# A Deep Neural Network-based Method for Estimation of 3D Lifting Motions

**Rahil Mehrizi[1], Xi Peng[2], Xu Xu [5], Shaoting Zhang[6], Kang Li [2,3,4]**

[1]Department of Industrial & Systems Engineering, [2]Department of Computer Science,

[3]Department of Biomedical Engineering, Rutgers University, Piscataway, New Jersey

[4]Department of Orthopaedics, Rutgers New Jersey Medical School, Newark, New Jersey

[5] Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina

[6]Department of Computer Science, University of North Carolina, Charlotte, North Carolina

*Keywords*: 3D pose estimation, machine learning, deep neural network, lifting, biomechanics

*Corresponding author. Tel.: +1-773-701-1722; fax: +1-732-445-5467

*E-mail address*: KL419@rutgers.edu (K. Li)

Word count:  3456

1

# Abstract:

The aim of this study is developing and validating a Deep Neural Network (DNN) based method for 3D pose estimation during lifting. The proposed DNN based method addresses problems associated with marker-based motion capture systems like excessive preparation time, movement obstruction, and controlled environment requirement. Twelve healthy adults participated in a protocol and performed nine lifting tasks with different vertical heights and asymmetry angles. They lifted a crate and placed it on a shelf while being filmed by two camcorders and a synchronized motion capture system, which directly measured their body movement. A DNN with two-stage cascaded structure was designed to estimate subjects' 3D body pose from images captured by camcorders. Our DNN augmented Hourglass network for monocular 2D pose estimation with a novel 3D pose generator subnetwork, which synthesized information from all available views to predict accurate 3D pose. We validated the results against the marker-based motion capture system as a reference and examined the method performance under different lifting conditions. The average Euclidean distance between the estimated 3D pose and reference (3D pose error) on the whole dataset was 14.72±2.96 mm. Repeated measures ANOVAs showed lifting conditions can affect the method performance e.g. 60° asymmetry angle and shoulder height lifting showed higher 3D pose error compare to other lifting conditions. The results demonstrated the capability of the proposed method for 3D pose estimation with high accuracy and without limitations of marker-based motion capture systems. The proposed method may be utilized as an on-site biomechanical analysis tool.

## 1. Introduction:

Work-related Musculoskeletal Disorders (WMSDs) are commonly observed among workers involved in material handling tasks such as lifting (Kuiper et al. 1999, da Costa et al. 2010). To improve work place safety and decrease the risk of WMSD, it is necessary to analyze biomechanical risk exposures associated with these tasks by capturing the body pose and assessing joints kinematics and critical joints stress.

The most common motion capture technique is using marker-based motion capture systems. These systems use reflective markers and a set of synchronized cameras to track the body movements: reflective markers are attached near the subject's joints and the 3D position of each joint is estimated using 3D coordinates of the reflective markers. Marker-based motion capture systems are considered as a reliable and accurate system but their widespread use is limited due to its drawbacks. First, they require a controlled environment to acquire high-quality data; second, attaching markers to the subject's body is time consuming and can also obstruct subject's activities.

Therefore, image-based motion capture techniques, have gained an increasing interest during the past decades and a variety of computer vision and machine learning approaches have been proposed for 3D human motion tracking and pose estimation (Bo et al. 2010, Amin et al. 2013, Zhou et al. 2016). Despite the success of these approaches, they suffer from the fact that they utilize hand crafted image features e.g. HOG (Dalal et al. 2005), SIFT (Müller et al. 2010), etc. With the emergence and advances of deep learning techniques, approaches that employ Deep Neural Networks (DNN) to learn high-level and semantic image features, have become the standard in the domain of vision tasks. DNNs consist of several hidden layers between the input and output layers and are capable of modeling complex non-linear relationships by learning high-level and semantic features from the data. They have achieved growing attention recently due to their high performance for several vision tasks such as face recognition (Daneshzand et al. 2018, Iranmanesh et al. 2018, Iranmanesh et al. 2018), human activity recognition (Baccouche et al. 2011, Yang et al. 2015), and human pose estimation (Peng et al. 2018, Tang et al. 2018, Zhao et al. 2018). Therefore, success of DNNs justifies investigation in other fields such as biomechanical analysis.

