3D Human Pose Reconstruction for Ergonomic Posture Analysis

Wenjing Chu

A Thesis in

The Department of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Science (in Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

October 2018

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Wenjing Chu

Entitled: 3D Human Pose Reconstruction for Ergonomic Posture Analysis

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Civil Engineering

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Dr. M. Nik-Bakht Chair

Dr. A. Hammad Examiner (External)

Dr. F. Nasiri Examiner

Dr. Z. Zhu Thesis Supervisor(s)

Dr. S. H. Han

Approved by _____

Date _____

ABSTRACT

The rapid development of the modular construction industry has produced the social concerns about workers' health and safety in the factory-controlled construction processes. According to the reports from Association of Workers' Compensation Boards of Canada (WCBC), approximately 2 in 100 workers are injured due to their awkward and improper postures and motions in the modular construction industry in Canada. The occurrence of injuries and accidents not only reduces the productivity but also increases the project cost. In this respect, the ergonomic posture built upon the self-report, manual observations, direct measurement or computer vision, is essential to identify, mitigate and prevent these postures for safety and productivity improvement. Advanced computer vision technologies have made the vision-based ergonomic posture analysis cost-effective in real workplaces. So far, several vision-based methods have been created to obtain the anthropometry data, such as joint coordinates and body angles, which are required for the ergonomic posture analysis. However, there are still some challenges like occlusions and lack of accuracy in complex working environments to reduce the reliability and robustness of these vision-based methods in practice.

This research proposes a novel framework that acquires the body joint angles for ergonomic posture analysis by reconstructing the 3D worker body with the 2D videos recorded from a monocular camera. The framework consists of (1) human tracking in the

given videos; (2) 2D body joints and body parts detection using the tracking results; (3) 2D pose refining based on integrating the 2D joints detection with the body parts detection; (4) 3D body model generation and body angle calculation; and (5) ergonomic posture analysis based on the obtained body angles. The proposed framework has been tested on the videos in real factories and the test results were compared with the motion data captured by the IMU-based suit. The results showed that the average 3D pose difference was 17.51 degrees in terms of joint angles and the lowest joint angle difference was around 4 degrees.

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who helped me during the writing of this thesis. My deepest gratitude goes first and foremost to my supervisors Dr. Zhenhua Zhu and Dr. Sang Hyeok Han, for their constant encouragement and guidance. They have walked me through all the stages of the writing of this thesis and offered me valuable suggestions in the academic studies.

Besides my supervisor, many thanks go to my current and former lab colleagues which are Ms. Chen Chen, Mr. Xiaoning Ren, Mr. Yusheng Huang, Mr. Amr Amer and Mr. Bingfei Zhang. They make the lab a convivial place to work and have helped, supported and corrected me in my research. Special thanks to my parents, my boyfriend and my faithful friends. They have provided the endless support throughout my journey of graduate study.

In addition, I would like to thank my examiners, Dr. Amin Hammad, Dr. Fuzhan Nasiri and Dr. Mazdak Nik-Bakht for their time, advice and effort in reviewing my thesis. Lastly, I would like to show my gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

This research aims to propose a framework using vision-based human motion capture system to obtain the anthropometry data such as joint coordinates and body angles on the video recorded by a single-color camera. The anthropometry data can be further used as the input for the ergonomic analysis. The following sections in this chapter introduce the research background, motivation, objectives, methodology, contributions, and the organization of this dissertation.

1.1 Background and Motivation

The modular construction (prefabricated and off-site construction) is a factory-controlled process which generates less material waste and site disturbances, and faster construction schedules with the mitigation of weather delays. It is supposed to be safer and smarter than conventional on-site construction. Based on these benefits, the modular construction industry has developed significantly in recent years. The gross revenue in the modular construction industry in 2016 was roughly $3.3 billion in North America which represents a 61.8% increase from 2015 (Modular Building Institute, 2016). However, in a view of safety improvement, the Association of Workers' Compensation Boards of Canada (WCBC) reported that the manufacturing and construction industries had the second and third highest number of lost-time claims due to injuries in 2015, accounting for 14% and

11%, respectively, of total workplace injury claims (232,629) (Workers' Compensation Board, 2018). In the United States, around 11% and 7% of all nonfatal occupational injuries and illnesses happened in the manufacturing and construction industries respectively (Bureau of Labour Statistics, 2016). The Occupational Health and Safety Institute in Alberta, Canada also illustrated that there were 5,216 number of injury claims in the manufacturing industry and 8,771 number of injury claims in the construction industry, which means approximately 2.3 in 100 workers were injured in total in the modular construction industry (OHS, 2016). As shown in Figure.1.1, The overexertion that happens when the worker executes a task that is beyond his or her physical strength, was the top one cause of worker injuries in both the construction industry (20%) and the manufacturing industry (22%). In addition, the bodily reaction was the third top cause of workers' injuries in the manufacturing industry, accounting for 15%. The bodily reaction is the injuries or illnesses resulting from a single awkward motion which poses stress or strain on some parts of the body (Canadian Standards Association, 2003). The awkward postures often occur when any joint of the body bends or twists excessively or any muscles stretch over beyond a comfortable range of motion.

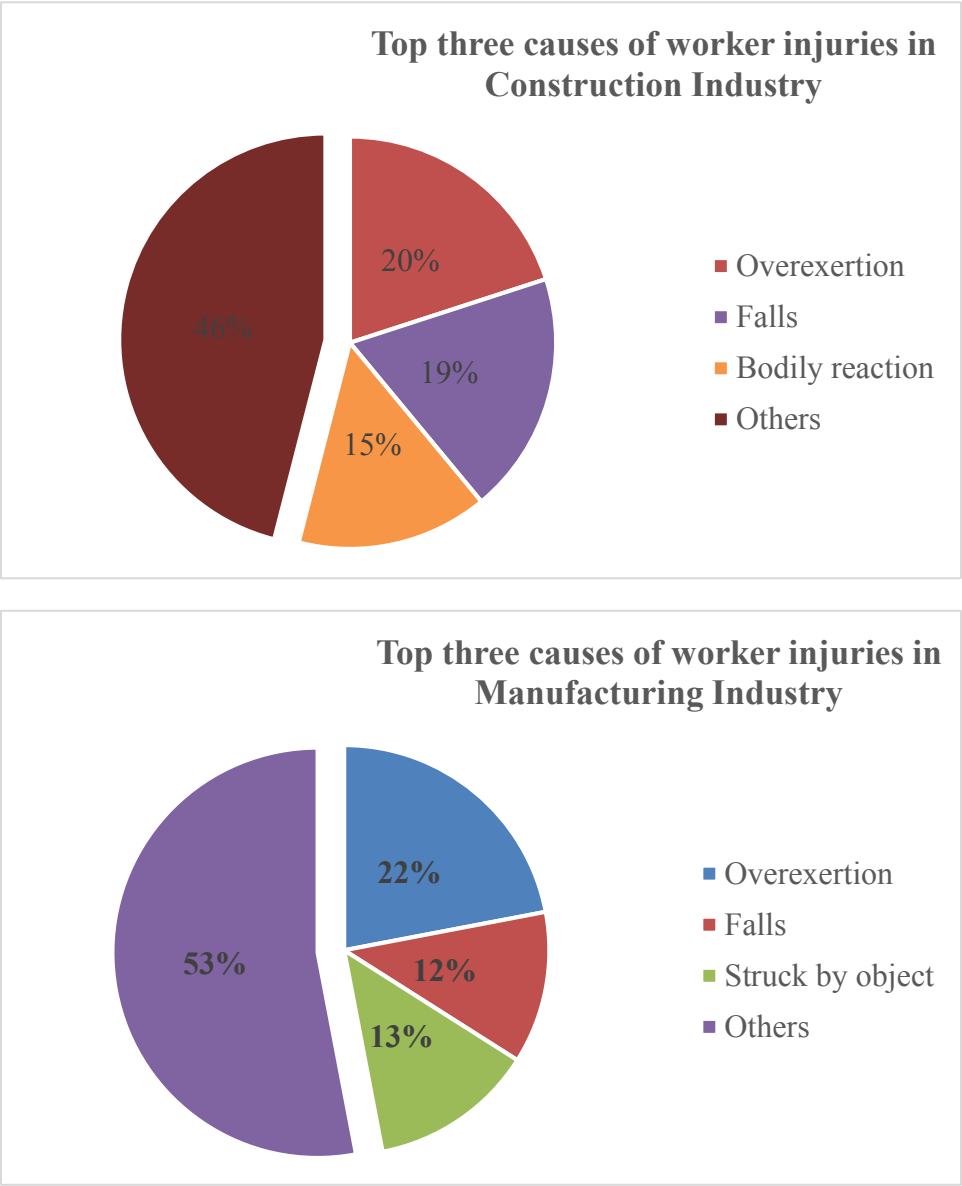

Figure 1.1 The causes of workplace injuries in Alberta

To address the issue of overexertion and awkward postures or motions of the workers, researchers and companies have focused on ergonomic posture analysis which requires 3D body angles as the input. These angles can be acquired by the following ways: (1) self-report, (2) expert observation, (3) direct measurement, and (4) vision-based methods

(Wang et al., 2015). The self-report and expert observation methods are easily to conduct with minimal disruption to workers' behaviour. However, they are highly error-prone due to the subjective judgement either by the workers themselves or the experts. The direct measurement methods use additional tools such as goniometers, force sensors, accelerometers, markers and sensors to directly capture human motion data such as 3D joint coordinates and body angles. Although the direct measurement methods are capable of providing detailed and accurate motion data, the high equipment cost, constant interference with the normal work and sophisticated instrumentation and laboratory (indoor) environmental setting preclude their widespread use in real workplaces. On the contrary, the vision-based methods collect the human motion data by marker-less sensors such as RGB-D cameras and normal colour cameras. This sort of methods does not require workers to be attached with any markers or signal receivers. Among the vision-based methods, the RGB-D camera-based methods need to be applied in the restricted range and applied in the indoor environments since their sensitivity of sunlight. Moreover, the cost of RGB-D cameras, even the most affordable Microsoft Kinect camera, is still much more than the normal colour cameras. On the other hand, the human motion capture data can be obtained effectively and efficiently from normal camera-based methods with the rapid development of related computer vision algorithms in recent years (Kale & Patil, 2016). In this respect, the normal colour camera-based human motion capture systems, as one of the most

promising computer vision-based methods, have been high attention to provide the inputs of ergonomic posture analysis. However, only few literature has researched the integration of the normal camera-based human motion capture system with the ergonomic posture analysis. In particular, the field of providing anthropometry data for ergonomic analysis by a single colour camera is largely unstudied. More details can be found in Chapter 2.

1.2 Methodology and Results

In order to provide more efficient and accurate anthropometry data for the ergonomic posture analysis which can further help reduce the workers' injuries and accidents, this research proposes a novel framework that acquires the 3D joint coordinates and body angles by reconstructing 3D human pose according to the 2D video recorded from a monocular camera. The framework contains five procedures: (1) the worker is visually tracked in the video to generate the region of interests (ROIs) on each video frame; (2) the 2D body joints and body parts of the worker in the ROIs are then detected; (3) the detected body joints and body parts are integrated to obtain the refined 2D pose; (4) the corresponding 3D human body pose and shape could be generated based on the refined 2D pose using a 3D generative model. Then the 3D body angles are computed based on the 3D generated pose; and (5) the calculated 3D body angles are used to perform the ergonomic analysis.

The proposed framework has been implemented in the Python 2.7 environments. The

real-recorded video about constructing masonry walls that provided by the (Alwasel et al., 2017) was selected to test the effectiveness of the framework. Alwasel et al. used the IMU-based motion capture suit which consists 17 sensors attached at each body segment to capture the motion data and this data was served as the ground truth that can evaluate the accuracy of the test results. The test results shows that the 3D human pose can be well generated by the proposed novel framework and the 3D body angles can be obtained effectively.

Overall, the lowest joint angle difference between the body angles generated from the proposed framework and the angles calculated using ground truth is around 4 degrees and the average joint angle difference is 17.51 degrees. It is also worth noting that the 2D pose refinement procedure in the proposed framework plays an important role on obtaining the accurate 3D model and anthropometry data. The 3D pose similarity calculated from the pose generated with the refinement procedure can be improved by 7.5%, compared with the 3D pose generated without the refinement procedure.

1.3 Contribution

The main goal of this research is to propose a novel framework to generate 3D body angles and coordinates from 2D videos recorded by the single color camera to facilitate the ergonomic analysis. The conventional ergonomic posture analysis methods are either inaccurate (self-reports and observation methods) or time-consuming and costly (direct

measurement methods). There is few literature in the area of single color-camera based ergonomic posture analysis. The results of the proposed framework show the potentials of this research direction.

