Predicting 3D Lower-back Joint Load in Lifting: A Deep Pose Estimation Approach

R. Mehrizi, X. Peng, D. Metaxas, X. Xu, S. Zhang, and K. Li*

***Abstract*— *Goal:* Lifting is a common manual material handling task performed in the workplaces. It is considered as one of the main risk factors for Work-related Musculoskeletal Disorders (WMSDs). An important criterion to identify the unsafe lifting task is the values of the net force and moment at L5/S1 joint. These values are mainly calculated in a laboratory environment, which utilizes marker-based sensors to collect 3D information and force plates to measure the external forces and moments. However, this method is usually expensive to setup, time-consuming in process, and sensitive to the surrounding environment. In this study, we propose a Deep Neural Network (DNN) based framework for 3D pose estimation, which address aforementioned limitations and we employ the results for L5/S1 moment and force calculation. *Methods:* At the first step of the proposed framework, full body 3D pose is captured using a DNN, then at the second step, estimated 3D body pose along with the subject's anthropometric information is utilized to calculate L5/S1 join's kinetic by a top-down inverse dynamic algorithm. *Results:* To fully evaluate our approach, we conducted experiments using a lifting dataset consists of twelve subjects performing various types of lifting tasks. The results are validated against a marker-based motion capture system as a reference. The grand mean±SD of the total moment/force absolute errors across all the dataset was 9.06 ±7.60 Nm/4.85±4.85 N. *Conclusion:* The proposed method provides a reliable tool for assessment of the lower back kinetics during lifting and can be an alternative when the use of marker-based motion capture systems is not possible.**

***Index Terms* Deep neural network, Lower back loading, Lifting, Occupational biomechanics**

I. Introduction

Work-related Musculoskeletal Disorders (WMSDs) are commonly observed among the workers involved in material handling tasks such as occupational lifting. In an epidemiology study by Manchikanti et. al. [1] it was found that heavy lifting is a predictors of future back pain. Kuiper et. al. [2] and Da Costa et. al. [3] also showed with reasonable evidences that lifting is one of the main risk factors for lower back, hip and knee WMSD.

To improve work place safety and decrease the risk of WMSDs, it is necessary to analyze biomechanical risk exposures associated with these tasks by capturing the body pose and assessing the critical joint stresses in order to compare the result with the limit of a person's capacity. In recent years, several systems were developed to capture the 3D body pose and assess the movement of workers. These systems can roughly be categorized into two groups: direct measurement and observational systems [4].

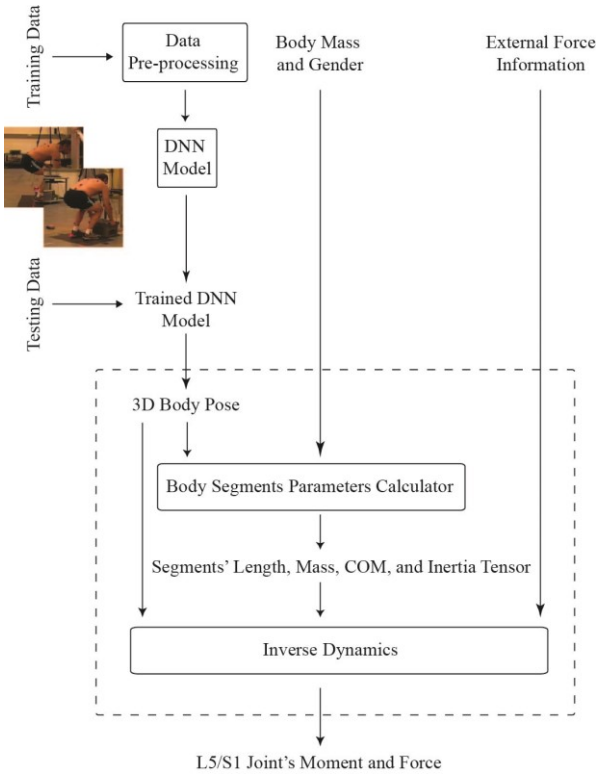


Fig. 1. Workflow of the proposed DNN based method.

Numerous studies have investigated lower back stress using direct measurement methods. There are reported values for a variety of tasks like lifting [5-10], balance recovery movement [11], and gait [12]. Direct measurement systems require markers or sensors attachment on the subject's body and are performed in a laboratory environment. They can provide reliable and accurate estimation of the joints kinematics and are considered as the established state-of-the-art for human motion analysis [13]. However, these methods are limited since they require expensive equipment, controlled environment, and can

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".

R. Mehrizi and D. Metaxas are with Rutgers, The State University of New Jersey, Piscataway, NJ, USA. , X. Peng is with Binghamton University, Vestal, NY, USA. X. Xu is with North Carolina State University, Raleigh, NC, USA. S. Zhang is with the University of North Carolina, Charlotte, NC, USA. *K. Li is with New Jersey Medical School, Rutgers, The State University of New Jersey, Newark, NJ, USA and University of Electronic Science and Technology of China, Chengdu, China (correspondence e-mail: kl419@rutgers.edu).

obstruct the subject's natural patterns of movements due to interference with musculoskeletal structures.

Observational systems like video-based coding system, on the other hand, use recorded videos of the subject and extract a few key frames from them. Then, raters estimate the body pose by making an optimal fit of a predefined digital manikin to the selected video frames. Finally, using the estimated body pose data and time information extracted from the videos, joints trajectory is generated for the entire task by applying a motion pattern prediction algorithm [14]. Xu et. al. [15] presented a video-based coding system to estimate 3D L5/S1 joint moments of lifting task on the basis of videos clips. Coenen et. al. [16] validated two versions of video analysis methods for estimation of the peak moment of the back load during lifting tasks. The results accuracy of the observational systems method rely on the experience of the rater, especially when joints angle become close to the posture boundaries [17]. Furthermore, they can easily become laborious as the number of key frames increases [18].

Advances in the field of computer vision, offer marker-less motion capture systems to overcome the limitations of the direct measurement and observational methods for biomechanical analysis. Even though, marker-less methods are considered as a potential substitute for the traditional marker-based method, they are not widely studied for biomechanical and clinical applications, which require higher accuracy and robustness in comparison with the other applications [13, 19]. There are few studies which proposed marker-less methods for biomechanical and clinical applications. In particular, [20] used the Microsoft Kinect for assessment of joints angles and calculating stride time for gait analysis. In another study by [21], the Microsoft Kinect depth sensor was used for assessing spinal loading during twenty different actions. Despite the acceptable accuracy of these methods, there are two major disadvantages of using the Microsoft Kinect for workplace activity assessment. First, depth sensors can be only used in a short range of distance from the depth sensors, which may not allow them to be used in large space workplaces [22]. Second, depth sensors are sensitive to the environment illumination and would be difficult to use in outdoor environments [23]. In another study by [24] a marker-less framework was proposed to estimate human pose from RGB images and without depth information. They employed a discriminative method to learn a mapping from image features to the 3D body posture and improved the results accuracy by adding morphological constraints. The results were then employed for lower back loads estimation [25]. Using neural network for physical effort assessment in manual material handling task was proposed in a work by Zurada et. al. [26] They used a neural network method that takes several input variables included number of lift per hour, peak moment, etc. and classifies industrial jobs as low or high risk for low back disorders. Davis et. al. [27], employed a hybrid neuro-fuzzy system [28] to estimate spine loads during sagittal lifting. They used a fuzzy technique [29] to identify key input variables of the models and then fed the input variables to a neural network model to translate them into EMG signals. Hou et. al. [30] proposed a recurrent fuzzy neural network to predict spine forces directly from kinematics data and without EMG measurements. These studies demonstrate the feasibility of computer vision and neural network approaches for the biomechanical analysis. However, they are limited to a few types of motions and lifting as one of the most common motions in the workplaces and as an important risk factor for WMSD is not well studies. Additionally, deep learning, which is considered as the state of art approach in the domain of the vision tasks is not studied for the field of biomechanical application.