Previous literatures explored computer vision and machine learning algorithms and proposed image-based methods for biomechanical analysis. In particular, Corazza et. al. (2006) and Sandau et. al. (2014) developed a generative method to fit a predefined 3D body model to a visual hull constructed from eight cameras. The fitting process was formulated as an optimization problem and they used body part segmentation and least-squares optimization to estimate the joint center positions. The same idea was taken to develop an underwater motion capture system for the analysis of arm movements during front crawl swimming (Ceseracciu et al. 2011). Despite the high accuracy of these methods, they critically rely on background subtraction, which requires a controlled environment and lighting conditions. Furthermore, large number of cameras is needed to construct a precise visual hull surface, which is not always practical in the workplace. Drory et al. (2017), proposed a discriminative method to find a mapping directly from the image features i.e. HOG (Dalal et al. 2005) to body pose parameters by utilizing training data. Their method was tested for full body kinematics estimation of a cyclist and it was shown that it is capable of estimating 2D pose accurately. However; the method performance was not tested for the 3D body pose estimation. In another study by Greene et al. (2018), a video tracking method was developed for classifying lifting postures automatically from video. They used a simple low-level image feature i.e. height and width of the silhouette obtained from background subtraction and applied regression tree algorithm to classify the lifting postures. Their method achieved high classification accuracy for three different lifting postures including squatting, stooping, and standing, however; it was not able to track the body pose over the whole video frames. These studies demonstrate the feasibility of computer vision and machine learning approaches for biomechanical analysis, but it remains unknown if deep learning as the state-of-the-art approach in the vision domain can be employed for this field. Therefore, the primary aim of this study is to investigate the possibility of deep learning network employment for an image-based 3D human pose estimation during lifting tasks. The secondary aim of this study is to determine how lifting conditions can affect the accuracy of the results.

## 2. Methods and Materials:

A lifting dataset comprises of videos and corresponding 3D body joint annotations during various symmetrical lifting tasks were derived in our previous studies (Mehrizi et al. 2017, Mehrizi et al. 2018). The experimental setup of those studies, along with the newly added lifting tasks (asymmetrical lifting), is described in this section.

### 2.1. Participants:

The dataset consists of 12 healthy males (age 47.50±11.30 years; height 1.74±0.07 meters; weight 84.50±12.70 kg) performing various lifting tasks in a laboratory at self-selected speed while being filmed by both camcorders and a synchronized motion tracking system that directly measured the body movement. They lifted a plastic crate (39 × 31 × 22 cm) weighing 10 kg and placed it on a shelf without moving their feet. All lifting trials started with participants standing in front of a plastic crate. The initial horizontal distance of the plastic crate and the lifting speed were chosen by lifters without constraint. They performed three vertical lifting heights ranges from floor to knuckle (FK), knuckle to shoulder (KS) and floor to shoulder (FS). Each vertical lifting range was combined with three asymmetry angles (0˚, 30˚ and 60˚), which was defined as the angle of the end position relative to the starting position of the crate (Fig. 1). A total of 9 lifts (3 lifting heights × 3 asymmetry angles) were performed by each participant in a full-factorial design with random sequence. For each combination of lifting tasks, two repetitions were performed and we used repetition one as training dataset and repetition two as testing dataset. Because two video clips were missed during the data collecting (repetition two of FK lifting with 0˚ and 30˚ asymmetry angles for subject 9), they were excluded from the dataset.

### 2.2. Data Acquisition:

24 Reflective markers were attached to the lifters` body segments (Cappozzo et al. 1995) and 3D positions of markers during the lifting tasks were measured by a motion tracking system with a sampling rate of 100 Hz. The raw 3D coordinate data were filtered with a fourth-order Butterworth low-pass filter at 8 Hz. Two digital camcorders with 720×480 pixel resolution, synchronized with the motion tracking system also recorded the lifting from two views, 90 degree (side view) and 135 degree positions (Fig 1).

----------------------------------------

Figure 1

----------------------------------------

### 2.3. Data Processing:

Videos are down-sampled from 30 fps to 15 fps for both training and testing sets to reduce redundancy and images are extracted from down-sampled videos. All images are cropped to a fixed size (256×256 pixels) and are adjusted such that the subject is located at the center.

