Toward Marker-free 3D Pose Estimation in Lifting: A Deep Multi-view Solution

Conference Paper · February 2018

CITATIONS
0

READS
44

6 authors, including:

Xi Peng
Rutgers, The State University of New Jersey
30 PUBLICATIONS **144** CITATIONS
SEE PROFILE

Dimitris N. Metaxas
Rutgers, The State University of New Jersey
718 PUBLICATIONS **20,772** CITATIONS
SEE PROFILE

Kang Li
Rutgers, The State University of New Jersey
60 PUBLICATIONS **264** CITATIONS
SEE PROFILE

Toward Marker-free 3D Pose Estimation in Lifting: A Deep Multi-view Solution

Rahil Mehrizi*, Xi Peng†, Zhiqiang Tang†, Xu Xu¶, Dimitris Metaxas† and Kang Li*†§

*Department of Industrial & Systems Engineering
Rutgers University, Piscataway, New Jersey, USA
Email: rahil.mehrizi@rutgers.edu
†Department of Computer Science, Rutgers University, Piscataway, New Jersey, USA
§Department of Orthopaedics, Rutgers New Jersey Medical School, Newark, New Jersey, USA
¶Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA

***Abstract*—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. To improve work place safety, it is necessary to assess musculoskeletal and biomechanical risk exposures associated with these tasks, which requires very accurate 3D pose. Existing approaches mainly utilize marker-based sensors to collect 3D information. However, these methods are usually expensive to setup, time-consuming in process, and sensitive to the surrounding environment. In this study, we propose a multi-view based deep perceptron approach to address aforementioned limitations. Our approach consists of two modules: a "view-specific perceptron" network extracts rich information independently from the image of view, which includes both 2D shape and hierarchical texture information; while a "multi-view integration" network synthesizes information from all available views to predict accurate 3D pose. To fully evaluate our approach, we carried out comprehensive experiments to compare different variants of our design. The results prove that our approach achieves comparable performance with former marker-based methods, i.e. an average error of 14.72 ± 2.96 mm on the lifting dataset. The results are also compared with state-of-the-art methods on HumanEva-I dataset [1], which demonstrates the superior performance of our approach.**

***Keywords*-markerless 3D human pose estimation; deep neural network; lifting**

I. INTRODUCTION

Work-related musculoskeletal disorders (WMSD) are commonly observed among the workers involved in material handling tasks such as lifting. 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 the joints kinematic and critical joints stress. In recent years, several systems were developed to capture the 3D body pose and assess the movement of workers, which roughly can be categorized into two groups: direct measurement and observational systems [2], [3].

Direct measurement systems require motion capture equipment and attachment of the reflective markers on the subject's body to capture the 3D coordination of the body joints. They are considered as a relatively reliable and accurate system for estimation of the joints kinematics. However, the wide spread use of direct measurement systems are limited due to its limitations. First, they require expensive motion capture equipment; second, attaching markers to the subject's body is time consuming and can obstruct the subject's activities.

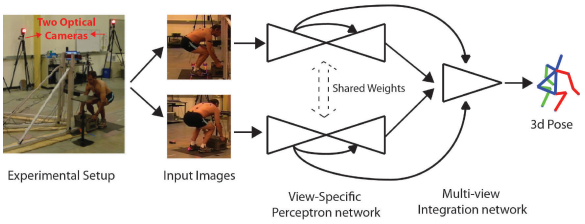

Fig. 1. An overview of our approach. The "view-specific perceptron" network extracts both shape and hierarchical texture from different views; while the "multi-view integration" network synthesizes information from all available views to estimation 3D pose. Hierarchical skip connections are not only shared locally inside the first network, but also shared globally between two networks for efficient and effective feature embedding.

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 [4], joints trajectory is generated for the entire task by applying a motion pattern prediction algorithm [5]. Observational systems are not as accurate as direct measurement systems and the result accuracy rely on the experience of the rater, especially when joints angle become close to the posture boundaries [6].