In this study, we propose and validate a Deep Neural Network (DNN) based framework to estimate 3D L5/S1 joint kinetic (i.e. force and moment) during lifting. The workflow of the proposed method is summarized in Fig. 1. The proposed method uses advanced computer vision approaches, in particular DNN, to estimate the 3D body pose from a two-view image taken by optical cameras. The estimated 3D body pose along with the subject's anthropometric information is then utilized to calculate joins' kinetic by an inverse dynamic algorithm. Since our proposed method eliminates the need of attaching markers onto the subjects' body segments or hiring raters to estimate the pose, it can overcome the limitations of direct measurement and observational systems. The results of our proposed method were compared with results obtained from a marker-based motion capture system as a reference and it was shown that the proposed method achieves promising results and can open new possibilities of deep learning application for biomechanical analysis with the aim of reducing WMSD in workplaces. To summarize our contributions are:

- We propose a DNN-based method to estimate accurate 3D pose from multi-view images.
- The proposed method is validated for L5/S1 joint kinetics estimation for well controlled symmetrical and asymmetrical lifting tasks.

II. MATERIALS

A. Participants and Procedure

A group of 12 healthy males (age 47.50±11.30 years; height 1.74±0.07 meters; weight 84.50±12.70 kg) participated in the experiment. Each participant performed various symmetric and asymmetrical lifting trials in a laboratory while being filmed by both camcorder and a synchronized motion tracking system that directly measured the body movement. All the participants wore black shorts. They lifted a plastic crate (39 × 31 × 22 cm) weighing 10 kg and placed it on a shelf without moving their feet. All the lifting trials started with the subjects standing in front of a plastic crate. The initial horizontal distance of the plastic crate and the lifting speed were chosen by the lifters without constraint. They performed three vertical lifting ranging from floor to knuckle (FK), knuckle to shoulder (KS) and floor to shoulder (FS) heights. Each vertical lifting range was combined with three end-of-lift angles (0, 30 and 60 degree), which is defined as the angle of the end position relative to the starting position of the box. A total of 9 lifts (3 lifting heights × 3 end-of-lift angles) were performed by each participant in a full-factorial design with random sequence.

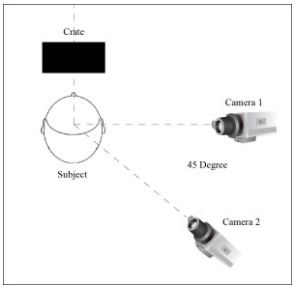


Fig. 2. Cameras position in the experiment setup

B. Data Acquisition

45 Reflective markers were attached to the lifters` body segments and 3D coordination of markers during the lifting tasks were measured by a motion tracking system (Motion Analysis, Santa Rosa, CA) 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 (GR-850U, JVC) with resolution 720×480 pixel, synchronized with the motion tracking system also recorded the lifting from two views, 90 degree (side view) and 135 degree positions (fig. 2). For asymmetrical lifting trials, participants turned away from the side view camera.

III. METHODS

In this work, we aim to predict the 3D L5/S1 joint kinetic from the multi-view RGB images. We proposed a DNN based framework for this purpose whose inputs are videos taken from two different views around the subject, and the output is the L5/S1 joint's force and moment values. As shown in fig. 1, in the offline phase, the training dataset is preprocessed and used to train a DNN to estimate the 3D body posture i.e. 3D joints center coordination. In the online phase, testing dataset is introduced into the trained DNN, and estimated 3D body posture along with the subject's anthropometric information are utilized to calculate body segments parameters. Finally, L5/S1 joint kinetic is determined by a top-down inverse dynamic algorithm according to the estimated 3D body posture and body segments parameters.

A. Data Pre-processing

In order to prepare the data for the proposed deep learning method, images are extracted from videos. Each video includes 200 frames with 30 fps rate. We down-sampled the video from 30 fps to 15 fps for both the training and testing sets to reduce redundancy. All of the images are adjusted to 256×256 pixels and are cropped such that the subject is located at the center.

3D joints annotation are provided by a motion capture system. We selected 24 markers to define 15 joint centers including head, neck, left/right shoulder, left/right elbow, left/right wrist, left/right hip, left/right knee, left/right ankle, and L5/S1 joint and only used the trajectory of these joints for training the network. The coordination of each joint is normalized from zero to one over the whole dataset. Given the camera parameters, 2D joints coordination are also calculated for each image. After pre-processing, the data structure consists of the cropped images and corresponding 2D joints annotation and normalized 3D joints annotation.

B. DNN Model

With the emergence and advances of deep learning techniques, approaches that employ deep convolutional neural networks to learn the image features, have become the standard in the domain of the vision tasks. DNN approaches have achieved the highest performance for several vision tasks such as human activity recognition [31, 32], face recognition[33, 34], and human pose estimation [35, 36]. In this study, the aim of the DNN model is to predict the 3D body pose (3D coordination of the body joints) from multi-view RGB images. Fig. 3 shows the architecture of the proposed DNN model, which consists of two networks: a "2D pose estimator" network and a "3D pose generator" network. The first network extracts both shape (2D pose) and hierarchical texture feature map independently from each view, while the second network synthesizes these information from all available views to generate the 3D pose. The DNN model has been reported in detail elsewhere [35]. To familiarize the reader, we will briefly explain each network.

1) 2D Pose Estimator Network

The 2D pose estimator network takes the RGB images as input and estimates its corresponding 2D pose for each view, independently. The 2D body pose is represented by J heatmaps, where J is the number of joints of the body. Each value in the heatmaps represents the probability of observing a specific joint at the corresponding coordination (fig. 4). The advantage of the heatmaps over direct regression of joint coordination is that it handles multiple instances in image and represents uncertainty.

We use Hourglass network [36], which has achieved state-of-the-art performance on large scale human pose datasets for 2D pose estimation. As shown in fig. *3*, Hourglass network [36] 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. In order to prevent the loss of high resolution information in the encoder, the feature maps before each pooling layer, which shares hierarchical texture feature, are directly added to the counterpart in the decoder. More details about the network architecture can be found in the corresponding paper [36].