3D joints annotation are provided by a motion capture system. We define 14 joint centers including head, neck, and left/right shoulders, elbows, wrists, hips, knees, and ankles, and only use the trajectory of these joints for training the network. The coordinates of each joint is normalized from zero to one over the whole dataset in order to ensure equal weightings across joints.

2D joints annotation are provided by registering 3D joints annotation in motion capture coordinate system, into image coordinates system. If x represents 3D annotation of joint j in motion capture coordinate system and y represents the 2D annotation of the same joint in image, coordinate system, then the following relation holds:

$$x = Cy$$

, where C is the camera matrix. In order to calculate the camera matrix, first for a few images we find 2D joints annotation manually. Then, having the corresponding 3D annotation for the same joints, we solve the above equation and find matrix C. Finally, for the rest of the images, 2D joints annotation can be found using calculated camera matrix (C) and 3D joints annotation available from motion capture system. We refer the reader to this work (Zhang 2000) for more information about the camera matrix calculation.

After data processing, the data structure consists of about 21,000 cropped images and corresponding normalized 3D joints annotation, and 2D joints annotation. Fifty percent of the data are used as the training data and the other fifty percent are saved for the testing data.

### 2.4. Network Architecture

The aim of our proposed DNN is to predict the 3D body from multi-view RGB images. Figure 2 shows an overview of the proposed network, which consists of two subnetworks: a "2D pose estimator" subnetwork and a "3D pose generator" subnetwork. The first subnetwork extracts 2D pose from the input image independently from each view. The estimated 2D poses are concatenated across all the views and are fed into the "3D pose generator” subnetwork to infer the 3D pose.

Inferring a 3D body pose from 2D body pose as the only intermediate supervision is inherently ambiguous. This ambiguity comes from the fact that there are usually multiple 3D poses corresponded to a single 2D pose. In order to overcome this challenge, we propose to apply skip connections from 2D pose estimator subnetwork to 3D pose generator subnetwork (Fig. 2). The idea of skip connection was first introduced by He et. al. (2016) for image recognition. They showed that in a very deep network, the gradients of the network's output with respect to the parameters in the lower layers become very small and as a result the network cannot learn the parameters effectively (gradient vanishing problem). The idea of skip connections is to provide inputs of a lower layer available for a higher layer by adding a shortcut connection and letting the network go back to an earlier time to pick up some information. We apply the same idea in our network by adding several skip connections between two subnetworks, which provide more intermediate cues for 3D pose inference and decrease ambiguity (Mehrizi et al. 2018).

----------------------------------------

Figure 2

----------------------------------------

### *2.4.1. 2D Pose Estimator Subnetwork*

The 2D pose estimator subnetwork takes RGB images as the input and estimates their corresponding 2D pose for each view, independently. The 2D body pose is represented by J heatmaps, where J is the number of body joints. Each value in heatmaps represents the probability of observing a specific joint at the corresponding coordinates (Fig. 3). The advantage of heatmaps over direct regression of joint coordinates is that it handles multiple instances in image and represents uncertainty.

We use Hourglass network (Newell et al. 2016), which has achieved state-of-the-art performance on large scale human pose datasets for 2D pose estimation. Hourglass network comprises of encoder and decoder. The encoder processes the input image with convolution and pooling layers to generate low resolution feature maps and the decoder processes low resolution feature maps with up-sampling and convolution layers to construct the high resolution heatmaps for each joint. More details about the network architecture can be found in the corresponding paper (Newell et al. 2016).

----------------------------------------

Figure 3

----------------------------------------

### *2.4.2. 3D Pose Generator Subnetwork*

The 3D pose generator subnetwork integrates information from multiple views to synthesize 3D pose estimation. The input of this subnetwork is concatenation of the 2D pose estimator subnetwork's outputs for two different views and the output is the 3D pose.

3D pose generator subnetwork is designed as an encoder, which includes convolutional and max-pooling layers (Fig. 4). The network starts with a convolutional layer with 256 channels and filter size of 1 and is followed by a series of max-pooling layers and residual modules. Each residual module consists of three convolutional layers and Rectified Linear Units (ReLUs) are used as the activation function in between each layer. The network ends with a fully-connected layer, which outputs the 3D coordinates of each joint. As mentioned in the previous section, in order to provide more intermediate cues for 3D pose generator subnetwork, we leverage skip connections between two subnetworks by adding the feature maps before each pooling layer of the 2D pose estimator subnetwork directly to the counterpart of the 3D pose generator subnetwork (Fig. 4).

