

**Video-Based Human Motion Capture and Force Estimation
for Comprehensive On-Site Ergonomic Risk Assessment**

by

Meiyin Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Civil Engineering)
in the University of Michigan
2019

Doctoral Committee:

Professor SangHyun Lee, Chair
Assistant Professor Jia Deng
Professor Vineet R. Kamat
Associate Professor Carol C. Menassa

Meiyin Liu

meiyin@umich.edu

ORCID iD: 0000-0002-9584-3929

© Meiyin Liu 2019

Dedication

To my parents

Acknowledgements

Firstly, I would like to thank my advisor, Dr. SangHyun Lee, for all his help and guidance that he has given me during my PhD study over the past six years, and I would like to thank my PhD committee members, Dr. Jia Deng, Dr. Vineet R. Kamat, and Dr. Carol C. Menassa, for their thoughtful advice for my dissertation. I would also like to thank the scholars who have directly influenced my PhD, Dr. SangUk Han and Dr. JoonOh Seo. I have been very fortunate to have been able to discuss my research with these great scholars.

Next, I would like to thank my colleagues and friends. My colleagues who I spent countless hours with at the University of Michigan discussing our research in general: Dr. Byungjoo Choi, Dr. Houtan Jebelli, Dr. Kwonsik Song, Daeho Kim, Gaang Lee, Sehwan Chung, Kai Fang, Yixin Jin, Kaiqin Yin, Zhiyun He, Chunan Ye, and Xinghui Xu. Their assistance, cooperation, and experience were essential for the completion of my PhD study.

Additionally, I would like to thank the organizations and programs that supported my research: National Science Foundation (NSF), NSF INTERNS, TOYOTA Motor Manufacturing, Humantech Inc., and Power Construction.

Last but not least, I would like to thank my family for their unwavering support and patience. This dissertation would not have been completed without their support.

Table of Contents

Dedication	ii
Acknowledgements	ii
List of Tables	vi
List of Figures.....	vii
Abstract.....	ix
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Remote Sensing-Based Ergonomic Risk Assessment.....	4
1.3 Knowledge Gaps.....	8
1.4 Research Objectives and Approaches	9
1.5 Structure of the Dissertation	13
Chapter 2 Video-Based 2D Human Motion Capture for Posture and Repetition Estimation	15
2.1 Introduction.....	15
2.2 Literature Review.....	18
2.3 Method	22
2.3.1 Human Localization.....	23
2.3.2 Human Pose Estimation.....	25
2.3.3 Optimization	26
2.4 Field Testing	29
2.4.1 Testing Condition and Tasks.....	29
2.4.2 Evaluation Metrics	30
2.4.3 Testing Results.....	31

2.5 Discussion	33
2.6 Conclusion	37
Chapter 3 Video-Based 3D Human Motion Capture for Posture and Repetition Estimation	
.....	40
3.1 Introduction.....	40
3.2 Literature Review.....	42
3.3 Method	46
3.3.1 Frame-wise 3D Human Pose Estimation	46
3.3.2 Optimization	47
3.4 Laboratory Testing.....	49
3.4.1 Testing Conditions	50
3.4.2 Evaluation Metric.....	51
3.4.3 Testing Result	53
3.5 Discussion	53
3.6 Conclusion	54
Chapter 4 Applications of Video-Based Human Motion Capture on Ergonomic Postural	
Analysis	56
4.1 Introduction.....	56
4.2 REBA (Rapid Entire Body Assessment)	57
4.1.1 REBA	57
4.1.2 Method	58
4.1.3 Testing Condition.....	59
4.1.4 Testing Result	62
4.2 Snook's Tables.....	65
4.2.1 Summary of Snook's Tables	65
4.2.2 Method	66
4.2.1 Testing Condition.....	70
4.3.2 Testing Result	70
4.3 NIOSH Lifting Equation.....	71
4.3.1 Summary of NIOSH Lifting Equation	71
4.3.2 Method	73
4.3.3 Testing Result	75

4.3.4 Discussion.....	75
4.4 Conclusion	76
Chapter 5 Vision-Based Hand Push Force Estimation from 3D Motion Capture	78
5.1 Introduction.....	78
5.2 Literature Review.....	80
5.3 Method	82
5.3.1 Kinematic Model for Force Estimation.....	83
5.3.2 Physics-based Force Reconstruction.....	85
5.3.3 Force Optimization	85
5.4 Lab Testing	87
5.4.1 Testing Condition.....	87
5.4.2 Measure of Accuracy	89
5.4.3 Testing Result	90
5.5 Conclusion	91
Chapter 6 Application of 3D Motion Capture and Force Estimation in Biomechanical	
Analysis	94
6.1 Introduction.....	94
6.2 Computerized Biomechanical Analysis Tools.....	95
6.3 Method	97
6.4 Lab Testing	98
6.5 Testing Result	99
6.6 Conclusion	100
Chapter 7 Conclusion and Recommendation.....	102
7.1 Summary of Research.....	102
7.2 Future Research	105
Bibliography	107