Given an input RGB image for view i ($\boldsymbol{x}^i \in \mathbb{R}^{W\times H\times 3}$), then 2D pose estimator network (f) for i-th view is a mapping as follow:

$$\left(\left\{h_1^i, \dots, h_J^i\right\}, \left\{t_1^i, \dots, t_S^i\right\}\right) = f(x^i), \quad (1)$$

where $t_s^i \in \mathbb{R}^{W_s\times H_s\times L_s}$ $\{s = 1, \dots, S\}$ is s-th texture feature map for view i, and $h_j^i \in \mathbb{R}^{W_h\times H_h\times L}\{j = 1, \dots, J\}$ is j-th joint heatmap for view i. The network parameters are learned by minimizing the loss function defined by a pixel-wise heatmap loss:

$$\mathcal{L}_{2d}^i = \frac{1}{J}\sum_{j=1}^{J}\left\|h_j^i - \hat{h}_j^i\right\|, \quad (2)$$

where $\|.\|$ is Euclidean distance and h_j^i is rendered from the ground truth 2D pose through a Gaussian kernel with mean equal to the ground truth and variance one.

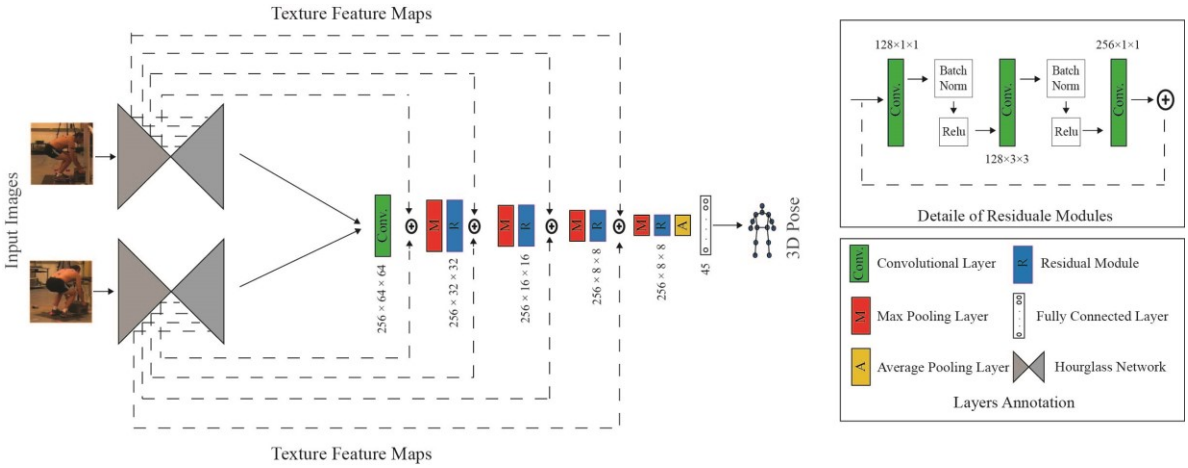


Fig. 3. **Left:** Deep neural network architecture: input images go through 2D pose estimator network and turn into 2D joint heatmaps and hierarchical texture feature maps. . 2D pose estimator architecture is similar to Hourglass network [36] and compromises of encoder and decoder. 2D joints heatmaps are processed in the 3D pose generator network and hierarchical skip connections are summed at specific layers of 3D pose generator network. 3D pose generator architecture is similar to encoder part of the Hourlass network [36], which includes max-pooling layers and residual learning modules. The number on 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. **Right:** Residual learning modules design. The number on each layer illustrate number of channels × filter size. Stride is equal to one in the whole residual modules.

Since the hierarchical texture feature maps of the network share useful information in different scales, they allow for a richer gradient signal and can provide more 3D cues compare to using only heatmaps [37]. So, we propose to employ them for a more efficient 3D generating by feeding them along with the 2D heatmaps to 3D pose generator network (fig. 3).

2) 3D Pose Generator Network

The purpose of the 3D pose generator network is to integrate information from multiple views to synthesize 3D pose estimation. The input of this network is the concatenation of the outputs of the 2D pose estimator network for N different views and the output is the 3D pose.

We propose a bottom up data driven method that directly generates the 3D pose skeleton from the outputs of the 2D pose estimator network. 3D pose generator network is designed as an encoder similar to the first part of the 2D pose estimator network, which includes max-pooling layers and residual learning modules [38] (fig. 3). Each 3D pose skeleton $p \in \mathbb{R}^{3\times J}$ is defined as a set of joints center coordination in 3D space. So 3D pose generator network (g) is a mapping as follow:

$$(\hat{p}) = g(\mathrm{C}(h_1^{\mathrm{i}}, \dots, h_{\mathrm{J}}^{i})_{i=1}^{N}, C\left(t_1^{\mathrm{i}}, \dots, t_{\mathrm{S}}^{i}\right)_{i=1}^{N}), \quad (3)$$

where $C(.)_{i=1}^{N}$ shows the concatenation across the views.

Knowing that 3D joints center coordination are available for the training dataset by means of a maker-based motion capture system, network parameters are learned by minimizing the loss between the available 3D joints center coordination and the corresponding estimated values as:

$$\boldsymbol{L_{3d}} = \mathbf{1}/\boldsymbol{J} \sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} \|\boldsymbol{p_j} - \hat{\boldsymbol{p}}_{\boldsymbol{j}}\|, \quad \mathbf{(4)}$$

where p_j and $\hat{p}_j$ are ground truth and estimated 3D coordination of joint center j respectively.

3) Training Strategy

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 fine-tuned the 2D pose estimator network on our lifting dataset with learning rate of 0.00025 for five epochs (750 iterations per epoch). At the second stage, 3D pose generator network was trained from scratch on our lifting dataset by using two-view images and corresponding normalized 3D pose skeleton. The models were trained with learning rate of 0.0005 for 50 epochs (900 iterations per epoch). In both stages, all lifting trials of subjects 1 to 10 were used as training dataset and all lifting trials of subjects 11 and 12 as testing dataset.

C. Body Segments Parameters Calculator

We define the human body with 11 body segments including head, trunk, pelvis, upper arms, forearms, thighs, and shanks. Distal and proximal joints of each segment are defined based on the approaches proposed by [39]. Given 3D coordination of the joints center, subject's gender, and total body mass, all of the body segment parameters including segments length, mass, position of the center of mass (COM), and inertia tensor are calculated based on the suggested values by [39].