----------------------------------------

Figure 4

----------------------------------------

### 2.5. Training Strategy:

The deep learning platform used in this study is Pytorch and training and testing are implemented on a machine with NVIDIA Tesla K40c and 12 GB RAM. The network is trained in a fully-supervised way with L2 loss function and using Adaptive Moment Estimation (Adam) (Kingma et al. 2014) as the optimization method.

We propose a two-stage training strategy that we found more effective instead of an end-to-end training for the whole network from the scratch. At the first stage, we use pre-trained Hourglass model (Newell et al. 2016) and fine-tune it on our lifting dataset with learning rate of 0.00025 for five epochs. We utilize data augmentation i.e. scaling (0.8-1.2), and rotation (+/- 20 degrees) to add variation into the training dataset and prevent overfitting. After data augmentation, each original image in the training dataset is replaced by its scaled and rotated version and the total number of images (dataset size) is kept fixed. Fine-tuning of this stage takes about 4000 seconds per epoch (20,000 seconds total).

At the second stage, 3D pose generator subnetwork is trained from scratch on our lifting dataset by using two-view images and corresponding normalized 3D pose skeleton. The network is trained with learning rate of 0.0005 for 50 epochs. Training of this stage takes about 800 seconds per epoch (40,000 seconds total).

### 2.6. Data Analysis

The performance of the proposed DNN based method is validated against the marker-based method as a reference. 3D pose error is calculated based on the average Euclidean distance between estimated 3D joints coordinates and corresponding ground-truth data obtained from a motion capture system for all of the joints. Let $\hat{P}_j = [\hat{x}_j, \hat{y}_j, \hat{z}_j]$ and $P_j = [x_j, y_j, z_j]$ represent estimated and ground-truth 3D coordinates of joint j, respectively, then 3D pose error is calculated as follow:

$$Error = {}^{1}\!/_{J} \sum_{j=1}^{J} \|\hat{P}_j, P_j\|$$

, where J is number of body joints and ‖. ‖ shows Euclidean distance.

Furthermore, in order to examine the effect of lifting conditions on accuracy of results, a repeated-measures analysis of variance (ANOVA) test is conducted. We perform a two-way repeated measures ANOVAs with type of lifting condition (height and asymmetry angle) as within subject factors and 3D pose error as dependent variables.

## 3. Results:

Results of applying the proposed DNN based method is given in this section. The accuracy of the estimated 2D pose landmarks obtained from the fine-tuned Hourglass model is around 97.55 % using the standard metric PCKh@0.5 (Andriluka et al. 2014). PCKh@0.5 metric defines a candidate landmark to be correct if it falls within a pixel threshold (50% of the head segment length) of the ground-truth landmark. The accuracy of the estimated 3D pose is measured by comparing the results with those are obtained from marker-based method in terms of 3D pose error. Averaged 3D pose error is 14.72±2.96 mm on the whole dataset (table 1). For qualitative results, we have provided representative 3D poses predicted by the proposed DNN based method in figure 5. It can be seen that even for posture with self-occlusion, the proposed method is able to predict the pose accurately.

In order to examine the effect of lifting conditions on the result accuracy, we have conducted an ANOVA test (table 2). Overall 3D pose error are significantly different between lifting conditions (height and asymmetry angle), but there is not a significant interaction between height and asymmetry angle. Among three different asymmetry angles, 60° has the highest 3D pose error and among lifting heights, the highest error is corresponded to FS (table 1).

Finally, we evaluate the performance of the proposed DNN based method for different body joints separately. As shown in figure 6, head, elbows, and wrists have higher error compare to other joints.

