Taylor & Francis
Taylor & Francis Group

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tciv20

Video-Based 3D pose estimation for residential roofing

Ruochen Wang, Liying Zheng, Ashley L. Hawke, Robert E. Carey, Scott P. Breloff, Kang Li & Xi Peng

To cite this article: Ruochen Wang, Liying Zheng, Ashley L. Hawke, Robert E. Carey, Scott P. Breloff, Kang Li & Xi Peng (2023) Video-Based 3D pose estimation for residential roofing, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11:3, 369-377, DOI: 10.1080/21681163.2022.2072394

To link to this article: https://doi.org/10.1080/21681163.2022.2072394

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
2023, VOL. 11, NO. 3, 369–377
https://doi.org/10.1080/21681163.2022.2072394

Taylor & Francis
Taylor & Francis Group

Video-Based 3D pose estimation for residential roofing

Ruochen Wang[a], Liying Zheng[b], Ashley L. Hawke[b], Robert E. Carey[b], Scott P. Breloff[b], Kang Li[c] and Xi Peng[a]

[a]Department of Computer & Information Science, University of Delaware, Newark, DE, USA; [b]Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV, USA; [c]Department of Orthopaedics, Rutgers New Jersey Medical School, Newark, NJ, USA

ABSTRACT

Residential roofers are often exposed to awkward postures and motions in a prolonged time, which may not only reduce their body stability and increase fall potential, but also increase the risk of musculoskeletal disorders (MSDs). To assess their risks of fatal and musculoskeletal injuries, it is crucial to capture 3D body poses of workers during roofing tasks. In this paper, we proposed a novel two-stage motion estimation approach based on a convolution neural network to estimate residential roofer's body poses using three-view video data. Our approach includes two stages: (1) use of an offline multi-view model to estimate the 3D pose in a single frame; (2) use of a multi-frame model to apply temporal convolutions to refine the multi-view outputs. The performance of the approach was evaluated by comparing our estimation with the gold-standard marker-based 3D human pose during one of the common residential roofing tasks – shingle installation. The evaluation results show that the proposed multi-frame model can effectively improve the accuracy of the coordinate sequence. Moreover, these results prove that the proposed video-based motion estimation approach can efficiently and accurately locate 3D body joints and pave the way for future onsite motion analysis during roofing activities.

ARTICLE HISTORY
Received 21 November 2021
Accepted 26 April 2022

KEYWORDS
3D human pose estimation; video-based motion prediction; residential roofing; deep learning

1. Introduction

Residential roofers typically spend more than 75% of their work time in stooping, crouching, kneeling and crawling postures, which require constant bending and twisting of workers' bodies (Dong et al. 2008). Prolonged exposures to such awkward postures and motions may not only reduce body stability and increase fall potential, but also increase the risk of musculoskeletal disorders (MSDs). As a result, the prevalence of MSDs among roofers is substantially higher than other construction worker populations – in 2013, the roofing trade has the 2nd highest incident rate of work-related MSDs among all construction sectors (Bureau of labor statistics 2013).

To assess the risks of fatal and musculoskeletal injuries in residential roofers, it is crucial to capture 3D body positions of workers during roofing tasks. Traditional motion capture systems often use optical cameras to capture locations of reflective markers placed on the human body. These systems need to attach a large number of markers on the human body, and require multiple dedicated cameras. It is not only time-consuming to place the large number of markers on the workers' bodies, but also infeasible to apply in realistic working conditions as they may affect workers' activities. Moreover, markers often fall off or become occluded, degrading the quality of the data. These motion capture systems appear to be expensive and impractical for use at residential roofing construction sites.

With the advancement in artificial intelligence technology, various marker-free human pose estimation methods have emerged in recent years. These methods make use of economical cameras and process the image features of the subject. In other terms, the 2D pose can be recognised directly from the image and then the 3D pose can be calculated by combining these 2D poses with camera parameters. Since no equipment other than cameras are required, the cost for capturing human motions is greatly reduced. At the same time, there is no need to attach markers on workers, which makes these methods practical and less time-consuming for workplace application.

Based on the input data, previous pose-estimation models can be divided into three categories: single-view, multi-view, and video-based human pose estimations.

1.1. Single-view

Estimating human posture from a single-view image in 2D space has been a common technique in the domain, as many deep learning-based approaches have recently been proposed to directly estimate 3D pose from single RGB images. Pavlakos et al. (2017) discretised the 3D space around the subject and predicts the voxel-based representation likelihood for each joint in a coarse-to-fine manner. Moreno-Noguer (2017) represented both 2D and 3D poses as $N \times N$ matrices containing Euclidean distances of pair-wise joints and formulate the problem as distance matrix regression. Zhou et al. (2017) introduced weakly-supervised learning towards data 'in the wild' and 3D geometric constraints to obtain a more accurate 3D pose through regularisation. After all, self-occlusion remains a challenge for single-view pose estimation that could hinder the overall performance. For example, if a limb on the opposite side is not visible, then the model suffers from estimating the exact position of that limb.

CONTACT Liying Zheng lzheng2@cdc.gov Health Effects Laboratory Division, National Institute for Occupational Safety and Health, MS 2027, 1095 Willowdale Road, Morgantown, WV, 26505 USA

1.2. Multi-view

To estimate more accurate 3D body posture and alleviate the problem of self-occlusion, many human pose models have migrated into leveraging capturing from multiple views of same subject. Martinez et al. (2017) used linear layers with short connections to upgrade existing 2D human pose coordinates to 3D space. Panoptic Studio (Joo et al. 2015) used a pre-trained model to estimate the 2D pose and then projected the key points into the volume and calculated the 3D body coordinates. Iskakov et al. (2019) further improved end-to-end training for via learnable algebraic and volumetric triangulation. Qiu et al. (2019) proposed a recursive Pictorial Structure Model that can recover the 3D pose from the estimated multi-view 2D pose heat maps, which greatly improved the accuracy without affecting the real-time performance. However, these approaches treated image information on each frame separately without considering the continuity of the adjacent frames, inevitably leading to unstable sequential output.

