system; considered as the reference system for the experiment, as illustrated in the figure 4. The 200Hz resulting data were downsampled to 60Hz, similarly to the video data used to train Opencap. The input 3D keypoints were estimated using the described method in Supplementary material section 7.1. The resulting data (estimated 3D keypoints and ground truth anatomical markers) were used to retrain the models, and perform the quantitative comparison between predicted and actual joint angles and anatomical landmark positions.

The denoted "Picking dataset" consists in handling and picking tasks. It involves 12 participants (3 women and 9 men, age: 32.6 ± 10 years, height: 1.73 ± 0.079 m, weight: 76 ± 16 kg). Participants were filmed (to use real HPE system) and equipped with the XSens inertial motion capture system. Once the skeleton model of each subject was calibrated, the XSens software (Roetenberg et al., 2009) simulated skin marker positions, including also the ones following the recommendations of the International Society of Biomechanics. The objective of this experiment was to emulate real work conditions: bimanual handling and picking tasks. Bimanual handling involved picking up a box (dimensions: $39 \times 29.5 \times 19$ cm) from a three-tiered shelf and placing it on another shelf, repeating this process 5 times following a specified order on the shelves. Picking task required picking up and replacing a small cubic object (dimensions: $5 \times 5 \times 5$ cm) at 16 different locations arranged on a table in front of the participant, following a specific order. The participants had to perform picking in ascending and then descending order, using their right and left hands, respectively. In total, each participant performed 4 picking actions. In the context of this paper, these tasks are interesting to challenge the HPE system, as they involve external occlusions (with the box and the table) and self-occlusion depending on different measurement viewpoints. Consequently, it may affect the quality of the HPE outputs before estimating the anatomical markers using the Opencap augmenter models. Whereas the Opencap system required multiple calibrated cameras, we used a single RGB camera, placed facing right during the experiment (see figure 4). To process the unique RGB camera, we used the KIMEA Cloud solution developed by Moovency. The video and XSens files were synchronized using a clapping signal at the start and end of each trial. Spatial alignment involved removing translational and orientation information from the resulting 3D pose data, ensuring that each 3D pose captured only the execution of motion, independent of location or viewpoint, as detailed in previous studies (Yasin et al., 2023, 2020; Chen and Koskela, 2013).

Before the training phase, we expressed the 3D positions of anatomical markers relatively to a root marker, specifically the midpoint of the hip keypoints. Additionally, we normalized these 3D positions based on the subject's height. The data was standardized to have zero mean and unit standard deviation, before being used for retrained the Opencap marker augmenter models.

(a). Asymmetric handling tasks **(b). Handling and picking tasks**

Figure 4: The experimental protocol for the asymmetric handling tasks (denoted lifting tasks) is presented in sub-figure **(a)**, as detailed in the study by (Muller et al., 2019a). The protocol for the handling and picking tasks (denoted picking tasks) is presented in sub-figure **(b)**.

2.5. Inverse kinematics

Opencap augmenter models are used to estimate the anatomical markers that are necessary to calibrate a biomechanical model and perform inverse kinematics, to compute the joint angles.

The whole body biomechanical model consisted of eighteen segments: *upper/lower trunk,left/right clavicle left/right arm, left/Right forearm, left/right thigh, left/right shank, left/right foot, and pelvis*. This model was articulated with 42 degrees of freedom, comprising a 6 degrees of freedom (DoFs) mobile base and 43 anatomical joint angles in accordance with recommendations from the International Society of Biomechanics (Wu et al., 1995) and summarized in table 1. Considering the output markerset delivered by the Opencap marker augmenter models, head and hands segments were removed from the initial model (no available head and hands markers in the markerset).

We used the Custom software (Muller et al., 2017, 2019b; Puchaud et al., 2020) to perform the geometrical calibration of the model according to the estimated anatomical landmark positions. This calibration was formulated into an optimization problem trying to minimize the distance between the experimental markers and the corresponding anatomical points of the model, by adjusting the segment lengths. This method was applied with ground truth motion capture data, and anatomical markers estimated by the Opencap augmenter models.

Once the calibration was performed, Custom was again used to perform inverse kinematics: a penalty method for constrained multibody kinematics optimization using the Levenberg-Marquardt algorithm (Livet et al., 2023). This method aimed to determine the joint angles $\boldsymbol{q}$ according to the position of the 3D anatomical markers.

2.6. Evaluation methodology

The goal of this work was to evaluate the performance of the Opencap augmenter models to predict the position of dense anatomical markers based on sparse 3D keypoints. As described in subsection 2.4, we used two types of datasets:

- Lifting dataset: the asymmetric handling task, composed of ground truth optical motion capture data,
- Picking dataset: the handling and picking tasks, composed of ground truth XSens motion capture data and RGB videos.

Let us consider now the evaluation metrics and implementation details used for this work.

2.6.1. Evaluation metrics

As the markersets are different in all the datasets, compared to the one used by the Opencap pretrained models, we only compared markers with similar definitions. The **Body Model** inputs are composed of 15 video keypoints, and the subject's Height and Weight. The outputs consist in 35 outputs markers, with 19 of them corresponding to anatomical markers that also exist in our two markersets. Thus, we did not use the following markers from the original **Body Model** to carry-out the comparisons: [r_thigh1_study, r_thigh2_study, r_thigh3_study, L_thigh1_study, L_thigh2_study, L_thigh3_study, r_sh1_study, r_sh2_study, r_sh3_study, L_sh1_study, L_sh2_study, L_sh3_study,

Joint	Corresponding exertion	Joint	Corresponding exertion
Lumbar Spine F/E	Lumbar Spine Flexion/Extension	**R/L Hip F/E**	R/L Hip Flexion/Extension
Lumbar Spine LF/LE	Lumbar Spine Lateral flexion/extension	**R/L Hip A/A**	R/L Hip Abduction/Adduction
Lumbar Spine I/E	Lumbar Spine Axial Rotation	**R/L Hip I/E**	R/L Hip Internal/External rotation
Thoracic Spine F/E	Thoracic spine Flexion/Extension	**R/L Knee F/E**	R/L Knee Flexion/Extension
Thoracic Spine LF/LE	Thoracic Spine Lateral Flexion/extension	**R/L Ankle F/E**	R/L Ankle Flexion/Extension
Thoracic Spine I/E	Thoracic Spine Axial Rotation	**R/L Ankle I/E**	R/L Subtalar Inversion/Eversion
R/L Clavicle P/R	R/L Clavicle Protraction/Retraction	**R/L Elbow F/E**	R/L Elbow Flexion Extension
R/L Clavicle D/E	R/L Clavicle Depression/Elevation	**R/L Forearm P/S**	R/L Forearm Pronation/Supination
R/L Clavicle I/E	R/L Clavicle Axial Rotation	**R/L Glenohumeral PoE**	R/L Glenohumeral Plane of Elevation
R/L Glenohumeral D/E	R/L Glenohumeral Depression/Elevation	**R/L Glenohumeral nPoE**	Negative Glenohumeral plane of elevation
R/L Glenohumeral I/E	Glenohumeral Internal/External rotation		

Table 1: Biomechanical model depicting joint angles with the following notations: **R/L** indicates Right/Left, **F/E** denotes Flexion/Extension, **LF/LE** represents Lateral Flexion/Lateral Extension, **I/E** stands for Internal/External, **P/R** refers to Protraction/Retraction, **D/E** signifies Depression/Elevation, **PoE** is Plane of Elevation, **nPoE** denotes Negative Plane of Elevation, **A/A** stands for Abduction/Adduction, **I/E** indicates Inversion/Eversion, and **P/S** represents Pronation/Supination.

RHJC, LHJC]. The **Arm Model** input data consist in 7 video keypoints, and the subject's weight and height. Its outputs are composed of 8 anatomical markers, similar to our anatomical markers.

Hence, for each motion clip of the datasets, we can compare the landmark position and joint angles estimated by Opencap augmenter models (using either motion capture or video input data) to ground truth values. For each trial, we can compare the results of direct ***Inference*** of the Opencap data augmenter models, without retraining, to those obtained with the ***Fully*** and ***Outputs*** fine tuning strategies.