List of Tables

Table 2.1 Accuracy of 2D Joint Location Estimation.....	32
Table 2.2 Accuracy of 2D Joint Angle Estimation	34
Table 3.1 Accuracy of 3D Joint Angle Estimation	52
Table 4.1 Selected Postures for Performance Evaluation on REBA	60
Table 4.2 Information on Participating Ergonomists for Performance Evaluation	61
Table 4.3 Performance Comparison of Frequency Estimation on REBA: Video1 (side-view) ...	63
Table 4.4 Performance Comparison of Duration Estimation on REBA: Video1 (side-view).....	63
Table 4.5 Performance Comparison of Frequency Estimation on REBA: Video2 (diagonal-view)	64
Table 4.6 Performance Comparison of Duration Estimation on REBA: Video2 (diagonal-view)	64
Table 4.7 Performance Comparison of Task Variables Estimation for Snook's Tables.....	69
Table 4.8 Performance Comparison of Task Variables Estimation on NIOSH Lifting Equation	74
Table 5.1 Performance Comparison of Force Estimation.....	90
Table 6.1 Accuracy of Biomechanical Analysis on Low Back Compression Estimation.....	100

List of Figures

Figure 2.1 Framework of Video-based 2D Human Motion Capture	23
Figure 2.2 Pipeline of Human Localization	24
Figure 2.3 2D Human Pose Estimation.....	26
Figure 2.4 State Update Scheme for Optimization	27
Figure 2.5 Examples of Raw Video Frames for 2D Human Motion Capture	30
Figure 2.6 Illustration for Evaluation Metric: PCKh	31
Figure 2.7 PCKh of Different Body Joints	31
Figure 2.8 Examples of Estimated 2D Human Pose.....	33
Figure 2.9 Accuracy for 2D Body Joint Location: All Joints VS. Visible Joints	35
Figure 2.10 Accuracy of 2D Body Joint Location: Construction Dataset VS. Daily Activity Dataset.....	36
Figure 2.11 Accuracy of 2D Joint Angle Estimation Regarding Body Joints	37
Figure 3.1 Testing Layout of 3D Human Motion Capture	51
Figure 3.2 Accuracy of 3D Joint Angle Estimation.....	52
Figure 4.1 Snapshots of Test Videos for Performance Evaluation on REBA	59
Figure 4.2 Example of Data Collection Form for Specialists.....	61
Figure 4.3 Performance Comparison of Frequency and Duration on REBA	65
Figure 4.4 3D Human Model Scaling (left) and World Coordinate Rectification (right)	67
Figure 4.5 Pipeline of Assessing Lifting Tasks with Snook's Tables.....	68

Figure 4.6 Examples of Recorded Lifting Tasks	68
Figure 4.7 Example of Lifting Task and Captured Motion Data	71
Figure 4.8 Pipeline of Assessing Lifting Tasks with NIOSH Lifting Equation	73
Figure 5.1 Framework of Force Estimation from Vision-based Human Motion Capture	83
Figure 5.2 Recurrent Neural Network Structure for Force Optimization	87
Figure 5.3 Test Setup for Force Estimation	88
Figure 5.4 Examples of Force Profile: Estimation VS. Ground Truth	89
Figure 5.5 Bland-Altman Plot of Estimation and Groundtruth Agreement on Force Estimation	92
Figure 6.1 Screenshot of Biomechanical Analysis Tool 3D SSPP™	96
Figure 6.2 Coordinate Systems Involved for Biomechanical Analysis	98

Abstract

Construction is one of the most hazardous industries with high non-fatal injuries because it involves physically demanding tasks performed in an unstructured and dynamic environment. Work-related musculoskeletal disorders (WMSDs) are the major cause of non-fatal injuries. Various methods, such as self-report, observation, and direct measurement, are used for assessing the risk level of WMSDs by quantifying the ergonomic risk factors (e.g., posture, repetition, and force). However, they are either time consuming or error-prone (e.g., self-report and observation) or invasive (e.g., direct measurement).

The recent advancement of computer vision allows for rapid, accurate, and non-invasive motion capture only using ordinary cameras. Key challenges remain for applying to assess jobs' ergonomic risks: 1) long-lasting occlusion in a construction site creates an obstacle to enforcing kinematic and temporal consistency between frames to estimate posture's frequency and repetition; 2) as a critical risk factor for ergonomic risk assessment, force is very challenging to non-invasively estimate, which hinders field applications; and 3) little effort has been made for comprehensive ergonomic risk assessment.

These knowledge gaps were addressed by three research objectives: 1) develop and validate a video-based human motion capture framework to quantify ergonomic risk factors of posture and its repetition by extracting continuous 2D/3D human model with enforced kinematic and temporal consistency; 2) develop and validate a video-based hand push force estimation framework; and 3) apply the risk factors estimated by videos to comprehensive ergonomic risk assessment tools including postural and biomechanical analysis.

Results yielded around 11.6 and 7.5 degrees of joint angle estimation error for 2D and 3D motion captures, respectively, despite prevalent occlusions. Also, resultant frequency and duration comparison with experienced ergonomists' observation demonstrates a great potential to robustly quantify jobs' ergonomic risk factors of posture and repetition. Lab-based testing shows an accurate peak force occurrence time and peak force magnitude estimation, suggesting a potential to quantify critical variables of push force exertion only from videos. By applying the collected risk factors comprehensively to several ergonomic risk assessment tools, it demonstrates a promising level of risk assessment accuracy compared with expert observation and sensor-based measurement. The proposed video-based motion capture and force estimation frameworks for comprehensive ergonomic risk assessment are expected to greatly reduce the time and effort of on-site data collection and increase the number of evaluated jobs with higher frequency thereby providing a better opportunity to understand and control WMSDs.