The contributions of this research are listed as follows:

1. Provide the motion capture data of workers without any wearable marker or sensor attached to the human body.

2. Propose a novel single camera-based ergonomic posture analysis which almost has not been researched yet.

3. Propose an integration method to improve the 2D body pose estimation based on the 2D body parts detection result.

4. Propose a framework for ergonomic posture analysis which can help to easily identify risk of worker's injury.

1.4 Dissertation Organization

The background and motivation, methodology and results, and contributions behind this research have been introduced in this chapter. The remaining chapters in the dissertation are organized as follows:

Chapter 2 is the background literature review. It first provides a holistic view on current available techniques of ergonomic posture analysis, then followed by the overview of the representative methods of 2D and 3D human pose estimation in the field of vision-

based human motion capture system. This chapter ends up with a summary of the limitations and issues of previous works this research going to solve.

Chapter 3 presents the methodology of this research. This research proposes a novel framework to capture the anthropometry data for ergonomic analysis from the single color camera-recorded videos. The framework is separated into five main procedures. The first human tracking procedure is for locating workers on each video frame. The second procedure is responsible for the initial data acquisition task including the 2D body joint detection and the 2D body part detection. The third procedure is to refine the 2D pose based on the 2D body part detection results. The fourth procedure is for reconstructing the 3D pose and calculating the body angles which are utilized for ergonomic analysis in the last procedure.

Chapter 4 describes the implementation details, such as the hardware configuration, the validation methods and the implemented environments, with the test results of the proposed framework. Also, the discussion about the analysis of the cause of the angle errors and some research findings can be found in this chapter.

Chapter 5 is the conclusion of the whole research. The background, motivation, methodology and the test results are quickly reviewed. At the end of this chapter, future research directions about single-camera based motion capture system for the ergonomic analysis are listed.

CHAPTER 2: LITERATURE REVIEW

This chapter first provides a holistic view on available techniques of ergonomic posture analysis and their limitations and gives a conclusion that the vision-based human motion capture system has been high attention. It is then followed by the introduction of the existing, representative methods of both 2D and 3D human pose estimation in the field of vision-based human motion capture system. The 2D pose estimation methods are reviewed since the 2D pose estimation is closely related to the 3D pose estimation.

2.1 Ergonomic posture analysis

Ergonomics is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance (Waldemar, 2006). Current available techniques for ergonomic analysis can be categorized into self-report, observation, direct measurement, and vision-based (optical) methods.

2.1.1 Self-report, observation & direct measurement methods

Self-report can be used to collect data by the methods such as worker diaries, interview and web-based questionnaires (Dane et al., 2002; Inyang et al., 2012). These

methods have the advantages of being straightforward to use, applicable to a wide range of working situations, and appropriate for surveying large numbers of workers at low cost. However, the worker perceptions have been found to be imprecise and unreliable (David et al., 2005). Further, the various levels of work literacy, comprehension and question interpretation may increase the difficulty of using self-report methods (Spielholz et al., 2001).

Observation is to record human postures in the workplace by experienced observers. A number of observational tools have been developed, including Ovako Working Posture Analyzing System (OWAS) (Karhu et al., 1977), Rapid Upper Limb Assessment (RULA) (McAtamney & Corlett, 1993) and Rapid Entire Body Assessment (REBA) (Hignett & McAtamney, 2004), allowing experts to record and evaluate the worker pose in relation to the evaluation of risk factors. Similar to the self-report methods, although direct observation methods produce minimal disturbance to worker's behavior, allowing for assessment of tasks in various environments, the methods are error-prone due to the influence by subjective judgement.

Direct measurement is often used to assist or replace expert observation in order to improve the accuracy of risk assessment. A wide range of direct measurement methods have been developed that require sensors or markers to be attached directly to the workers. For instance, Lumbar motion monitor (LMM) (Marras & Granata, 1997) was developed to

assess workers' risk of low back injury. The electromyography (EMG) (Ning et al.. 2010; 2014) is used primarily in studying muscle exertions by attaching a group of sensors on the skin over the sampling muscles. The optical marker-based direct measurement tool, Vicon, uses retroreflective markers attached on the human body and multiple infrared cameras to track 3D motion of joints and body segments (Richards et al., 1999). Although the direct measurement methods can provide detailed and accurate motion data, most of them require complicated instrumentation and laboratory environments to capture human motion. The body-attached markers or sensors also affect the workers' behavior and undermine the performance of regular activities.

2.1.2 Vision-based methods

Vision-based methods concern marker-less biomechanics in which the RGB-D cameras or multiple cameras are often used to capture human motions. The RGB-D cameras can produce the depth images which can infer the point cloud to estimate the 3D human body pose. The value of each pixel in the depth images indicates the calibrated distance between the camera and the scene. One of the most commonly used RGB-D cameras, Microsoft Kinect, consists of an infrared transmitter and an infrared camera and this system is able to determine the depth of each image pixel inside the device to its corresponding location in the scene which can simplify the 3D pose reconstruction process. For instance, (Diego-Mas & Alcaide-Marzal, 2014) used a computerized OWAS to permit

the data acquisition from Kinect and data processing in order to identify the risk level of each recorded postures. (Ray & Teizer, 2012) used a predefined set of rules to categorize the captured body posture information by Kinect as ergonomic or non-ergonomic. Both of them solely focused on simple posture classification such as lifting and crawling in indoor environments. This type of methods do not need workers to be attached with marker or sensors so that they are viable to be applied in real factories (Coenen et al., 2011). But they can only be used in the short distance (i.e., less than 4 meters) and in the place without sunlight. The depth camera might produce holes in the depth images when the shooting area cannot be seen by both the projector and RGB camera. In addition, the post-processing classification and recognition are only applied to relatively simple postures and motions in a restricted range currently (Wang et al., 2015).

A typical vision-based method using multiple cameras can be found in the work of (Han & Lee, 2013). The method extracted features from 2D images and estimated correspondences on images taken from two cameras. Then the 3D skeleton can be extracted through triangulation and the pre-trained motion templates and skeleton models were compared with the extracted models to detect unsafe actions. In this method, the accuracy of extracted skeleton model and the trained model and the criterion of comparison can all affect the analysis performance. The lots of computational time required to process the data,

and the presence of errors in calculation of corresponding points which may decrease its robustness are the other two limitations of this type of methods (Seo et al., 2015).

There is few vision-based methods for ergonomic posture analysis based on a single colour camera. The latest paper proposed by (Zhang et al., 2018) used 3D view-invariant features from a single 2D camera to recognize the unsafe postures. They estimated the 3D skeletons by lifting 2D coordinates into 3D using a multi-stage convolutional neural network and a probabilistic 3D joint estimation model (Tome et al., 2017). Three posture classifiers regarding arms, back and legs are trained, so that the generated 3D pose can be classified as safe or unsafe. Their work shows the potentials in ergonomic posture analysis to improve worker's safety and healthy but the joint loadings and tissue are not considered in their work.

The biomechanical models can be introduced to estimate tissue and joint loadings, which are highly associated with worker injuries, based on the 3D anthropometry data such as both left and right joint coordinates and joint angles (Wang et al., 2015). Biomechanical models are usually applied for the post-processing of 3D human motion data captured by the direct measurement and vision-based methods. For instance, the computerized software packages such as 3D Static Strength Prediction Program (3DSSPP) (Chaffin et al., 2006), OpenSim (Delp et al., 2007), and Visual 3D (C-Motion, 2013) are available to compute joint loadings and are proven to have the potentials of using 3D motion data captured by

vision-based methods (Ray & Teizer, 2012). Thus, the better results of ergonomic analysis can be obtained using 3D motion data through the above tools.

In summary, the vision-based methods for ergonomic posture analysis should be considered as the most important research thrust due to their cost-effectiveness and their applicability in real factories. Among these methods, the RGB-D camera-based methods can produce relative accurate data since they use the sensor inside the camera to capture the additional depth information. But they are usually unable to collect data accurately in the presence of sunlight and have relatively short applicable range. Also, they are still expensive compared with the normal colour cameras. The multiple camera-based methods require plenty computational time to process the data and they may lack of robustness due to the results can be affected by many factors such as the association of view-points, the calculation of correspondent points and the reconstruction process. To the authors' knowledge, the ergonomic analysis based on the 2D videos recorded from a monocular camera has almost not been researched. Thus, the objective of this paper is to provide a novel vision-based framework for ergonomic analysis from a single colour camera.

2.2 2D human pose generation

In the field of single camera-based 2D human pose generation, the classical approach is to adopt the pictorial structures (PS) to represent the spatial relationship between various parts of the human body. Various PS models have been proposed over the years, such as

the original simple appearance model requiring background subtraction (Felzenszwalb & Huttenlocher, 2005), the cardboard model modelling appearance of body parts as rigid templates (Sapp et al., 2010) and the novel Beyond Trees model (Lan & Huttenlocher, 2005). In addition to the works concentrating on exploring better model structures, the approach proposed by (Andriluka et al., 2009) integrated a strong human body part detector with the PS model, achieving better results. Also, (Ukita, 2012) focused on using the extra contour cues to evaluate parts connectivity. However, they have one major limitation that the limbs of human must be seen clearly in the image for successfully detection of body poses. Also, the PS models are prone to characteristic errors due to the lack of the important information about human body poses, such as the relationship between body parts beyond kinematic constraints and the balance or coordination constraints. Even though the introduction of the mixtures of deformable parts model (DPM) (Felzenszwalb et al., 2010) significantly extended the application scope of the PS model, achieving more efficient and accurate results, it did not completely overcome the inherent limitations of the PS model. The works built upon the DPM did not show very promising results at the expense of substantial computational pressure (Pishchulin et al., 2013; Yang et al., 2011).

The increasingly advanced human pose estimation system was indeed boosted by the Convolutional Neural Network (CNN) which is the hierarchical feature extractor belonging to the Deep Learning technique. The first deep learning-based method for human pose

estimation which called DeepPose proposed in 2014, outperformed all the previous work in terms of the accuracy (Toshev & Szegedy, 2014). The DeepPose can result in high precision pose estimates, while it is much simpler to formulate than PS model methods since the features representations, detectors for body parts and the model topology designed in PS model methods can all be automatically learned by CNNs. In the DeepPose, the pose estimation was formulated as a joint regression problem in which the location of each body joint was directly regressed through a 7-layered convolutional network given a full image as input. Similar to the DeepPose, (Carreira et al., 2016) also directly regressed the body joint coordinates but expanded the convolutional network to encompass both input and output spaces using iterative error feedback. Nevertheless, the regression-based pose estimation methods have two common limitations: (1) their accuracy can be quite low, especially when the human body has a large deformation compared with the normal state like up-right standing; and (2) it is difficult to calculate the pose of multiple individuals shown in the same image. Thus, the regression-based methods were quickly replaced by the subsequent heatmap detection-based methods in which the result is a response-map indicating a per-pixel likelihood for each key joint location on the human skeleton. The work of (Tompson et al., 2014) proved, for the first time, that the heatmap detection-based convolutional network can obtain better results compared with the regression-based network. In their work, a convolutional network which utilized a multi-resolution feature

representation with overlapping receptive fields, was combined with a graphical model which learned typical spatial relationships between the joints. The design of the popular Hourglass network (Newell et al., 2016) largely builds off of their work, exploring the way of capturing information across scales and combing features across different resolutions. The difference is that Hourglass network do not use any graphical model to achieve superior performance but designs the network in which repeatedly using top-down to bottom-up to infer the location of the human body. Every top-down to bottom-up structure is an hourglass module. (Pfister et al., 2015) replaced the use of graphical model with additional convolutional layers in the proposed Flow Convolutional network (Flow ConvNets) to enable learning an implicit spatial model. Moreover, the Flow ConvNets worked on the videos so that it can benefit from temporal context by combining information across the multiple frames using optical flow, resulting in a significantly improved performance of 2D human pose generation. However, the method was limited to estimate only the pose of the upper body. The convolutional pose machines (CPM) proposed by (Wei et al., 2016) also proved that sequential CNN is capable of learning a spatial model for the body pose by communicating increasingly refined uncertainty-preserving beliefs between stages in the network. But the CPM method requires a massive amount of training data to achieve good estimation results.

The methods described above have been used solely for single-person pose estimation. As for the multi-person pose estimation, there are mainly two approaches which are the top-down approach and bottom-up approach. The top-down approach is to detect people at first and then estimate body pose independently. In this approach, the multi-person pose estimation task is transformed into single-person pose estimation task. For instance, (Yang & Ramanan , 2011) proposed a flexible mixture-of-parts model for human detection and pose estimation and then performing non-maximum suppression on the multiple-pose hypotheses corresponding to various root part positions. But the most of top-down approaches are solely suitable for the cases that do not have overlapping body parts. The bottom-up approach is to directly search for the joint candidates on the whole image using different feature extractor and associated them to individual people. A typical work in bottom-up approach is Part Affinity Fields (PAF) method proposed by (Cao et al., 2016). The PAF method uses an architecture of two-branch CNN to jointly predict confidence maps for body part detection and part affinity fields for part association. It can achieve real-time performance (approximately 5 millisecond per image), but its detection results are not stable. Since the PAF method adopts a greedy search algorithm to detect candidate body joints, the non-joint points in the image are likely to be detected as joint points. These extra detected joint points directly affect the accuracy of pose detection. Moreover, (Pishchulin, et al., 2016) proposed an approach named DeepCut which achieved state-of-the-art results

for both single-person and multi-person pose estimation. It adopted the bottom-up approach, first finding out and clustering all candidate joint points of an image and then determining which joint point belongs to which person by an optimization formula. The computational cost was huge in DeepCut since it utilized the adaptive fast R-CNN architecture and Integer Linear Program (ILP) at the same time which are both computationally intensive. Thus, its accelerated version, DeeperCut (Nsafutdinov et al., 2016) was introduced to adapt to the newly proposed residual net for body part extraction and achieved higher accuracy. Figure.2.1 shows the comparison result between the PAF method and the DeeperCut method. The exact location of error detection using the PAF method has been highlighted by the red rectangle boxes in the figure. Hence, the DeeperCut has been selected for the research due to its pose detection stability and the wide-range applicability not only for single pose estimation but also can be used for the multi-person pose estimation.