For this comparison, we computed the average Root Mean Square Error RMSEm and the corresponding standard deviations (ρm) to quantify the disparities between measured and estimated 3D positions of anatomical markers. Additionally, we estimated the 95% confidence interval (CI) to further assess the precision of the measurements (Simundic et al., 2008).

Similarly, for the resulting joint angles, after IK, the average root mean squared error (RMSE_{jc}) and corresponding standard deviation (ρ_{jc}) were computed, along with the 95% confidence interval (CI) for these measures were computed, to compare joint angles obtained from ground truth marker position and augmented models ones. For **Mean Error (All joint angles)**, we considered all the angles of the biomechanical model. For **Mean Error (OpenCap joint angles)**,

we only considered the following angles: **[R/L Hip F/E, R/L Hip A/A, R/L Hip I/E, R/L Knee F/E, R/L Ankle F/E, R/L Ankle I/E, Lumbar Spine F/E, Lumbar Spine LF/LE, Lumbar Spine F/E]**. All the pelvis degrees of freedom (rotation/translation) were removed from the computation as they represent the position and orientation of the pelvis in the global coordinate system and vary depending on the experimental setup.

The evaluation was conducted using a Leave-One-Out procedure by subject. In this method, one subject is systematically removed from the training set, iteratively, and the model is tested on that removed subject. This procedure was repeated 5 times, leading to 5 subsets of training and testing sets randomly selected among the available subjects. For the lifting task, we had 11 subjects, with an average of 120000 samples in the training set and 12000 samples in the test set. For the picking task, we had 13 subjects, with an average of 160000 samples in the training set and 18000 samples in the test set.

2.6.2. Implementation details

We implemented our learning algorithms using Keras/Tensorflow (Géron, 2022), and used an NVidia RTX A3000 GPU for training and tests. The optimal model was achieved using an early stopping technique: training concluded when the loss failed to decrease with a minimum delta of 1×10^{-4} and a patience value of 10 epochs. The Adam optimization algorithm was used with a batch size of 64 and a learning rate (α) set to 6×10^{-6}. In order to train the learning algorithms, we adopted the standard mean squared error (MSE) loss function, after processing 64 training samples: the model updates its parameters based on the average errors calculated over these 64 samples.

3. Results

In this section, we present the performance of the two fine tuning strategies compared to using the pretrained Opencap marker augmenter models. Firstly, we compare the accuracy of the two strategies for predicting the 3D anatomical markers (see subsection 3.1). Secondly, we evaluate the impact of joint angle estimation while using the Custom inverse kinematic framework (see subsection 3.2).

3.1. 3D anatomical markers positions

The table 2 presents the average RMSE (RMSE_m) in millimeters along with their corresponding standard deviations (ρ_m) and 95% confidence interval **(CI)** for ***Inference*** and both fine tuning strategies, namely ***Fully*** and ***Outputs***, for the Lifting task. For the **Body Model**, the results show an important decrease of the prediction error from 39 ± 2 mm down to 15 ± 2 mm and and 16 ± 1 mm for the ***Fully*** and ***Outputs*** fine tuning strategies respectively. More important error decreases were observed for the **Arm Model**.

Model	**Movement**	**Data type**	***Inference [mm]***	***Fully [mm]***	***Outputs [mm]***
			$\text{RMSE}_\text{m} \pm \rho_\text{m}$ (CI)	**$\text{RMSE}_\text{m} \pm \rho_\text{m}$(CI)**	**$\text{RMSE}_\text{m} \pm \rho_\text{m}$(CI)**
Body	Lifting	MoCap	$39 \pm 2(38, 41)$	$15 \pm 2(13, 17)$	$16 \pm 1(14, 17)$
Arm	Lifting	MoCap	$31 \pm 4(29, 34)$	$9 \pm 1(8, 11)$	$11 \pm 1(10, 11)$
Body	Picking	RGB	$104 \pm 14(96, 112)$	$26 \pm 1(24, 28)$	$46 \pm 4(42, 50)$
Arm	Picking	RGB	$160 \pm 41(137, 182)$	$95 \pm 13(84, 106)$	$97 \pm 6(91, 103)$

Table 2: Prediction error of **Body Model** and **Arm Model** marker augmenter models for asymmetric handling movements (Lifting task) and industrial handling and picking movements (Picking task). Average RMSE (RMSE_m) and corresponding standard deviations (ρ_m) and 95% confidence interval **(CI)** are given in millimeters. Prediction error is given when using ***Inference***, and ***Fully*** and ***Outputs*** fine tuning strategies.

For the Picking task, we obtained similar important decrease of the prediction error when using fine tuning compared to directly applying the pretrained model. For the **Body Model**, the pretrained models leaded to 104 mm and 160 mm errors for the **Body Model** and **Arm Model** respectively. Let us recall here that the Picking task involved real video and HPE as input of the system, and that the ground truth was obtained with Xsens sensors. These data may differ from those obtained to evaluate the Opencap marker augmenter models for Lifting task. For the **Body Model**, this error decreased down to 26 mm and 46 mm for the ***Fully*** and ***Outputs*** fine tuning strategies respectively. For the **Arm Model**, the error decreased from 160 mm to 95 mm and 97 mm for the ***Fully*** and ***Outputs*** fine-tuning strategies, respectively. However, even with this reduction, the fine-tuned **Arm Model** still exhibits relatively high prediction errors under these experimental conditions.

Let us consider now the computing performance of the two fine tuning strategies in the various experimental conditions. Table 3 reports the training time (in minutes), the number of Epochs used to converge, and the amount of parameters that were trained by both the ***Fully*** and the ***Outputs*** fine tuning strategies. As the ***Fully*** strategy retrains all the layers of the model, it leads to adapt a large amount of parameters compared to the ***Outputs*** strategy.

Model	Movement	Data type	Time	Epoch	Params	Data
			Fully			
Body	Lifting	MoCap	78	166	504394	120000
Arm	Lifting	MoCap	193	241	607832	120000
Body	Picking	RGB	35	68	504394	160000
Arm	Picking	RGB	16	26	607832	160000
			Outputs			
Body	Lifting	MoCap	34	162	19587	120000
Arm	Lifting	MoCap	73	300	3696	120000
Body	Picking	RGB	47	141	19587	160000
Arm	Picking	RGB	24	64	3696	160000

Table 3: Performance indicators of the training process in all the test conditions: training time in minutes, number of epochs, number of trained parameters, and training data size for different tested fine-tuning strategies.

Figure 5 illustrates the comparison of the estimated **RSHO** (right acromion) position during inference stage and both fine tuning strategies, and compares it to the ground truth, during Lifting Task.

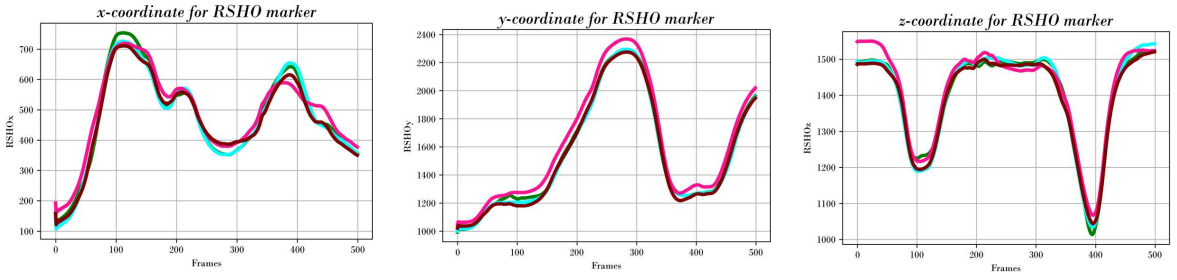

Figure 5: Estimated 3D trajectory (in mm) of the **RSHO** (right acromion) anatomical landmarker using ***Inference*** (fuchsia) and the two fine-tuning strategies for Lifting Task: ***Fully*** (green) and ***Outputs*** (maroon). Ground truth value is depicted in cyan.