Chapter 1 Introduction

1.1 Background

Construction is a labor-intensive industry, requiring labor force as one of the major resources in projects. In the U.S., 11.2 million employees worked in the construction industry, which accounted for about 7% of the overall U.S. workforce in 2018 (BLS 2018). From an economic perspective, labor cost also often forms 33-50% of the total project cost in construction (Hanna et al., 2001; Siriwardana and Ruwanpura, 2012). In this regard, efficient planning, monitoring, and controlling of the onsite employees' performance is key to the success of construction projects. Particularly, the non-fatal injury rate of construction workers has still ranked as the fourth-highest among U.S. industries (BLS 2015a; BLS 2015b).

Construction workers are frequently exposed to forceful and repetitive exertions with awkward postures, which leads to work-related musculoskeletal disorders (WMSDs) (Boschman et al. 2012; Everett 1999). WMSDs are a major cause of non-fatal injuries across industries. The rate of WMSDs in construction is 25% higher than the rate for all other industries combined in 2013 (BLS 2014).

In contrast to many occupational diseases that develop from specific hazardous agents, most WMSDs are found to be multifactorial and are the results of exposure to concurrent risk factors (Van Der Beek & Frings Dressen 1998). Scientific research studies have investigated the level of risk to develop WMSDs by identifying the associations between quantified exposure to various risk factors and the incidence or prevalence of WMSDs (Bernard 1997). Based on the

findings, ergonomic principles could be proposed by providing specific measures of exposure to risk factors to prevent or limit the risks (Ringleberg & Voskamp 1996). The well accepted primary risk factors include, but are not limited to, awkward postures, repetition (frequency and duration), and exerted force (NIOSH 2014).

To control WMSDs, ergonomic evaluation methods are widely applied in research and practice to evaluate the jobs regarding quantified exposure to risk factors. The methods fall into three major categories of self-reports, observational methods, and direct measurement (David 2005).

Self-reports from workers include diaries, interviews, and questionnaires where the subjects report relevant information. The content mainly includes symptoms such as pain located in a specific body part as well as postural discomfort. Self-Estimated exposure to risk factors and demographic information could also be collected. Examples of questionnaires include, but are not limited to, Nordic Musculoskeletal Questionnaires (Cheraghi et al. 2018; Kuorinka et al. 1987; Saha et al. 2017), Borg Rating of Perceived Exertion (RPE) Scale (Borg 1990; Jebelli and Lee 2019), and Job Requirements and Physical Demands Survey (JRPDS). Self-reports have been widely applied mainly due to their simplicity. Despite their accessibility, such methods lack objectivity and remain error-prone.

Observational-based methods consist of simpler observational techniques (David 2005), or pen and paper-based observational methods (Beheshti et al. 2016), and advanced observational techniques (David 2005). Simpler observational techniques usually require an expert to observe and record the exposure to risk factors in real-time by following a pre-defined check-list based on an ergonomic risk assessment tool, such as Rapid Entire Body Assessment (REBA), Rapid Upper Limb Assessment (RULA), and Ovako Working Posture Assessment

System (OWAS). These tools generally require information about exposure to primary risk factors, including posture, repetition, and force. Based on the required level of detail, the input data could be as rough as “back straight” vs. “back bent” as in OWAS, or as detailed as “trunk bending forward within 0° - 20°” vs. “trunk bending forward within 20° - 60°” as in REBA and RULA. Advanced observational techniques enable the utilization of a video recording device, so the observer can playback and pause the video to observe multiple body parts simultaneously. Without the real-time constraint, a goniometer can be used to measure the body joint angles on a paused video (Fransson-Hall et al. 1995), and even a body-attached marker for location tracking to estimate rough velocity and acceleration from videos (Mahyuddin et al. 2011). These methods do not attach sensors to the human body and collect data. For additional information on these techniques, please see the systematic review in David (2005). Observation-based methods provide more subjective and accurate data than self-report but still, remain time-consuming and error-prone.

With the advancement of sensing technology, a wide range of more direct measurement methods was developed, with sensors attached directly to the human body (David 2005). Sensing technology has gained extensive attention due to its accuracy, objectiveness, and versatility, compared with self-report and observation-based methods. Acceleration-based sensors were a commonly used type mounted on the subjects’ key body part to analyze the risk factors concerned. Yan et al. (2017) augmented personal protective equipment (PPE) by mounting inertial measurement units (IMUs) on the worker’s head and trunk, and measured both body parts’ postures in angles including flexion, lateral, and axial. Nath et al. (2017) measured the trunk and shoulder flexion by mounting smartphones on the subject’s waist and arm, respectively. To directly obtain joint angle measurements for ergonomic evaluation, Alwasel et

al. (2011) applied magneto-resistive sensors to evaluate the shoulder posture. To collect the full-body postures simultaneously, commercial human motion capture devices, such as IMU (Cho et al. 2018) and optical marker-based system, e.g. VICONTM, OptoTrakTM, and QualysisTM (Alwasel 2017; Han and Lee 2013; Seo 2016) were explored. Apart from posture, force exertion was also measured or estimated. Jahanbanifar and Akhavian (2019) predicted the subject's hand push force from acceleration data collected by a smartphone that affixed on the arm. Jacobs and Ferris (2015) applied pressure sensors on subjects' shoe insoles to measure the ground reaction force. With the static or continuous force profile with directional information, the biomechanical analysis could be conducted to estimate the internal loads on major body joints. Direct measurement can collect high-quality data but has the drawback of being invasive and not very time-saving considering the device setup and testing.