In this study, we propose a Deep Neural Network (DNN) based multi-view perceptron framework for maker-free 3D motion capture. Fig. 1 illustrates the experimental setup along with the overview of our approach. Our method consists of two networks: a "view-specific perceptron" network extracts both 2D shape and hierarchical texture information from different views [7]; while a "multi-view integration" network synthesizes information from all available views to provide accurate 3D pose. It will be shown that sharing hierarchical texture information globally between the two networks in addition to a locally use inside the first network can significantly improve

the accuracy. It makes the method suitable for biomechanical analysis, which requires higher accuracy compare to other applications. Since our proposed method eliminates the need of attaching markers onto the subject's body segments or hiring raters to estimate the pose, it can overcome the limitations of both direct measurement and observational systems. To summarize, our contributions are:

- We propose a novel DNN-based method to estimate accurate 3D pose from multiple 2D views for biomechanical analysis.
- We propose hierarchical skip connections to share rich texture information in different scales, which is proved to be crucial for efficient and accurate 3D inference.
- Comprehensive experiments are performed to evaluate different variants of our design, which proves the superior performance of our approach in various aspects.
- The results on a real-world multi-view lifting dataset prove that our approach can meet the high-level accuracy requirement in workplace biomechanical analysis.

II. BACKGROUND

We review related works in two categories. The first category is an overview of the previous work in marker-less posture estimation for biomechanical application, which is important since the focus of this study is on this specific application. The second category, is a summary of recent human pose estimation methods using deep learning.

A. Pose Estimation for Biomechanical Application

Human pose estimation is important for biomechanical analysis and preventing WMSD. Even though, marker-less pose estimation 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 [8]. There are few studies which explored the field of computer vision and proposed marker-less methods for biomechanical and clinical applications. In particular, [9], [10] proposed a computer vision based method for estimation of 3D pose estimation and lower back loads in the symmetrical lifting tasks. In another study by [11], a Levenberg-Marquardt minimization scheme over an iterative closest point algorithm was employed to estimate human motion through a marker-less motion capture system. Goffredo et. al. [12] proposed a marker-less framework to estimate human pose for a sit-to-stand task by means of a maximum likelihood approach carried out in the GaussLaguerre transform domain. These studies demonstrate the feasibility of computer vision 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 fully studied. Additionally, deep learning, which is considered as the state-of-the-art approach in the domain of the vision tasks is not studied for the field of biomechanical application. In this study, we propose a deep learning based framework for maker-less 3D motion capture. It will be shown that using the proposed framework can achieve very high accuracy, which is suitable for the biomechanical analysis.

B. Deep Learning for Human Pose Estimation

Earlier computer vision based approaches for 3D human pose estimation used a discriminative or generative method to learn a mapping from the image features to the 3D human pose. All of these approaches suffer from the fact that they utilize hand crafted image features e.g. HOG [13], SIFT [14], etc. Approaches based on the hand crafted image features are not able to handle heterogeneous or complex datasets [15], [16]. With the emergence and advances of deep learning techniques, approaches that employ deep convolutional neural networks to learn the image features [17], have become the standard in the domain of the vision tasks. DNN approaches have achieved the highest performance for several vision tasks such as visual recognition [18], [19], image generation [20], [21], and human pose estimation [22], [23].