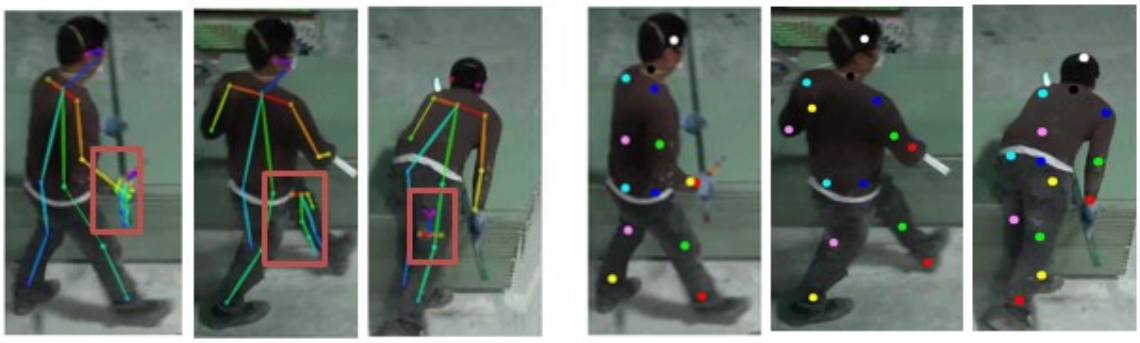

Figure 2.1 Comparison between PAF (left three) and DeeperCut (right three)

2.3 3D human pose generation

Single camera-based 3D human pose generation is a popular but much more challenging task. The challenges mainly come from the depth ambiguities, high-dimensional representation of human pose, self-occlusion and observation ambiguities (Hen & Paramesran, 2009). The loss of the depth information during the projection from 3D world to 2D image causes the depth ambiguities. Observation ambiguities means that one single 2D pose can be mapped into more than one possible 3D human pose. According to how the researchers interpret the structure of the human body, the 3D human pose generation techniques can be divided to discriminative (model-less) approaches and generative (model-based) approaches.

2.3.1 Discriminative approaches

The discriminative approaches can be further categorized into learning-based and example-based approaches (Sarafianos et al., 2016). Some learning-based works focus on learning the mapping between the low-level image observations, such as the silhouettes and edges, with 3D human body pose. For instance, (Elgammal & Lee, 2004) introduced a method to reconstruct 3D body pose based on the silhouettes information using a single camera. The mapping between the 3D pose spaces to silhouette spaces has been learned by local linear embedding (LLE) (Roweis & Saul, 2000). The work of (Atrevi et al., 2017)

also utilized the silhouette information to generate the 3D human pose. The difference is that they trained a prior to learn correspondence between silhouettes and the simulated human skeletons using Blender software. Then the detected silhouettes and the simulated silhouettes were matched using the geometrics invariant moments such as Krawtchouk Moments (Yap et al., 2003). Another branch of this approach relies on the convolutional network. The 2D pose estimation acts a crucial role for 3D pose generation in these works, such as Vnect (Mehta et al., 2017) and Hourglass (Newell et al., 2016). The former regressed 2D and 3D poses jointly by a CNN-based pose prior with Kinematic skeleton fitting and the latter generated the 3D pose using its special "stacked hourglass" network based on the successive steps of pooling and up-sampling. Both methods can only get the skeletal 3D pose of a human without the shape information and the problem of observation ambiguities was still serious.

In example-based approaches, a series of exemplars with the corresponding pose descriptors are stored and the final pose is estimated by interpolating the candidates obtained from a similarity search. For instance, (Huang & Yang, 2009) represented each test sample as a compact linear combination of training samples. By this way, the occluded test images can be recovered with a sparse linear combinations of correctly un-occluded training images. The main advantage of discriminative approaches is the fast speed and robustness due to the employed models have fewer dimensions. However, the generative

approaches are capable of inferring poses with better precision since they can handle complicated human body configurations with clothing and accessories (Sarafianos et al., 2016).

2.3.2 Generative approaches

A strong model of body shape learned from thousands of people and captured the kinematic constraints can help reduce the ambiguities, making the 3D pose generation task easier. Most of the generative (model-based) approaches consist two stages: modeling and estimation (Poppe, 2007). Modeling is the construction of the likelihood function, taking the camera model, the image descriptor, human body model structure and kinematic motion constraints into account. Estimation is responsible for searching out the most likely pose according to the likelihood function and image observations. The PS models used for the 2D human pose estimation can be extended here for 3D pose estimation and the example can be found in the work of (Zuffi & Black, 2015) and the work of (Belagiannis et al., 2014). Another popular trend is to combine the discriminative approach with the generative approach. For instance, (Hasler et al., 2010) fitted a parametric body model to the silhouettes but they required known segmentation of silhouettes and some manually provided correspondences. (Sigal et al., 2008) also proposed a method to generate 3D pose by using the generative model, SCAPE, to fit the image silhouettes. However, their method can perform well only if the silhouettes are prepared in advance. (Zhou et al., 2010) also

fitted a parametric model of body shape and pose to a cleanly segmented silhouette through significant manual intervention, but they did not evaluate quantitative accuracy.

(Bogo et al., 2016) described the first way to automatically estimate the 3D pose of the human body as well as the 3D body shape from a single unconstrained image and achieved the state-of-the-art results. Their work is enabled by the use of SMPL model. SMPL (Loper et al., 2015) model defines how joint locations are related to the 3D surface of the body, enabling inference of shape from joints. The 3D shape can help model interpenetration, avoiding impossible poses.

In summary, the 3D human pose generation techniques can be categorized into the discriminative approaches and the generative approaches. The discriminative approaches tend to train a mapping between the low-level image features such as silhouettes and edges and the 3D pose or use the convolutional network to directly learn the matching. The generative approaches basically have two stages which are modelling and estimation and their common feature is the use of a strong model of human body. The combination of these two approaches is the mainstream in recent years. This type of methods can make use of both the low-level image features and also the model information to reduce the ambiguities.

2.4 Research gaps and objective

Generating the anthropometry data for the ergonomic posture analysis based on the single color camera is supposed to help to increase the efficiency, cost-effectiveness and the range of use. However, only few literature researched in the area of single color camera-based ergonomic posture analysis. On the other hand, although the existing methods for the 2D and 3D human pose estimation have shown promising results, the current pose estimation methods still need to improve the accuracy of body joint detection which is directly related to the body angle calculation for ergonomic analysis. The high-enough pose accuracy should be considered as the most important factor for the effective and accurate ergonomic posture analysis. Therefore, any efforts made towards increasing the accuracy of the 2D and 3D pose estimation should be considered as the contribution.

In addition, it should be noted that there is a deficiency of positioning 2D body joints in the most of heatmap regression-based 2D pose estimation methods which are currently in widespread use. When determining the position of each 2D body joint by the heatmap detection-based methods, the pixel with the highest confidence value in the corresponding heatmap is often selected as a final joint point. In fact, this criterion may not always be correct. Take the DeeperCut method for example, one output sample image with the detected 14 body joints that denoted by assorted colors is on the left while the heatmaps for all joints are shown on the right in Figure.2.2. The left shoulder joint is completely ill-

detected, and its detected position is even near the right shoulder joint. It is due to the fact that the pixel with the highest confidence value in the left shoulder's heatmap locates in the right shoulder part.

In this respect, this research study aims to fill the research gaps and propose a novel framework to generate more accurate body joint angles from 2D videos recorded by the single color camera. The objective is achieved by the proposed framework which can mitigate the deficiency of the 2D heatmap detection-based methods for reliable 3D body pose. The proposed framework is expected to function in the real modular construction industry, and the improved anthropometry data such as body joint angles can be beneficial to the ergonomic posture analysis which can further help reduce the awkward and improper postures of workers in modular construction.

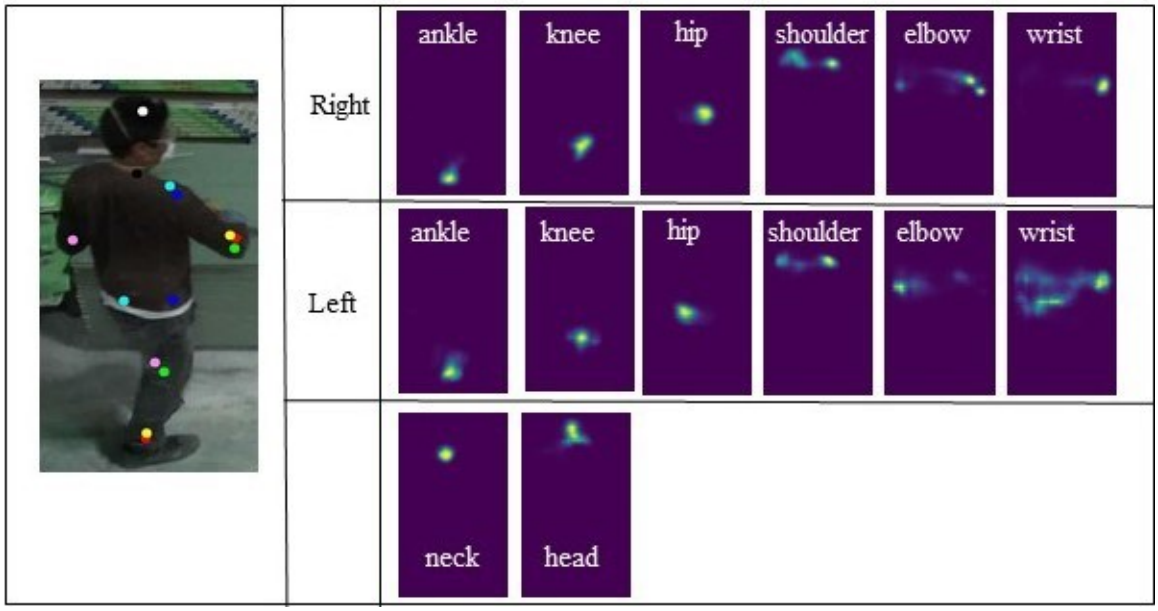

Figure 2.2 An example image for initial pose and the heatmaps for its 14 joints

CHAPTER 3: METHODOLOGY

This chapter describes the details of the proposed framework for generating the 3D anthropometry data for ergonomic posture analysis. As shown in Figure.3.1, the overall framework consists five procedures. Firstly, the worker is tracked in the video frames. The tracking result is represented as a set of rectangular windows containing the worker which can serve as the Region of Interests (ROIs) for the subsequent tasks. The second procedure is responsible for initial data acquisition task including the 2D initial joints detection and the 2D body parts detection. Next, the 2D initial joints detection is integrated with the body parts detection to obtain the refined 2D pose. Finally, the 3D pose is reconstructed according to the 2D refined pose and the 3D body angles are calculated based on the 3D joint coordinates which can serve as the input for the further ergonomic posture analysis.