The length of the segment i (l_i) is calculated as the Euclidean distance between its corresponding distal and proximal joint centers. Let M be the subject's total mass, and m_i be the segment i mass, then:

$$m_i = \bar{r}_i^m \times M, \quad (5)$$

where $\bar{r}_i^m$ is the mean relative mass of the segment i, given in the literatures [39]. The 3D position of the segment i's COM (com_i) is located on the line that connects its corresponding distal $(p_{ds(i)})$ and proximal $(p_{pr(i)})$ joint center and can be calculated based on the mean longitudinal distance of the COM from its proximal joint center $(\bar{r}_i^{cm})$[39], as follow:

$$com_i = p_{pr(i)} + \bar{r}_i^{cm} \times (p_{ds(i)} - p_{pr(i)}). \quad (6)$$

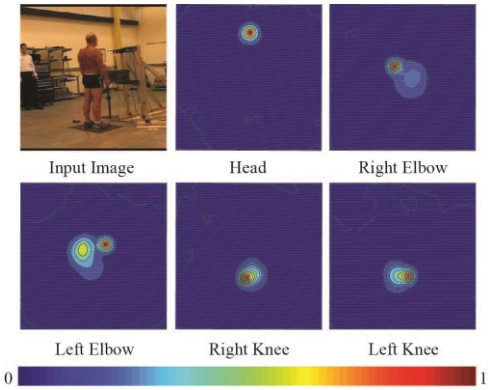


Fig. 4. The input image and corresponding heatmaps for five selected joints. Each value in the heatmaps presents the probability of observing a specific joint at the corresponding coordination.

Finally, the inertial tensor of the segment i (I_i), can be calculated as follow:

$$I_i = m_i \times (l_i \times \bar{r}_i)^2, \tag{7}$$

where $\bar{r}_i = [\bar{r}_i^x, \bar{r}_i^y, \bar{r}_i^z]$ is the mean relative radius of gyration of the segment i about each axis [39].

D. Inverse Dynamics

To calculate the joints kinetic information from the estimated joints kinematic information (position, velocity, and acceleration), a top-down inverse dynamics model [40] was used. A global equation of motion was applied to estimate the net forces (F_{L5S1}) and moments (M_{L5S1}) at L5/S1 joint in the global coordinate system, as described by [40]:

$$F_{L5S1} = -F_r - \sum_{i=1}^{k} m_i g + \sum_{i=1}^{k} m_i a_i \tag{8}$$

$$M_{L5S1} = -(r_r - r_{L5s1}) \times F_r - \sum_{i=1}^{k} [(r_i - r_{L5s1}) \times m_i g] + \sum_{i=1}^{k} [(r_i - r_{L5s1}) \times m_i a_i] + \sum_{i=1}^{k} (I_i \propto_i), \tag{9}$$

where r_r and r_{L5S1} are the vectors to the position of the external force and L5/S1 joint respectively, and F_r is the external force vector. r_i is the vector to the COM of segment i, k is the number of segments of the upper body up to L5/S1 joint (i.e. head, trunk, upper arms, and forearms), and a_i and $\propto_i$ are the linear and angular acceleration vectors of the COM of segment i, respectively. As it can be seen in the (8) and (9), in order to calculate F_{L5S1} and M_{L5S1}, external force information are required. In the top-down model, external forces information can be calculated based on the mass and acceleration of the box. In bottom-up model, on the other hand, force plates data can be used to measure the external forces, external moments and their points of application. So using a top-down model instead of a bottom-up model for the inverse dynamics process seems more practical for an on-site biomechanical analysis, since it removes the need for the force plates [25].

IV. DATA ANALYSIS

A. Validation

The performance of our proposed method is validated against the reference in terms of the accuracy of estimated 3D L5/S1 joint moment and force values. The validation is performed by calculating Root Mean Squared Error (RMSE) and Pearson's correlation coefficient (R). Furthermore, for each of the lifting trial, absolute peak values over the whole lifting cycle was extracted from estimated L5/S1 moment and force series and was compared to the corresponding values obtained by the reference using RMSE and R. Finally, for absolute peak values of all lifting trials together, intra class correlation coefficients (ICC) were calculated. For all of the ICC calculation, ICCs less than 0.40 were assumed poor, ICCs between 0.40 to 0.75 were good and ICCs greater than 0.75 were considered excellent [41]

B. Lifting Cycle Normalization

To evaluate the performance of the proposed method, independent of the subjects, estimated forces and moments were normalized with respect to the body mass and body mass × stature, respectively [12]. However, in order to make the kinetic values more clinically-meaningful, normalized kinetic values were multiplied by mean body mass and mean body × stature mass across subjects [42]. Finally, all kinetic values were time-normalized to a 100% of a lifting cycle. The lifting cycle is defined as the time that a subject grabs the box to the time that the box is left on the shelf.

V. RESULTS

A. L5/S1 Joint Moment Time Series

Results show a good agreement between the estimated L5/S1 joint moments in each of the three planes and the references. The grand mean (±SD) of the total moment absolute errors across all the subjects and trials was 9.06 (±7.60) Nm. Fig. 5 presents a typical example of a lifting trial, showing the L5/S1 joint moment time series calculated based on the proposed DNN based method and the reference. For dominant moment component (sagittal moment), R coefficient for all lifting trials were high (mostly above 0.95) and RMSE were small (mostly below 20 Nm) (table 1). For non-dominant L5/S1 moment components (lateral and rotation moment) on the other hand, R values were lower than dominant moment component. However, the RMSE were also small (less than 10 Nm). This likely happens due to a smaller moment in lateral and rotation planes during lifting, which leads to a small moment variance in this plane

Figure 6 presents the average and standard deviation of the total moment across the subjects for each of the nine lifting tasks. It shows a good fit of the proposed method with the reference for all of the lifting tasks with no evidence of systematic overestimation or underestimation. Standard deviation across the subjects are also in a good agreement by the reference.

TABLE 1
COMPARISON OF THE L5/S1 JOINT'S KINETICS BETWEEN THE PROPOSED DNN BASED METHOD AND THE REFERENCE FOR EACH LIFTING TRIAL, SUBJECT, AND PLANE SEPARATELY. LAT.= LATERAL, SAG.= SAGITTAL, ROT.= ROTATION, ANT-POST= ANTERIOR POSTERIOR, MED-LAT= MEDIOLATERAL, VERT.= VERTICAL. LIFTING TRIALS ARE SHOWN AS THEIR "VERTICAL LIFTING RANGE _ END OF LIFT ANGLE". RMSE= ROOT MEAN SQUARED ERROR, SD=STANDARD DEVIATION OF THE ERROR. R= PEARSON'S CORRELATION COEFFICIENT VALUES. S11: SUBJECT11, S12: SUBJECT 12.