1.3. Video-based

The outputs of the image-based human pose model are frequently unstable during video processing, since slight perturbations between nearby frames can cause huge discrepancies in the neural network's output. In order to make maximum use of the information in the video, some models use a multi-frame technique to estimate human position to tackle this difficulty. Heat maps paired with light flow across neighbouring frames were utilised by Pfister et al. (2019) to predict a 2D human position. Katircioglu et al. (2018) and Lin et al. (2017) employed long short-term memory, the most prevalent network for processing sequence information, to produce 3D poses predicted from a single observation. Hossain and Little (2018) introduced Seq2seq (Sutskever et al. 2014) to estimate a 3D human posture from a 2D pose sequence in human pose estimation. Pavllo et al. (2019) introduced a fully convolutional architecture that uses semi-supervised learning to train the model and performs temporal convolutions over 2D keypoints for accurate 3D posture prediction in video. The video-based sequential data appear to be able to improve and stabilise the prediction of 3D joint positions.

In this paper, we propose a new deep-learning-based approach for 3D pose estimation during residential roofing tasks. Instead of pursuing end-to-end training, we split our model into two stages: a multi-view model and a multi-frame model. Therefore, the model can be trained on a single graphics processing unit (GPU) board. The multi-view model deploys an algebraic triangulation neural network, which is pre-trained offline using the Human3.6 M – the largest existing public dataset of 3D human poses (Ionescu et al. 2014), to estimate the 3D pose in a single frame. The multi-frame model aggregates the results from video-based continuous frames and applies temporal convolutions to refine the multi-view model's outputs. The performance of this two-stage approach is evaluated by comparing the intermediate and final outputs from these two stages with the gold-standard marker-based 3D human poses estimations.

2. Methods

2.1. Data acquisition

2.1.1. Experimental data collection

Seven healthy male subjects participated in this study and performed one of the most common roofing tasks – the shingle installation task. The study was approved by the Institutional Review Board of the National Institute for Occupational Safety and Health (16-HELD-01XP). The informed consent from all the subjects was obtained before the testing. The whole-body marker data (total 79 makers placed on the subject) were collected using a motion capture system with 14 optical cameras (Vantage V16, Vicon Motion System Ltd., Oxford, UK). Additionally, three video cameras were used to record the subject's movement from three perspectives simultaneously. The original video resolution was 1280×720, and the frame rate was 100 Hz. Each subject was asked to perform three trials of the shingle installation task. The duration of each trial varied from 11 seconds to 17 seconds. As a result, the data set contained 63 videos of 7 subjects (each subject has 3 trials and 3 perspectives).

2.1.2. Data pre-processing

In general, the marker positions recorded by the Vicon system were different from the joint positions defined by common human pose models, so they could not be directly used for training pose detection models. A musculoskeletal modelling software, OpenSim (Delp et al. 2007) was used to estimate the ground-truth joint coordinates for training the neural network. In OpenSim, we first scaled the OpenSim generic model to generate a subject-specific model by matching the Vicon-recorded markers in a static standing trial and the virtual markers on the generic model. We then applied the inverse-kinematics method to predict joint kinematics (e.g. joint angles and positions) by matching the recorded markers and the virtual markers on the scaled model for the roofing trials. These predicted joint centres serve as the ground-truth labels for the neural network's supervision. In our model, the joint definitions were similar to those of the Human3.6 M (Ionescu et al. 2014); the only difference is that we did not generate the coordinates of the nose. Thus, the number of joints in our experiment was 16 instead of 17 in Human3.6 M. Figure 1 shows the Vicon-recorded markers and the joint centres (shown in the stick models) estimated via OpenSim from three different perspectives. The input images extracted from the videos are cropped to a size of 384×384 before passing into the neural network. For our deep learning-based approach, we used the data from subjects 1 to 3 for training and those of subjects 4 to 7 for testing.

2.2. The neural network

In the present study, we present a two-stage marker-free 3D human pose estimation method for roofing workers which consisted of a multi-view pose estimation model and a multi-frame pose model. The multi-view model, based on an existing 3D pose estimation model (Iskakov et al. 2019), generates roofing worker poses from three different views. The multi-frame model was used to fuse temporal information to further improve the precision and stability of the multi-view model result. Figure 2 shows an overview of our model architecture.

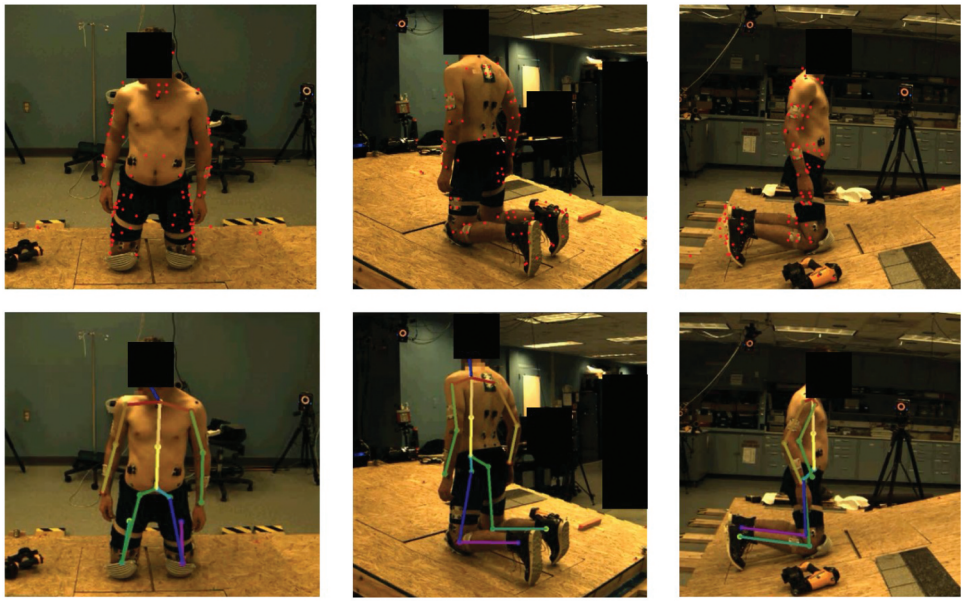

Figure 1. The Vicon-recorded markers (the top row) and the OpenSim-predicted joint centres shown in the stick models (the bottom row) from three different views.