To evaluate the impact of the quantity of training data that were used for training, on the fine tuning results, we also trained the **Body Model** and the **Arm** with less data. For the Lifting Task, the data from 5 subjects among the 11 wad used for training, and the data from one of the remaining 11-5=6 subjects was used for testing. These tests were carried-out on with the ***Fully*** and the ***Outputs*** fine tuning strategies. The results (see table 4) show an increase of error when using this lower quantity of training data, for all the strategies, and all the models.

3.2. Joint angles estimation

The estimated anatomical markers of the lifting tasks data were used to compute the joint angles of a biomechanical model, using the Custom Software. Table 5 reports the average RMSE ($RMSE_{jc}$) and corresponding standard deviation ρ_{jc}, between the predicted joint angles and the one obtained with ground truth anatomical markers. ($RMSE_{jc}$) is given for the ***Inference***, ***Fully*** and ***Outputs*** fine tuning strategies. Overall, both fine tuning strategies showed improvements

Model	***Fully [mm]***		***Outputs [mm]***	
	50% dataset	100% dataset	50% dataset	100% dataset
Body	$19 \pm 3(16, 22)$	$15 \pm 2(13, 17)$	$31 \pm 2(29, 33)$	$16 \pm 1(14, 17)$
Arm	$13 \pm 2(11, 16)$	$9 \pm 1(8, 11)$	$25 \pm 1(24, 27)$	$11 \pm 1(10, 11)$

Table 4: Prediction error of **Body Model** and **Arm Model** marker augmenter models for asymmetric handling movements (Lifting Task) when training with all the data or half of the dataset.

over the inference method in most joint estimations. For instance, in the **Right Hip F/E** joint, ***Fully*** reduced ($RMSE_{jc}$) from 8.2° down to 6.9°. ***Outputs*** strategy further decreased ($RMSE_{jc}$) down to 7.3°. Similar trends were observed across other joints. Figure 6 provides a visual representation of joint angles over time in all the conditions. In this figure, the right and left hip, ankle, and elbow joints are depicted.

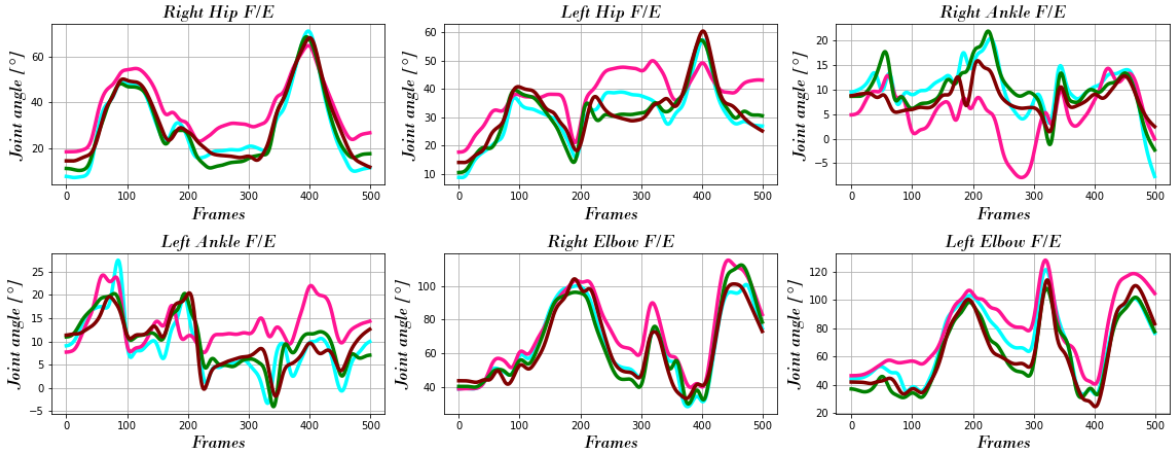

Figure 6: The estimated joint angles (in degrees) for the **Right Hip F/E**, **Left Hip F/E**, **Right Ankle F/E**, **Left Ankle F/E**, **Right Elbow F/E**, and **Left Elbow F/E** joints are presented using three different methodologies: ***Inference*** (represented in fuchsia), and two fine-tuning strategies for the Lifting Task: ***Fully*** (green) and ***Outputs*** (maroon). The ground truth values are depicted in Cyan.

4. Discussion

The aim of this paper was to evaluate two strategies to generalize the Opencap pre-trained marker augmenter models for the body and the arms. The two major challenges are 1) the new type of motions that was not present in the initial Opencap training set, and 2) the differences in output markersets. Subsection 4.1 proposes a discussion about the results obtained when predicting the position of markers which do not completely fit the original Opencap markerset. The impact of the joint angles computation using inverse kinematics is discussed in subsection 4.2.

4.1. 3D anatomical markers

Considering the lifting tasks, based on Mocap data, the **Body Model** consistently demonstrated higher error values across all tested strategies compared to the **Arm Model**. Additionally, our results indicated that the ***Fully*** strategy yielded the lowest root mean square error values for both models, compared to the ***Outputs*** strategy. It is consistent with the fact that the ***Outputs*** strategy has less parameters to tune and, thus, less possibilities to find an accurate solution. However, the ***Fully*** strategy requires a huge amount of parameters (>500K parameters for both the Arm and the Body models), which may lead to overfitting as the training dataset is not big enough.

For both strategies, the error (between 9 mm and 16 mm) yielded in similar range compared to those observed after an inverse kinematics step, in classical motion capture analyses (Begon et al., 2018; Puchaud et al., 2020; Lund et al.,