---------------------------------------

Table 1 & 2 & Figure 5 & 6

---------------------------------------

## 4. Discussion:

In this study, we developed and validated a DNN based method for 3D pose estimation during lifting. In agreement with the primary aim of this study, the results showed that the proposed method is capable of estimating 3D body pose with high accuracy from only a multi-view image and without attaching any markers on the subject's body. It makes our proposed method an alternative solution to the marker-based motion capture methods without being constrained to an expensive laboratory with controlled environment conditions or obstructing subject movement by attaching markers. The most important reason for the success of the DNNs is the ability of the network to learn semantic and high level image features from the input data, compare to traditional machine learning algorithms, which require hand crafted image features as an input. DNN has been successfully applied in other biomechanical analysis. For example, Eskofier et. al. (2016) and Camps et. al. (2017) employed a DNN to assess Parkinson's movement disorder. In another study by Hu et. al. (2018), a DNN was utilized for surface and age detection from walking pattern. Our work was completely different with these studies, since they used IMU data, while we used RGB images as the network input. Furthermore, our network design and domains of application of our work was significantly different.

The second aim of this study was investigating the effect of lifting conditions i.e. vertical height and asymmetric angle, on the proposed DNN based method performance. ANOVA results revealed that there is a significant difference in 3D pose error between lifting conditions. Floor to shoulder height lifting and 60˚ asymmetry angle showed the highest 3D pose error. This is likely happening due to the higher pose variation for these lifting tasks. Moreover, most part of the movement in lifting task happens in the sagittal plane, while for 60˚ asymmetry angle lifting, there are small movements in frontal and rotation planes as well. Estimating body joints coordinates in these planes is more difficult considering the position and number of the cameras. It is worth noting that although the error difference from lifting conditions is significant, the magnitude of the error was small for all of the lifting conditions (table 1).

Moreover, we evaluated the performance of our proposed method across different body joints. Among all of the joints, left and right wrists exhibited the highest error (fig. 6). This is attributed to the more self/object occlusion incidence for these joints during lifting e.g. when the subject placed the box on the shelf, lower arms could be blocked either by the shelf or by the torso. This is in agreement with other studies (Drory et al. 2017) showing that it is more challenging to estimate the pose when a human-object interaction exists due to the occlusion. This problem can be addressed by increasing number of cameras, which makes it less possible for a joint to be occluded in all the views.

Besides the advantages of the proposed method there are several limitations that have to be addressed. First, the effect of number and position of cameras was not explored. Camera number and placement can highly influence the accuracy of results, especially in case of self or object occlusion presence. It is likely that using more cameras placed all around the subject could provide higher accuracy for arm joints, which are mostly blocked by the box or torso in the current setup. Second, our study focused on lifting task. Generalization of these results should be done with caution as it is unknown whether and how well this method can be extended for other tasks. However, considering the strength of DNNs and present results, we believe that our proposed method has the ability of generalization and it might only need a simple fine tuning for a new task. Third, the presence of markers on the body may alter the natural appearance of the body and might make the network to be trained to detect only the markers. One option to address this limitation could be covering the markers locations by a pixel mask. Fourth, results showed that large asymmetric lifting angle (60˚) leads to higher 3D pose error, so we cannot exclude the larger error might happen in other asymmetrical liftings with higher trunk rotation. It suggests investigating the proposed method for other tasks in the follow up studies. Finally, one important aspect of the biomechanical analysis for different activities including lifting, is the measurement of internal-external joint rotation. Since in the proposed method, each segment is represented by only two single points (distal and proximal joints), it may not be enough for the measurement of internal-external joint rotation. As a future work, we plan to extend our proposed method to estimate full 3D body mesh, which represents the entire shape of the body with point clusters instead of a small number of single points and makes this measurement possible. The ultimate goal of our research is to provide an on-site biomechanical analysis tool by taking advantages of DNNs and the present study can be considered as a starting point of the research along this direction.

## 5. Acknowledgment

This work was supported in part by NSF (CMMI 1334389, IIS 1451292, IIS 1555408, and IIS

1703883). The lifting data collection was conducted at Liberty Mutual Research Institute for Safety when one of the authors was in Harvard School of Public Health - Liberty Mutual postdoctoral program.

## Conflict of Interest Statement

The authors have no any financial or proprietary interests in the materials described in the article.