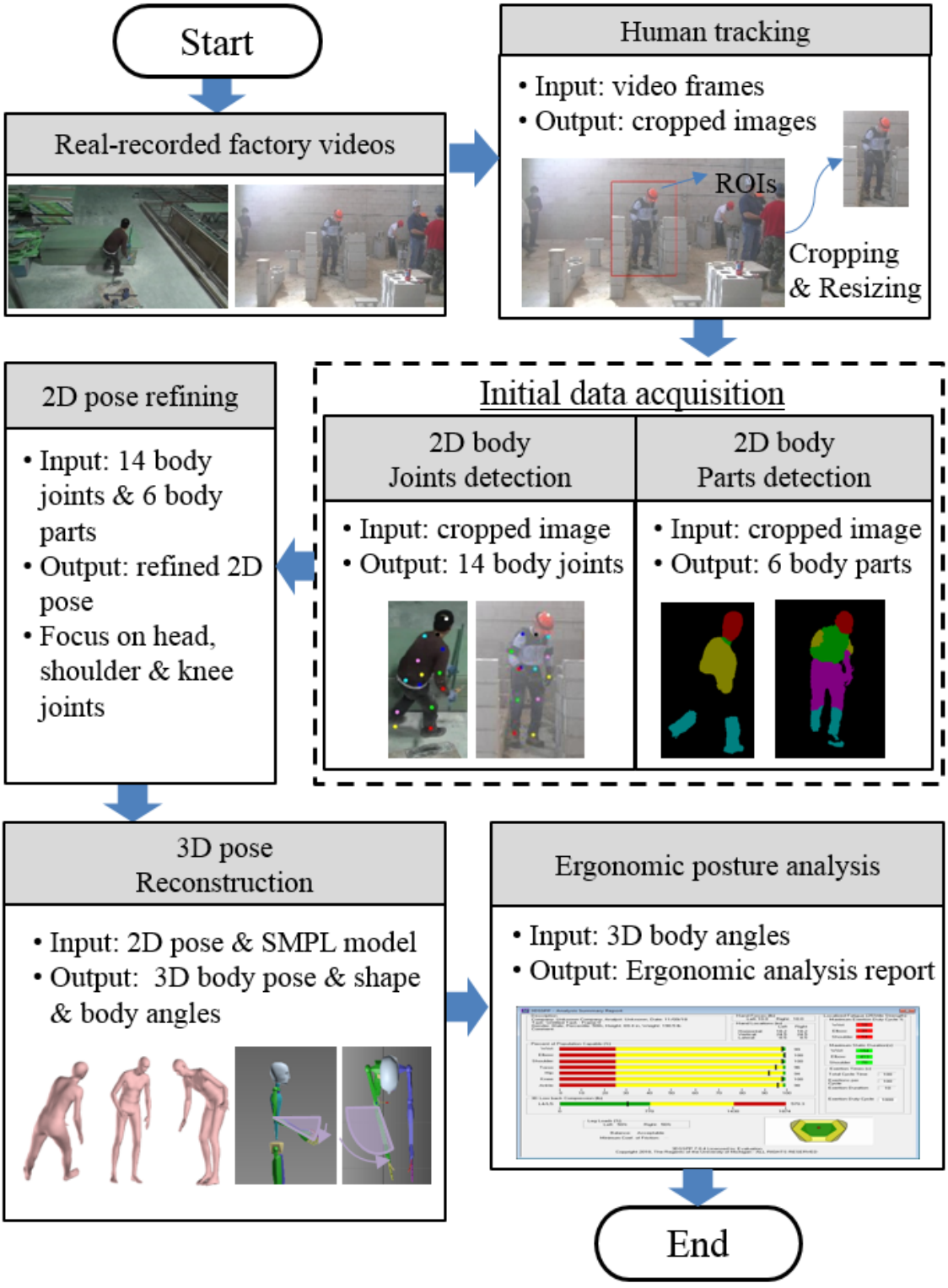

Figure 3.1 Flowchart of overall framework

3.1 Human tracking

Human tracking is to determine regions, typically the smallest rectangular bounding boxes that enclose the same human in the video sequence. A lot of the vision-based human motion capture systems adopt the human tracking as the first step to generate the ROIs when the input is video. This is due to: (1) most of the convolutional networks for human pose estimation have limits on the size of the input image. For instance, the size of the input image cannot be greater than a certain value; (2) original video frames always contain plenty irrelevant background information which can distract the pose detection result; and (3) processing larger images demands higher computer memory and computational power which can be time-consuming. The above issues can all be solved or mitigated by the human tracking since it can filter the irrelevant background information and provide the ROIs that focus mainly on the human body. The smaller size of image (the size is manually initiated with a rectangular window in the first video frame) can be obtained by cropping the ROIs from the video frames in this procedure.

The proposed framework adopts the human tracking as the first procedure as the others. The CNN-based tracking algorithm MDNet (Nam & Han, 2016) has been selected since it achieved the state-of-the-art performance on the human-tracking challenge competition (Kristan et al., 2015). Some sample video frames processed by MDNet are shown in Figure.3.2. It is worth noting that the cropped images according to the ROIs from

the tracking process need to be further resized for the subsequent pose estimation procedure. The details about how to choose the image size and the effect of image size on the pose estimation results will be described in the discussion session in Chapter 4.

Figure 3.2 Sample video frames of human tracking results

3.2 Initial data acquisition

This procedure is responsible for the initial data acquisition which comprises the 2D body joint detection and the 2D body parts detection. These data are integrated together to obtain the refined 2D pose in the next procedure.

3.2.1 2D body joint detection

The CNN-based DeeperCut algorithm has been selected to perform the body joint detection due to its stable performance and high-precision results (see Chapter 2 for details). As shown in Figure.3.3, 2D body pose in the DeeperCut is in the form of 14 body joints: head, neck and right/left ankle, knee, hip, shoulder, elbow, wrist. DeeperCut adapted the modified extremely deep Residual Network for human body joint detection. It can effectively regress the heatmap of each joint, which can indicate the reliable probability of each pixel in the image to be the joint. The heatmap is considered relatively reliable overall even though the final joint detection results may be biased. Hence, the heatmap of each joint will be used later to evaluate whether the results of human body part detection are reliable, and which detected body part can be used to refine the initial detected pose.

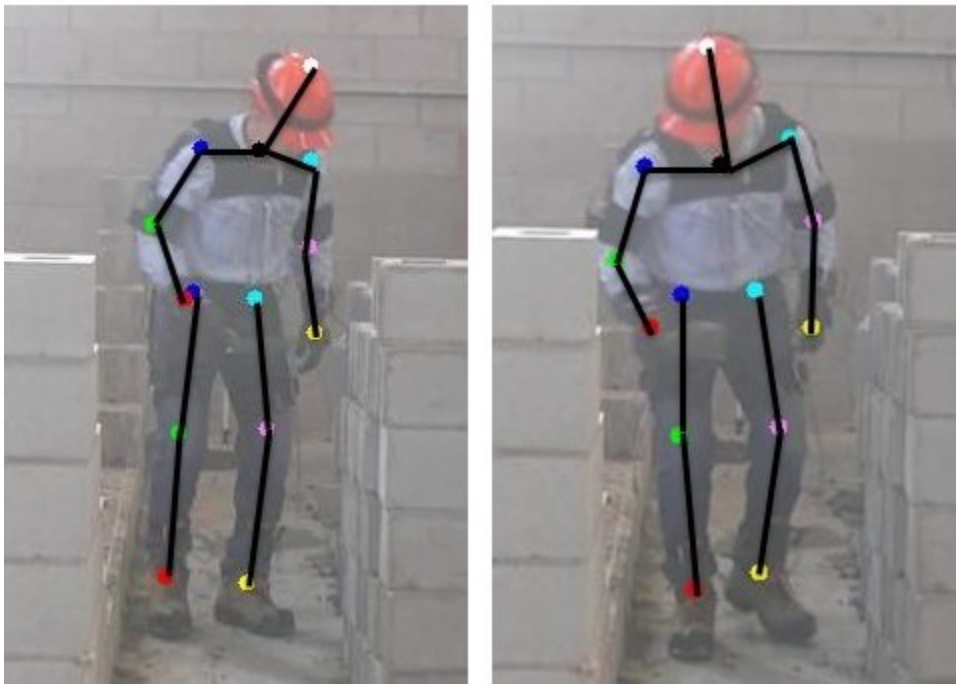

Figure 3.3 Schematic diagram of the joint locations in DeeperCut

3.2.2 2D body part detection

As for the 2D body part detection, this thesis adopts the Deeplab v2 method (Chen et al., 2018), which achieved 87.4% accuracy on person segmentation in PASCAL VOC Challenge (Everingham et al., 2015). One modification made here is to merge the 24 detailed human body part annotations in the Pascal-Person-Part dataset (Chen et al., 2014) into 6-part classes which are head, torso, upper /lower arm and upper/lower leg. The reliability of these detected body parts is then evaluated by the heatmaps produced in the initial body joint detection. In the heatmap, the value of each pixel ranges from zero to one. The higher the value of the pixel, the more likely the pixel is the joint point. The detected parts can be regarded reliable when its region contains the pixel with a high value (larger than 0.2 in this study) in the corresponding joint's heatmap. The value 0.2 is determined according to the experimental test. It has been proved that the selected value can achieve reliable results and the selection is scientific since the value 0.2 is the maximum probability in the heatmap of most occluded joints and the near-minimum value in the bright area of the well-detected joints' heatmap. Take the evaluation of head part as an example, if the detected head part region contains a pixel with the value higher than 0.2 in the corresponding head heatmap, the head part will be considered reliable. Figure 3.4 shows a sample image of human body part detection result with the heatmap of the head joint to illustrate the part evaluation process. The red box denotes the location of the head part, and

the label next to the heatmap illustrates the correspondence between the brightness and confidence value of the pixel. It is clear to be seen that the head part region contains the pixel with the confidence value higher than 0.2. Thus, the head part should be considered reliable. The evaluation of other detected parts can be implemented in the same way.

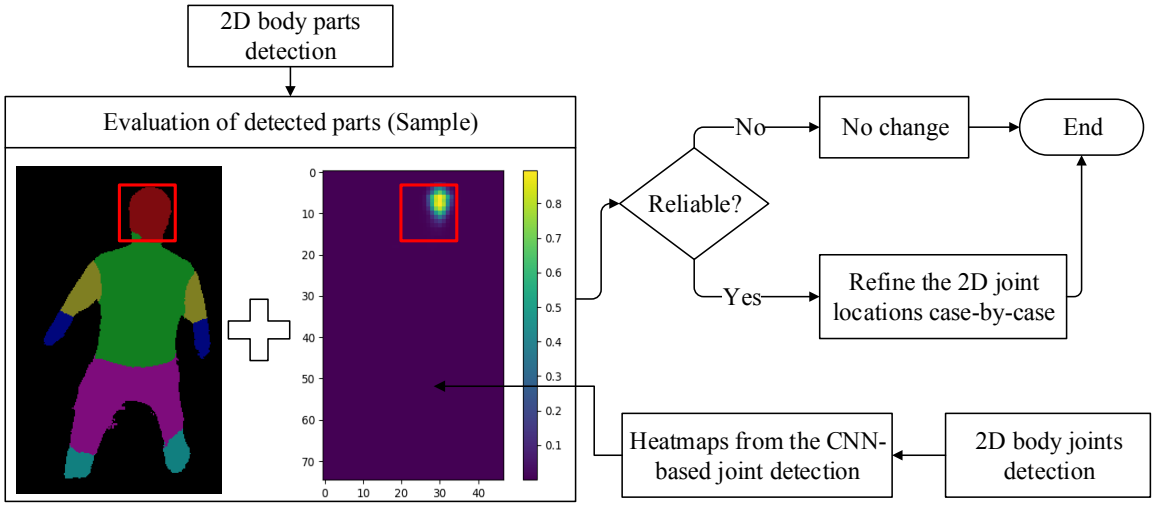

Figure 3.4 Sample image of body part detection results and the heatmap of head joint

3.3 2D refined pose generation

The third procedure is to refine the initial 2D pose based on the corresponding reliable part detection results. The deficiency of the heatmap detection-based methods, which is that the pixel with the highest confidence value in the heatmap will always be selected as the joint point, have been discussed in Chapter 2. In order to address the deficiency, the refinement which is built on a case-by-case basis is performed to reduce the detection errors in terms of joint locations.

3.3.1 Head joint refinement

Assume the detected head part in previous procedure is reliable, whether the initial detected head joint locates in the head part region needs to be first examined. If the initial detected head joint is not in the reliable head part region, then the pixel with the highest confidence value in the head part region should be selected as the head joint. If the detected head joint locates in the reliable head part region, the height and the width of the head part region should be compared in the next step. It is found that the main error of the head detection is that the detected location may lower than the real location of the head joint through a lot of observation. By examining the aspect ratio of the head part region, the head orientation can be roughly determined. In other words, the refinement can be processed only in the condition that the direction of the head is roughly vertical which means the height of the head part should be larger than the width of the head part region. Hence, if the height of the head part is larger than the width of the head part, the pixel with the highest confidence value in the top ¼ of the head part region should be selected as the head joint. The corresponding refinement process of head joint is illustrated in Figure.3.5.

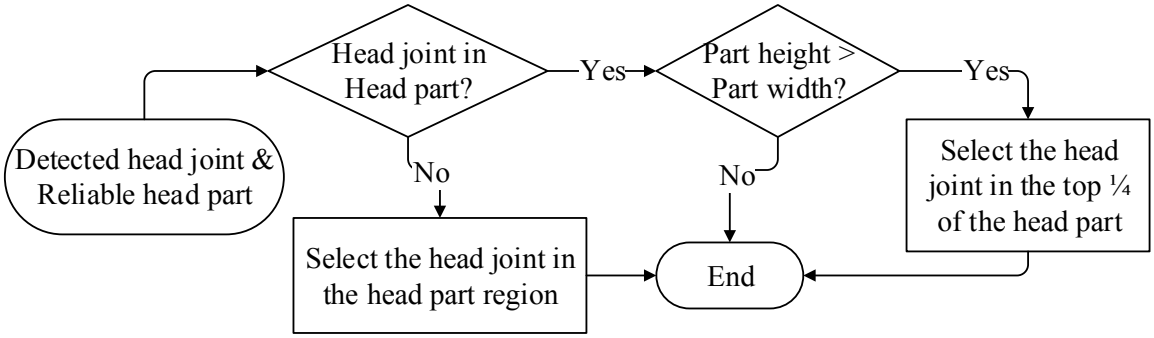

Figure 3.5 Refinement flow chart of head joint

3.3.2 Shoulder joint refinement

As for the shoulder joints, the number of the reliable upper arm part should be first checked. If there is only one reliable upper arm part, then the part should be checked to determine whether it belongs to the left arm or the right arm. For instance, if the part only contains the pixel with the confidence value larger than 0.2 in the left shoulder's heatmap, the part is identified as the left upper arm. Then the pixel with the highest confidence value in the part region will be selected as the left shoulder joint. The right shoulder joint can be refined in the same way when the part is identified as the right upper arm. If the detected part region contains the pixel with the confidence value larger than 0.2 in both the left and right shoulder's heatmap, the shoulder joints would not be refined since it is difficult to determine the way to do the refinement.