2.2.1. The multi-view model

The multi-view model estimated the 3D workers pose from three different views for each frame. In this case, we employed the algebraic model (Iskakov et al. 2019), which is one of the state-of-art multi-camera 3D human pose models. This network uses ResNet-152 (He et al. 2016) as the backbone to obtain heat maps of 2D human pose and computes the softmax across the spatial axes to get the 2D positions of the joints. The calculation formula of 2D position is as follows:

$$H'_{c,j} = exp(a h_\theta(I_c)_j) / (\sum_{r_x=1}^{W} \sum_{r_y=1}^{H} exp(a h_\theta(I_c)_j))$$
$$x_{c,j} = \sum_{r_x=1}^{W} \sum_{r_y=1}^{H} r(H'_{c,j}(r)) \quad (1)$$

where I_c is the cropped images of human, h_θ is the 2D backbone. W and H correspond to width and height of the heatmap output by model. a is the softmax inverse parameter. To strengthen the model's confidence in the 2D joint heatmaps, we use softargmax to calculate the 2D positions of the joints in their corresponding heatmaps and get the 2D joint keypoints $x_{c,j}$ as the output. The 2D positions of three views with the confidences are then passed to the algebraic triangulation module, which outputs the 3D worker pose.

The 3D coordinates of the algebraic model results were normalised for each subject, because different subject positions were slightly different. These variations did not have a significant impact on the image based human pose estimation model, because the convolution network was translation independent. However, the

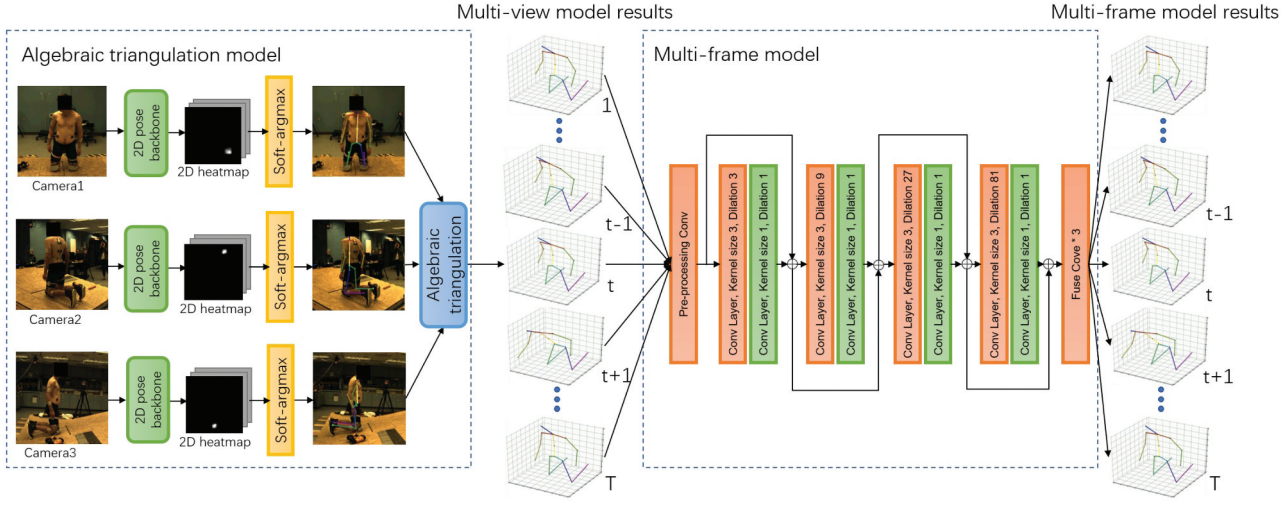

Figure 2. An overview of the proposed two-stage model – the multi-view model and the multi-frame model. The multi-view model is to estimate the 3D body pose based on a single frame; the multi-frame model further improves the accuracy and stability of estimations by aggregating the multi-frame information.

input of the multi-frame model were coordinate sequences instead of images captured during the task, which was important to normalise the coordinates of all subjects.

2.2.2. The multi-frame model

The multi-frame model is to improve the precision and stability of a 3D pose sequence by combining multi-frame information. The details of the network structure are shown in Table 1, which was based on the temporal convolution (Pavllo et al. 2019) with four residual blocks. Temporal convolution has been widely used in audio processing (Rethage et al. 2018), machine translation (Gehring et al. 2017), and other sequential processing tasks. Compared with the Recurrent Neural Network (RNN)-based model, the Convolutional Neural Network (CNN)-based model is more suitable for parallel processing. The gradient path of back propagation will not increase with the input sequence length, which mitigates gradient vanishing and exploding when processing the sequence. Therefore, we chose to use temporal convolution to build our model.

2.2.2.1. Network architecture.

The input of the multi-frame model is a 3D coordinate sequence of length T, and each time step contains 3D coordinates of J joint positions (×, y, z). We used a convolution layer of kernel size 3 to preprocess the coordinates and then sent them into four residual blocks surrounded by a skip connection. The residual network (He et al. 2016) was chosen since our model was used to adjust the input coordinates but not to recalculate the joint coordinates. Each residual block contained two convolution layers whose kernel sizes were 3 and 1. Dilation (Yu and Koltun 2016) is used to change the size of the receptive field, which can effectively expand the size of the receptive field of CNN without a pooling layer. Compared to pooling, dilation convolution has learn-able parameters which effectively avoid the loss of detail information. This advantage makes dilation convolution widely used in semantic segmentation and other tasks that need to retain detailed information. With the accumulation of residual blocks, the size of the receptive field was expanded due to the increase in the dilation values. In our model, the dilation values of the four residual blocks were 3, 9, 27, and 81, resulting in a network with the largest receptive field (245 frames) under the current structure. In the experimental section, this set of dilation values was indicated as the optimal choice for our task. Finally, we used three ordinary convolution layers to sort out the output of the residual network and made the output data dimension equal to the input data.