## List of Tables:

Table 1- average 3D pose error (mm) for each subject and lifting condition. The first row shows the lifting vertical height and second row presents the asymmetric angle. NA: video clips were missed during the experiment. Standard deviations are shown in the parenthesis.

| Subject | FK | | | KS | | | FS | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0° | 30° | 60° | 0° | 30° | 60° | 0° | 30° | 60° |
| 1 | 13.8<br>(5.6) | 16.3<br>(7.8) | 14.0<br>(4.2) | 14.7<br>(3.2) | 9.5<br>(2.9) | 16.3<br>(8.7) | 13.4<br>(4.4) | 10.6<br>(2.8) | 14.6<br>(4.7) |
| 2 | 12.5<br>(4.4) | 10.4<br>(3.5) | 14.2<br>(6.0) | 10.2<br>(3.4) | 16.8<br>(7.2) | 14.8<br>(8.5) | 16.9<br>(4.1) | 17.0<br>(4.4) | 19.7<br>(8.4) |
| 3 | 13.0<br>(3.5) | 14.6<br>(5.5) | 19.2<br>(7.7) | 15.5<br>(8.3) | 14.7<br>(4.4) | 14.7<br>(3.4) | 24.3<br>(5.2) | 14.7<br>(3.7) | 18.0<br>(5.7) |
| 4 | 17.4<br>(3.0) | 15.6<br>(4.6) | 15.0<br>(4.4) | 20.8<br>(6.5) | 14.5<br>(3.0) | 19.1<br>(6.0) | 19.8<br>(7.3) | 16.8<br>(7.0) | 17.2<br>(4.8) |
| 5 | 13.6<br>(5.4) | 16.0<br>(7.6) | 15.9<br>(8.3) | 11.1<br>(2.1) | 12.2<br>(4.2) | 16.6<br>(6.7) | 12.8<br>(3.4) | 14.7<br>(6.1) | 19.0<br>(8.0) |
| 6 | 12.6<br>(5.1) | 11.0<br>(3.7) | 15.0<br>(3.3) | 15.8<br>(10.6) | 13.7<br>(3.1) | 14.8<br>(5.5) | 15.4<br>(4.8) | 14.2<br>(5.1) | 17.6<br>(4.8) |
| 7 | 15.9<br>(7.6) | 14.3<br>(3.9) | 16.4<br>(7.2) | 9.9<br>(2.8) | 14.6<br>(6.7) | 19.0<br>(10.5) | 12.9<br>(4.2) | 14.2<br>(5.0) | 18.8<br>(7.9) |
| 8 | 12.4<br>(3.7) | 13.4<br>(6.1) | 14.6<br>(4.2) | 10.8<br>(2.3) | 14.8<br>(7.2) | 15.2<br>(6.5) | 13.6<br>(4.6) | 15.1<br>(5.1) | 17.0<br>(6.8) |
| 9 | NA | NA | 16.3<br>(6.5) | 13.0<br>(3.4) | 14.4<br>(3.0) | 16.4<br>(4.8) | 12.2<br>(5.4) | 20.4<br>(8.0) | 21.4<br>(4.8) |
| 10 | 10.8<br>(3.8) | 10.8<br>(3.8) | 13.0<br>(5.6) | 13.1<br>(3.8) | 7.7<br>(2.6) | 10.3<br>(5.1) | 13.6<br>(6.8) | 11.5<br>(5.5) | 12.5<br>(7.9) |
| 11 | 15.3<br>(3.3) | 15.3<br>(3.3) | 14.2<br>(5.2) | 11.9<br>(1.8) | 10.5<br>(2.6) | 11.3<br>(5.9) | 12.6<br>(4.5) | 12.5<br>(4.9) | 14.4<br>(5.4) |
| 12 | 11.1<br>(2.5) | 11.1<br>(2.5) | 18.0<br>(7.5) | 12.8<br>(2.0) | 11.5<br>(3.0) | 16.9<br>(6.2) | 17.4<br>(7.5) | 17.7<br>(5.6) | 14.5<br>(5.2) |
| Average | 13.5 | 13.5 | 15.5 | 13.3 | 12.9 | 15.4 | 15.4 | 14.9 | 17.1 |

Table 2- Outcomes of a two-way repeated measure ANOVAs test for effect of lifting conditions on 3D pose error. Bold numbers indicate significant differences ($p<0.05$). SS= Sum of Squares, DF= Degree of Freedom, MS= Mean square.