If there are two reliable upper arm parts and one detected shoulder joint is in one part while another detected shoulder joint is not in any part, the pixel with the highest

confidence value in the no-joint part region should be selected as the second shoulder joint. If there are two reliable upper arm parts and two detected shoulder joints locate in the same part, the shoulder joint with lower confidence value in its heatmap should be moved to the no-joint region. The pixel with the highest confidence value in the no-joint part region should be selected as the new shoulder joint. The above two conditions are the common error in the initial joint detection, the other cases are not discussed not only because of the complicated situation, but also the low occurrence probability. The corresponding refinement process of shoulder joint is illustrated in Figure.3.6.

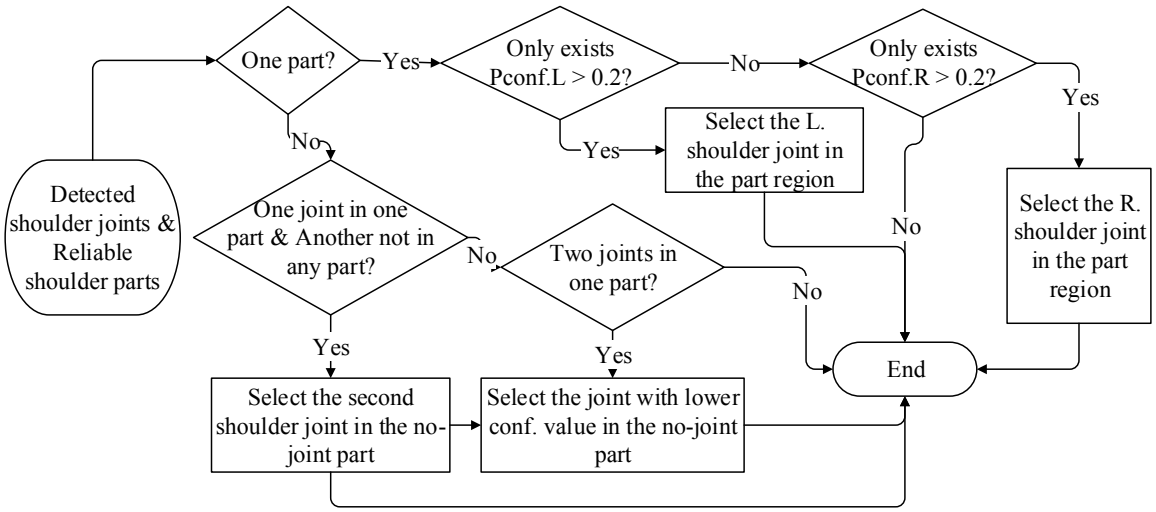

Figure 3.6 Refinement flow chart of shoulder joints

3.3.3 Knee joint refinement

When it comes to the refinement of knee joints, the number of the reliable upper-leg part and lower-leg part need to be both checked at first. This is owing to that the position

of the knee joint is related to both the upper-leg and lower-leg parts. The refinement of knee joints is applicable in this thesis when there is only one reliable upper-leg part, or one reliable lower-leg part or one reliable upper-leg part with one reliable lower-leg part. In other words, the situation that the left and right knee joints are on the same horizontal line can be suitable for refinement. In this situation, the worker's legs get close to each other and initial joint detection method is error prone in this case. As shown in Figure.3.6, for both left and right knee joint, if there is only one reliable upper-leg part, the pixel with the highest confidence value in the lower 1/3 of the upper-leg region in their respective heatmap will be selected as the knee joints. This is to ensure that the position of the knee joints would not be too high compared to the real position. The refinement process is the same when there is only one detected reliable lower-leg part. The only difference is that the searching region changes to the top 1/4 of the lower-leg region. This is due to the height of the lower-leg part is normally higher than the height of the upper-leg part. Adopting the 1/4 as the dividing standard in the lower-leg part can better restrict the height of the knee joints.

When one reliable upper-leg part and one reliable lower-leg part have been detected at the same time, the height of the upper-leg part and the height of the lower-leg part should be compared at first. If the height of the upper-leg part is twice higher than the height of the lower-leg part, the refinement is exactly the same with the refinement process when

only one reliable upper-leg part has been detected. If the height of the lower-leg part is twice higher than the height of the upper-leg part, the refinement is the same as the refinement when there is only one reliable lower-leg part. The above two situations are designed to some special cases when the detected difference between the height of the upper leg part and lower leg part is large. Otherwise, the searching region will be determined by the boundary line between the upper-leg part and the lower-leg part. The top and the bottom of the restricted searching region is five pixels above the boundary line and five pixels below the boundary line respectively. The knee joints are selected according to the pixel with the highest confidence value in the border area in the corresponding heatmaps. The corresponding refinement process of knee joint is illustrated in Figure.3.7.

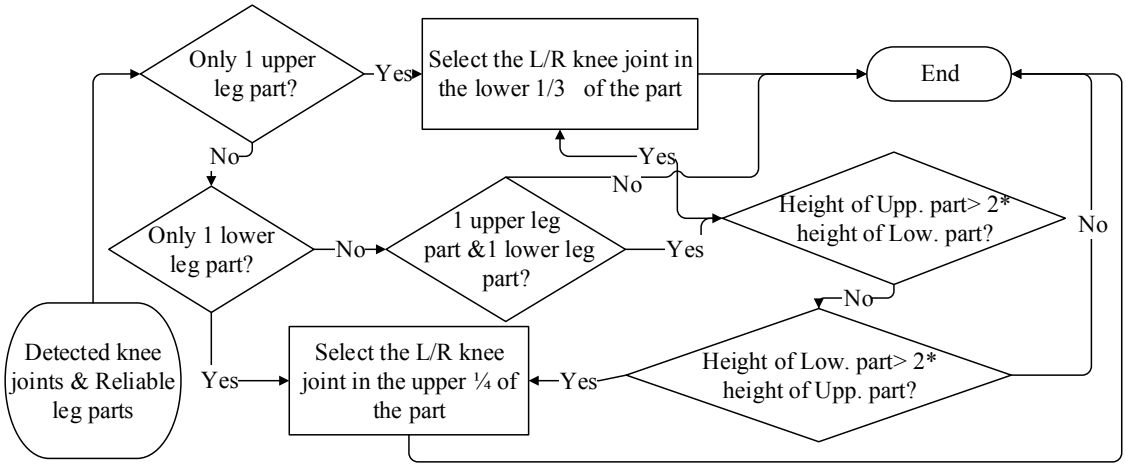

Figure 3.7 Refinement flow chart of knee joints

3.4 3D pose reconstruction and body angle calculation

The last procedure is responsible for inferring the corresponding 3D human pose based on the 2D refined body pose and then calculating the body angles for ergonomic posture analysis. This thesis adopts the 3D generative Skinned Multi-Person Linear (SMPL) model which created by (Loper et al., 2015) to perform the reconstruction process. The 3D SMPL model is a skinned vertex-based model that has been trained from thousands of 3D scans and hence can accurately represent a wide variety of body joints and shapes in natural human poses. Following the workflow of (Bogo et al., 2016), the 3D human pose is reconstructed by reducing the error between the projected 3D SMPL joints and the 2D refined joints generated in the previous procedure. In this way, the 3D poses and shape that optimally match the 2D joints are obtained.

Then the human body angles are calculated accordance with the instruction in (3DSSPP, 2017). In the 3DSSPP, each body segment can be described by two angles: a horizontal angle and a vertical angle. The horizontal angles are measured between the body segment and the X-axis while looking down onto the figure from overhead, while the vertical angels are simply measured between the body segments with the X-Y plane. There are seven body segments which are clavicle, upper arm, lower arm, hand, upper leg, lower leg and foot, thus in total 14 body angles should be obtained. Figure.3.8 shows the difference between the horizontal angles and vertical angles and the definition of some

angles can be seen as well.

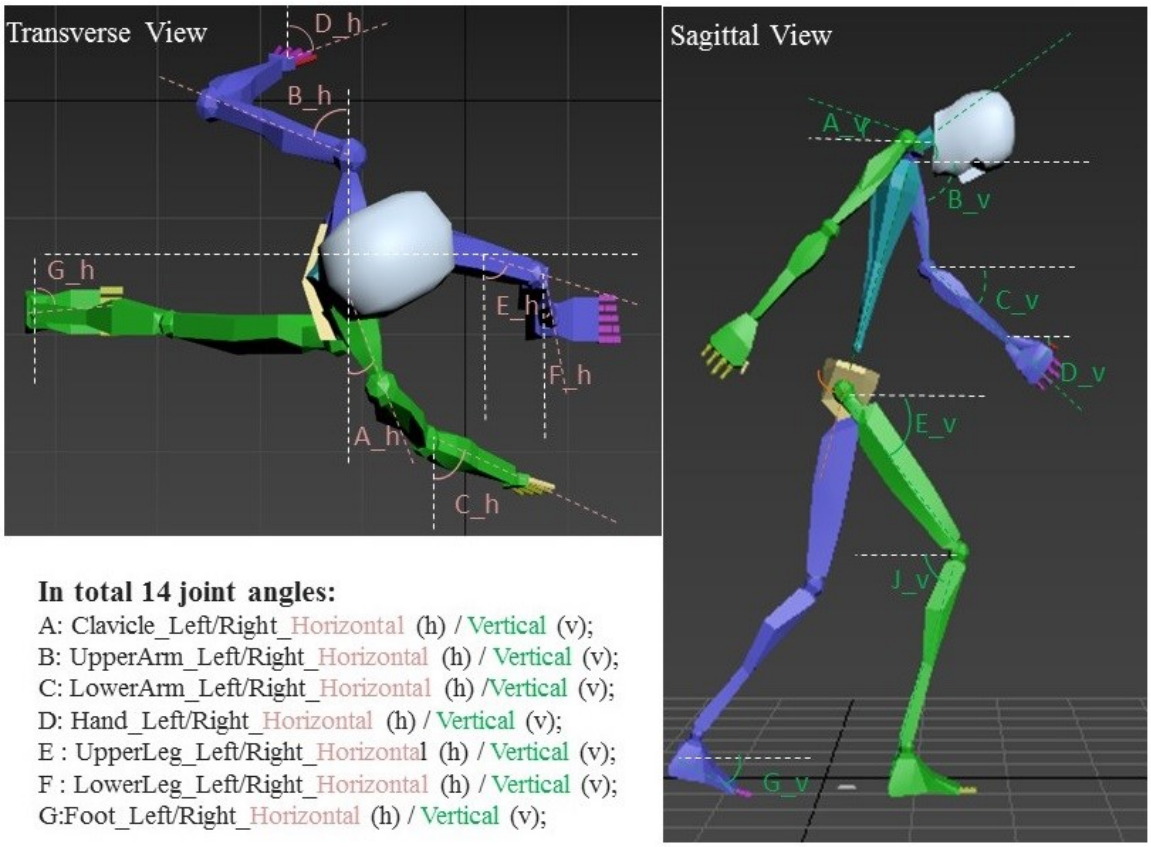

Figure 3.8 Definition of various body angles (Modified in (Li et al., 2017))

3.5 Ergonomic posture analysis

The calculated 3D body angles can be inputted into 3DSSPP software developed by (Chaffin et al., 2006) for further ergonomic analysis. The 3DSSPP software can assess the risk factors that may produce excessive physical loads on the worker's body through a biomechanical analysis. The analysis requires three types of input which are 3D joint angles, external loads and anthropometry such as the gender, age, height and weight. The

joint angles can be obtained by the proposed framework and the others can be simply obtained by observing the task. And then the software can provide more than 15 detailed analysis reports in different aspects, such as the localized fatigue report, 3D low back analysis report, strength capabilities report, and joint forces report. Figure.3.9 shows the analysis summary report that contains five areas of information: Hand Forces, Low Back Compression, Percent Capable, Balance, Coefficient of Friction and Localized Fatigue. These information can help the manager detect the unsafe motions easily and provide the direction for the redesign of the workplace.