More recent DNN approaches for 3D human pose estimation tend to learn an end-to-end DNN to regress directly from the images to the 3D joints coordination [24]–[27]. Other DNN approaches, on the other hand, have studied frameworks that employ 2D pose estimation as an intermediate step and leverage this information to infer the 3D pose from it. Chen et. al. [28] suggests that 2D pose is a useful intermediate representation and can aid the 3D pose estimation. While [28]–[30] represents intermediate 2D pose as 2D coordination of the joints, [31]–[33] define it by a set of heatmaps that encode the probability of observing a specific joint at the corresponding image location. Tome et. al. [32] proposes multi-stage DNN architecture combined with a probabilistic knowledge of 3D human pose, which estimates 2D joints heatmap and 3D pose simultaneously to improve both tasks. Pavlakos et. al. [31] trains a DNN with 2D joints heatmaps as an intermediate representation to predict per voxel likelihood for each joint in the 3D space instead of directly regressing the 3D joints coordination. They uses a coarse-to-fine technique to overcome the high dimensionality problem of the volumetric representation. They also suggests combining 2D joints heatmaps with image features for the intermediate representation, to take advantage of both the image cues along with the reliably detected heatmaps. The same combination is applied in a study by [33] in which 2D joints heatmaps and images are fed into a two-stream architecture, Then the combination of these two streams are then fed into a fusion stream at a specific layer to obtain the final 3D human pose estimate.

In this study, we propose a novel deep learning based method to estimate 3D human pose from multi-view images. Instead of using the 2D heatmap as the only intermediate supervision, we propose to share both shape and hierarchical texture locally and globally for efficient 3D inference. In contrast to the recent work in 3D human pose estimation whose focus are on single view and challenging setting, the proposed network is designed to handle multi-view images, which is a common setup for the biomechanical analysis experiments.

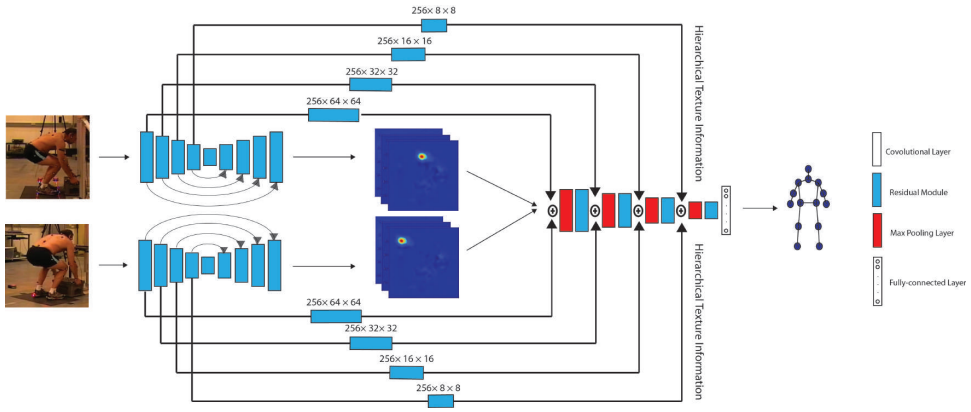

Fig. 2. An overview of our approach. The "view-specific perceptron" network extracts both shape and hierarchical texture from different views; while the "multi-view integration" network synthesizes information from all available views to estimation 3D pose. Note that the hierarchical skip connections are not only shared locally inside the first network, but also shared globally between two networks for efficient and effective feature embedding.

III. METHODS

In this work, we aim to predict the 3D body pose from multi-view RGB images. We proposed a deep learning based method for this purpose whose inputs are totally N different viewpoints around the subject and the output is the 3D coordination of the body joints, which define the pose. Fig. 2 shows an illustration of our approach for N=2, which consists of two networks: a "view-specific perceptron" network and a "multi-view integration" network. The first network extracts both shape (2D pose) and hierarchical texture information independently from each view, while the second network synthesizes these information from all available views to infer the 3D pose.

A. View-specific Perceptron Network

View-specific perceptron network extracts rich information independently from each view, which includes not only 2D shape but also hierarchical texture information for 3D inference in the next step. Each 2D body pose is represented by J heatmaps, where J is the number of body joints. Let $x^i \in \mathbb{R}^{W\times H\times 3}$ be the input RGB image for view i, $t_s^i \in \mathbb{R}^{W_s\times H_s\times L_s}(s=1,...,S)$ be s-th texture feature map for view i, and $h_j^i \in \mathbb{R}^{W_h\times H_h\times L}(j=1,...,J)$ be j-th joint heatmap for view i. Then, view-specific perceptron network (f) for i-th view is a mapping as follow:

$$((h_1^i,...,h_J^i),(t_1^i,...,t_S^i)) = f(x^i).$$

The intermediate supervision is performed by pixel-wise heatmap loss:

$$\mathcal{L}_{2d}^i = 1\ /J\sum_{j=1}^{J}||h_j^i - \widehat{h_j^i}||,$$

where $||.||$ is the Euclidean distance and $\widehat{h_j^i}$ is rendered from the ground truth 2D pose through a Gaussian kernel with mean equal to the ground truth and variance one.
We use Hourglass architecture [22], which has achieved state-of-the-art performance on large scale human pose datasets. Hourglass network [22] comprises of encoder and decoder. Encoder processes the input image with convolution and pooling layers to generate low resolution feature maps and the decoder processes the low resolution feature maps with up-sampling and convolution layers to construct the high resolution heatmaps for each joint. One of the key components of the Hourglass network [22] is the skip connections, the feature maps before each pooling layer, which are directly added to the counterpart in the decoder in order to prevent the loss of high resolution information in the encoder. These hierarchical skip connections of the network share rich texture information in different scales. So, we propose to employ them for a more efficient 3D inference by feeding them to the multi-view integration network. We will show soon that they allow for a richer gradient signal and can provide more 3D cues compare to using only heatmaps or a combination of heatmaps and unprocessed input images.

B. Multi-view Integration Network

The multi-view integration network integrates information from multiple views to synthesize 3D pose estimation. The input of this network is the concatenation of the outputs of the view-specific perceptron network for N different views and the output is the 3D pose. Each 3D pose skeleton $p \in \mathbb{R}^{3\times J}$ is defined as a set of joints coordination in 3D space. So multi-view integration network (g) is a mapping as follow:

$$(\hat{p}) = g(concat(h_1^1...,h_1^N)...,concat(h_J^1...,h_J^N) \\ ,concat(t_1^1,.....,t_1^N),.....,concat(t_S^1,...,t_S^N)).$$

By assuming that 3D joints annotations are available for training dataset, the loss function can be defined as

$$\mathcal{L}_{3d}^i = 1\ /J\sum_{j=1}^{J}||p_j^i - \widehat{p_j^i}||,$$

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

We propose a bottom up data driven method that directly generates the 3D pose skeleton from the outputs of the view-specific perceptron network. The multi-view integration network is designed as an encoder. We tested two types of encoders: first, an encoder consists of a series of convolutional layers with kernel and stride size of 2 in which the resolution of the feature maps are half at each layer; second, an encoder similar to the first part of the Hourglass network [22], which includes max-pooling layers and standard convolutional layers are replaced by a stack of residual learning modules [34]. In the rest of this paper, we call the first and second network architectures as simple encoder and half-hourglass network, respectively. For both network architectures, the encoder output is then forwarded to a fully-connected layer with output size of $3 \times J$ for estimating 3D pose skeleton and measuring the loss function for training. Fig. 3 shows the schematic comparison of simple encoder and half-hourglass architecture in a simplified setting. It will be shown that, half-hourglass network that benefits from residual modules and periodically insert of max-pooling layer can provide more accurate 3D pose compare than the simple encoder network.

C. Hierarchical Skip Connections

Inferring a 3D pose from joints heatmap as the only intermediate supervision, which is a widely used strategy in previous studies [31], [32], is inherently ambiguous. This ambiguity comes from the fact that usually exist multiple 3D poses corresponded to a single 2D pose. In order to overcome this challenge in 3D pose estimation, joints heatmaps can be combined with either input image or its lower-layer features [33]–[35] as the intermediate supervision. While taking the input image into account can provide more information compare to only joints heatmap, combining hierarchical texture information, learnt at the view-specific perceptron network, extract additional cues [35]. As a result, we propose to leverage skip connections of the Hourglass network [22] to multi-view integration network. In our proposed framework, each of the four skip connections produced in the encoder part of the Hourglass network [22], is processed with residual modules and summed with the counterpart in the half-hourglass network (fig. 2). In order to handle multi-view setup, each skip connection should be concatenated across the views before being provided as inputs for the network.