| **Factor** | **SS** | **DF** | **MS** | **F** | **Prob>F** |
|---|---|---|---|---|---|
| Vertical height | 76.74 | 2 | 38.37 | 5.38 | **0.0061** |
| Asymmetry angle | 102.35 | 2 | 51.17 | 7.18 | **0.0012** |
| Vertical height × Asymmetry angle | 1.91 | 4 | 0.48 | 0.07 | 0.9916 |
| Error | 691.31 | 97 | 7.13 | | |

## List of Figures:

Figure 1- **Left:** Starting and end position of the crate for the floor to shoulder height lifting task. Top row shows the starting position of the crate and second to fourth rows show the end position of the crate for 0°, 30°, and 60° asymmetric angles, respectively. **Right:** Experimental setup for the simulated lifting tasks. Black dots on the subject's body represents markers which were used for capturing ground-truth motion data. Three of ten used digital cameras of motion tracking system can be seen in this picture. One of two used digital camcorders which was installed on the side view is also shown.

Figure 2- overview of the proposed deep learning network. "2D pose estimator" subnetwork extracts 2D pose from the input image independently from each view. The estimated 2D poses are concatenated across all the views and are fed to the "3D pose generator" subnetwork to infer the 3D pose. Skip connections are applied between "2D pose estimator" subnetwork and "3D pose generator" subnetwork to provide more cues for 3D pose inference and decrease ambiguity.

Figure 3- The input image and corresponding estimated heatmaps for five selected joints. Each value in the heatmaps presents the probability of observing a specific joint at the corresponding coordinates. White dots show the ground-truth 2D joints annotation obtained by registration of 3D joints annotation to image coordinates (section 2.3). The upper left of the image is the origin and the first and second number in the parenthesis show the pixel coordinates in the horizontal and vertical axis, respectively.

Figure 4- Network architecture of the proposed deep learning based method. "3D pose generator" subnetwork starts with a convolutional layer, followed by a series of max-pooling layers and residual modules, and ends with a fully-connected layer. The numbers inside each layer illustrate the corresponding size of the feature maps (number of channels × resolution) for convolutional layers and residual modules and the number of neurons for fully connected layers. Detailed design of residual modules and layer annotations are shown in the right column.

Figure 5- Representative 3D poses predicted by the proposed deep learning based method. Each dashed box represents a scenario; Left: multi-view images, Middle: corresponding estimated 3D pose, Right: corresponding ground-truth 3D pose obtained from the motion capture system.

Figure 6- Average of the individual 3D pose error for different body joints over the whole dataset. Bars show standard deviation.

# References:

Amin, S., M. Andriluka, M. Rohrbach and B. Schiele, 2013. Multi-view Pictorial Structures for 3D Human Pose Estimation. Bmvc, Citeseer.

Andriluka, M., L. Pishchulin, P. Gehler and B. Schiele, 2014. 2d human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on computer Vision and Pattern Recognition.

Baccouche, M., F. Mamalet, C. Wolf, C. Garcia and A. Baskurt, 2011. Sequential deep learning for human action recognition. International Workshop on Human Behavior Understanding, Springer.

Bo, L. and C. Sminchisescu, 2010. "Twin gaussian processes for structured prediction." International Journal of Computer Vision 87(1-2): 28.

Camps, J., A. Samà, M. Martín, D. Rodríguez-Martín, C. Pérez-López, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo and J. Cabestany, 2017. Deep Learning for Detecting Freezing of Gait Episodes in Parkinson's Disease Based on Accelerometers. International Work-Conference on Artificial Neural Networks, Springer.

Cappozzo, A., F. Catani, U. Della Croce and A. Leardini, 1995. "Position and orientation in space of bones during movement: anatomical frame definition and determination." Clinical biomechanics 10(4): 171-178.

Ceseracciu, E., Z. Sawacha, S. Fantozzi, M. Cortesi, G. Gatta, S. Corazza and C. Cobelli, 2011. "Markerless analysis of front crawl swimming." Journal of biomechanics 44(12): 2236-2242.

Corazza, S., L. Muendermann, A. Chaudhari, T. Demattio, C. Cobelli and T. P. Andriacchi, 2006. "A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach." Annals of biomedical engineering 34(6): 1019-1029.

da Costa, B. R. and E. R. Vieira, 2010. "Risk factors for work-related musculoskeletal disorders: a systematic review of recent longitudinal studies." American journal of industrial medicine 53(3): 285-323.

Dalal, N. and B. Triggs, 2005. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE.