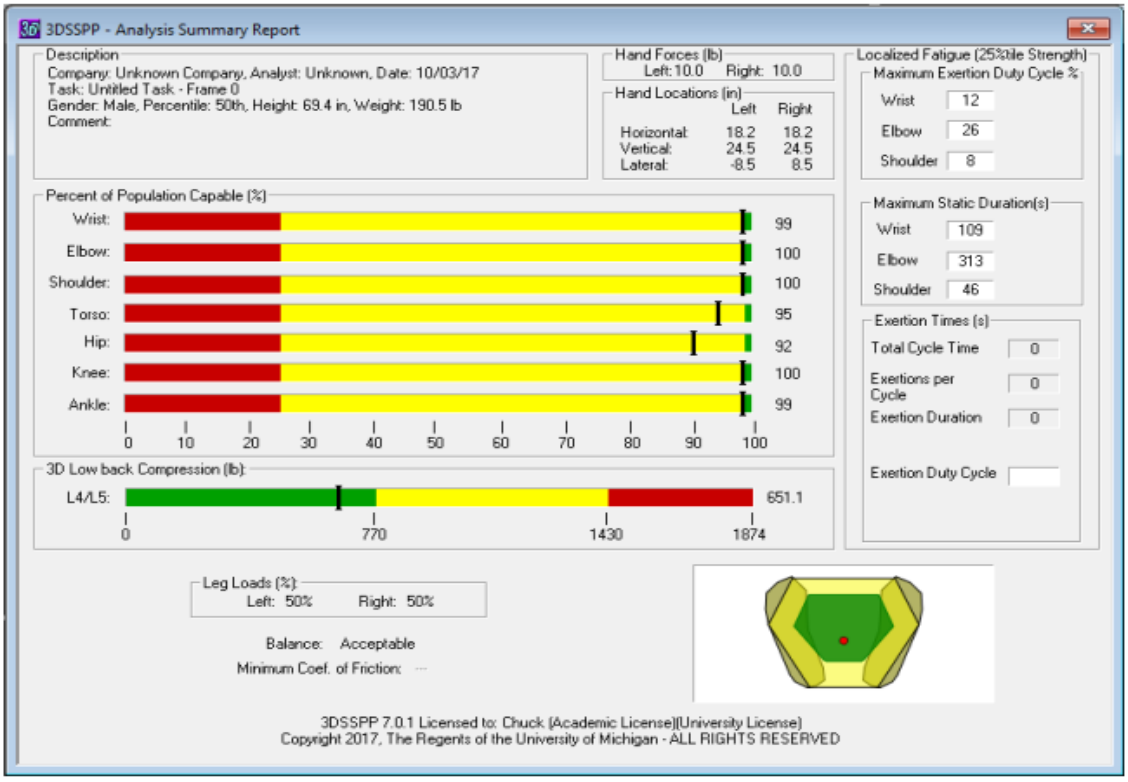

Figure 3.9 Analysis summary report of 3DSSPP

CHAPTER 4: RESULTS AND DISCUSSION

In order to validate the effectiveness of the proposed framework, the proposed framework has been implemented in the Python 2.7 environment, Ubuntu 16.04 LTS system. The hardware configuration for the implementation was listed as follows: An Intel® CoreTM I7-4820k CPU @ 3.70 GHz * 8, a 31.4 gigabytes memory, and an NVIDIA Titan X GPU.

4.1 Implementation and results

To measure the performance of the proposed framework for reconstructing the 3D human pose to further facilitate the ergonomic posture analysis, the results of implementation on one test video have been evaluated. The test video with the ground truth motion data was collected by the (Alwasel et al., 2017). They used the state-of-the-art IMU-based motion capture suit (Xsens, 2016) which consists 17 sensors attached at each body segment including the head, shoulder, elbow, hip, knee, foot, and so on. Each sensor is comprised of a triaxial accelerometer, a triaxial gyroscope, and a magnetometer. The suit has 125 Hz sampling rate while the test video was converted to frames at a rate of 25 Hz. Hence, the new ground truth motion data was formed by extracting one frame every five in the original ground truth file and then the frame-to-frame comparison can be performed.

The test video mainly recorded when the worker was constructing concrete masonry walls, as shown in Figure.4.1. The blocks used are standard concrete masonry units weighing 10 kg with dimensions of 390 × 190 × 100 mm. Since the whole test video mainly recorded 40 times of repetitive block laying activity, the first block laying period that lasted for 20 seconds, has been selected as a sample to test the effectiveness of the framework. Figure.4.2 shows the experimental results on the 2nd, 83rd, 216th and 301st frames. The reconstructed 3D models are placed next to the worker in the frames for better visual comparison.

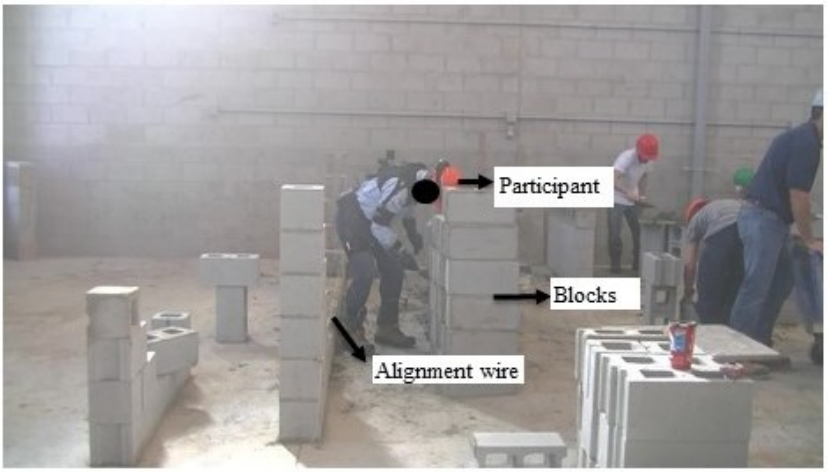

Figure 4.1 Example of test video frame

Figure 4.2 Video frames with the reconstructed 3D pose

To evaluate the accuracy of the reconstructed 3D pose, the 3D joint coordinates need to be extracted from both the predicted 3D pose and the ground-truth motion data. Then the Random Sample Consensus (RANSEC) algorithm (Fischler & Bolles, 1981) has been used to align this two 3D coordinate systems. RANSEC is a resampling technique that generates candidate solutions by using the minimum data points required to estimate the underlying model parameters. It uses the smallest set possible and proceeds to enlarge the set with consistent data points while the conventional sampling techniques use as much of the data as possible to obtain an initial solution and then proceed to prune outliers. According to the requirements of this work, the minimum number of required points are set to 3 and the iteration will be ended when the error stops deceasing for a while.

After that the 3D body angles including the horizontal angles and the vertical angles

in both sets are calculated separately according to the angle definition in 3DSSPP. Figure.4.3 shows some results of body angle comparison for all the 500 frames in the 20-second video (Frame rate 25 f/s). The light blue line represents the body angles calculated using the ground truth while the dark gold line represents the body angles generated form the proposed framework. To make it clearer, Table 1 lists both the horizontal and vertical angle differences on average for each body segment. The results showed that the overall average angle error is 17.51. The lowest angle error which is the angle difference of lower arm part is only around 4.5 degrees. In addition, the ergonomic analysis results on the 26th frame can be found in Figure 4.4. The 3D generated pose by the proposed framework and the pose in 3DSSPP can both be seen in the figure. According to the report, the pressure resulted from the current posture is in an acceptable range.

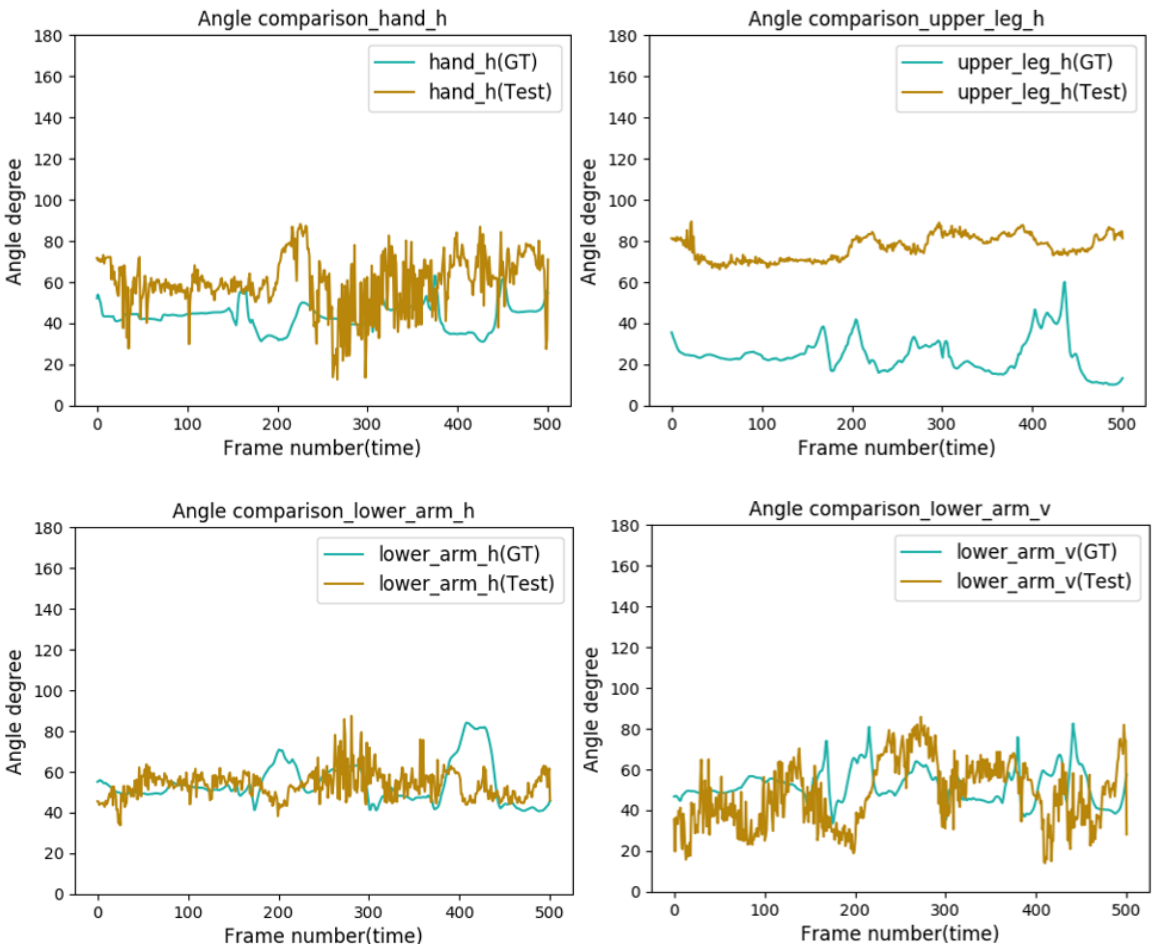

Figure 4.3 Example for angle comparison

Table 1 Summary of angle error for each body part. (Error in degrees)

Angle error	Pre_shoulder	Upp_arm	Low_arm	Hand	Upp_leg	Low_leg	Foot
Horizontal	11.6624	15.8567	**4.5499**	14.2342	45.2209	14.197	19.4261
Vertical	13.1649	10.018	10.8425	17.0877	20.2033	20.7261	28.0379

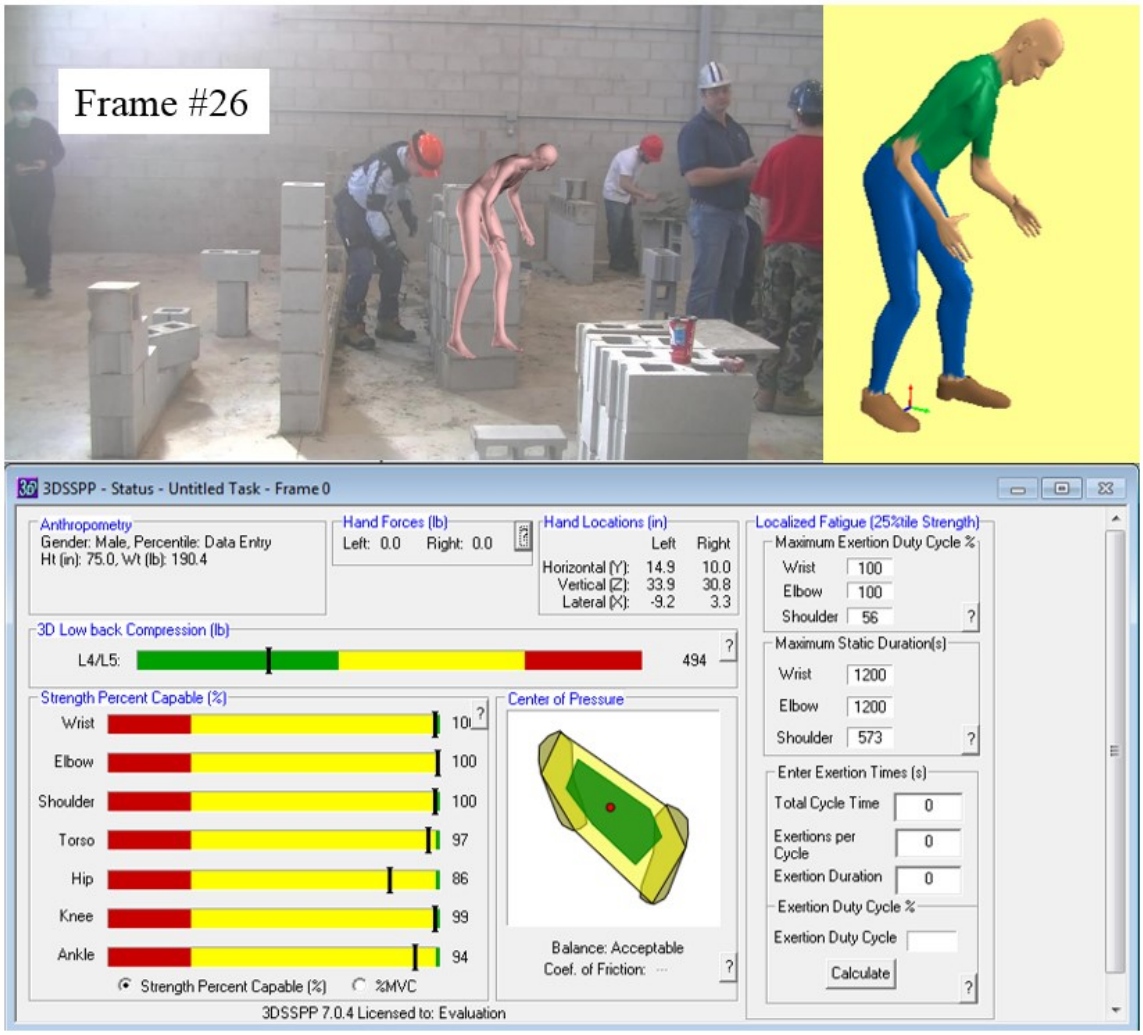

Figure 4. 4 Example of ergonomic posture analysis

4.2 Discussion

The discussion section consists of two parts: analysis of the cause of angle errors and the research findings part. The first part discusses the possible reasons of the angle errors and the second part proves that the 3D reconstructed human pose is highly dependent on the 2D generated pose and the size of the input image for the CNN can significantly affect the detection results.