The three types of methods: 1) self-report, 2) observation, and 3) direct measurement, are listed in order of increasing accuracy of the data collected from and invasiveness to the worker being evaluated (David 2005). While observation-based methods are more applied to evaluate on-going work for its moderate invasiveness, direct measurement is mainly used on closely assessing simulated jobs in a controlled environment due to its high accuracy. In practice, there remains a lack of a non-invasive yet rapid and accurate tool to evaluate ergonomic risks of on-site construction workers' jobs.

1.2 Remote Sensing-Based Ergonomic Risk Assessment

Aiming at a non-invasive yet rapid and accurate ergonomic risk assessment tool for on-site application, remote-sensing technologies were explored, towards collecting data as a body-attached sensor can, and quantify the risk factors as an observation-based tool requires.

Range and image/video sensors are the primary remote sensing devices to capture human motion coded from the location of major body joints (Wang et al. 2015). Time-of-flight 3D sensors generate the point cloud of a given scene. With a proper segmentation of the subject from the background, human skeleton or posture could be extracted with proper algorithm or software (Diraco et al. 2013). The stereo camera also provides point cloud data similar to time-of-flight 3D sensors by commercialized software, e.g., Bumblebee XB3™ (Seo et al. 2017; Starbuck et al. 2014). To mitigate the demand for expertise in algorithm development, the RGB-D sensor, Microsoft Kinect (Redmond, Washington), was applied to extract and visualize a human skeleton in user-friendly software. It can collect the subject's joints location by converting the directly measured depth from the particular image pixel to the corresponding point in the 3D scene (Wang et al. 2015). It further provides pixel color, thus including appearance information compared to pure 3D sensors. With the publicly available software development kit (SDK), a human skeleton in 20 joints could be extracted and visualized, and easily estimate needed risk factors in an ergonomic risk assessment tool e.g. OWS (Diego-Mas and Alcaide-Marzal 2014; Dzung et al. 2017; Seo 2016) or for customized risk factor analysis of postures (Ray and Teizer 2012; Seo et al. 2017). As the RGB-D sensor can only be operated in an indoor environment and possesses a limited range of 4 meters (Seo et al. 2017), however, its on-site deployment on an outdoor construction site is not feasible.

In pursuit of rapid and non-invasive remote sensing technology with on-site accessibility, camera-based approaches augmented by computer vision algorithm attracted noticeable research efforts in recent years. The major challenge for camera-based approaches focuses on their robustness under various viewpoints, illumination, and occlusion condition, which relies on an advanced computer vision algorithm to handle and generate quality data to quantify ergonomic

risk factors. There are two categories of methods to quantify human motion from 2D images/videos for worker's postural analysis.

One type of methods is to directly estimate the posture categories from an image or video frame. The most detailed data that an observation-based ergonomic risk assessment tool requires is posture categories, by partitioning a body joint angle's range (e.g., shoulder) into several categories of certain increments (e.g., $<30^\circ$). It does not necessitate estimating the accurate numerical angle value, thus many studies apply machine learning algorithms to train a posture classifier and directly recognize the posture categories of each body joint from an image. Some earlier studies (Gong et al. 2011; Liu et al. 2016) showed promising potential of such methods by recognizing postures associated with the performed tasks (e.g., traveling, transporting, and ladder-climbing) including posture with ergonomic risk (e.g., bending and squat-lifting). Inspired by such ideas, Seo et al. (2016) extracted human silhouette from videos and recognized among several common postures including back-bending, arm-reaching, and knee-bending, defined by an ergonomic risk assessment tool (OWAS). This study suggested the potential to automate the assessment tool to supplement human observation. Instead of recognizing postures of different body parts, Greene et al. (2019) further recognized among different lower-limb postures: squatting, stooping and standing, by extracting partitioning information from the human silhouette. The other type of methods first captures human motion in an articulated model with explicit information such as joint locations in Cartesian coordinate. Subsequently, such information could be used to calculate body joint angles and then be converted to common posture categories as a key input of assessment tools. Liu et al. (2017) demonstrated the potential for 2D human motion capture of on-site construction workers with smartphone video. To mitigate the distortion brought from 2D video, Yan et al. (2017b)

developed a view-invariant feature with 2D human motion capture to classify postures following the ergonomic evaluation tool, i.e. OWAS. (Dai and Ning (2013) estimated 3D human motion from the monocular camera by known information of the camera's location and pose. Yu et al. (2019) introduced the study that captured 3D human motion from monocular/2D images and automated posture classification following assessment tool (i.e. REBA). The stated limitation of this study was the lack of quantifying the repetition, which is a major risk factor for assessment tools. In addition, the human model was trained from images with single subject thus hard to generalize to the general population with diverse stature.

Comparing with posture recognition, fewer studies explored the potential to estimate force exertion non-invasively. Gaddam et al. (2016) demonstrated the potential of estimating ground reaction force from the spine's frame-wise location collected by RGB-D sensor (Microsoft Kinect). Pham et al. (2015) used the same sensor to estimate hand contact force by capturing the hand pose. To apply ubiquitous visual sensing device, Sartison et al. (2018) developed a machine learning-based approach to estimate finger grip force from RGB frame sequence with visual markers on fingers. These studies showed the potential of estimating force from visual and motion data, but only focused on a single body part. To address this issue and explore the potential to generalize to every individual body joint and various tasks, Pham et al. (2018) demonstrated the feasibility of estimating hand contact force of multiple tasks from whole-body motion data captured from IMU sensors. The method was developed for arbitrary body joints and tasks which showed great potential to estimate hand forces for construction tasks that involve whole-body movement. A thrilling question arises whether human motion data captured from a video could generate hand force estimation that most ergonomic risk assessment tools require as a key input.