Joint Angles	***Inference [°]***	***Fully [°]***	***Outputs [°]***
Metric	**RMSE$_{jc}$ $\pm$ ρ_{jc}(CI)**	**RMSE$_{jc}$ $\pm$ ρ_{jc}(CI)**	**RMSE$_{jc}$ $\pm$ ρ_{jc}(CI)**
Right Hip F/E	8.2 $\pm$ 1.6	6.9 $\pm$ 2.0	7.3 $\pm$ 0.9
Right Hip A/A	5.4 $\pm$ 0.5	3.4 $\pm$ 0.4	5.3 $\pm$ 1.0
Right Hip I/E	15.6 $\pm$ 1.0	7.3 $\pm$ 0.4	11.0 $\pm$ 2.5
Right Knee F/E	8.5 $\pm$ 1.0	4.0 $\pm$ 1.7	7.9 $\pm$ 1.4
Right Ankle F/E	5.2 $\pm$ 0.4	4.0 $\pm$ 1.1	5.1 $\pm$ 1.2
Right Ankle I/E	16.5 $\pm$ 2.8	8.5 $\pm$ 1.0	11.1 $\pm$ 2.2
Right Clavicle P/R	22.5 $\pm$ 2.1	20.4 $\pm$ 2.7	22.9 $\pm$ 2.6
Right Clavicle D/E	13.4 $\pm$ 2.6	11.1 $\pm$ 1, 8	13.0 $\pm$ 1.6
Right Clavicle I/E	36.5 $\pm$ 3.4	34.7 $\pm$ 3.4	37.6 $\pm$ 2.2
Right Glenohumeral PoE	132.9 $\pm$ 26.3	118.3 $\pm$ 44, 8	132.7 $\pm$ 63.2
Right Glenohumeral D/E	57.5 $\pm$ 12.8	55.0 $\pm$ 23.1	57.8 $\pm$ 22.6
Right Glenohumeral nPoE	132.9 $\pm$ 26.3	118.3 $\pm$ 44.8	132.7 $\pm$ 63.2
Right Glenohumeral I/E	20.2 $\pm$ 2.3	15.4 $\pm$ 0.4	17.6 $\pm$ 2.4
Left Hip F/E	8.6 $\pm$ 1.6	7.0 $\pm$ 2.5	7.4 $\pm$ 1.8
Left Hip A/A	4.8 $\pm$ 0.3	3.2 $\pm$ 0.4	5.0 $\pm$ 1.2
Left Hip I/E	11.9 $\pm$ 2.9	6.5 $\pm$ 1.1	10.8 $\pm$ 2.8
Left Knee F/E	9.0 $\pm$ 1.3	3.9 $\pm$ 1.8	7.7 $\pm$ 1.0
Left Ankle F/E	6.3 $\pm$ 0.6	4.1 $\pm$ 1.4	5.2 $\pm$ 1.5
Left Ankle I/E	14.6 $\pm$ 2.6	7.1 $\pm$ 1.3	10.0 $\pm$ 2.0
Left Clavicle P/R	22.6 $\pm$ 4.0	21.8 $\pm$ 3.5	22.5 $\pm$ 2.5
Left Clavicle D/E	12.9 $\pm$ 1.3	10.9 $\pm$ 1.1	11.6 $\pm$ 1.0
Left Clavicle I/E	38.2 $\pm$ 5.4	34.5 $\pm$ 5.7	35.9 $\pm$ 4.5
Left Glenohumeral PoE	132.7 $\pm$ 27.3	136.3 $\pm$ 47.8	150.4 $\pm$ 32.4
Left Glenohumeral D/E	50.0 $\pm$ 15.4	47.1 $\pm$ 11.4	49.3 $\pm$ 13.0
Left Glenohumeral nPoE	132.7 $\pm$ 27.3	136.3 $\pm$ 47.8	150.4 $\pm$ 32.4
Left Glenohumeral I/E	20.6 $\pm$ 2.6	14.7 $\pm$ 0.7	17.1 $\pm$ 1.4
Right Elbow F/E	14.6 $\pm$ 1.9	6.8 $\pm$ 0.8	8.1 $\pm$ 0.5
Right Forearm P/S	23.6 $\pm$ 6.5	14.6 $\pm$ 2.9	15.8 $\pm$ 2.9
Left Elbow F/E	13.9 $\pm$ 2.0	7.5 $\pm$ 1.6	8.0 $\pm$ 0.6
Left Forearm P/S	26.2 $\pm$ 2.4	14.9 $\pm$ 5.1	14.8 $\pm$ 4.8
Lumbar Spine F/E	16.4 $\pm$ 4.1	15.2 $\pm$ 1.7	15.6 $\pm$ 1.3
Lumbar Spine LF/LE	9.2 $\pm$ 1.0	8.4 $\pm$ 0.4	9.4 $\pm$ 0.7
Lumbar Spine I/E	51.6 $\pm$ 4.3	53.2 $\pm$ 6.4	52.3 $\pm$ 5.0
Thoracic Spine F/E	13.2 $\pm$ 0.9	13.2 $\pm$ 0.6	13.7 $\pm$ 0.7
Thoracic Spine LF/LE	13.4 $\pm$ 0.2	12.7 $\pm$ 0.6	13.9 $\pm$ 1.4
Thoracic Spine I/E	49.2 $\pm$ 4.0	49.5 $\pm$ 4.7	47.8 $\pm$ 3.9
Mean Error (All joint angles)	32.5 $\pm$ 5.6(29, 35)	28.8 $\pm$ 7.7(22, 35)	31.8 $\pm$ 7.9(24, 38)
Mean Error (OpenCap joint angles)	12, 8 $\pm$ 1, 7(11, 13)	9.5 $\pm$ 1.6(8, 10)	11.4 $\pm$ 1.8(9, 13)

Table 5: The average error in joint angles estimation using ***Inference***, ***Fully*** and ***Outputs*** conditions for Lifting tasks. Average RMSE (RMSE$_{jc}$) and corresponding standard deviation (ρ_{jc}) and 95% confidence interval **(CI)** are expressed in degrees.

2015; Muller et al., 2015) (errors ranging from 4 mm to 40 mm). These results suggest that such models can be used to generate inputs for classical inverse kinematics methods, leading to a similar level of uncertainty on the joint angles. One can think that weighting markers showing the highest (RMSE$_{jc}$) may be a good way to minimize their impact on the inverse kinematics outputs (Livet et al., 2023).

For the Picking tasks, based on RGB input videos, the **Body Model** better estimated the anatomical markers, compared to the **Arm Model**, for all learning strategies. The two fine tuning strategies enhanced performance, but the ***Fully*** strategy obtained better improvements: $RMSE_m$ decreased from almost 75% for the **Body Model**, and 40% for the **Arm Model**. However the residual error was still high (26 to 95 mm) compared to the results obtained with the Mocap data (Lifting Task). A first explanation leads to the use of a different HPE in this work compared to the original Opencap paper. Hence, the first input layers have been trained with a slightly different definition and distribution of 3D keypoints. Only the ***Fully*** strategy can retrain the input layer, which is supported by clearly more accurate marker prediction than the ***Outputs*** strategy. Hence, it would be interesting to evaluate a new strategy, denoted ***Inputs*** strategy, that would introduce a new input layer, and retrain the two resulting input layers while freezing the remaining of the architecture. Moreover, in the Picking dataset, the reference data was obtained from XSens motion capture clips, which estimated surface anatomical markers based on inertial sensor data and a calibrated skeleton. Consequently, the way these markers were estimated was different from original motion capture data used by Opencap, to train the models. In addition, similarly to the Mocap data, the nature of the motion itself may lead to a different distribution of input-output samples, and may need additional data to properly handle this new distribution.

To conclude, although fine tuning enables us to significantly decrease the estimation error compared to using inference directly, the results obtained from RGB data for the Picking task do not seem usable as inputs for classical inverse kinematics methods, with average errors going up to 97 mm.

Although we found a larger error in the ***Outputs*** strategy compared to the ***Fully*** strategy in all cases, the number of trainable parameters was much bigger for the ***Fully*** strategy (see table 3). As there is an unbalance between the high number of parameters and the small size of training data, the ***Fully*** strategy may fall to overfitting (Goodfellow et al., 2016). In Alwosheel et al. (2018), authors suggest that a dataset of about 10 times the number of parameters could be enough for classical deep learning training to decrease the risk of overfitting. In our case, this rule was not respected for the ***Fully*** approach whereas it was the case for the ***Outputs*** one. This should be balanced by the fact that fine-tuning does not impact the loss in a similar manner as a full training.

We also have demonstrated that decreasing the size of the training dataset, using only 5 subjects among 11 for training, leaded to more important errors. This drop of performance is especially true for the ***Outputs*** strategy, which is tuning less parameters than the ***Fully*** strategy. This suggests that the number of trial data actually significantly affect the fine tuning performance, and should be considered for future use of this approach.

4.2. Joint angles

For the Lifting tasks based on optoelectronic motion capture data, the results show that the angle prediction error for **Right Hip I/E** reduced more compared to the same angle on the other body side **Left Hip I/E**. Ankle joints estimation error showed small improvements after fine tuning. The estimation error for the forearm P/S exhibited the highest $RMSE_{jc}$ values, indicating greater difficulty in accurately estimating these angles. Since the **Arm Model** was responsible for predicting the elbow and wrist markers, there was a need for targeted improvements in this specific model. Indeed, the input of the marker augmenter lacks of information about the hand and explains the poor accuracy in the estimation of the forearm joint angles. Similarly, the lack of information about the head position is an issue that may impact the lumbar and thoracic joint angles prediction. This issue suggest the development of more advanced HPE methods, able to track additional anatomical markers on the head and the hands of the subjects, which are very relevant information in ergonomics.

Figure 6 illustrate the resulting joint angles obtained from augmented data. Both ***Fully*** and ***Outputs*** show better results than the ***Inference*** method. Compared to previous works, we observed varying levels of accuracy among different joint angle estimation approaches, particularly in the context of walking and bipedal locomotion tasks. Previous work (Kanko, 2020) reported that Theia system has a mean angular error of 6.4°, with a range spanning from 3.3° to 11°. In contrast, Pose2Sim exhibited a mean error of 4.9°, with confidence intervals between 3.1° and 6.6° (Pagnon et al., 2021), indicating a more consistent accuracy compared to Theia, with less estimated degrees of freedom. Similarly, (Needham et al., 2022) reported a mean error of 4.9°, with slightly tighter confidence intervals from 2.9° to 6.0°, underscoring the reliability of their system. Opencap, as described by (Uhlrich et al., 2023), matched Needham's system in mean error (4.9°) and confidence intervals (2.9°–6.0°). In the current study, the joint mean error on the same set of joints was $9.6 \pm 1.6°$ showing a slightly less accurate result but still acceptable in ergonomics for posture assessment (Plantard et al., 2017b; Rodrigues et al., 2022).