4.2.1 Analysis of the cause of errors

It should be noted that there are some reasons that may affect the final angle comparison results: (1) the number of the body joints used in the ground truth motion data is 28 while the number of joints used in the SMPL model is 24. The joints closest to the definition of SMPL joints were selected from the ground truth to do the comparison and this may cause some additional errors; (2) the occlusion that occurs from the 234th frame to 388th frame and in the last 15 frames can affect the detection accuracy. It can be seen form Figure.4.3 that predicted body angles fluctuate widely in the range of the occlusion.

In addition, the test results denote that the accuracy of upper leg part reconstruction is much worse than the other parts. It may be explained by the fact that the detected 2D hip joints were not used when matching the projected 3D joints with the 2D detected joints in the 3D SMPL model fitting process. Moreover, it can be seen that in the lower right figure in Figure.4.3, the curve of the predicted upper leg angle basically has the same trend as the ground truth curve. This indicates that the movement of the predicted pose tends to be correct with the offset from the ground truth and the error is not disordered.

4.2.3 Research findings

The effectiveness of the 2D pose refinement was evaluated in order to prove that the 3D reconstructed human pose is highly dependent on the 2D generated pose. To implement

the assessment, the 3D pose similarity was calculated on the 3D pose generated with the refinement process and the pose generated without the refinement process respectively. The 3D pose similarity is calculated based on the 2D pose similarity algorithm used in (Bilgeckers, 2017). Specifically, the 3D predicted pose is aligned to the ground-truth pose via the Procrustes superimposition on every frame. Then the 3D pose is projected onto the X-Y plane and the X-Z plane separately to obtain the horizontal pose and vertical pose so that the 3D pose similarity can be transformed to the 2D pose similarity problem. Figure.4.5 shows the pose similarity results from the proposed framework with the refinement process and without the refinement process. It can be seen that the poses with the refinement process is much superior to the poses without the refinement process. As a result, the 2D pose refinement process introduced in the framework is critical.

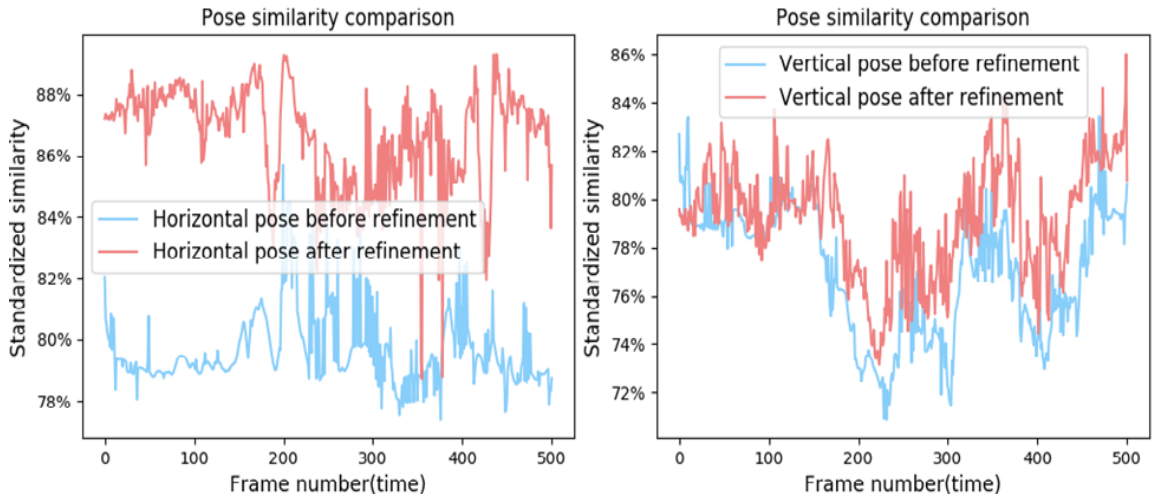

Figure 4.5 The results of pose similarity

Another finding is that the size of the input image for the CNN can significantly affect the detection results. The previous work found that scaling images to a standing height of around 340 pixels performs best (Nsafutdinov et al., 2016). Two image sizes, around 322 * 674 pixels and around 150 * 350 pixels have been selected to validate the impact of input image size on 3D pose accuracy. The large input image size leads to the very bad detection results of both 2D and 3D pose. Figure.4.6 selected the 17th frame as an example to show the influence of input size on 2D body pose detection. The left two figures in Figure.4.6 show the generated 2D and 3D pose when the input image size is 322 * 674 pixels and the input image size for the right two figures is 150 * 350 pixels. Moreover, the comparison of 3D pose accuracy in terms of joint angle errors can be seen in Table 2. The angle errors of each joints when the input image size is 322 * 674 pixels were much larger than the angles errors when the input image size is 150 * 350 pixels. The difference of the average joint angle errors can be around 6 degrees. The significantly impact of input image size on 3D reconstruction results can be illustrated.

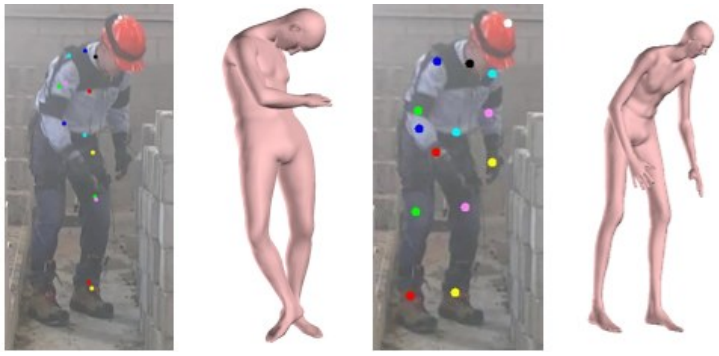

Figure 4.6 The different reconstructed 2D and 3D poses based on the same frame with different input image size

Table 2 The angle errors based on the same frame with different input image size.

(Error in degrees)

	Image size	Pre_shoulder	Upp_arm	Low_arm	Hand	Upp_leg	Low_leg	Foot
Hor_	322*674	22.7162	24.3586	11.9026	22.9731	52.0793	17.7061	27.0076
angle	150*350	11.6624	15.8567	4.5499	14.2342	45.2209	14.197	19.4261
Ver	322*674	20.2734	15.9707	16.5903	23.1951	28.5399	24.5428	23.3036
angle	150*350	13.1649	10.018	10.8425	17.0877	20.2033	20.7261	28.0379

CHAPTER 5: CONCLUSIONS

In this chapter, the conclusions of this research are presented. After that, the recommendation and future works are discussed in the end. With the rapid development of the modular construction industry, the workers' safety has also received widespread social attention. According to the reports from Occupational Health and Safety Institute, the awkward and unsafe postures of worker can be considered as one of the main causes of the workers' injuries and accidents. In order to detect these awkward and unsafe postures, this research proposed a novel framework to obtain the 3D motion data required for the ergonomic posture analysis from the single-color camera recorded videos. The framework contains five main procedures which are the human tracking procedure, the initial data acquisition procedure, the 2D refined pose generation procedure, the 3D pose reconstruction with body angle calculation procedure and the ergonomic posture analysis procedure. In the human tracking procedure, the worker has been tracked in the video and the tracking results are extracted to detect the initial 2D pose and the human body parts in the second procedure. Then, the 2D initial pose is refined based on the results of the body part detection. At last, a parameterized 3D body model SMPL is fitted to the refined 2D pose to obtain the 3D body pose and shape and the 3D body angles can be calculated based on the 3D joint coordinates. At last, the 3D body angles can serve as the input for the ergonomic posture analysis which can further help reduce the awkward and improper

postures and motions in the workplace.

The framework has been tested on the real-recorded factory video in which the worker was responsible for constructing the concrete masonry walls. The test results showed that the accuracy of the 3D human pose reconstruction and the generated body angles could reach a high level. The lowest joint angle difference between the body angles generated from the proposed framework and the angles calculated based on the ground truth was around 4 degrees. Moreover, the test results of 3D pose similarity indicated that the 2D pose refinement procedure is essential, and it can greatly increase the accuracy of the 3D pose reconstruction.

Since the field of ergonomic posture analysis using the single camera remains to be researched and developed, there are many directions for future work. For instance, the body shape and pose can benefit from other features such as the silhouettes. The multiple camera views may mitigate the problem of occlusion to some extent. Additional facial detector or the hand detector may improve the detection results of head joint and wrist joint. In addition, a personalized pose estimator can make it possible to establish motion files for different workers and provide special recommendations for the postures of the workers according to each worker's file.

REFERENCES

3DSSPP. (2017). *3D Static Strength Prediction ProgramTM Version 7.0.0.* Retrieved October 2018, from https://c4e.engin.umich.edu/assets/3DSSPP-Version-7.0.0-Manual.pdf

Alwasel , A., Sabet, A., Nahangi, M., Haas, C. T., & Abdel-Rahman, E. (2017). Identifying poses of safe and productive masons using machine learning. *Automation in Construction, 84*, 345-355.

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. *Computer Vision and Pattern Recognition*, (pp. 1014-1021).

Atrevi , D. F., Vivet, D., Duculty , F., & Emile, B. (2017). A very simple framework for 3D human poses estimation using a single 2D image: comparison of geometric moments descriptors. *Pattern Recognition, 71*, 389-401.

Belagiannis , V., Amin, S., Andriluka, M., Schiele, B., Navab, N., & Ilic, S. (2014). 3D pictorial structures for multiple human pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , (pp. 1669-1676).

Bilgeckers . (2017). Retrieved October 2018, from Single Pose Comparison—a fun application using Human Pose Estimation: https://becominghuman.ai/single-pose-comparison-a-fun-application-using-human-pose-estimation-part-2-4fd16a8bf0d3

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Remero, J., & Black, M. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape form a sinlge image. *European Conference on Computer Vision*, (pp. 561-578). Springer, Cham.

Bureau of Labour Statistics. (2016). Retrieved April 2018, from Nonfatal occupational injuries and illnesses requiring dyas away from work: http://www.bls.gov/news.release/pdf/osh2.pdf

Canadian Standards Association. (2003). Coding of work injury or disease information. *Canadian Standards Association*. Retrieved April 2018, from http://www.alberta.ca/ohs-statics.aspx

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. *Computer Vision and Pattern Recognition*, (p. arXiv:1611.08050).

Carreira , J., Agrawal, P., Fragkiadaki, K., & Malik, J. (2016). Human pose estimation with iterative error feedback. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4733-4742).

Chaffin, D. B., Andersson, G., & Martin, B. J. (2006). Occupational biomechanics. New York: Wiley.

Chen, L. C., Papandreous , G., Kokkinos, L., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution,

and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40*(4), 834-848.

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., & Yuille, A. (2014). Detect what you can: Detecting and representing objects using holistic models and body parts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1971-1978).

C-Motion. (2013). *Visual 3D:3D analysis toolkit for 3D biomechanics modeling, analysis, and reporting.* Retrieved October 2018, from www.c-motion.com/products/visual3d/

Coenen, P., Kingma, I., Boot, C. R., Faber, G. S., Xu, X., Bongers, P. M., & Dieen, J. v. (2011). Estimation of low back moments from video analysis: A validation study. *Journal of biomechanics, 44*(13), 2369-2375.