		L5/S1 Joint Moment									L5/S1 Joint Force								
Plane		Lat.	Sag.	Rot.	Lat.	Sag.	Rot.	Lat.	Sag.	Rot.	ant-post	med-lat	Vert.	ant-post	med-lat	Vert.	ant-post	med-lat	Vert.
Lifting Trial		FK_00			FK_30			FK_60			FK_00			FK_30			FK_60		
RMSE	S11	3.76	23.38	3.50	5.65	21.63	4.23	4.88	16.76	3.31	8.84	7.67	20.00	12.02	10.46	22.20	8.26	7.69	11.97
	S12	7.71	13.09	2.68	10.67	13.21	2.14	18.28	10.59	2.11	8.97	6.16	18.25	7.29	6.43	13.10	6.74	6.50	16.38
SD	S11	1.22	9.63	2.22	2.97	11.64	2.34	2.46	9.77	2.17	5.30	4.08	13.47	6.32	5.94	14.96	4.91	5.09	8.17
	S12	3.67	7.53	1.74	4.70	5.49	1.19	4.90	5.43	1.37	4.93	4.05	10.59	5.27	3.93	6.61	4.01	3.70	10.97
R	S11	0.92	0.94	0.65	0.98	0.97	0.54	1.00	0.99	0.82	0.80	0.05	0.90	0.54	0.46	0.88	0.63	0.82	0.91
	S12	0.41	0.85	0.52	0.99	0.96	0.58	1.00	0.99	0.90	0.78	0.70	0.85	0.78	0.85	0.90	0.79	0.85	0.89
Lifting Trial		KS_00			KS_30			KS_60			KS_00			KS_30			KS_60		
RMSE	S11	3.40	5.22	0.96	5.34	6.12	1.17	5.19	4.04	1.23	7.99	5.19	6.33	7.36	5.58	8.84	7.38	3.92	7.62
	S12	3.03	4.15	1.14	7.69	5.89	1.39	4.61	4.45	1.33	7.01	5.19	12.15	8.30	8.19	16.31	6.09	6.77	8.43
SD	S11	1.62	2.87	0.67	3.56	3.34	0.69	2.68	2.51	0.79	5.01	3.60	3.44	3.55	3.50	5.60	3.98	2.63	5.12
	S12	1.67	2.79	0.63	3.33	3.51	0.77	2.39	2.69	0.77	4.54	2.96	6.35	4.36	5.16	11.17	3.47	4.40	4.98
R	S11	0.84	0.98	0.54	1.00	0.94	0.87	1.00	0.99	0.94	0.74	0.67	0.95	0.87	0.69	0.90	0.89	0.95	0.92
	S12	0.69	0.97	0.55	0.99	0.91	0.81	0.99	0.98	0.96	0.81	0.33	0.82	0.88	0.77	0.73	0.90	0.78	0.89
Lifting Trial		FS_00			FS_30			FS_60			FS_00			FS_30			FS_60		
RMSE	S11	4.93	16.53	2.58	5.12	13.03	3.38	5.47	17.29	2.37	7.47	6.63	12.45	11.36	9.43	14.45	11.01	6.48	19.58
	S12	6.18	6.83	3.45	8.84	11.27	2.91	8.04	9.54	2.06	6.21	8.90	17.32	8.26	9.37	21.14	9.68	8.65	19.56
SD	S11	3.35	9.90	1.91	3.05	7.89	2.49	3.40	11.80	1.54	5.27	4.11	9.84	6.73	5.06	9.35	5.27	3.66	12.34
	S12	2.96	4.52	2.70	3.52	8.41	1.91	3.75	7.51	1.33	3.71	6.15	11.28	5.09	5.36	12.23	7.57	4.34	10.90
R	S11	0.76	0.99	0.90	0.99	0.99	0.82	1.00	0.99	0.84	0.70	0.17	0.96	0.64	0.60	0.68	0.64	0.79	0.91
	S12	0.95	0.98	0.45	0.97	0.99	0.51	0.99	0.99	0.64	0.88	0.26	0.51	0.87	0.73	0.86	0.52	0.79	0.79

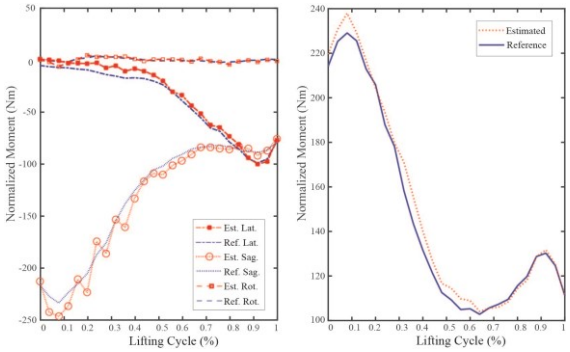


Fig. 5. Estimated versus reference L5/S1 joint moment for floor-to-knuckle-height and 60 degree end-of-lift angle lifting trial (left). The total moment is the vector summation of the L5/S1 moments at each three planes (right).

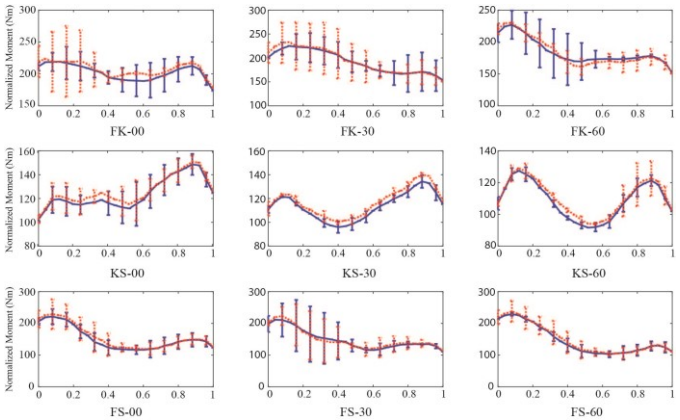


Fig. 6. Average estimated (dashed line) versus reference (solid line) L5/S1 joint total moment across the subjects for each lifting task. The vertical bars show the standard deviation for every 8 percent of the lifting cycle.

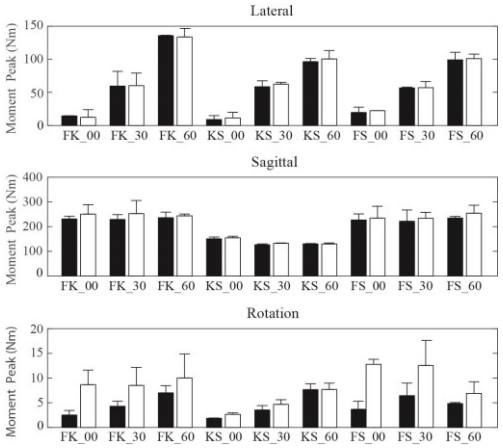


Fig. 7. Average of the peak L5/S1 joint moment across the subjects obtained from the reference (black) and the proposed DNN based method (white) for each of the lifting trial and plane separately. Lifting trials are shown as their "vertical lifting range _ end of lift angle". Standard deviations are shown by error bars.