Inverse kinematics, as expressed in the current paper, is affected by several factors: soft tissue artifacts (STA), kinematic mismatch due to limited degrees of freedom (DoFs) in the model, experimental marker misplacement,

geometrical calibration of the model, and measurement noise. In addition, marker augmentation through Opencap generates additional uncertainty: the learned augmented anatomical positions are inaccurate, and may be affected by postures far from the ones used to train the model. In our case, these issues may explain that the highest joint angle differences are reached for internal/external rotations, that are the most affected by small uncertainties on the marker positions.

The biomechanical model should be questionned as well. First, the shoulder joint angle errors are very high, but this result should be considered with caution: the glenohumeral joint is modeled with a redundancy of the plane of elevation to avoid Gimbal lock issues that generates an infinity of solutions to get the proper orientation of the humerus with regard to the thoraco-scapular complex. Thus, the reconstruction error remains low, but the algorithm proposes an alternative angle sequence to place the humerus. As well, the outputs of the marker augmentation gives a limited set of information out of the sagittal plane for the trunk, leading in particularly high errors in joint angles quantifying internal/external rotations. All of those restrictions are confirmed by the fact that Opencap was evaluated using mainly lower limbs joint angles.

The calibration of the model should be taken with caution as well. Indeed, the calibrated model is based on the marker augmentation that suffers from the inaccuracy of the segment lengths, issued from the joint centers estimation. Therefore, the calibrated model may be far from the one obtained directly from the motion capture data.

4.3. Applicability in ergonomics and perspectives

Calibration-free approaches for ergonomic assessments (Plantard et al., 2017b) rely on skeletal data that lack the precision necessary for accurate joint angle computation according to ISB standards. These methods also exhibit significant errors during occlusions. However, they offer real-time implementation. In contrast, methods such as Opencap and Pose2Sim, which require calibration, although potentially longer to use, due to the need for precise calibration processes, can better incorporate biomechanical constraints, leading to more accurate assessments. The trade-off between speed and accuracy must be carefully considered when selecting an approach for real-time ergonomic assessment. Furthermore, deploying these systems in industrial contexts requires careful consideration of factors such as the number of camera views and robustness to occlusions. At a minimum, two camera views are recommended to ensure comprehensive coverage and reliability (Uhlrich et al., 2023).

When dealing with real conditions, such as cluttered environments, occlusions, clothes, lighting conditions…, capturing the operator's motion generally leads to sparse and noisy data. In the same way, according to the complexity of the task, the operator biomechanical model may or not have some simplifications. This variability of experimental and modelling conditions may complicate the task of the pre-trained DL Opencap marker augmenter. It may also lead to important errors that may not be compatible with traditional inverse kinematics and dynamics frameworks, such as OpenSim (Delp et al., 2007), Anybody (Damsgaard et al., 2006), or Custom (Muller et al., 2019b). In this paper, the tested markersets were different, and the studied tasks mostly involved upper-limb movements, contrary to the original data used to train the Opencap marker augmenter. To exploit this approach to a new output markerset, the idea supported in this paper is to add a new output layer which contains as neurons as the 3D coordinates of the studied markerset (i.e. 3 times the number of markers). In this paper, in the ***Outputs*** strategy, we proposed to re-train the two output layers (the original and the additional ones), which leads to 3696 parameters for the **Arm Model**, and 19587 parameters for the **Body Model**. (Alwosheel et al., 2018) consider that the ratio between the number of observations and the number of weights of an artificial neural network should be higher than 10 to limit the risk of overfitting. It means that 36960 and 195870 poses should be required to retrain the **Arm Model** and the **Body Model** respectively. Hence, for a new type of motions, it suggests to collect similar ground truth and accurate data in laboratory conditions, using for example IMU-bases or optoelectronic systems. Once these data are collected, the Opencap marker augmenter can be re-trained offline before being used with new on-site Mocap data. We also quantified the decrease of accuracy when using a much smaller set of data for training, demonstrating an important limitation of this DL based approach. For companies which develop such RGB-based ergonomic tools, it involves regularly collecting new data, with ground truth motion capture, to improve their models, or adapt to specific needs of their customers.

The results reported in this paper tend to show that input 3D keypoints obtained with computer vision systems lead to less accurate results compared to using reference Mocap systems. Future works would be needed to evaluate the relevance of applying the same strategy for input data: adding a new input layer which is re-trained according to the new types of inputs. However, this would also require to jointly capture these 3D keypoints with a reference and the on-site systems concurrently. To take the on-site conditions into account (such as occlusions or lighting problems), it would require to move the reference system on-site, which might be difficult. Future works will explore how to optimize the

re-training strategies in this condition. By retraining the input keypoints, we could expect an increase of the accuracy, as improvements observed for the output layers.

Computation time needed for training with such a dataset leads to 16 to 193 minutes according to the conditions and the fine tuning strategy. However, this computation is performed offline, which does not affect the inference computation time used to exploit the re-trained Opencap marker augmenter.

5. Conclusion

This study highlighted the potential of using DL-based methods, such as Opencap, for estimating joint angles from sparse 3D keypoints obtained in industrial conditions. While these methods showed promising results in enhancing sparse 3D video keypoints for inverse kinematics analysis, their generalization capabilities across different types of tasks and markersets remains difficult. The main contribution of this paper is to propose and evaluate methods to retrain Opencap to new experimental conditions, including new poses and new markerset. It provides companies and researchers with guidelines to efficiently adapt Opencap to their motion capture protocols and methods. Our findings indicated that while pretrained models, such as Opencap, could provide valuable insights, they might require fine tuning on task-specific datasets to achieve optimal performance. However, it is important to notice that fine tuning comes with its own set of limitations, such as the risk of catastrophic forgetting (Arora et al., 2019), where the model might lose previously learned information when adapting to new tasks. We showed that retraining the very last output layers only, provides very promising results, with a limited set of examples for training. We also showed that the accuracy of such marker augmenter decreases when using real RGB data and HPE as inputs, compared to reference Mocap data. It opens new questions about the interest of applying the same fine tuning strategy to retrain the first input layers, in order to adapt to new HPE specifications. However, this is more difficult to handle, especially for collecting relevant training data with video. The ability to accurately estimate reliable joint angles from on-site RGB videos opens up new opportunities for research and practical applications to exploit on-site RGB videos to estimate joint torques and forces using standard inverse dynamics framework. Further exploration of fine-tuning techniques and expansion of training datasets could enhance the reliability and applicability of these methods in diverse real-world scenarios.

6. Acknowledgments

This work was partially funded by the Cifre convention N° 0936/2021, Moovency company. This work was supported by French government funding managed by the National Research Agency under the Investments for the Future program (PIA) with the grant ANR-21-ESRE-0030 (CONTINUUM project).

7. Supplementary material

7.1. Inputs and Outputs of Marker Augmenter Models (PDF).

In this section, we describe the 3D keypoints used as inputs in **Opencap, lifting, and picking tasks**, along with their corresponding anatomical markers as outputs.

References

Abobakr, A., Nahavandi, D., Hossny, M., Iskander, J., Attia, M., Nahavandi, S., Smets, M., 2019. RGB-D ergonomic assessment system of adopted working postures. Applied ergonomics 80, 75–88.

Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2018. Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. Journal of choice modelling 28, 167–182.

Arora, G., Rahimi, A., Baldwin, T., 2019. Does an lstm forget more than a cnn? an empirical study of catastrophic forgetting in nlp, in: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, pp. 77–86.

Barone, A.V.M., Haddow, B., Germann, U., Sennrich, R., 2017. Regularization techniques for fine-tuning in neural machine translation. arXiv preprint arXiv:1707.09920 .

Begon, M., Andersen, M.S., Dumas, R., 2018. Multibody kinematics optimization for the estimation of upper and lower limb human joint kinematics: a systematized methodological review. Journal of biomechanical engineering 140, 030801.

Bordes, F., Balestriero, R., Garrido, Q., Bardes, A., Vincent, P., 2022. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. arXiv preprint arXiv:2206.13378 .