Dane, D., Feuerstein , M., Huang, D. G., Lennart, D., Danielle, A., & Andrew, L. (2002). Measurement Properties of a Self-Report Index of Ergonomic Exposures for Use in an Office Work Environment. *Journal of Occupational and Environmental Medicine, 44*(1), 73-81.

David, G. C. (2005). Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. *Occupational medicine, 55*(3), 190-199.

Delp, S. L., Anderson , F. C., Arnold, A. S., Loan, P., Habib, A., John, C. T., . . . Thelen, D. G. (2007). OpenSim: open-source software to create and analyze dynamic simulations of movement. *IEEE transactions on biomedical engineering, 54*(11), 1940-1950.

Deutscher , J., & Reid , I. (2005). Articulated body motion capture by stochastic search. *International Journal of Computer Vision, 61*(2), 185-205.

Diego-Mas, J. A., & Alcaide-Marzal, J. (2013). Using Kinect™ sensor in observational methods for assessing postures at work. *Applied Ergonomics, 45*(4), 976-985.

Elgammal, A., & Lee, C. S. (2004). Inferring 3D body pose from silhouettes using activity manifold learning. *In Computer Vision and Pattern Recognition*, *2*, pp. II-II.

Everingham, M., Eslami, S. A., Van, G. L., John Winn, W., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision, 111*(1), 98-136.

Felzenszwalb , P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence, 32*(9), 1627-1645.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International journal of computer vision, 61*(1), 55-79.

Fischler, M., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*(6), 381-395.

Han, S. U., & Lee, S. H. (2013). A vision-based motion capture and recognition framework for behavior-based safety management. *Automation in Construction, 35*, 131-141.

Hasler , N., Ackermann, H., Rosenhahn, B., Thormahlen, T., & Seidel, H. P. (2010). Multilinear pose and body shape estimation of dressed subjects from image sets. *Computer Vision and Pattern Recognition (CVPR)* (pp. 1823-1830). IEEE.

Hen, Y. W., & Paramesran, R. (2009). Single camera 3d human pose estimation: A review of current techniques. *Technical Postgraduates (TECHPOS)* (pp. 1-8). 2009 International Conference.

Hignett, S., & McAtamney, L. (2004). Rapid entire body assessment. *Handbook of Human Factors and Ergonomics Methods*, 97-108.

Hofmann, M., & Gavrila, D. M. (2009). Multi-view 3d human pose estimation combining single-frame recovery, temporal integration and model adaptation. *Computer Vision and Pattern Recognition*, (pp. 2214-2221).

Huang, J. B., & Yang, M. H. (2009). Estimating human pose from occluded images. *Asian Conference on Computer Vision*, (pp. 48-60). Springer,Berlin, Heidelberg.

Inyang, N., Al-Hussein, M., El-Rich, M., & Al-Jibouri, S. (2012). Ergonomic analysis and the need for its integration for planning and assessing construction tasks. *Journal of Construction Engineering and Management, 138*(12), 1370-1376.

Kale, G. V., & Patil, V. H. (2016). A Study of Vision based Human Motion Recognition and Analysis. *International Journal of Ambient Computing and Intelligence (IJACI) , 7*(2), 18.

Karhu, O., Kansi, P., & Kuorinka, I. (1977). Correcting working postures in industry: a practical method for analysis. *Applied ergonomics, 8*(4), 199-201.

Kristan , M., Matas, J., Leonardis , A., Felsberg, M., Cehovin, L., Fernandez, G., . . . Montero, A. S. (2015). The visual object tracking vot2015 challenge results. *Proceedings of the IEEE international conference on computer vision workshops*, (pp. 1-23).

Lan, X., & Huttenlocher, D. P. (2005). Beyond trees: Common-factor models for 2d human pose recovery. *Computer Vision. 1*, pp. 470-477. Tenth IEEE International Conference on Computer Vision.

Li, X., Han, S. H., Gul, M., Al-Hussein, M., & El-Rich, M. (2017). 3D Visualization-Based Ergonomic Risk Assessment and Work Modification Framework and Its Validation for a Lifting Task. *Journal of Construction Engineering and Management, 144*(1), 04017093.

Loper , M., Mahmood, N., Romero , J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG), 34*(6), 248.

Marras , W. S., & Granata, K. P. (1997). Spine loading during trunk lateral bending motions. *Journal of biomechanics, 30*(7), 697-703.

McAtamney, L., & Corlett, E. N. (1993). RULA: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics, 24*(2), 91-99.

Mehta , D., Sridhar , S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., . . . Theobalt, C. (2017). Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG), 36*(4), 44.

Modular Building Institute. (2016). *Permanent Modular Construction: Annual Report.* Retrieved April 2018, from http://www.modular.org/documents/document_publication/mbi_sage_pmc_2017_reduced.pdf

Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 4293-4302).

Newell , A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. *European Conference on Computer Vision*, (pp. 483-499). Springer, Cham.

Ning, X., & Mirka, G. A. (2010). The effect of sinusoidal rolling ground motion on lifting biomechanics. *Applied Ergonomics, 42*(1), 131-137.

Ning, X., Zhou, J., Dai, B., & Jaridi, M. (2014). The assessment of material handling strategies in dealing with sudden loading: The effects of load handling position on trunk biomechanics. *Applied ergonomics, 45*(6), 1399-1405.

Nsafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *European Conference on Computer Vision*, (pp. 34-50). Springer, Cham.

OHS. (2016). Retrieved April 2018, from Occupational Injuries and Diseases in Alberta reports: http://www/alberta.ca/ohs-statics.aspx

Pfister, T., Charles, J., & Zisserman, A. (2015). Flowing convnets for human pose estimation in videos. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 1913-1921).

Pishchulin , L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., & Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 4929-4937).

Pishchulin , L., Jain, A., Andriluka, M., Thormahlen , T., & Schiele , B. (2012). Articulated people detection and pose estimation: Reshaping the future. *Computer Vision and Pattern Recognition* (pp. 3178-3185). 2012 IEEE Conference.

Pishchulin, L., Andriluka, M., Gehler, P., & Schiele, B. (2013). Strong appearance and expressive spatial models for human pose estimation. *Proceedings of the IEEE international conference on Computer Vision*, (pp. 3487-3494).

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer vision and image understanding*, *108*, pp. 4-18.

Ray, S. J., & Teizer, J. (2012). Real-time construction worker posture analysis for ergonomics training. *Advanced Engineering Informatics, 26*(2), 439-455.

Rhodin, H., Robertini, N., Richardt, C., Seidel, H.-P., & Theobalt, C. (2015). A versatile scene model with differentiable visibility applied to generative pose estimation. *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 765-773).

Richards , J. G. (1999). The measurement of human motion: A comparison of commercially available systems. *Human movement science, 18*(5), 589-602.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323-2326.

Sapp, B., Toshev, B., & Taskar, B. (2010). Cascaded models for articulated pose estimation. *European conference on computer vision*, (pp. 406-420). Springer, Berlin, Heidelberg .

Sarafianos , N., Boteanu, B., lonescu, B., & Kakadiaris , I. A. (2016). 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, *152*, pp. 1-20.

Seo, J. O., Han, S. U., Lee, S. H., & Kim, H. (2015). Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics, 29*(2), 239-251.

Sigal, L., Balan, A., & Black, M. J. (2008). Combined discriminative and generative articulated pose and non-rigid shape estimation. *Advances in neural information processing systems*, (pp. 1337-1344).

Sminchisescu , C. (2008). 3D human motion analysis in monocular video: techniques and challenges. *In Human Motion* , (pp. 185-211). Springer, Dordrecht.

Spielholz, P., Silverstein , B., Morgan, M., Checkoway, H., & Kaufman, J. (2001). Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. *Ergonomics, 44*(6), 588-613.

Stoll, C., Hasler, N., Gall, J., Seidel, H.-P., & Theobalt, C. (2011). Fast articulated motion tracking using a sums of Gaussians body model. *Computer Vision (ICCV)*, (pp. 951-958).

Tome, D., Russell, C., & Agapito, L. (2017). Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2500-2509).

Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, (pp. 1799-1807).

Toshev , A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 1653-1660).

Trucco, E., & Verri, A. (1998). Introductory techniques for 3-D computer vision. *Englewood Cliffs: Prentice Hall*.

Ukita, N. (2012). Articulated pose estimation with parts connectivity using discriminative local oriented contours. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3154-3161).

Waldemar, K. (2006). *The International Ergonomic Association (IEA)*. doi:10.1201/9780849375477.ch29

Wang, D., Fei, D., & Xiaopeng, N. (2015, June). Risk Assessment of Work-Related Musculoskeletal Disorders in Construction: State-of-the-Art Review. *Journal of Construction Engineering and Management, 141*(6).

Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 4723-4732).

Workers' Compensation Board. (2018). Retrieved April 2018, from Employers home: Modified work: http://www.wcb.ab.ca/return-to-work/

Xia, L., Chen, C. C., & Aggarwal, J. k. (2012). View invariant human action recognition using histograms of 3d joints. *Computer vision and pattern recognition workshops (CVPRW)*, (pp. 20-27).

Xsens. (2016). Retrieved October 2018, from Xsens: https://www.xsens.com/

Yang, Y., & Ramanan , D. (2011). Articulated pose estimation with flexible mixtures-of-parts. *Computer Vision and Pattern Recognition* (pp. 1385-1392). 2011 IEEE Conference .

Yap, P. T., Paramesran, R., & Ong, S.-H. (2003). Image analysis by Krawtchouk moments. *IEEE Transactions on Image Processing*, *12*, pp. 1367-1377.

Zhang, H., Yan, X., & Li, H. (2018, October). Ergonomic posture recognition using 3D view-invariant features from single ordinary camera. *Automation in Construction, 94*, 1-10.

Zhou, S., Fu, H., Liu, L., Cohen-Or, D., & Han, X. (2010). Parametric reshaping of human bodies in images. *ACM Transactions on Graphics (TOG), 29*(4).

Zuffi, S., & Black, M. J. (2015). The stitched puppet: A graphical model of 3D human shape and pose. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3537-3546).

APPENDIX

This appendix mainly describes the system requirements of this research and the details of the implementation. The whole proposed framework consists of five procedures which are human tracking, initial data acquisition, 2D pose refining, 3D pose reconstruction and ergonomic posture analysis. Each procedure has its own attached folder containing the corresponding codes and data. All the codes are tested on 64 bit Linux (Ubuntu 16.04 LTS) and Python 2.7 environments.

1. Human tracking (py-MDNet)

Prerequisites: PyTorch and its dependencies

Usage: cd ../py-MDNet/dataset/OTB/;

build test_video;

cd test_video;

build img;

cd img;

Put all the test images under the '/img' folder and create a txt file named 'groundtruth_rect.txt' to contain the coordinates of the ROIs in at least first image.

cd ../py-MDNet/tracking

python run_tracker.py –s test_video [-d (display fig)] [-f (save fig)]

After that the images with the tracked ROIs can be generated and then the images are cropped according to the location of ROIs.

2. Initial data acquisition

The initial data acquisition procedure contains two parts which are the 2D initial pose detection (DeeperCut) and the 2D body parts detection (Deeplab v2).

(1) DeeperCut part detectors

Prerequisites: Caffe

Usage: make all pycaffe

pip install click (required for demo only)

cd /deepcut-cnn/models/deepercut; ./download_model.sh

cd /deepcut-cnn/python/pose

python ./pose_demo.py [test_image_folder]--out_name=[output folder]

(2) Deeplab v2

Prerequisites: Caffe

Usage: cd ../deeplab_v2/pascal_person_part/

cd test.sh

Modify the path of the Caffe and the test image folder and so on.

cd list

Create a txt file containing each test image path. (Such as '/JPEGImage/0001.png')

sh ./run_pascal.sh

3. 2D pose refining

The 2D pose refining is performed according to the 2D part detection results. After obtaining the 2D part detection results, the path of the folder containing the part detection results needs to be declared in the 'deepcut-cnn/python/pose/estimate_pose.py'. The codes about how to evaluate the body parts and how to refine the 2D pose based on the results of body part detection have been written in the estimate_pose.py. Hence, the only thing needs to do is to change the 24th line in the 'pose_demo.py'. Instead of importing the 'estimate_pose_BACK', the 'estimate_pose' should be imported.

4. 3D pose reconstruction (Smplify)

Prerequisites: numpy>=1.11.0, scipy>=0.17, chumpy, opendr, matplotlib, OpenCV.

Usage:

Wrap the location information of 2D detected joints into an .npz file and put the file under the directory 'smplify_public/results'.

Put the test image folder under the directory 'smplify_public/images'.

In a new terminal window, navigate to the 'smplify_public/code/' directory.

Modify the path of the image folder, data folder and the output folder.

python fit_3d.py ../ --viz

Then the 3D model can be obtained and the 3D coordinates of the joints are obtained. As for the body angle calculation for the ergonomic analysis, run the script 'validation_angle.py'.

5. Ergonomic analysis

The 3DSSPP has been selected to perform the ergonomic posture analysis. The Run Batch File feature of the software enables automatic analysis of tasks specified in a data file. As for the single posture, the software enables inputting angles manually. After inputting the body angle information, a large number of ergonomic reports can be exported.