Materials and Methods

Seven healthy male subjects participated in this study and performed one of the most common roofing tasks—a shingle installation task. Each person was asked to perform three trials of the shingle installation task. We collected whole-body marker data (total 79 makers placed on each subject) using a Vicon 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 1280x720, and the frame rate was 100 Hz. The video data set contains 63 videos of 7 subjects, and each with 3 trials and 3 perspectives. The whole-body joint positions of the subjects were estimated from three sources—the "ground truth" estimation based on the marker data, and two outputs from a two-stage deep-learning-based motion estimation using the video data (where the first stage used a multi-view model to estimate the 3D pose in a single frame and the second stage used a multi-frame model to apply temporal convolutions to refine the multi-view outputs).

The ground truth estimations of the joint positions were based on the marker data recorded by the Vicon system, which is currently the gold standard method to obtain the joint kinematics. An open-source musculoskeletal modeling software—OpenSim—was used to estimate the joint positions based on the recorded marker locations. The joint kinematics were estimated by matching the virtual markers placed on the OpenSim musculoskeletal models to the recorded real markers on the human subjects, which was referred as the inverse kinematics approach. The missing marker issue was also resolved by the inverse-kinematics approach in OpenSim. In the present study, the joint definitions were similar as in the largest existing public dataset of 3D human poses—Human3.6M.

The multi-view model estimated the 3D joint locations from three different views for each frame. The algebraic model from "Learnable Triangulation of Human Pose" was served as the base model since it is one of the state-of-art multi-camera 3D human pose models. This network uses ResNet-152 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 algebraic model was pre-trained offline using the public Human3.6M video data and refined using the recorded roofing video data. Moreover, the 3D coordinates of the algebraic model results are normalized for each subject.

The multi-frame model further improved the precision and stability of a 3D pose sequence by combining multi-frame information. The model had several grouped temporal convolution layers and the dilation was used to change the size of the receptive field. The input of the multi-frame model was a 3D joint coordinate sequence. We used a convolution layer of kernel size 3 to preprocess the sequence and then sent it into four residual blocks surrounded by a skip connection. With the accumulation of residual blocks, the dilation values increased to expand the size of the receptive field. 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 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.