2.2.2.2. Loss function.

Mean Squared Error (MSE) has been used to calculate the error between the processed coordinate sequence and the ground truth as loss function. Similar to Rayat Imtiaz Hossain and Little (2018), we also impose a temporal smoothness constraint in loss function which calculates the mean of the L2 norm of the first order derivative of T frames 3D joint coordinate. This constraint can effectively improve the robustness of the model. The overall loss function is:

$$L(P_{tj}, \hat{P}_{tj}) = \frac{1}{T * J} \sum_{t=1}^{T} \sum_{j=1}^{J} \left(P_{t,j} - \hat{P}_{t,j}\right)^2 + \frac{\alpha}{(T-1)*J} \sum_{t-1}^{T-1} \sum_{j=1}^{J} \left(\hat{P}_{t+s,j} - \hat{P}_{t,j}\right)^2 \quad (2)$$

where $\hat{P}$ denotes the estimated 3D joint locations while P denotes the ground truth of 3D joint. The length of sequence is T, and the number of joints is J. The α is scalar hyper-parameters to control the significance of temporal smoothness constraint.

2.2.2.3. Padding.

The multi-frame model has a large receptive field (245 frames), and it can handle the input video that is shorter than the model's receptive field. For the video less than the model's receptive field (i.e. the video with 244 or less frames), the padding frames may be needed. We used replicate padding in the model, where the padded frames were considered as static frames. It is a reasonable assumption since most of the padded frames were needed at the beginning or the end of the video when the worker was often stationary.

2.2.2.4. Grouping.

Different from the role of group convolution in the previous studies (e.g. compressing the model (Krizhevsky et al. 2017) or increasing the calculation speed (Howard et al. 2017), we used group convolution to further improve the precision of the output 3D coordinates. In the ungrouped convolution layer, the features of each output channel are affected by the features of all input channels. This structure can process RGB images effectively, because there is a strong correlation between the different channels of the image. However, our multi-frame model takes the 3D coordinates of J joints as input, and the dimension of the input vector of each frame is J × 3. Using an ungrouped convolution layer means that each joint position is inevitably affected by the other J-1 joint positions. This outcome is not reasonable in our task (Figure 3). The figure clearly illustrates that the position of the feet keeps stationary while the wrists move rapidly. The relationship between the two joint positions is very weak. It is not a wise idea to use all input features to calculate each output feature. Therefore, we add groups to the network.

In our model, the pre-processing layer and four residual blocks are all divided into J groups, which correspond to J different joints of the model input. They calculate the position offset of each input joint independently without interference from other joint positions. As movement tends to change the three 3D coordinates of the joint simultaneously, a strong correlation exists between these 3D coordinates and no further grouping is required.

Table 1. The proposed multi-frame model architecture.

	Preprocessing Conv	Block 1	Block 2	Block 3	Block 4	Fuse Conv1	Fuse Conv2	Fuse Conv3
Kernel Number	4096	4096	4096	4096	4096	1024	1024	48
Kernel Size	3	3,1	3,1	3,1	3,1	3	1	1
Dilation	1	3,1	9,1	27,1	81,1	1	1	1
Group	16	16	16	16	16	1	1	1

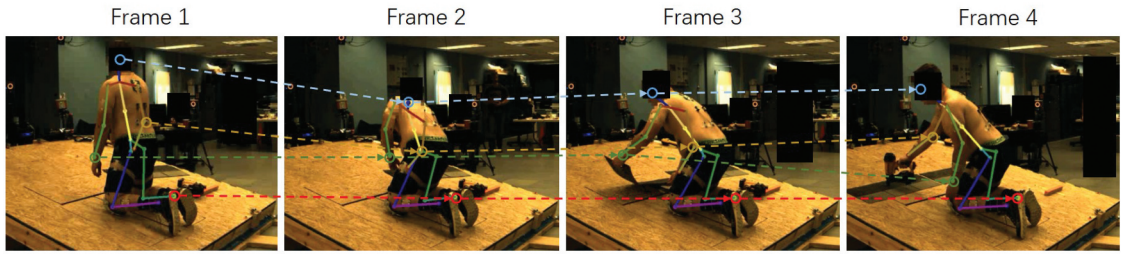

Figure 3. Motivation of grouping joints for accurate estimation. Different joints are marked with different colours. In this shingle-installation action, the lower body of the worker keeps stationary while his upper body moves frequently. Processing body joints in separate groups can effectively address issues of occlusions and asynchronous movements.

2.2.2.5. Training strategy. The multi-frame model is to improve the precision of the 3D coordinate sequence. Given the limited human pose data, training our model on the recorded pose data directly may not make the network output more precise coordinates, the output error is even greater than the input. To solve this problem, we pre-trained the network and made it output the same sequence as the input, and then tried to improve the performance of the network. In the pre-train stage, we used the input pose sequence with random noises (including translation and scaling) as labels to train the network (Figure 4). Random data is infinite, greatly reducing the need of our model for real data. The noises added to each frame within the same video were exactly the same to maintain the pose coherence among the frames with the assumption that the 'noisy' environment was consistent during the same trial. The loss function of the pre-training stage is shown as follows:

$$L(\hat{P}_{t,j}) = \frac{1}{T*J}\sum_{t=1}^{T}\sum_{j=1}^{J}\left(f(\lambda_j\hat{P}_{t,j}+\mu_j)-(\lambda_j\hat{P}_{t,j}+\mu_j)\right)^2 \quad (3)$$

where λ_j and μ_j are the random scaling and translation of the joint position. λ_j and μ_j are the random scaling and translation of the joint position. f is our multi-frame model. When the output sequence is similar to the input sequence, ground truth is used to train the network. The effects of this pre-training strategy on the performance of the model was evaluated using the experimental data.