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Chen, X., Koskela, M., 2013. Sequence alignment for rgb-d and motion capture skeletons, in: International Conference Image Analysis and Recognition, Springer. pp. 630–639.

Damsgaard, M., Rasmussen, J., Christensen, S.T., Surma, E., De Zee, M., 2006. Analysis of musculoskeletal systems in the anybody modeling system. Simulation Modelling Practice and Theory 14, 1100–1111.

Delp, S.L., Anderson, F.C., Arnold, A.S., Loan, P., Habib, A., John, C.T., Guendelman, E., Thelen, D.G., 2007. Opensim: open-source software to create and analyze dynamic simulations of movement. IEEE transactions on biomedical engineering 54, 1940–1950.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Egeonu, D., Jia, B., 2024. A systematic literature review of computer vision-based biomechanical models for physical workload estimation. Ergonomics , 1–24.

Falisse, A., Uhlrich, S.D., Hicks, J.L., Chaudhari, A.S., Delp, S.L., 2023. Marker data augmentation for robust markerless motion capture, in: XIX International Symposium on Computer Simulation in Biomechanics July 26th–28th 2023, Kyoto.

Fu, Z., Yang, H., So, A.M.C., Lam, W., Bing, L., Collier, N., 2023. On the effectiveness of parameter-efficient fine-tuning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12799–12807.

Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Han, Z., Gao, C., Liu, J., Zhang, S.Q., et al., 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608 .

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Kanko, R.M., 2020. Validation of a markerless motion capture system for human movement analysis. Master's thesis. Queen's University (Canada).

Li, D., Zhang, H., 2021. Improved regularization and robustness for fine-tuning in neural networks. Advances in Neural Information Processing Systems 34, 27249–27262.

Livet, C., Rouvier, T., Sauret, C., Pillet, H., Dumont, G., Pontonnier, C., 2023. A penalty method for constrained multibody kinematics optimisation using a levenberg–marquardt algorithm. Computer Methods in Biomechanics and Biomedical Engineering 26, 864–875.

Lu, T.W., O'connor, J., 1999. Bone position estimation from skin marker co-ordinates using global optimisation with joint constraints. Journal of biomechanics 32, 129–134.

Lund, M.E., Andersen, M.S., de Zee, M., Rasmussen, J., 2015. Scaling of musculoskeletal models from static and dynamic trials. International Biomechanics 2, 1–11.

Mao, Y., Deng, Z., Yao, H., Ye, T., Kawaguchi, K., Zou, J., 2023. Last-layer fairness fine-tuning is simple and effective for neural networks. arXiv preprint arXiv:2304.03935 .

Muller, A., Germain, C., Pontonnier, C., Dumont, G., 2015. A simple method to calibrate kinematical invariants: application to overhead throwing, in: ISBS-conference proceedings archive.

Muller, A., Pontonnier, C., Dumont, G., 2017. Uncertainty propagation in multibody human model dynamics. Multibody System Dynamics 40, 177–192.

Muller, A., Pontonnier, C., Dumont, G., 2019a. Motion-based prediction of hands and feet contact efforts during asymmetric handling tasks. IEEE Transactions on Biomedical Engineering 67, 344–352.

Muller, A., Pontonnier, C., Puchaud, P., Dumont, G., 2019b. Custom: a matlab toolbox for musculoskeletal simulation. Journal of Open Source Software 4, 1–3.

Needham, L., Evans, M., Wade, L., Cosker, D.P., McGuigan, M.P., Bilzon, J.L., Colyer, S.L., 2022. The development and evaluation of a fully automated markerless motion capture workflow. Journal of Biomechanics 144, 111338.

Pagnon, D., Domalain, M., Reveret, L., 2021. Pose2sim: an end-to-end workflow for 3d markerless sports kinematics—part 1: robustness. Sensors 21, 6530.

Plantard, P., Muller, A., Pontonnier, C., Dumont, G., Shum, H.P., Multon, F., 2017a. Inverse dynamics based on occlusion-resistant kinect data: Is it usable for ergonomics? International Journal of Industrial Ergonomics 61, 71–80.

Plantard, P., Shum, H.P., Le Pierres, A.S., Multon, F., 2017b. Validation of an ergonomic assessment method using kinect data in real workplace conditions. Applied ergonomics 65, 562–569.

Puchaud, P., Sauret, C., Muller, A., Bideau, N., Dumont, G., Pillet, H., Pontonnier, C., 2020. Accuracy and kinematics consistency of marker-based scaling approaches on a lower limb model: a comparative study with imagery data. Computer Methods in Biomechanics and Biomedical Engineering 23, 114–125.

Rodrigues, P.B., Xiao, Y., Fukumura, Y.E., Awada, M., Aryal, A., Becerik-Gerber, B., Lucas, G., Roll, S.C., 2022. Ergonomic assessment of office worker postures using 3d automated joint angle assessment. Advanced Engineering Informatics 52, 101596.

Roetenberg, D., Luinge, H., Slycke, P., et al., 2009. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV, Tech. Rep 1, 1–7.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, H., 2024. Continual learning of large language models: A comprehensive survey. arXiv preprint arXiv:2404.16789 .

Simundic, A.M., et al., 2008. Confidence interval. Biochemia Medica 18, 154–161.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Uhlrich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L., 2022. Opencap: 3d human movement dynamics from smartphone videos. biorxiv 1, 1 (2022), 1–48.

Uhlrich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L., 2023. Opencap: Human movement dynamics from smartphone videos. PLoS computational biology 19, e1011462.

Wu, G., Cavanagh, P.R., et al., 1995. Isb recommendations for standardization in the reporting of kinematic data. Journal of biomechanics 28, 1257–1262.

Yasin, H., Ghani, S., Krüger, B., 2023. An effective and efficient approach for 3d recovery of human motion capture data. Sensors 23, 3664.

Yasin, H., Hussain, M., Weber, A., 2020. Keys for action: an efficient keyframe-based approach for 3d action recognition using a deep neural network. Sensors 20, 2226.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Advances in neural information processing systems 27.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proceedings of the IEEE 109, 43–76.

Appendices

A. Inputs and Outputs of Marker Augmenter Models

This section explores the inputs and outputs associated with marker augmenter models, as depicted in figures 1 and 2. Initially, in section A.1, we examine the original inputs and outputs defined by Opencap (Uhlrich et al., 2023). Subsequently, section A.2 presents our adapted inputs and outputs for lifting tasks. Lastly, section A.3 discusses the adapted inputs and outputs for picking tasks. For both lifting and picking tasks, our outputs focus on a subset of the OpenCap marker set. Specifically, markers such as `r_calc`, `r_thigh1`, `r_thigh2`, `r_thigh3`, `L_thigh1`, `L_thigh2`, `L_thigh3`, `r_sh1`, `r_sh2`, `r_sh3`, `L_sh1`, `L_sh2`, `L_sh3`, `RHJC`, and `LHJC` are excluded from both inference error estimation and fine-tuning training and error estimation phases.

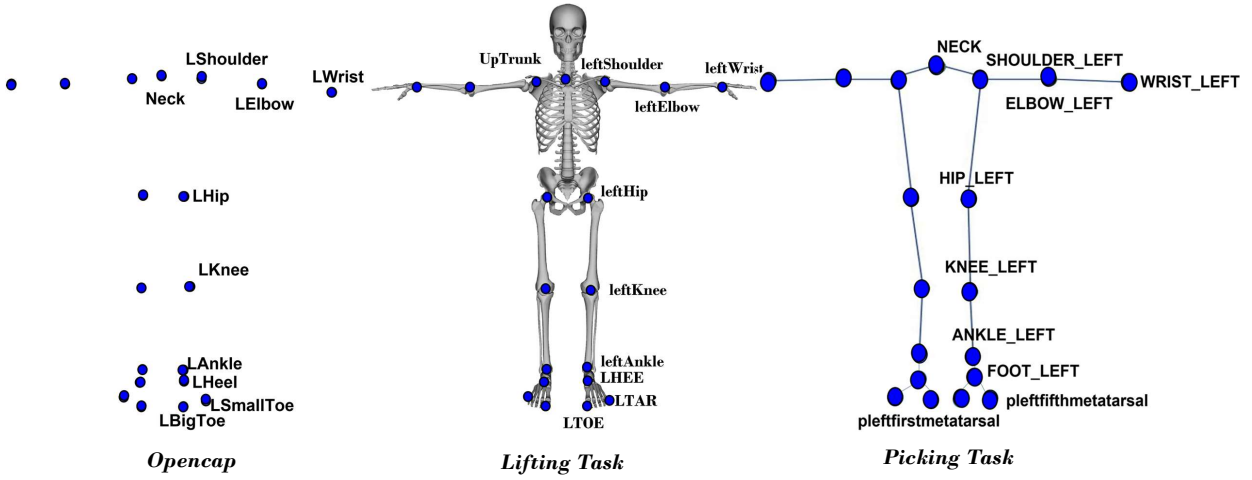

Figure 1: OpenCap, Lifting Task (MoCap data), and Picking Task (RGB data) are compared in terms of their 3D keypoints.