Daneshzand, M., M. Faezipour and B. D. Barkana, 2018. "Towards frequency adaptation for delayed feedback deep brain stimulations." Neural regeneration research 13(3): 408.

Drory, A., H. Li and R. Hartley, 2017. "A learning-based markerless approach for full-body kinematics estimation in-natura from a single image." Journal of Biomechanics 55: 1-10.

Eskofier, B. M., S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken and T. Kautz, 2016. Recent machine learning advancements in sensor-based mobility analysis: deep learning for Parkinson's disease assessment. Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE.

Greene, R. L., Y. H. Hu, N. Difranco, X. Wang, M.-L. Lu, S. Bao, J.-H. Lin and R. G. Radwin, 2018. "Predicting Sagittal Plane Lifting Postures From Image Bounding Box Dimensions." Human factors: 0018720818791367.

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

Hu, B., P. Dixon, J. Jacobs, J. Dennerlein and J. Schiffman, 2018. "Machine learning algorithms based on signals from a single wearable inertial sensor can detect surface-and age-related differences in walking." Journal of Biomechanics.

Iranmanesh, S. M., A. Dabouei, H. Kazemi and N. M. Nasrabadi, 2018. "Deep cross polarimetric thermal-to-visible face recognition." arXiv preprint arXiv:1801.01486.

Iranmanesh, S. M., H. Kazemi, S. Soleymani, A. Dabouei and N. M. Nasrabadi, 2018. "Deep Sketch-Photo Face Recognition Assisted by Facial Attributes." arXiv preprint arXiv:1808.00059.

Kingma, D. P. and J. Ba, 2014. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980.

Kuiper, J. I., A. Burdorf, J. H. Verbeek, M. H. Frings-Dresen, A. J. van der Beek and E. R. Viikari-Juntura, 1999. "Epidemiologic evidence on manual materials handling as a risk factor for back disorders: a systematic review." International Journal of Industrial Ergonomics 24(4): 389-404.

Mehrizi, R., X. Peng, Z. Tang, X. Xu, D. Metaxas and K. Li, 2018. "Toward Marker-free 3D Pose Estimation in Lifting: A Deep Multi-view Solution." arXiv preprint arXiv:1802.01741.

Mehrizi, R., X. Peng, X. Xu, S. Zhang, D. Metaxas and K. Li, 2018. "A Computer Vision Based Method for 3D Posture Estimation of Symmetrical Lifting." Journal of Biomechanics.

Mehrizi, R., X. Xu, S. Zhang, V. Pavlovic, D. Metaxas and K. Li, 2017. "Using a marker-less method for estimating L5/S1 moments during symmetrical lifting." Applied ergonomics 65: 541-550.

Müller, J. and M. Arens, 2010. Human pose estimation with implicit shape models. Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams, ACM.

Newell, A., K. Yang and J. Deng, 2016. Stacked hourglass networks for human pose estimation. European Conference on Computer Vision, Springer.

Peng, X., Z. Tang, F. Yang, R. S. Feris and D. Metaxas, 2018. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Sandau, M., H. Koblauch, T. B. Moeslund, H. Aanæs, T. Alkjær and E. B. Simonsen, 2014. "Markerless motion capture can provide reliable 3D gait kinematics in the sagittal and frontal plane." Medical Engineering and Physics 36(9): 1168-1175.

Tang, Z., X. Peng, S. Geng, L. Wu, S. Zhang and D. Metaxas, 2018. Quantized densely connected U-Nets for efficient landmark localization. European Conference on Computer Vision (ECCV).

Yang, J., M. N. Nguyen, P. P. San, X. Li and S. Krishnaswamy, 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. IJCAI.

Zhang, Z., 2000. "A flexible new technique for camera calibration." IEEE Transactions on Pattern Analysis and Machine Intelligence 22.

Zhao, L., X. Peng, Y. Tian, M. Kapadia and D. Metaxas, 2018. Learning to forecast and refine residual motion for image-to-video generation. European Conference on Computer Vision, Springer, Cham.

Zhou, X., M. Zhu, S. Leonardos, K. G. Derpanis and K. Daniilidis, 2016. Sparseness meets deepness: 3D human pose estimation from monocular video. Proceedings of the IEEE conference on computer vision and pattern recognition.