IV. MATERIALS

A. Participants and Data Acquisition

Our lifting dataset consists of 12 subjects. Each subject performed various symmetric and asymmetrical lifting tasks in a laboratory at a self-selected speed. A motion tracking system was used to capture 3D coordination of body joints. Two digital camcorder (GR-850U, JVC, Japan) with resolution 720×480 pixel, synchronized with the motion tracking system also recorded the tasks from 90 degree (side view) and 135 degree view positions. Subjects lifted a plastic crate ($39 \times 31 \times 22$ cm) weighing 10 kg and placed it on a shelf without moving their feet. They performed three vertical lifting

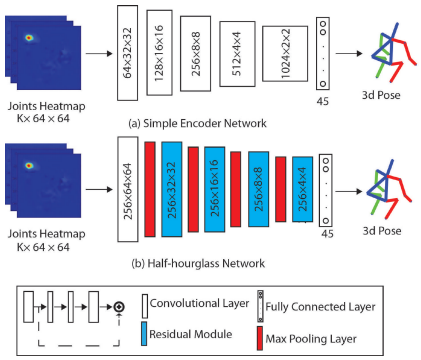

Fig. 3. Architecture comparison of (a) simple encoder and (b) half-hourglass design for multi-view integration network. The numbers inside each layer illustrate the corresponding size of the feature maps (number of channels resolution) for convolutional layers and residual modules [36] and the number of neurons for fully connected layers. The architecture on the bottom represents residual module that is used throughout the network.

ranging from floor to knuckle height (FK), knuckle to shoulder height (KS) and floor to shoulder height (FS). 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. For each combination of the lifting task, two repetitions were performed, providing a total of 18 lifts ($3 \times 3 \times 2$).

B. Data Pre-processing

To prepare the training images, we follow [36] to down sample images from videos. Each video includes 200 frames with 30 fps rate, where only odd frames are employed in this study to prevent overfitting. 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 23 markers to define 14 joints including head, neck, left/right shoulder, left/right elbow, left/right wrist, left/right hip, left/right knee, and left/right ankle 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. After pre-processing, the data structure consists of the cropped images and corresponding normalized 3D joints annotation for every odd frame of the videos.

C. 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 used the Hourglass model [22] for MPII [37] and fine-tuned it on our lifting dataset with learning rate of 0.00025 for five epochs. At the second stage, multi-view integration model 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.

In order to evaluate the performance of the network for both single-view and two-view setups, we ran two experiments: first, the network was trained for a single-view setup using

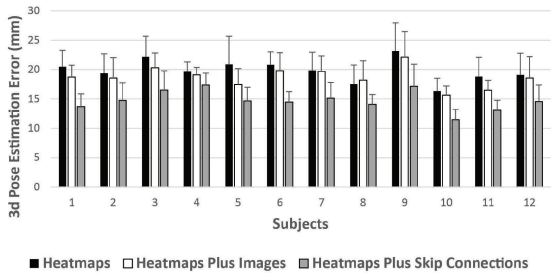

Fig. 4. Average of the 3D pose estimation error of different subjects for three variants of multi-view integration inputs. Bars show the variance.

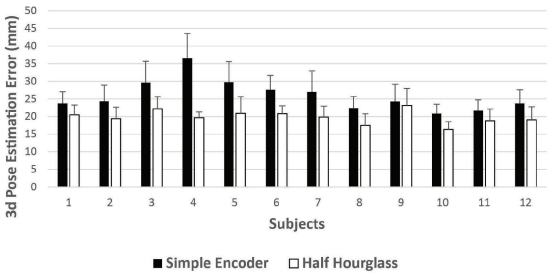

Fig. 5. Average of the 3D pose estimation error of different subjects for simple encoder and half-hourglass architecture. Bars show the variance.

both 90 and 135 degree view separately; second, the network was trained for a two-view setup utilizing both views together as inputs for the network. In all of the experiments, repetition one of all the subjects and lifting tasks were used as training dataset and repetition two as testing dataset. It will be shown that the proposed network is robust and can achieve high performance in both experiments.