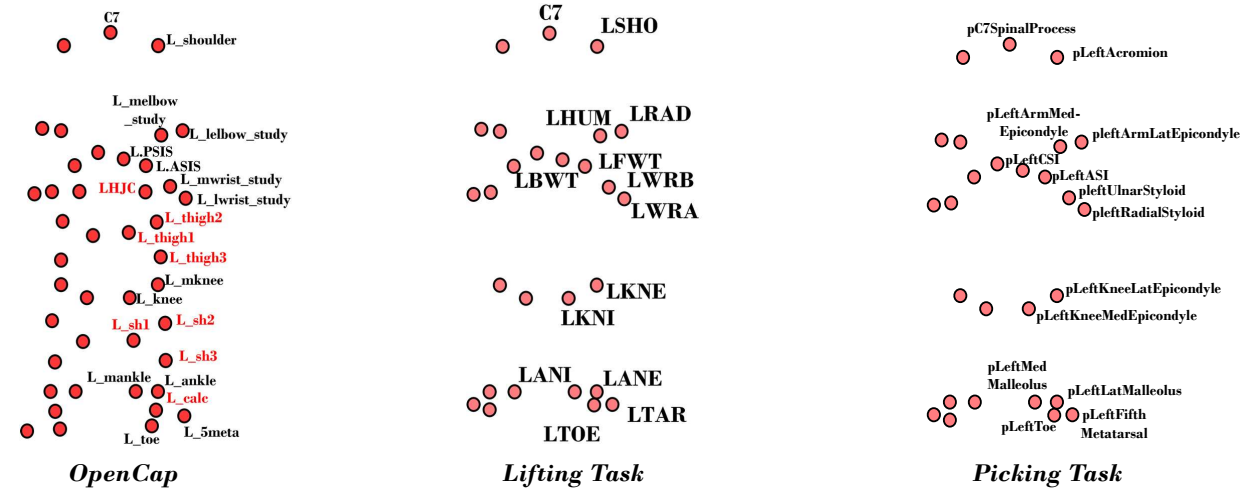

Figure 2: Anatomical markers in **OpenCap, Lifting Task (MoCap data), and Picking Task (RGB data)** are compared. Markers in the OpenCap marker set highlighted in red are excluded from inference error estimation and are not considered during fine-tuning training or error estimation.

A.1. Opencap original inputs/outputs

The method detailed by (Uhlrich et al., 2023) employs a synthetic technique to create datasets by matching 3D video keypoints with corresponding 3D anatomical markers, as depicted in figures 1 and 2. These datasets are derived from 108 hours of motion capture data, which had previously been processed using OpenSim software (Seth et al., 2018) and compiled from various published biomechanics studies. Opencap emphasizes generating 3D anatomical markers using this synthetic approach, based on the same motion capture data and biomechanics studies processed with OpenSim software. For the **Body Model**, the markers included r.ASIS, L.ASIS, r.PSIS, L.PSIS, r_knee, r_mknee, r_ankle, r_mankle, r_toe, r_5meta, r_calc, L_knee, L_mknee, L_ankle, L_mankle, L_toe, L_5meta, L_calc, r_shoulder, L_shoulder, C7, r_thigh1, r_thigh2, r_thigh3, L_thigh1, L_thigh2, L_thigh3, r_sh1, r_sh2, r_sh3, L_sh1, L_sh2, L_sh3, RHJC, and LHJC. For the **Arm Model**, the markers include r_lelbow_study, r_melbow_study, r_lwrist_study, r_mwrist_study, L_lelbow_study, L_melbow_study, L_lwrist_study, and L_mwrist_study.

A.2. Lifting tasks adapted inputs/outputs

To emulate the 3D video keypoints from MoCap data, we implemented the following two steps:

1. Transforming the MoCap data from the world reference frame to the pelvis reference frame.
2. Estimating the 3D keypoints in the Opencap global reference frame (See the figure 1).

1. **Transforming the MoCap data from the world reference frame to the pelvis reference frame:** Our local reference frame was represented by a transformation matrix that converts coordinates from the world reference frame to the pelvis reference frame. The following sub-steps outline the process for defining a local reference frame:

   (a) **Identify the anatomical landmarks:** We used the pelvis as an anatomical landmark to define a local reference frame. This pelvic reference is established by an anatomically accurate local reference frame centered on the pelvis. The anatomical markers **RFWT** (right anterior superior iliac spine), **LFWT** (left anterior superior iliac spine), **RBWT** (right posterior superior iliac spine), and **LBWT** (left posterior superior iliac spine) are utilized as anatomical landmarks. To ensure a consistent local reference frame using anatomical markers that may move during motion analysis, the local reference frame is determined in each frame.

   (b) **Define the local axes:** The X-axis vector is defined as $\mathbf{x} = 0.5\,((\mathbf{LFWT} + \mathbf{RFWT}) - (\mathbf{LBWT} + \mathbf{RBWT}))$. For the Y-axis vector, we first compute $\mathbf{z'} = \mathbf{RBWT} - \mathbf{LBWT}$ and then $\vec{y} = \vec{z'} \wedge \vec{x}$. The Z-axis vector is determined by $\vec{z} = \vec{x} \wedge \vec{y}$. The origin is given by $O = 0.25 \times (\mathbf{LFWT} + \mathbf{RFWT} + \mathbf{RBWT} + \mathbf{LBWT})$. The figure 3 below illustrates the local reference frame details.

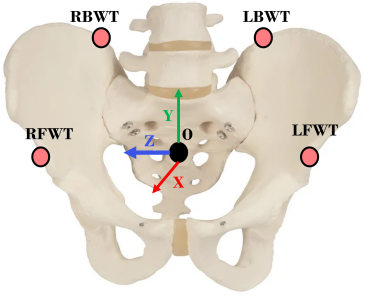

Figure 3: Local reference frame details

   (c) **Calculate the orthonormal vectors:** The transformation matrix can be constructed using the orthonormal vectors that define the pelvis reference frame. Orthonormal vectors indicate the direction of the X, Y, and Z axes within the pelvis reference frame relative to the global reference frame. Our transformation matrix

can be constructed by placing these vectors as columns of a 3×3 matrix. The vectors $\mathbf{f_x} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$, $\mathbf{f_y} = \frac{\mathbf{y}}{\|\mathbf{y}\|}$, $\mathbf{f_z} = \frac{\mathbf{z}}{\|\mathbf{z}\|}$ represent the orthonormal basis vectors for the pelvis reference frame.

(d) **Create the transformation matrix:** To convert marker positions from the global reference frame to the pelvis reference frame, we need to define the transformation matrix that relates the two coordinate systems. The transformation matrix is:

$$ {}^{0}\mathrm{T}_{\mathrm{P}} = \begin{bmatrix} f_{z0} & f_{x0} & f_{y0} & O_z \\ f_{z1} & f_{x1} & f_{y1} & O_x \\ f_{z2} & f_{x2} & f_{y2} & O_y \\ 0 & 0 & 0 & 1 \end{bmatrix} $$

(e) **Apply the transformation matrix:** We applied the transformation matrix ${}^{0}T_{P}$ to convert coordinates from the world reference frame to the pelvis reference frame, represented by ${}^{P}K = {}^{P}T_{0}\,{}^{0}K$. Here, ${}^{P}K$ denotes the marker position in the pelvis reference frame, ${}^{P}T_{0}$ is the inverse of ${}^{0}T_{P}$, and ${}^{0}K$ represents the marker position in the world reference frame.