1.3 Knowledge Gaps

Summarized from existing studies, monocular camera-based approaches are promising for on-site ergonomic evaluation of being rapid, non-invasive, and effort-saving. However, as it relies on advanced computer vision and machine learning algorithm to extract quality and versatile input for accurate and comprehensive ergonomic risk assessment, there remain several key issues to be addressed.

First, there lacks an approach to estimate the repetition of posture. The repetition of a posture is expressed by the frequency (number of occurrences) and duration that a specific body joint angle (e.g., shoulder angle) falls into a pre-defined range (e.g., $30^\circ - 60^\circ$). The knowledge gap of estimating the repetition of postures from videos, however, would not be easily addressed by independently capturing the human model in a frame-wise manner, as the estimated postures may not be accurate enough to reserve temporal smoothness between consecutive frames, especially under frequent self- and external occlusions in an unstructured and dynamic construction site.

Second, there lacks a vision-based force estimation approach for non-invasive assessment of forceful exposure. Exerted force was demonstrated viable to be estimated from 3D motion data collected by body-attached devices. The motion data has a full degree of freedom (DOF) that can generate all needed kinematic parameters of the human body to formulate the physics-based equations and optimization process. However, vision-based motion capture uses a simplified human model with less DOF. For example, a forearm's axial rotation motion can be captured by sensors attached to its surface, while a vision-based approach cannot as represented by a non-volumetric line segment. Fewer DOF results in an insufficient number of kinematic parameters to formulate equations. Consequently, this knowledge gap could be further expressed

as: the potential of estimating hand force with the simplified human model of reduced DOF via vision-based motion capture has not been explored.

Further, little effort has been made to apply the collected comprehensive risk factors (i.e., posture, repetition, and force) to quantify risk factors in ergonomic risk assessment tools. This is mainly because existing efforts tried to address a limited number of pieces in vision-based ergonomic risks assessment (e.g., frame-by-frame motion capture). The proposed motion capture framework in this thesis explicitly estimates body joint location and use it to calculate the joint angle and quantify required risk factors (i.e., posture, repetition, and force) for ergonomic risk assessment tools. Among angle-based postural analysis tools like REBA, only posture was quantified while repetition was not incorporated. Little attempt was made for distance-based postural analysis tools like NIOSH Lifting Equation and Snook's Tables, for which posture is estimated from body joint location. Biomechanical analysis that requires continuous whole-body joint angle and tri-axial hand force was not explored with vision-based motion and force data collection. All of them are widely used in practice, which will a great benefit from the proposed vision-based framework.

1.4 Research Objectives and Approaches

To address the aforementioned issues, this research proposes a video-based approach (e.g., smartphone's built-in camera) with developed human motion capture and a hand push/pull force estimation framework to enable automated ergonomic risk assessment for on-site construction jobs. Specifically, several research objectives are identified, as listed below. The corresponding approaches are briefly described following each objective.

1. Develop and validate a video-based human motion capture framework to quantify ergonomic risk factors of posture and its repetition by extracting continuous 2D/3D human model with enforced kinematic and temporal consistency while handling long-lasting occlusion in a construction site.

The proposed framework addresses the first knowledge gap, by developing a video-based motion capture framework with two major modules. The first module leverages the state-of-the-art 2D/3D human pose estimation method for still images and obtains initial motion data independently in a frame-wise manner. It is followed by an optimization module embedding kinematic constraints between adjacent joints and temporal constraints between consecutive frames. The constraints are developed from anthropometry data and knowledge about human motion (e.g., range of arms reach). Further, the human pose estimation models (2D and 3D) are trained with large dataset incorporating a diversity of subject's stature, appearance, and surroundings. It enables the model to handle subject variation and occlusion. In addition, as the framework is designed specifically for ergonomic risk factor estimation, the joint angle and body part length are directly constrained, instead of only joint locations which were commonly applied in the computer vision community. A 2D-based framework is developed as an initial step. Subsequently, the framework replaces the 2D pose estimation module with a 3D approach and optimization module with one incorporating corresponding modification for 3D motion data.

To validate the proposed framework for 2D human motion capture, body joint location and angle estimation accuracy are evaluated against human annotation on in-field videos captured for 10 construction jobs. The framework with 3D motion capture was validated regarding body joint angle estimation accuracy that is evaluated by a marker-based motion capture system in a simulated lab setting.

2. Propose a hand push/pull force estimation framework with the simplified human model of reduced DOF via the proposed video-based 3D human motion capture. The lifting and lowering tasks were widely studied as the exerted hand force could be assumed as the object's load weight (Fings-Dresen et al. 2000). Tasks like pushing and pulling, which are also prevalent in construction jobsites, usually rely on the direct measurement of force with a specialized device. To test the feasibility of estimating hand force for these tasks, a video-based hand force estimation framework was proposed to evaluate tasks without prior knowledge about the force exertion. As the hand force cannot be directly calculated from motion due to its indeterminacy issue, this research proposes a two-module framework that first calculates the physically plausible hand force from whole-body motion and then estimates actual exertion with an artificial neural network. The 3D human motion data captured from proposed video-based approach has reduced DOF with an insufficient number of kinematic parameters. This research formulates the physics-based equations and optimization process with limited parameters and shows the potential of estimating push/pull force with simplified human model.