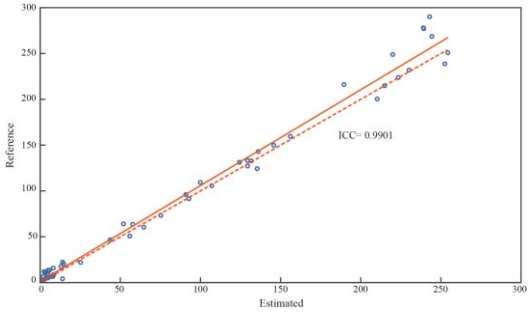


Fig. 8. Scatter plot shows the relation between peak moments estimated by the proposed DNN based method and the reference. Data are pooled over the whole testing dataset. The solid line is the linear regression line fits trough the data points and the dashed diagonal line is the identity line. ICC indicates the intra-class correlation between the reference and estimated peak moments.

B. L5/S1 Joint Moment Peaks

Absolute peak values extracted from the moment time series of the proposed method are compared to corresponding values of the reference across the whole lifting trials (fig. 7). The RMSE and R coefficient of the peak total moment were 6.14 Nm and 0.96 respectively. Finally, ICCs of peak moments over all pooled video dataset (2 subjects, 9 lifting trials, and 3 planes) were about 0.99 between the reference and the proposed method, which is considered as excellent [41] (fig. 8).

C. L5/S1 Joint Force Time Series

For all of the lifting trials, the correspondence between 3D L5/S1 joint force obtained from the reference and estimated from the proposed method was good. For dominant force component (vertical force), R values were mostly above 0.80 and RMS mostly below 20 N (table 1 and fig. 9). The grand mean (±SD) of the total force absolute errors across all the subjects and trials was 4.85 (±4.85) N. For non-dominant L5/S1 force components (anterior-posterior and mediolateral force), both R values and RMSE were mostly smaller than dominant force component.

Figure 10 presents the average and standard deviation of the total force across the subjects for each of the nine lifting tasks. It shows a good fit of the proposed method with the reference for all of the lifting tasks with no evidence of systematic overestimation or underestimation.

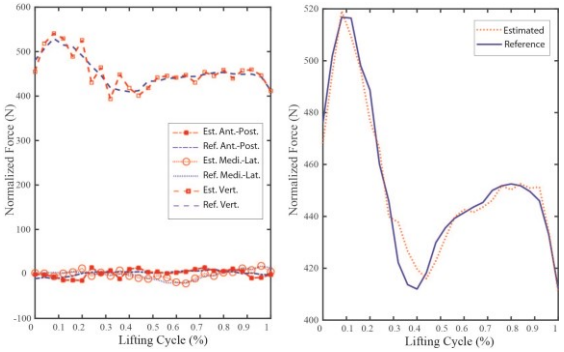


Fig. 9. Estimated versus reference L5/S1 joint force for floor-to-knuckle-height and 60 degree end-of-lift angle lifting trial (left). The total force is the vector summation of the L5/S1 forces at each three planes (right).

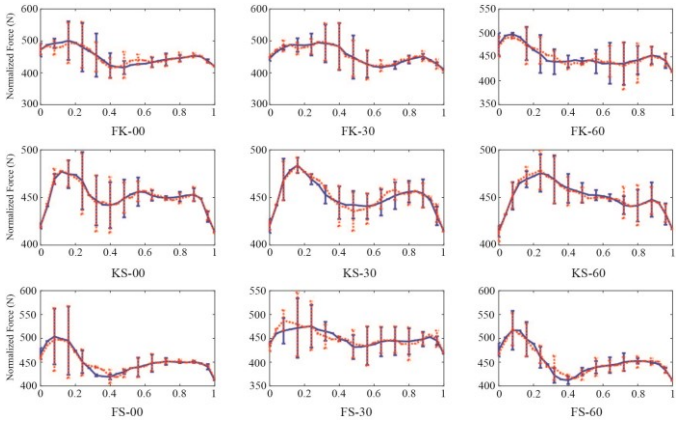


Fig. 10. Average estimated (dashed line) versus reference (solid line) L5/S1 joint total force across the subjects for each lifting task. The vertical bars show the standard deviation for every 8 percent of the lifting cycle.

D. L5/S1 Joint Force Peaks

Absolute peak values extracted from the force time series of the proposed method were compared to the corresponding values of the reference across the whole lifting trials (figure 11). The RMSE and R coefficient of the peak total force were 4.45 N and 0.99 respectively. Finally, ICCs of the peak forces over whole pooled video dataset (2 subjects, 9 lifting trials, and 3 planes) were about 0.99 between the reference and the proposed method, which is considered as excellent [41] (fig. 12).

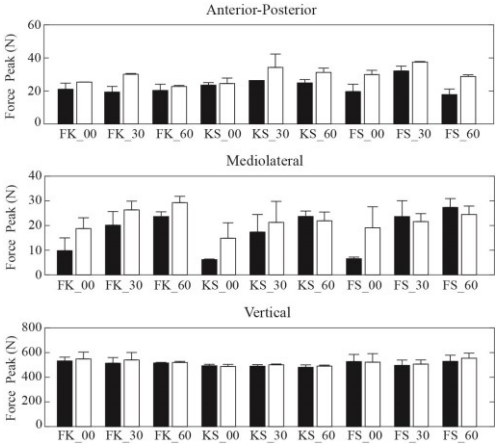


Fig. 11. Average of the peak L5/S1 joint force across the subjects obtained from the reference (black) and the proposed DNN based method (white) for each of the lifting trial and plane separately. Lifting trials are shown as their "vertical lifting range _ end of lift angle". Standard deviations are shown by error bars.

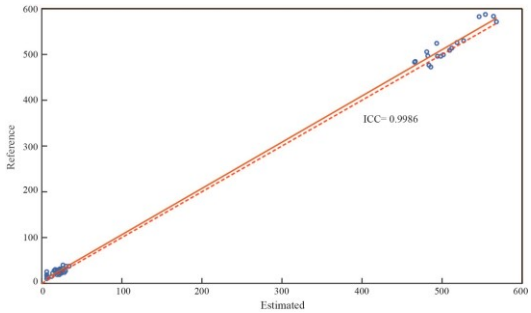


Fig. 12. Scatter plot shows the relation between peak forces estimated by the proposed DNN based method and the reference. Data are pooled over the whole testing dataset. The solid line is the linear regression line fits trough the data points and the dashed diagonal line is the identity line. ICC indicates the intra-class correlation between the reference and estimated peak forces.

VI. Discussion

In this work, we presented a DNN-based method for 3D human pose estimation and validated the results for L5/S1 joint kinetic estimation. We validated our method by comparing the results with the reference obtained from a marker-based motion capture system. The results show a strong correspondence between the methods for estimated L5/S1 joint kinetic during the whole lifting cycle as well as estimated peak kinetic values. The performance of the proposed method for L5/S1 joint moment estimation is comparable or better than the performance reported in previous studies using a video-based coding system [16] (Mean±SD of peak total moment of 12.13±9.67 Nm compare to 28.27±4.49 Nm and 27.84±2.41

Nm for two different video systems) and Inertial sensors [43] (Mean±SD of extension moment of 7.0±7.1 Nm compare to 11.5±7.4 Nm to 31.0±16.6 Nm for different lifting styles). To the best knowledge of the authors, the present study is the first work using deep learning for L5/S1 joint moment estimation.