3. Results

3.1. Evaluation metrics

Mean per joint position error (MPJPE) is the most common 3D human pose evaluation metric for calculating the average Euclidean distance between estimated 3D joint coordinates and corresponding ground truth. In a video, MPJPE is defined as follows:

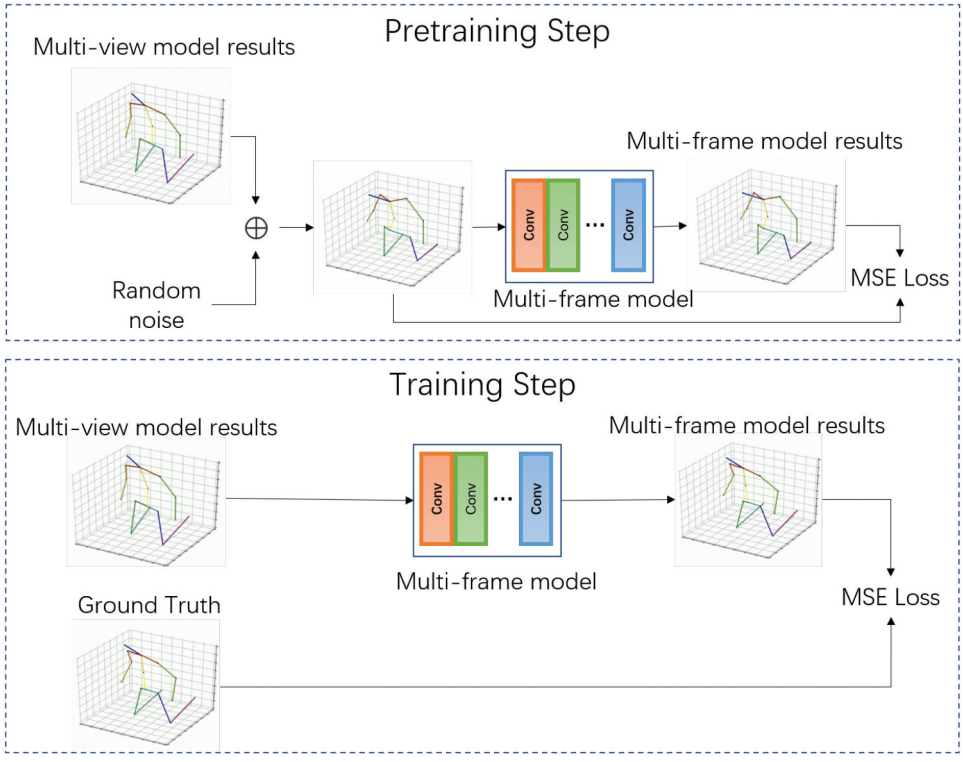

Figure 4. The proposed two-step model training. We first pre-trained the model by applying random noise to multi-view model outputs; then fine-tuned the pre-trained model with ground truth annotations.

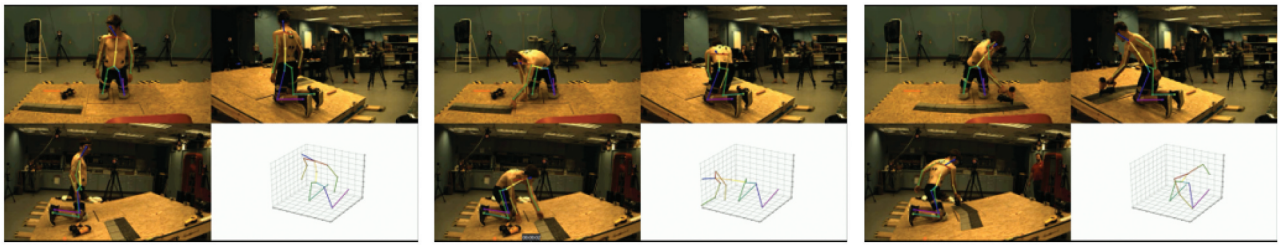

Figure 5. Demonstration of our model outputs: 2D pose estimation for each view and a uniform 3D pose estimation.

$$MPJPE = \frac{1}{T * J}\sum_{t=1}^{T}\sum_{j=1}^{J}(P_{tj} - \hat{P}_{tj})^2 \quad (4)$$

where T is the video length, J is the number of joints, P_{tj} is the ground truth 3D coordinate of joint j in frame t, and $\hat{P}_{tj}$ is the predicted 3D coordinate of joint j in frame t.

The MPJPE of the original algebraic model trained by Human3.6 m was 62.95 mm in our testing data set. We fine-tuned this model on our training data set and brought its MPJPE down to 27.93 mm in our testing data set. We used Mean Square Error (MSE) as the loss function, Adaptive Moment Estimation (Adam) (Kingma and Ba 2015) as the optimiser, and the learning rate was 10^{-5}. Finally, we used the fine-tuned algebraic model to estimate the 3D human pose coordinates for our entire data set. In the following steps, our multi-frame model took these coordinates as input and output more accurate coordinates of human pose by combining the information between different frames.

3.2. Evaluation results

The representative results on the roofing data set are shown in Figure 5, which shows the human pose images from three different perspectives as input, and the final output of the 3D human skeleton. Figure 6 shows the estimation errors (MPJPE) of different joints over time in our roofing data set. The MPJPE curve of the human pose coordinate sequence before reprocessing by multi-frame model is drawn by the green line, and the MPJPE curve of the processed sequence is drawn by the blue line. For most joints, our multi-frame model effectively improves the accuracy and stability of the multi-view model outputs.