To validate the force estimation framework, lab testing of hand push force estimation from video-based 3D motion data capture was conducted with its accuracy evaluated against force measured by a body-attached force transducer.

3. Apply the collected risk factors such as posture, repetition, and force to ergonomic risk assessment tools. To address the third knowledge gap based on how the previous two were addressed, this research evaluates the estimation accuracy of risk factors and overall risk in 4 different tools: 1) An angle-based postural analysis tool (e.g., REBA) that requires posture and its repetition. With the proposed 2D motion capture framework, this research validates REBA-defined posture's frequency and duration estimation compared with 27

ergonomists' observation in a lab testing, against body-attached IMU-based motion capture system as the baseline. 2) A distance-based manual material handling (MMH) assessment tool (e.g., Snook's Tables) that requires the horizontal distance from the hands to the front of the body and vertical distance of hands' displacement. Snook's Tables is commonly applied due to its simplicity and insensitivity to input data incremental changes. The parameters needed for Snook's Tables are estimated from 10 different lifting tasks with 3D motion capture in a field condition and validated against tapeline measurement for distance-related variables. 3) A distance and 3D angle-based MMH assessment tool (e.g., NIOSH Lifting Equation) that require similar risk factors with Snook's Tables with further precision and additive risk factors. Hands' horizontal and vertical distance from ankles, and hands' vertical travel distance are needed and be precise to inch. Additionally, the asymmetric angle is required, indicating the level of trunk twisting during a lifting, which should be 3D. NIOSH Lifting Equation is believed to be more conducive than Snook's Table as it requires more data (e.g., 3D asymmetric angle) with higher precision, but sensitive to input data variation (Russell et al. 2007). Similar validation protocol with that in 3) is applied to validate the estimated risk factors, except that estimated asymmetric angle is evaluated against manual observation. Further, as 3D motion data is shared for these two tools, the sensitivities of input variables, with different levels of precision, to the overall risk will be discussed. Laboratory-based testing is conducted with a subject performing pushing tasks. The risk level of the lower back is validated against that calculated from motion data collected by marker-based system and force data by a hand-attached force transducer.

1.5 Structure of the Dissertation

This dissertation is a compilation of studies to achieve the proposed research objectives. It consists of 7 chapters that Chapter 2 and 3 address the first knowledge gap, and Chapter 5 addresses the second while Chapter 4 and 6 address the third. Following is the list of the brief description of the chapters.

Chapter 1: Introduction. This chapter covers the background and motivation, current approaches, knowledge gaps, and research objectives with proposed approaches.

Chapter 2: Video-based 2D Human Motion Capture for Posture and Repetition Estimation. This chapter introduces a proposed framework to capture continuous 2D human motion from a video and demonstrates its robustness in quantifying ergonomic risk factors of posture and repetition. 2D body joint location and angle estimation accuracy are validated by manual annotation on images for on-site construction jobs.

Chapter 3: Video-based 3D Human Motion Capture for Posture and Repetition Estimation. This chapter adopts the framework from the prior chapter with modification to incorporate 3D human motion capture with corresponding modification. 3D body joint angle estimation accuracy is validated by marker-based motion capture system for lab-based simulated tasks.

Chapter 4: Applications of Video-based Human Motion Capture on Ergonomic Postural Analysis. This chapter consists of 3 parts. First, the feasibility of application in an angle-based postural analysis tool (REBA) is demonstrated. The estimation of frequency and duration for REBA-defined postures via the proposed 2D approach and specialists' observation are compared against that calculated from body-attached IMU-based motion capture system in a simulated lab testing. Second, the feasibility of application in a distance-based postural analysis

part from a risk evaluation tool (NIOSH Lifting Equation) is demonstrated with the proposed 3D approach. Field testing was conducted using tapeline measurement to collect distance-related variables. Additionally, the application to another distance-based tool (Snook's Tables) is also demonstrated in the same field condition.

Chapter 5: Video-based Hand Push/Pull Force Estimation. This chapter introduces the proposed hand force estimation framework from captured whole-body human motion data via a video recording. Lab testing was conducted to demonstrate the feasibility of estimating hand push force with video-based 3D motion capture. The estimated tri-axial force was evaluated against that collected from a 6 DOF force transducer attached to the hand.

Chapter 6: Application of 3D Motion Capture and Force Estimation in Biomechanical Analysis. This chapter validates the effectiveness of estimated hand push force by comparing the biomechanical analysis result from the video-based approach and directly measured hand push force from a force transducer and motion data from a marker-based motion capture system.

Chapter 7: Conclusions and Recommendations. This chapter provides a summary of the conclusions that can be drawn from the studies. Several recommendations for future research derived from this dissertation are also provided.

Chapter 2 Video-Based 2D Human Motion Capture for Posture and Repetition Estimation

2.1 Introduction

To assess the overall ergonomic risk of a job, an ergonomic risk assessment tool first identifies and quantifies the risk factors (e.g., posture, repetition, and force). Posture is commonly represented by a specific range of joint angle for a body joint (e.g., 20°- 60° of shoulder angle). Repetition, on the other hand, is represented by the frequency and duration of specific posture. Duration is straightforward to quantify, as to sum up the number of time frames where the posture concerned is identified. The frequency can be the number of occurrences of one posture per unit time in tools like REBA and RULA, or one posture for a specific period in tools like Snook's Table. Estimating frequency is essentially identifying each occurrence of posture.