The performance of the proposed DNN based method was evaluated for each plane separately. For the non-dominant components (lateral and rotation moment and anterior-posterior and mediolateral force), the R coefficient of time series were smaller than dominant components (sagittal moment and vertical force). However considering the smaller RMSE for non-dominant components, this is most likely happening due to smaller moment and force variances in these planes than less accurate results. Previous studies have reported similar performance comparison for the dominant and non-dominant L5/S1 moment using a video coding system [15] and Inertia sensors [44].

Furthermore, the results show a good fit of the proposed method with the reference for all of the lifting tasks. However, the average total L5/S1 joint moment and force difference across the whole dataset between the proposed method and the reference for KS lifting was smaller than FK and FS lifting (3.11 Nm compare to 9.43 Nm and 7.41 Nm for moment and 3.93 N compare to 6.21 N and 5.66 N for force). It may be caused by the insignificant movement of the lower body for grabbing the box from knuckle height level in comparison with floor level, which leads to higher accuracy of the joints kinematic estimation.

This study demonstrates the applicability of deep learning techniques in the context of biomechanical analysis and can be considered as a simple and relatively cheap solution for the drawbacks associated with the marker-based motion analysis methods. For future work, subjects with reported low back pain can be added to the dataset and a classification algorithm can be utilized to classify each lifting task as a safe or unsafe lifting based on the estimated L5/S1 kinetic values. This method can provide a reliable tool for detecting the risk of lower back injuries during occupational lifting.

The present study is a starting point of the research along this direction. There are three limitations about this study that should be further investigated in the future research. First, the proposed method was validated for lifting without moving feet. The proposed method requires parameter fine-tuning for new tasks and may or may not work as well for an unseen task. Whether and how well this method can be extended for more general lifting tasks would be worth to investigate. Second, for calculating the external force in inverse dynamics, we assumed an equal weight distribution of the crate between the both hands, which is not accurate. Finally, although our proposed method is capable of handling occlusion, but the performance may not be as well in case of highly occluded images or using monocular images.

VII. Conclusion

The current study shows the applicability of deep learning as a viable tool for assessment of lower back loads during occupational lifting. The accuracy of the method is comparable with the marker-based motion tracking systems without the limitations associate with these systems. This simple and relatively cheap method can be used for on-site biomechanical analysis in order to decrease the risk of lower back injuries in the workplaces.

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.

References

[1] L. Manchikanti, V. Singh, F. J. Falco, R. M. Benyamin, and J. A. Hirsch. (2014). Epidemiology of low back pain in adults," *Neuromodulation: Technology at the Neural Interface,* 17, pp. 3-10.

[2] J. I. Kuiper, 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, pp. 389-404.

[3] B. R. Da Costa, and Edgar Ramos Vieira. (2010). Risk factors for work‐related musculoskeletal disorders: a systematic review of recent longitudinal studies," *American journal of industrial medicine* 53.3, pp. 285-323.

[4] A. J. van der Beek and M. Frings-Dresen. (1998). Assessment of mechanical exposure in ergonomic epidemiology," *Occupational and environmental medicine,* 55, pp. 291-299.

[5] M. De Looze, I. Kingma, J. Bussmann, and H. Toussaint. (1992). Validation of a dynamic linked segment model to calculate joint moments in lifting," *Clinical Biomechanics,* 7, pp. 161-169.

[6] P. Desjardins, A. Plamondon, and M. Gagnon. (1998). Sensitivity analysis of segment models to estimate the net reaction moments at the L5/S1 joint in lifting," *Medical engineering & physics,* 20, pp. 153-158.

[7] I. Kingma, M. P. de Looze, H. M. Toussaint, H. G. Klijnsma, and T. B. Bruijnen. (1996). Validation of a full body 3-D dynamic linked segment model," *Human Movement Science,* 15, pp. 833-860.

[8] C. Larivière and D. Gagnon. (1998). Comparison between two dynamic methods to estimate triaxial net reaction moments at the L5/S1 joint during lifting," *Clinical Biomechanics,* 13, pp. 36-47.

[9] C. Larivière and D. Gagnon. (1999). The L5/S1 joint moment sensitivity to measurement errors in dynamic 3D multisegment lifting models," *Human movement science,* 18, pp. 573-587.

[10] A. Plamondon, M. Gagnon, and P. Desjardins. (1996). Validation of two 3-D segment models to calculate the net reaction forces and moments at the L 5 S 1 joint in lifting," *Clinical Biomechanics,* 11, pp. 101-110.

[11] T. Robert, L. Chèze, R. Dumas, and J.-P. Verriest. (2007). Validation of net joint loads calculated by inverse dynamics in case of complex movements: application to balance recovery movements," *Journal of biomechanics,* 40, pp. 2450-2456.

[12] B. D. Hendershot, and Erik J. Wolf. (2014). Three-dimensional joint reaction forces and moments at the low back during over-ground walking in persons with unilateral lower-extremity amputation," *Clinical Biomechanics* 29.3, pp. 235-242.

[13] L. Mündermann, S. Corazza, and T. P. Andriacchi. (2006). The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications," *Journal of NeuroEngineering and Rehabilitation,* 3, p. 6.

[14] S. M. Hsiang, G. E. Brogmus, S. E. Martin, and I. B. Bezverkhny. (1998). Video based lifting technique coding system," *Ergonomics,* 41, pp. 239-256.

[15] X. Xu, C.-C. Chang, G. S. Faber, I. Kingma, and J. T. Dennerlein. (2012). Estimating 3-D L5/S1 Moments During Manual Lifting Using a Video Coding System: Validity and Interrater Reliability," *Human factors,* 54, pp. 1053-1065.

[16] P. Coenen, I. Kingma, C. R. Boot, G. S. Faber, X. Xu, P. M. Bongers*, et al.* (2011). Estimation of low back moments from video analysis: A validation study," *Journal of biomechanics,* 44, pp. 2369-2375.

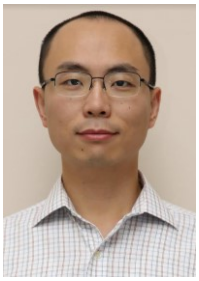
[17] P. Coenen, I. Kingma, C. R. Boot, P. M. Bongers, and J. H. van Dieën. (2013). Inter-rater reliability of a video-analysis method measuring low-back load in a field situation," *Applied ergonomics,* 44, pp. 828-834.