2. **Estimating the 3D keypoints in the Opencap global reference frame:** After expressing the 3D positions of anatomical markers in the pelvis reference frame, we used the following regression equations to estimate 3D keypoints:
Up trunk *according to* Reed et al. (1999):

$$\begin{aligned} \mathrm{UpTrunk}_z &= \mathrm{C7}_z \\ \mathrm{UpTrunk}_x &= \mathrm{C7}_x + \cos(8 \times \pi/180) \times 0.55 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \\ \mathrm{UpTrunk}_y &= \mathrm{C7}_y + \sin(8 \times \pi/180) \times 0.55 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \end{aligned}$$

Shoulders *according to* Reed et al. (1999):

$$\begin{aligned} \mathrm{rightShoulder}_z &= \mathrm{RSHO}_z \\ \mathrm{rightShoulder}_x &= \mathrm{RSHO}_x + \cos(11 \times \pi/180) \times 0.43 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \\ \mathrm{rightShoulder}_y &= \mathrm{RSHO}_y - \sin(11 \times \pi/180) \times 0.43 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \\ \mathrm{leftShoulder}_z &= \mathrm{LSHO}_z \\ \mathrm{leftShoulder}_x &= \mathrm{LSHO}_x + \cos(11 \times \pi/180) \times 0.43 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \\ \mathrm{leftShoulder}_y &= \mathrm{LSHO}_y - \sin(11 \times \pi/180) \times 0.43 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \end{aligned}$$

Ankles:

$$\begin{aligned} \mathrm{rightAnkle} &= (\mathrm{RANE} + \mathrm{RANI}) \times 0.5 \\ \mathrm{leftAnkle} &= (\mathrm{LANE} + \mathrm{LANI}) \times 0.5 \end{aligned}$$

Knees:

$$\begin{aligned} \mathrm{rightKnee} &= (\mathrm{RKNE} + \mathrm{RKNI}) \times 0.5 \\ \mathrm{leftKnee} &= (\mathrm{LKNE} + \mathrm{LKNI}) \times 0.5 \end{aligned}$$

Hips *according to* Leardini et al. (1999):

$$\mathrm{rightHip}_z = ((\mathrm{LFWT}_z + \mathrm{RFWT}_z) \times 0.5) + 0.38 \times \mathrm{norm}(\mathrm{RFWT} - \mathrm{LFWT})$$

Marker	Definition	Marker	Definition
RFWT	Right anterior superior iliac spine	**LKNI**	Medial condyle of the left femur
LFWT	Left anterior superior iliac spine	**LKNE**	Lateral condyle of the left femur
RBWT	Right posterior superior iliac spine	**LANI**	Left internal malleolus
LBWT	Left posterior superior iliac spine	**LANE**	Left external malleolus
RKNI	Medial condyle of the right femur	**LTOE**	Left acropodion
RKNE	Lateral condyle of the right femur	**LTAR**	Left Ankle I/E folding
RANI	Right internal malleolus	**RSHO**	Right acromion
RANE	Right external malleolus	**LSHO**	Left acromion
RTOE	Right acropodion	**C7**	Spinous process of the 7th cervical
RTAR	Right Ankle I/E folding	**RRAD**	Head of the right radius
RHUM	Medial epicondyle of the right humerus	**LRAD**	Head of the left radius
RWRA	Styloid process of the right radius	**LHUM**	Medial epicondyle of the left humerus
RWRB	Styloid process of the right ulna	**LWRA**	Styloid process of the left radius
LWRB	Styloid process of the left ulna		

Table 1
Definitions of anatomical markers used in MoCap data for lifting tasks.

$$\begin{aligned}
\text{rightHip}_x &= ((\text{LFWT}_x + \text{RFWT}_x) \times 0.5) - 0.31 \times \text{norm}[((\text{LFWT} + \text{RFWT}) \times 0.5) - ((\text{LBWT} + \text{RBWT}) \times 0.5)] \\
\text{rightHip}_y &= ((\text{LFWT}_y + \text{RFWT}_y) \times 0.5) - 0.096 \times [\text{norm}(\text{RANI} - \text{RKNE}) + \text{norm}(\text{RKNE} - \text{RFWT})] \\
\text{leftHip}_z &= ((\text{LFWT}_z + \text{RFWT}_z) \times 0.5) - 0.38 \times \text{norm}(\text{RFWT} - \text{LFWT}) \\
\text{leftHip}_x &= ((\text{LFWT}_x + \text{RFWT}_x) \times 0.5) - 0.31 \times \text{norm}[((\text{LFWT} + \text{RFWT}) \times 0.5) - ((\text{LBWT} + \text{RBWT}) \times 0.5)] \\
\text{leftHip}_y &= ((\text{LFWT}_y + \text{RFWT}_y) \times 0.5) - 0.096 \times [\text{norm}(\text{LANI} - \text{LKNE}) + \text{norm}(\text{LKNE} - \text{LFWT})]
\end{aligned}$$

Elbows:

$$\begin{aligned}
\text{rightElbow} &= (\text{RHUM} + \text{RRAD}) \times 0.5 \\
\text{leftElbow} &= (\text{LHUM} + \text{LRAD}) \times 0.5
\end{aligned}$$

Wrists:

$$\begin{aligned}
\text{rightWrist} &= (\text{RWRA} + \text{RWRB}) \times 0.5 \\
\text{leftWrist} &= (\text{LWRA} + \text{LWRB}) \times 0.5
\end{aligned}$$

After estimating the 3D keypoints, which were initially expressed in a local reference frame, we converted them into the world reference frame and subsequently into the Opencap global reference frame.

For the outputs, the **Body Model** included the following markers: **RFWT, LFWT, RBWT, LBWT, RKNI, RKNE, RANE, RANI, RTOE, RTAR, LKNI, LKNE, LANE, LANI, LTOE, LTAR, RSHO, LSHO, and C7**. For the **Arm Model**, the markers included are: **RRAD, RHUM, RWRA, RWRB, LRAD, LHUM, LWRA, and LWRB**. These are detailed in table 1 and illustrated in figure 2.

A.3. Picking tasks adapted inputs/outputs

To process the unique RGB camera, we utilized the KIMEA Cloud solution developed by Moovency. This enabled us to obtain the 3D keypoints, as illustrated in figure 1.

For the outputs, the markers for the **Body Model** included pRightASI, pLeftASI, pRightCSI, pLeftCSI, pRightKneeMedEpicondyle, pRightKneeLatEpicondyle, pRightLatMalleolus, pRightMedMalleolus, pRightToe, pRightFifthMetatarsal, pLeftKneeMedEpicondyle, pLeftKneeLatEpicondyle, pLeftLatMalleolus, pLeftMedMalleolus, pLeftToe, pLeftFifthMetatarsal, pRightAcromion, pLeftAcromion, and pC7SpinalProcess. For the **Arm Model**,

the markers included pRightArmLatEpicondyle, pRightArmMedEpicondyle, pRightRadialStyloid, pRightUlnarStyloid, pLeftArmLatEpicondyle, pLeftArmMedEpicondyle, pLeftRadialStyloid, and pLeftUlnarStyloid, as shown in figure 2.

References

Leardini, A., Benedetti, M., Catani, F., Simoncini, L., Giannini, S., 1999. An anatomically based protocol for the description of foot segment kinematics during gait. Clinical biomechanics 14, 528–536.

Reed, M.P., Manary, M.A., Schneider, L.W., 1999. Methods for measuring and representing automobile occupant posture. Technical Report. SAE Technical Paper.

Seth, A., Hicks, J.L., Uchida, T.K., Habib, A., Dembia, C.L., Dunne, J.J., Ong, C.F., DeMers, M.S., Rajagopal, A., Millard, M., et al., 2018. Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. PLoS computational biology 14, e1006223.

Uhlrich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L., 2023. Opencap: Human movement dynamics from smartphone videos. PLoS computational biology 19, e1011462.