Compared with the significant amount of studies focusing on automated posture recognition, frequency estimation, for quantifying repetition, is much less studied in non-invasive and effort-saving remote sensing-based approaches. While routinely given by manual input in other industries, including manufacturing, automated frequency estimation is vital for on-site application in construction. Specifically, construction jobs are much less cyclical and have larger variation between cycles, compared with industries like manufacturing. As remote sensing-based technologies, especially with ordinary cameras, relies on an algorithm to estimate measurements, the ability to quantify posture and its repetition for such approaches is vital for assessing ergonomic risk for on-site construction jobs.

On the other hand, the unstructured and dynamic environment in a construction site brings in various challenges to accurately estimate posture and repetition. For posture recognition, construction jobs have apparent bias and large variation in appearance compared with common human pose datasets collected from daily activities. Construction jobs require wearing personal protective equipment (PPE) that impose bias of appearance especially on body joints such as hands (with gloves), head (with hardhat) and trunk (with reflective vest). The variation of appearance focuses on the interaction with various tools and materials. Additionally, frequent occlusion creates a challenge to estimate the body joint location. For a camera-based approach, the ability to recognize posture needs to handle the appearance variation resulting from the site condition. For repetition estimation, the challenge focuses on that little effort was made to automate the quantification, prior to discussing the frequent, long-lasting occlusion that creates a further challenge for the algorithm to handle.

With the advancement of deep learning in the recent years, studies have shown the human pose estimator trained from daily activity images but on a large scale (e.g., 40K images in MPII human pose dataset) has a potential to yield reasonable result on images taken from a construction site. Liu et al. (2017) demonstrated the potential of applying a state-of-the-art convolutional neural network-based 2D human pose estimator trained from 40K images collected from daily activities, directly to video frames of 2 actual construction jobs with feasible accuracy on estimating body joint location. Yan et al. (2017) showcased several images from construction sites with super-imposed 2D skeleton captured by a convolutional neural network. Though without sufficient validation of the captured motion data for posture recognition, potential has been shown on a small number of static images.

While most studies still measure frequency and duration manually (Bao et al. 2006), a few studies made attempts to estimate the posture frequency from continuous human motion captured from videos. The key to estimate frequency is to identify the start and end time frame of the event concerned. By assuming the reliable quality of collected motion data, Chen et al. (2013) identified the time frames where the velocity of the wrist's pixel location reached its local minima and labeled them the start/end point of a job cycle. Akkas et al. (2017) advanced the frequency estimation method by introducing machine learning algorithms and use the first cycle to train the estimator, without assuming perfect motion data. While these methods were tested to be feasible for cyclical jobs that provide some prior knowledge about the motion pattern, they would not be a feasible alternative for construction jobs, which are dynamic and have significantly fewer cyclical movements and larger variation between cycles due to an unstructured and dynamic environment. To identify the frequency of a posture, the key is to identify the time frame when a joint angle value enters or exits the boundary of a given range. The challenge then focuses on the robustness of frame-wise angle estimation, while reserving its temporal consistency across frames.

Given recent advancements in human pose estimation algorithms in the computer vision community, angle estimation accuracy does not pose an excessive challenge to researchers. Temporal consistency, however, appears to be much less studied, especially for videos from construction jobsites cluttered with severe and long-duration occlusions. To the best of my knowledge, this challenge still remains unsolved.

2.2 Literature Review

Most studies addressing human motion capture on video frame sequences, essentially address two problems: spatial consistency within a frame and temporal consistency between frames. This section reviews the existing literature and has three major focal points: an explanation of the overarching framework to enforce temporal consistency while handling long-lasting occlusion, which is then followed by a review of its two key modules for spatial and temporal consistency respectively.

Framework of Human Pose Estimation on Video

As most of the studies for general human motion capture application favors to work towards real-time pose estimation, they process a video from the first to its next, without assuming the availability of frames after the one under processing (Achilles et al. 2016; Xiao and Zhu 2018; Yang et al. 2005). Consequently, the most commonly applied framework is detection-and-tracking-based approaches that estimate one current frame at a time, based on the estimated human pose estimated in the past frames (Dabral et al. 2018; Girdhar et al. 2018; Gkioxari et al. 2016). This type of methods incorporates two modules: detection module to enforce spatial consistency between body joint location within a frame and tracking module for temporal consistency between frames.

Among a smaller number of studies that assume availability of future frames (i.e., formulate the problem as offline human pose tracking in videos), a two-stage scheme is commonly applied in which the first module generates a frame-wise estimation of human pose independently on every frame and the second utilized global optimization to modify the body joint location by enforcing temporal constraints (Baradel et al. 2017). However, most of the

studies focused on addressing short-term occlusion/imprecision for 3 - 20 frames (Achilles et al. 2016; Fabbri et al. 2018; Gkioxari et al. 2016; Zhang and Shah 2015). For a normal video taken at 30 frames per second (fps), 20 frames represented less than one second, which is not sufficient for the long occlusion duration of videos captured from construction sites.