4. Discussion

4.1. Comparisons to other models

To further prove the effectiveness of our model, we compared the performance of our model with those of other common sequential processing models on our data set. We compared four different sequence processing methods (Table 2): the traditional methods (mean filter), RNN(Long Short-Term Memory (LSTM), ungrouped CNN, and our multi-frame model. Among these models, the window size of the mean filter is five frames, same as the experiment in (Rayat Imtiaz Hossain and Little 2018). Temporal Convolution (Temporal Conv) is the network structure used in (Pavllo et al. 2019), which does not have grouping step.The comparison between the Temporal Conv and our model shows the importance of grouping. The LSTM model has exactly the same structure as (Lin et al. 2017), which has 1024 hidden units and one fully-connected layer. Instead of estimating 3D human pose from a 2D pose sequence, we use this network to reprocess the existing 3D coordinate sequence. We also tried to train the LSTM with/without pre-train. The experimental results show that compared with other methods, our proposed model can improve the precision of 3D coordinate sequence more effectively. The traditional method can only reduce the sequence oscillation, but can't effectively improve the precision. The ungrouped temporal convolution network can improve the precision of 3D coordinate sequence, but the effect is very limited, further proof that grouping plays an important role in our task. Finally, the performance of the LSTM is significantly improved by pre-training, but its performance is still

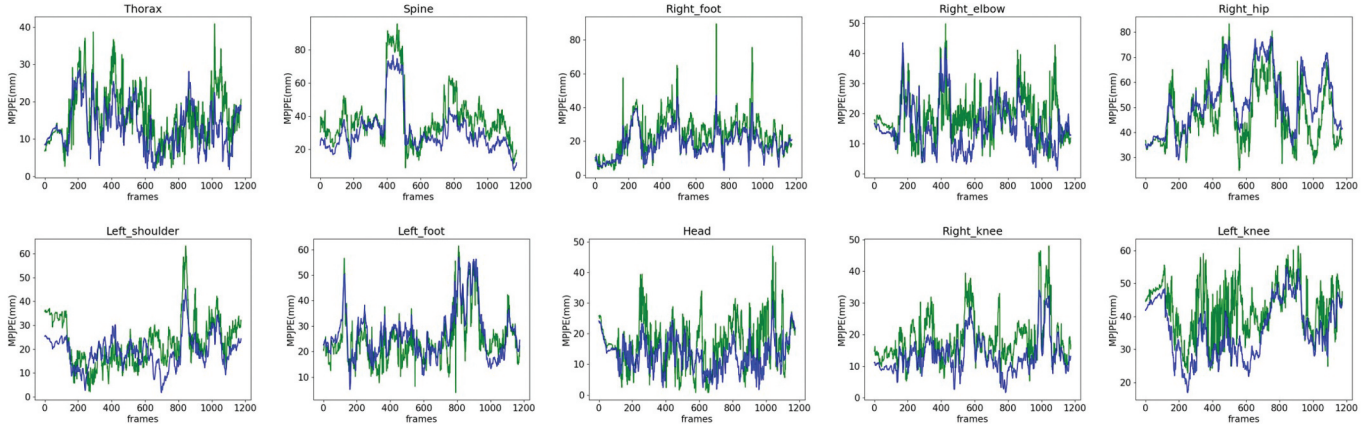

Figure 6. Estimation errors of different joints over time. green: the multi-view model outputs. blue: the multi-frame model outputs. The multi-frame model can consistently improve the accuracy and stability of the multi-view model outputs.

Table 2. Comparison between the proposed multi-fame model and other baseline sequential models.

	MPJPE(mm)
Multi-view result (baseline)	27.93
Mean filter	27.45
LSTM	39.56
LSTM+Pre-train	25.77
Temporal Conv	26.11
The proposed multi-frame model	24.81

Table 3. Ablation study of using different dilation sizes (corresponding to receptive sizes).

	Block 1 dilation	Block 2 dilation	Block 3 dilation	Block 4 dilation	Receptive field
Recept-21	1	1	3	3	21
Recept-37	1	3	3	9	37
Recept-85	1	3	9	27	85
Recept-101	3	9	9	27	101
Recept-137	3	9	27	27	137
Recept-245	3	9	27	81	245

lower than our grouped multi-frame model. Considering that Convolutional models enable parallelisation while RNNs (LSTM) cannot be parallelised, our model has great advantages in practical application.

4.2. Ablation study of the multi-frame model

In the context of machine learning (especially complex deep neural networks), we employed an ablation study to evaluate the effects of different network architecture and the pre-training on the model performance. The results show that the selected network structure and parameters in the present study are reasonable, and pre-training is crucial for model performance.

4.2.1. Different perceptive fields

Unlike processing images, the perceptive field plays an important role in the model because it indicates the number of frames acquired by the model. Therefore, we carried out experiments to see how the multi-frame model performs for different perceptive fields. By adjusting the dilation values, we get six networks with different perceptive fields, as shown in Table 3. Among these networks, Recept-245 corresponds to the max receptive field of our model, in which a larger dilation will lead to discontinuity of the receptive field. We then trained them using the ADAM optimiser with the same parameters. The results are shown in Figure 7, it can be found that for the pre-trained model, increasing the receptive domain can improve the model performance. A larger receptive field means that the model can get more frames to correct for coordinate errors. Hence, the results of this experiment are reasonable.

4.2.2. Different network sizes

We also investigated the impact of different network sizes on model performance, as shown in Figure 8. Considering that all layers of our model have the same number of convolution kernels except the fuse layers, we set five different numbers of kernels for all layers in the experiment. The experimental results show that the model performance will be improved slightly with the expansion of network scale. When the kernel number is 4096 and the reception field is 245, the network has the best performance.

4.2.3. Dilation and pooling

The performance of the network using pooling and dilation convolution were also investigated under different receptive fields, as shown in Table 4. The experimental results show that dilation convolution can retain more information than pooling, and achieve better results on the roofing data set.