[18] C.-C. Chang, S. Hsiang, P. G. Dempsey, and R. W. McGorry. (2003). A computerized video coding system for biomechanical analysis of lifting tasks," *International Journal of Industrial Ergonomics,* 32, pp. 239-250.
[19] M. Daneshzand, M. Faezipour, and B. D. Barkana. (2017). Computational Stimulation of the Basal Ganglia Neurons with Cost Effective Delayed Gaussian Waveforms," *Frontiers in computational neuroscience,* 11, p. 73.
[20] A. Pfister, A. M. West, S. Bronner, and J. A. Noah. (2014). Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis," *Journal of medical engineering & technology,* 38, pp. 274-280.
[21] X. Ning and G. Guo. (2013). Assessing spinal loading using the Kinect depth sensor: A feasibility study," *IEEE Sensors journal,* 13, pp. 1139-1140.
[22] I. T. Weerasinghe, J. Y. Ruwanpura, J. E. Boyd, and A. F. Habib, "Application of Microsoft Kinect sensor for tracking construction workers," in *Construction Research Congress 2012: Construction Challenges in a Flat World*, 2012, pp. 858-867.
[23] M. R. Andersen, T. Jensen, P. Lisouski, A. K. Mortensen, M. K. Hansen, T. Gregersen*, et al.* (2012). Kinect depth sensor evaluation for computer vision applications," *Technical Report Electronics and Computer Engineering,* 1,
[24] R. Mehrizi, 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,*
[25] R. Mehrizi, 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,*
[26] J. Zurada, W. Karwowski, and W. S. Marras. (1997). A neural network-based system for classification of industrial jobs with respect to risk of low back disorders due to workplace design," *Applied Ergonomics,* 28, pp. 49-58.
[27] K. G. Davis, Y. Hou, W. S. Marras, W. Karwowski, J. M. Zurada, and S. E. Kotowski, "Utilization of a hybrid neuro-fuzzy engine to predict trunk muscle activity for sagittal lifting," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2008, pp. 1064-1067.
[28] W. Lee, W. Karwowski, W. Marras, and D. Rodrick. (2003). A neuro-fuzzy modeling of myoelectrical activity of trunk muscles due to manual lifting tasks," *Ergonomics,* 16, pp. 285-309.
[29] Y. Hou, J. M. Zurada, W. Karwowski, W. S. Marras, and K. Davis. (2007). Identification of key variables using fuzzy average with fuzzy cluster distribution," *IEEE transactions on fuzzy systems,* 15, pp. 673-685.
[30] Y. Hou, J. M. Zurada, W. Karwowski, W. S. Marras, and K. Davis. (2007). Estimation of the dynamic spinal forces using a recurrent fuzzy neural network," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* 37, pp. 100-109.
[31] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International Workshop on Human Behavior Understanding*, 2011, pp. 29-39.
[32] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition," in *IJCAI*, 2015, pp. 3995-4001.
[33] S. M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi. (2018). Deep cross polarimetric thermal-to-visible face recognition," *arXiv preprint arXiv:1801.01486,*
[34] S. M. Iranmanesh, 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,*
[35] R. Mehrizi, X. Peng, Z. Tang, X. Xu, D. Metaxas, and K. Li, "Toward Marker-free 3D Pose Estimation in Lifting: A Deep Multi-view Solution," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, 2018, pp. 485-491.
[36] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016, pp. 483-499.
[37] R. Mehrizi, 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,*
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[39] P. De Leva. (1996). Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters," *Journal of biomechanics,* 29.9, pp. 1223-1230.
[40] A. L. Hof. (1992). An explicit expression for the moment in multibody systems," *Journal of biomechanics,* 25.10, pp. 1209-1211.
[41] J. L. Fleiss. (2011). *Design and analysis of clinical experiments,* Vol. 73. John Wiley & Sons,
[42] I. Shojaei, Vazirian, M., Croft, E., Nussbaum, M. A., & Bazrgari, B. (2016). Age related differences in mechanical demands imposed on the lower back by manual material handling tasks," *Journal of biomechanics,* 49, pp. 896-903.
[43] G. S. Faber, I. Kingma, and J. H. van Dieën. (2010). Bottom-up estimation of joint moments during manual lifting using orientation sensors instead of position sensors," *Journal of biomechanics,* 43, pp. 1432-1436.
[44] G. Faber, C. Chang, I. Kingma, J. Dennerlein, and J. van Dieën. (2016). Estimating 3D L5/S1 moments and ground reaction forces during trunk bending using a full-body ambulatory inertial motion capture system," *Journal of biomechanics,* 49, pp. 904-912.

Rahil Mehrizi received her B.S. in Mechanical Engineering from University of Tehran in 2009 and M.S. degree in Industrial Engineering in 2013. She is currently working towards the Ph.D. degree at the Department of Industrial and Systems Engineering, Rutgers University, New Jersey. Her research interests include computer vision, biomechanical modeling and analysis, and musculoskeletal image analysis.

Xi Peng received the Ph.D. degree from Rutgers University, NJ, USA, in 2018, the M.S. degrees from Chinese Academy of Sciences in 2011, and the B.E. degree from Beihang University, Beijing, China in 2008. He is currently an Assistant Professor in Department of Computer Science, Binghamton University - State University of New York, NY, USA. His research interest includes deep learning, machine learning, and intelligent data analytics in vision or language understanding.

Dimitris N. Metaxas received the B.E. degree from the National Technical University of Athens, Athens, Greece, in 1986, the M.S. degree from the University of Maryland, College Park, MD, USA, in 1988, and the Ph.D. degree from the University of Toronto, Toronto, ON, Canada, in 1992. He is a Professor at the Computer Science Department, Rutgers University, Piscataway, NJ, USA, and is directing the Computational Biomedicine Imaging and Modeling Center. His current research interests include the development of formal methods upon which computer vision, computer graphics, and medical imaging can advance synergistically.



Xu Xu is an assistant professor in Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests are mainly on human factors and ergonomics engineering, occupational biomechanics, optimization-based biomechanical modelling, data mining on human motion data, and musculoskeletal injury prevention. He earned a B.S. in industrial engineering from Tsinghua University in 2004, and an M.S and Ph.D. in industrial engineering from North Carolina State University in 2006 and 2008. He also completed a postdoctoral fellowship in

Department of Environmental Health at Harvard School of Public Health in 2010.

Shaoting Zhang received the BE degree from Zhejiang University in 2005, the MS degree from Shanghai Jiao Tong University in 2007, and the PhD degree in computer science from Rutgers in January 2012. He is an assistant professor in the Department of Computer Science, University of North Carolina at Charlotte. Before joining UNC Charlotte, he was a faculty member in the Department of Computer Science, Rutgers-New Brunswick (research assistant professor, 2012-2013). His research is on the interface of medical imaging informatics, large-scale visual understanding, and machine learning. He is a senior member of the IEEE.

Kang Li received his Ph.D. degrees from University of Illinois at Urbana-Champaign in 2009. He is an Associate Professor of the Department of Orthopaedics at Rutgers New Jersey Medical School (NJMS) and a graduate faculty member of the Department of Computer Science at Rutgers University. He is also a visiting professor of School of Mechatronics Engineering at UESTC. Before joining NJMS, he was a faculty member in the Department of Industrial and Systems Engineering at Rutgers. His research interests include AI in healthcare, musculoskeletal biomechanics, medical imaging, healthcare engineering, design and biorobotics, and human factors/ergonomics.