D. Evaluation Protocol

Following the evaluation protocol of the publicly available datasets [1] we calculate 3D human pose estimation error based on the average Euclidean distance between estimated 3D joints coordination and corresponding ground-truth data obtained from a motion capture system.

V. RESULTS

In this chapter we present experimental results on our lifting dataset. We used Pytorch interface in this work and training and testing have been performed on a machine with NVIDIA Tesla K40c and 12 GB RAM. We executed three experiments to study the effect of three different factors on the accuracy of results. First, three variants of multi-view integration inputs; including joint heatmaps, joints heatmaps plus input images, and joints heatmaps plus skip connections, are tested to assess how the accuracy changes by feeding more 3D cues to this network. Second, we tested two network structures for multi-view integration, namely simple encoder and half-hourglass, to evaluate the influence of using max-pooling and residual learning modules instead of standard convolutional layers on our dataset. Third, single-view and two-views training are performed to study the robustness of the method as a function of the number of cameras. Finally, we chose the method with the highest performance (fig. 2) and applied it on HumanEva-I dataset [1] and compared our results with other state-of-the-art method on this dataset.

A. Different Variants of Multi-view Integration Inputs

Fig. 4 illustrates the 3D pose estimation error of different subjects for three combinations of input for multi-view integration network. It can be seen that summing up skip connections with feature maps in-between residual modules can achieve the highest accuracy. The error reduction of combining input images with joints heatmaps is only %6 (19.82 ± 3.77 mm vs 18.69 ± 3.25 mm), compare to %26 (19.82 ± 3.77 mm vs 14.72 ± 2.96 mm) error reduction by combining skip connections and joint heatmaps as input to the multi-view integration network. While input images might provide noisy information for the network, these skip connection features can extract semantic information at multiple levels of 2D pose estimation and provide more cues for 3D pose inference.

B. Simple Encoder vs Half-hourglass

Fig. 5 illustrates the 3D pose estimation error of different subjects for simple encoder and half-hourglass architectures. The average error over the whole dataset is 25.98 ± 6.39 mm and 19.82 ± 3.77 mm for these networks, respectively. We found that using half-hourglass network that benefits from residual modules and periodically insert of max-pooling layer reduces the error by %24. This happens due to the fact that networks with residual modules gain accuracy from greatly increased depth and addressing the degradation problem [34]. In addition, inserting max-pooling layer in-between successive convolutional layers reduces the number of parameters and computation in the network, and control overfitting. For qualitative results, we have provided representative 3D poses predicted by our proposed method (half-hourglass architecture and using both heatmaps and skip connections as the inputs for multi-view integration network) in figure 6. It can be seen that even for posture with self-occlusion, our method is able to predict the pose accurately.

C. Single-view vs Two-view Training

Fig. 7 illustrates the 3D pose estimation error of different subjects for single-view and two-view setups. For single-view setup, the network is trained and tested on both views. As was expected, we found that feeding two-view images as inputs to the network increases the accuracy. Since self-occlusion is a main source of ambiguity in pose estimation and can be addressed by using multiple cameras. For biomechanical applications due to the need of very accurate estimated pose, using multiple cameras is crucial [8].

D. Compare with Other Work

We applied our method on HumanEva-I dataset [1] to be able to compare it with state-of-the-art multi-view methods. We followed the standard protocol of the dataset for evaluation

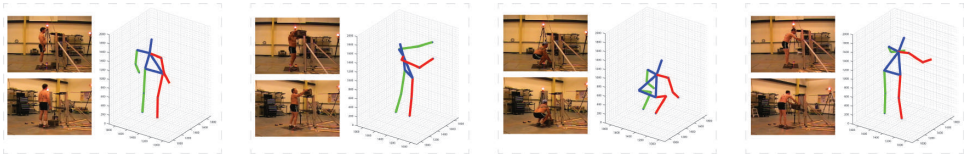

Fig. 6. Qualitative results for lifting datasets. Each dashed box represents a scenario: **Left-** multi-view images, **Right-** corresponding estimated 3D pose.