4.2.4. Pre-Training

Figure 7 also shows that pre-training greatly improves the performance of the model. Considering that the MPJPE of the input 3D coordinate sequence is 27.93 mm, the result of direct training is

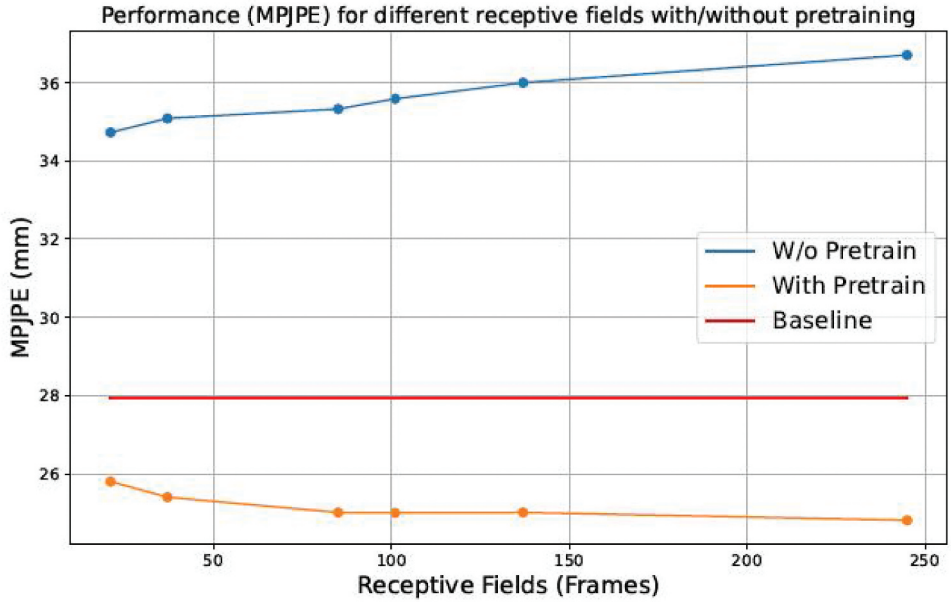

Figure 7. Performance (mpjpe) for different receptive fields, with and without pre-training.

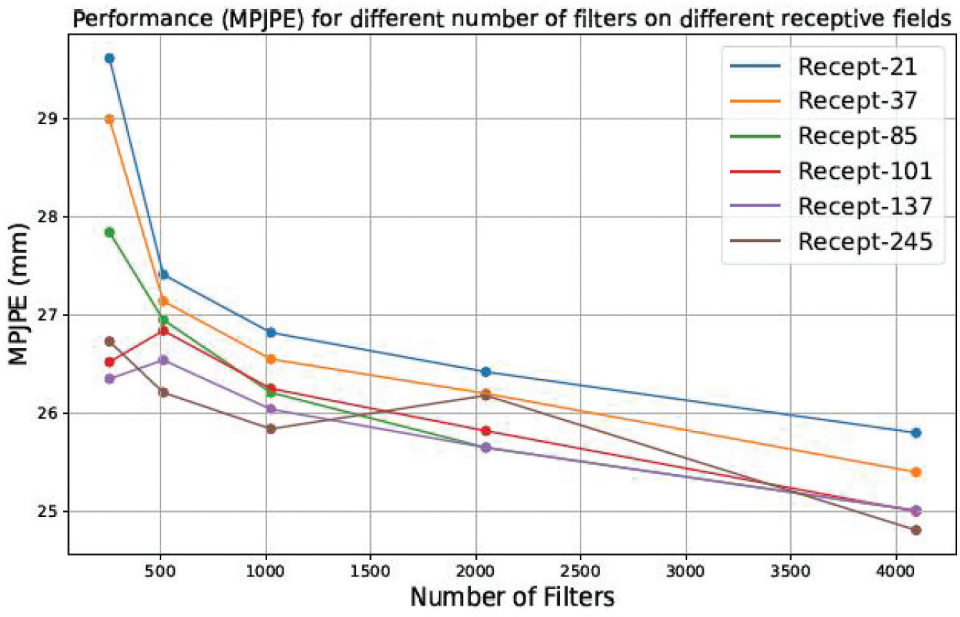

Figure 8. Ablation study on multi-frame model's number of convolutional filters on different receptive fields. Performance (mpjpe) for different number of filters on different receptive fields.

Table 4. Ablation study of using pooling layers and dilation convolution layers.

	Use pooling	Use dilation convolution
Recept-21	27.31	25.8
Recept-37	26.57	25.4
Recept-85	26.31	25.01
Recept-101	26.12	25
Recept-137	25.78	25.01
Recept-245	25.54	24.81

poor. The model without pre-training cannot improve the accuracy of the input coordinate sequence regardless of how large the perceptive field is used. The coordinate accuracy of the output is much lower than the input, and the model cannot maintain the quality of the input sequence. This outcome may be attributed to insufficient training data, given that human pose data are difficult to collect. Clearly, pre-training greatly improved the performance of our model, where the MPJPE of the output sequence decreased by more than 10% compared to that of the input sequence in our testing data set. Therefore, adding a pre-training process by using random data is an important step to increase the accuracy.

4.3. Contributions and limitations

Our two-stage motion estimation approach can effectively and efficiently estimate 3D worker poses in the roofing task. The multi-view model was pre-trained offline using the Human3.6 M, which greatly reduced the computational cost. Through the multi-view model, the estimated human body postures are more accurate and stable. The accuracy and stability are crucial to the human movement analysis in the biomechanical domain since they will largely affect the following musculoskeletal loading analysis in the future work.

It is worth noting that there are limitations in applications of our work. When collecting multi-view video data for the pose estimation, we need to prearrange multiple video cameras and require the camera configuration parameters. Acquiring synchronised camera video and configuration data often involves the specialised hardware and software support. The experimental part of the work may still require a certain amount of set-up time and effort. In addition, due to the limitation of the multi-camera coverage, the workers may not be able to move in a wider range as they prefer in a real job site.

