Video-Based 3D pose estimation for residential roofing-dataset

Introduction

To protect residential roofing construction workers from both fatal and musculoskeletal injuries, it is necessary to assess the musculoskeletal and biomechanical risks in residential roofing tasks. This undertaking requires accurate information of workers' 3D body positions to analyze kinematics and kinetics of the human body. In this study, we proposed a novel 2-stage motion estimation approach based on a convolution neural network to estimate residential roofer's body positions using three-view video data. Instead of pursuing end-to-end training, our approach includes two stages: (1) use a multi-view model to estimate the 3D pose in a single frame; (2) use 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 estimation ("ground truth"). The evaluation results show that our marker-free video-based approach can accurately capture the 3D posture of workers during the common roofing task and the proposed multi-frame model can effectively improve the precision of the coordinate sequence. The values of mean per joint position error of estimated human position before and after processing by the multi-frame model are 27.93 and 24.81 mm, respectively. These results prove that the proposed marker-free motion capture estimation approach can efficiently and accurately locate 3D body joints and pave the way for future onsite musculoskeletal motion analysis during roofing activities.

Methods

Ground Truth

- The whole-body marker data (total 79 makers placed on each subject) were recorded using a Vicon motion capture system with 14 optical cameras (Vantage V16, Vicon Motion System Ltd., Oxford, UK).
- An open-source musculoskeletal modeling software—OpenSim—was used to estimate the joint positions based on the recorded marker locations. The joint definitions were similar as in the Human3.6M (the largest existing public dataset of 3D human poses).

Multi-view model (stage 1)

- The learnable triangulation model was chosen for the multi-view model since it is one of the state-of-art multi-camera 3D human pose models. This network used ResNet-152 as the backbone to obtain heat maps of 2D human pose and computed the softmax across the spatial axes to get the 2D positions of the joints.
- The learnable triangulation model was pre-trained by the public Human3.6M video data and fine-tuned by the recorded roofing video data.
- 3D coordinates of the algebraic model results were normalized for each subject.

Multi-frame model (stage 2)

- Multi-frame model consisted of several grouped temporal convolution layers with the dilation.
- Before using the ground truth to train our model, the model was pre-trained using the

human pose data with random noise. Human pose data with noise has been used as both the input and the label of the network to train the network.

- Mean Squared Error (MSE) was used to calculate the error between the processed coordinate sequence and the ground truth as a loss function. A temporal smoothness constraint was used in the loss function by calculating the mean of the L2 norm of the first order derivative of the 3D joint locations.

Citations

Wang R, Zheng L, Hawke AL, Carey RE, Breloff SP, Li K, Peng X. Video-Based 3D pose estimation for residential roofing. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2022 May 14:1-9.

Acknowledgements

This work was supported by NIOSH (CAN:19390DSW).
Ruochen Wang, ruochen@udel.edu
Liying Zheng, lwf5@cdc.gov
Ashley L. Hawke, ptt3@cdc.gov
Robert E. Carey, ohn7@cdc.gov
Scott P. Breloff, lqz0@cdc.gov
Kang Li, kl419@rutgers.edu
Xi Peng, xipeng@udel.edu

Contact

For further information contact:
Physical Effects Research Branch, Health Effects Laboratory Division
National Institute for Occupational Safety and Health (NIOSH), Morgantown, WV, USA
Phone: (304)285-6121