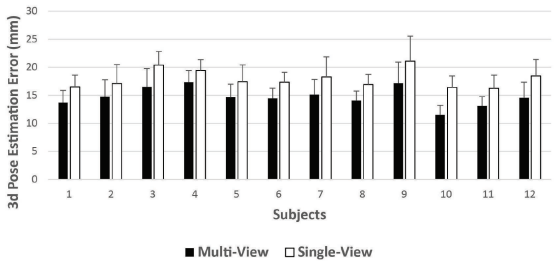

Fig. 7. Average of the 3D pose estimation error of different subjects for single-view and multi-view setups. Bars show the variance.

TABLE I
COMPARISON OF OUR METHOD WITH STATE-OF-THE-ART METHODS ON HUMANEVA DATASET [1]. NA INDICATES THAT RESULTS ARE NOT REPORTED FOR THE CORRESPONDING ACTION IN THE ORIGINAL PAPER.

Methods	Walking			Jogging		
	S1	S2	S3	S1	S2	S3
Elhayak et. al. [38]	66.5	NA	NA	NA	NA	NA
Amin et. al. [39]	54.5	50.2	54.7	NA	NA	NA
Sedai et. al. [40]	42.4	34.1	62.9	70.9	50.6	55.1
Zhang et. al. [41]	44.3	58.4	66.0	55.4	68.2	57.5
Tekin et. al. [25]	37.5	25.1	49.2	NA	NA	NA
Ours	**40.4**	**23.5**	**33.4**	**43.0**	**45.1**	**30.9**

and compared the performance of our method on walking and jogging sequences of subjects 1 to 3. As can be seen from the results in table 1, our proposed deep convolutional neural networks obtain the best result for both walking and jogging sequences.

VI. CONCLUSION

In this work, we presented a multi-view DNN-based method for 3D pose estimation. In part of this study, we introduced an approach to integrate hierarchical texture information with estimated joints heatmap to infer 3D pose. We tested several different network architectures to analyze the influence of various parameters on the accuracy of the results. With optimal network architecture, which consists of half-hourglass architecture for multi-view integration network combined with skip connections, we estimated the pose on our lifting dataset with 14.72 ± 2.96 mm error compared to marker-based motion capture system. The results on a publicly available dataset (HumanEva-I [1]) also shows the superior performance of our approach. This result demonstrates the applicability of deep learning techniques in the context of biomechanical analysis. For future work, we want to use the estimated 3D pose for further biomechanical analysis like calculating joints force and moment in order to automatically detect not-safe lifting in the workplaces with the aim of preventing injuries.

VII. ACKNOWLEDGMENTS

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. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *IJCV*, vol. 87, no. 1, pp. 4–27, 2010.

[2] G. David, "Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders," *Occupational medicine*, vol. 55, no. 3, pp. 190–199, 2005.

[3] P. Spielholz, B. Silverstein, M. Morgan, H. Checkoway, and J. Kaufman, "Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors," *Ergonomics*, vol. 44, no. 6, pp. 588–613, 2001.

[4] X. Peng, Q. Hu, J. Huang, and D. N. Metaxas, "Track facial points in unconstrained videos," in *BMVC*, 2016.

[5] S. M. Hsiang, G. E. Brogmus, S. E. Martin, and I. B. Bezverkhny, "Video based lifting technique coding system," *Ergonomics*, vol. 41, no. 3, pp. 239–256, 1998.

[6] P. Coenen, I. Kingma, C. R. Boot, P. M. Bongers, and J. H. van Dieën, "Inter-rater reliability of a video-analysis method measuring low-back load in a field situation," *Applied ergonomics*, vol. 44, no. 5, pp. 828–834, 2013.