5. Conclusion

We present a two-stage marker-free 3D motion estimation for the shingle installation task in residential roofing in this paper, which can accurately predict construction workers'pose based on video data acquired from three viewpoints. The first stage is to build a multi-view 3D human pose estimation model based on an algebraic model, and its responsibility is to calculate the worker's pose for each frame. The second stage is to incorporate a multi-frame human pose model that can improve the precision and stability of a 3D pose sequence by fusion temporal information. Compared to other marker-free methods, our approach fuse the information from multiple views and frames at the same time. Moreover, our model can be trained on a single GPU board. Concurrently to the camera data, the Vicon motion capture system collected 3D marker data of seven subjects. Compared to the gold-standard marker-based human pose estimation, the performance evaluation results show that our approach can estimate the worker pose efficiently and accurately. Moreover, the multi-frame model can further improve the accuracy of coordinate sequences by integrating temporal information from different frames. The values of MPJPE before and after reprocessing are 27.93 and 24.81 mm, respectively. This result indicates that this two-stage marker-free motion estimation approach based on deep learning is sufficient to accurately capture the posture of the workers and provides a basis for future musculoskeletal motion analysis. Future work will focus on using deep learning techniques to further analyse the relationship between workers' joint

kinematics and musculoskeletal loading which could provide more insights on work-related musculoskeletal injuries.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centersfor Disease Control and Prevention (NIOSH/CDC). Mention of any company or product does not constitute endorsement by NIOSH/CDC.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the National Institute for Occupational Safety and Health (NORA Small Grant 29390DSW) and partially supported by the US National Science Foundation (Grant IIS 1703883).

ORCID

Liying Zheng http://orcid.org/0000-0003-1610-3795
Robert E. Carey http://orcid.org/0000-0002-4648-4464

References

Bureau of labor statistics (BLS). 2013 6. Nonfatal cases involving days away from work: selected characteristics (2011 forward). [accessed 2013 Nov 9]. http://www.bls.gov/data/#injuries.

Delp SL, Anderson FC, Arnold AS, Loan P, Habib A, John CT, Guendelman E, Thelen DG. 2007. OpenSim: open-source software to create and analyze dynamic simulations of movement. IEEE Trans Biomed Eng. 54 (11):1940–1950. doi:10.1109/TBME.2007.901024

Dong X, Men Y, Fujimoto A, et al. 2008. The construction chart book. USA, CPWR-The center for Construction research and Training. 6.

Gehring J, Auli M, Grangier D, Yarats D, and Dauphin YN. 2017. Convolutional sequence to sequence learning. In: *International Conference on Machine Learning*; Sydney, AUSTRALIA. PMLR. p. 1243–1252.

He K, Zhang X, Ren S, and Sun J. 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; LAS VEGAS, US. p. 770–778.

Hossain MRI and Little J. 2018. Exploiting temporal information for 3d human pose estimation. *ECCV*.

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, and Adam H. 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Honolulu, Hawaii, US.

Ionescu C, Papava D, Olaru V, Sminchisescu C. 2014 . Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell. 36(7):1325–1339. doi10.1109/TPAMI.2013.248

Iskakov K, Burkov E, Lempitsky V, and Malkov Y. 2019. Learnable triangulation of human pose. In: *Proceedings of the IEEE International Conference on Computer Vision*; Seoul, Korea (South). p. 7718–7727.

Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Kanade T, Nobuhara S, and Sheikh Y. 2015. Panoptic studio: a massively multiview system for social motion capture. In: *2015 IEEE International Conference on Computer Vision (ICCV)*; Santiago, Chile. p. 3334–3342.

Katircioglu I, Tekin B, Salzmann M, Lepetit V, Fua P. 2018. Learning latent representations of 3d human pose with deep neural networks. Int J Comput Vis. 126(12):1326–1341. doi:10.1007/s11263-018-1066-6

Kingma DP and Ba J. 2015. Adam: Adam: A method for stochastic optimization. In: *The International Conference on Learning Representations*; San Diego, CA, US.

Krizhevsky A, Sutskever I, Hinton GE. 2017. Imagenet classification with deep convolutional neural networks. Commun ACM. 60(6):84–90. doi:10.1145/3065386

Lin M, Lin L, Liang X, Wang K, and Cheng H. 2017. Recurrent 3d pose sequence machines. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Honolulu, Hawaii, US. p. 5543–5552.

Lin M, Lin L, Liang X, Wang K, and Cheng H. 2017. Recurrent 3d pose sequence machines. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Honolulu, Hawaii, US. p. 810–819.

Martinez J, Hossain R, Romero J, and Little JJ. 2017 October. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; Venice, Italy.

Moreno-Noguer F. 2017. 3d human pose estimation from a single image via distance matrix regression. In : *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Honolulu, Hawaii, US. p. 1561–1570.

Pavlakos G, Zhou X, Derpanis KG, and Daniilidis K. 2017. Coarse-To-Fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Honolulu, Hawaii, US. p. 7025–7034.

Pavllo D, Feichtenhofer C, Grangier D, and Auli M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Long Beach, CA, US. pp. 7745–7754.

Pavllo D, Feichtenhofer C, Grangier D, and Auli M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Long Beach, CA, US. p. 7753–7762.

Pfister T, Charles J, and Zisserman A. 2019. Flowing convnets for human pose estimation in videos. In: *Proceedings of the IEEE International Conference on Computer Vision*; Seoul, Korea (South). p. 1913–1921.

Qiu H, Wang C, Wang J, Wang N, and Zeng W. 2019. Cross view fusion for 3d human pose estimation. In: *International Conference on Computer Vision (ICCV)*; Seoul, Korea (South).

Rayat Imtiaz Hossain M and Little JJ. 2018. Exploiting temporal information for 3d human pose estimation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; Munich, Germany. p. 68–84.

Rethage D, Pons J, and Serra X. 2018. A wavenet for speech denoising. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; Calgary, AB, Canada; IEEE. p. 5069–5073.

Sutskever I, Vinyals O, and Le QV. 2014. Sequence to sequence learning with neural networks. *NIPS*.

Yu F and Koltun V. 2016. Multi-Scale context aggregation by dilated convolutions. In: *The International Conference on Learning Representations*; Caribe Hilton, San Juan, Puerto Rico.

Zhou X, Huang Q, Sun X, Xue X, and Wei Y. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*; Venice, Italy. p. 398–407.