[7] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, and D. Metaxas, "From circle to 3-sphere: Head pose estimation by instance parameterization," *CVIU*, vol. 136, pp. 92–102, 2015.

[8] L. Mündermann, S. Corazza, and T. P. Andriacchi, "The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications," *Journal of NeuroEngineering and Rehabilitation*, vol. 3, no. 1, p. 6, 2006.

[9] R. Mehrizi, X. Xu, S. Zhang, V. Pavlovic, D. Metaxas, and K. Li, "Using a marker-less method for estimating l5/s1 moments during symmetrical lifting," *Applied ergonomics*, 2017.

[10] R. Mehrizi, X. Peng, X. Xu, S. Zhang, D. Metaxas, and K. Li, "A computer vision based method for 3d posture estimation of symmetrical lifting," *Journal of Biomechanics*, 2018.

[11] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi, "Markerless motion capture through visual hull, articulated icp and subject specific model generation," *IJCV*, vol. 87, no. 1, pp. 156–169, 2010.

[12] M. Goffredo, M. Schmid, S. Conforto, M. Carli, A. Neri, and T. D'Alessio, "Markerless human motion analysis in gauss–laguerre transform domain: An application to sit-to-stand in young and elderly people," *IEEE Trans Inf Technol Biomed*, vol. 13, no. 2, pp. 207–216, 2009.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[14] J. Müller and M. Arens, "Human pose estimation with implicit shape models," in *ARTEMI*, 2010.

[15] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *ARTEMI*, 2015.

[16] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Applied Sciences*, vol. 7, no. 1, p. 110, 2017.

[17] D. Li, X. Wang, and D. Kong, "Deeprebirth: Accelerating deep neural network execution on mobile devices," *CoRR*, vol. abs/1708.04728, 2017.

[18] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *ICCV*, 2017, pp. 1623–1632.

[19] X. Di and V. M. Patel, "Large margin multi-modal triplet metric learning," in *FG*, vol. 1, 2017, pp. 1–8.

[20] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces," *arXiv:1708.02681*, 2017.

[21] H. Zhang, V. Sindagi, and V. M. Patel, "Image deraining using a conditional generative adversarial network," *arXiv:1701.05957*, 2017.

[22] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[23] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[24] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *ACCV*, 2014.

[25] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct prediction of 3d body poses from motion compensated sequences," in *CVPR*, 2016.

[26] G. Rogez and C. Schmid, "Mocap-guided data augmentation for 3d pose estimation in the wild," in *Advances in Neural Information Processing Systems*, 2016.

[27] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *ECCV*, 2016, pp. 38–56.

[28] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," *arXiv:1612.06524*, 2016.

[29] S. Park, J. Hwang, and N. Kwak, "3d human pose estimation using convolutional neural networks with 2d pose information," in *ECCV*, 2016.

[30] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3d human poses from a single image," in *CVPR*, 2014.

[31] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *CVPR*, 2017.

[32] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," *arXiv:1701.00295*, 2017.

[33] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in *ICCV*, 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[35] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: A weakly-supervised approach," in *ICCV*, 2017.

[36] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas, "Piefa: Personalized incremental and ensemble face alignment," in *ICCVn*, 2015, pp. 3880–3888.

[37] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.

[38] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras," in *CVPR*, 2015.

[39] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view pictorial structures for 3d human pose estimation.," in *BMVC*, 2013.

[40] S. Sedai, M. Bennamoun, and D. Q. Huynh, "A gaussian process guided particle filter for tracking 3d human pose in video," *IEEE Trans. Image Process*, vol. 22, no. 11, pp. 4286–4300, 2013.

[41] W. Zhang, L. Shang, and A. B. Chan, "A robust likelihood function for 3d human pose tracking," *IEEE Trans. Image Process*, vol. 23, no. 12, pp. 5374–5389, 2014.