Among methods dealing with human pose estimation in videos with a specific focus on long-lasting occlusion handling, studies about offline videos were reviewed. Yang et al. (2005) first recognized the object's state of being before, during, or after occlusion, and then decided its location by assuming it during occlusion was similar to that during non-occluded states. This assumption might be valid, given the application scenario of human location tracking in a surveillance video. However, it is not viable for locating body joints with frequent movements. Similarly, Liang et al. (2018) detected the frames with occlusion, and then recovered the occluded object's trajectory based on the occluding object. If occluded by its container, the trajectory would align with the detected container while otherwise remaining still. Given the complex interaction between construction workers and related tools/materials, this assumption would be difficult to generalize. Kobayashi et al. (2018) applied the two-stage detection-and-correction scheme on human pose estimation and corrected the occluded body joints' location by assuming a linear translation between consecutive frames. While this scheme initially appeared promising, the linear translation assumption would not work beyond its 3-frame short sequence. The proposed method adopts the two-stage detection-and-correction scheme, while the detection and correction modules are developed differently according to the literature review on possible methods for two modules, respectively, as further explained below.

Frame-wise Human Pose Estimation

Convolution neural network (CNN or ConvNet) has attracted much attention in the computer vision domain, for its impressive performance on a range of visual tasks (Tompson et al. 2015; Zeiler and Fergus 2014), such as human pose estimation. CNN-based approaches share the advantage that there is no need to develop or select hand-crafted features, which saves significant human effort to explore. On the powerful computing resource developed recently, i.e., graphical processing unit (GPU), a huge-scale dataset could be trained and equip the model with strong ability to generalize across datasets.

Among the CNN-based approaches for 2D human pose estimation, early work (Toshev and Szegegy 2014) formulated the pose estimation as a regression problem (Papandreou et al. 2017) and directly estimate the coordinates of joint locations regarding image pixel. Later work found it more advantageous to generate joints location confidence map, and then calculate the optimal joints locations by finding the confidence maxima (Jain et al. 2014; Newell et al. 2016; Tompson et al. 2015).

Based on the widely accepted finding from the computer vision domain, the literature review focused on this genre of methods. The existing 2D human pose estimation work falls into either of two sub-categories: top-down or bottom-up approach (Papandreou et al. 2017). The top-down approach assumed only a single person presented in the image and bounded by a known rectangle (i.e., “bounding box”). Subsequently, body joints associated with the subject were localized within the box. The bottom-up approach assumed an unknown number of people appearing in the image and localized all body joints first. Then, the association between body joints and their subjects were identified. There remains no final conclusion of which category generally outperforms the other (Papandreou et al. 2017). However, top-down approaches are

more promising for on-site ergonomic evaluation for two major reasons: 1) it should be safe to assume single person presence for ergonomic evaluation as an observer tends to focus on one subject at a time and maximize the human figure in the recorded video to capture the most details; and 2) localizing body joints before identifying their association with the subject might be risky for construction jobs. For example, construction workers' hands are usually covered with gloves and co-occur with tools or materials. As the appearance of such joints is remarkably different from the dataset trained for the human pose estimator, localizing the joints first is expected to be challenging.

Based on the literature review and subjective hypothesis, preliminary testing was conducted by comparing the performance of representative methods from top-down and bottom-up approaches. The testing result suggests the top-down methods work better for the sample dataset from on-site conditions. The state-of-the-art method of the category of top-down approaches (Newell et al. 2016) was selected as a major module in this research. Additionally, as multi-person presence is possible in construction sites, to adopt a top-down approach that assumes a single person's presence, a human location module is added prior to pose estimation module.

Detection and Optimization of Occluded/Misdetected Joints

Regarding the detection of occluded or misdetected joints that need correction, a literature review was also conducted. Most of the approach focused on developing kinematic (or anatomical) and temporal constraints to evaluate the quality of raw detection (Kobayashi et al. 2018):

1) Confidence of joint location. As a general output from 2D human pose estimation methods, the confidence value of each joint suggests the quality of location estimation. From the empirical evaluation, occluded or misdetections usually come with a low confidence level.

2) Joints translation between frames. Despite that the joint location was estimated in 2D image's pixel coordinates, the projected translation (change of location) between two consecutive time frames would be small, given a normal video captured at 30 fps.

3) Change in bone length between frames. A single joint's (e.g., wrist) translation may result from movements of all linked ones (e.g., shoulder) and cannot be constrained too tightly. The constraint on the change of bone length (e.g., forearm) could work as a supplement to better focus on the quality of joints concerned (e.g., wrist/elbow).

As these ideas focused on the robustness of joints' location estimation, this research added several more to constrain the joint angles directly, as ergonomic risk assessment focus on the estimation of body joint angle. In addition, as anthropometry data provides a statistical foundation for human figure measurement (e.g., bone length ratio of the major population), relevant data was also included in this work.

2.3 Method

A two-stage detection-and-correction scheme was adopted for the proposed video-based 2D human motion capture approach. The detection phase was implemented by a top-down 2D human pose estimation module. To incorporate the fast advancement of the pose estimation in the computer vision community, the proposed framework considered it a replaceable module and aligned the rest modules with its common interface. The subsequent correction phase was formulated as an optimization module by enforcing kinematic and temporal constraints. As the

selected top-down pose estimator assumed a known bounding box around the human, an additional human localization module was developed prior to the human pose estimation module. The structure could be illustrated by the proposed framework, as shown in Figure 2.1.

The human localization and the human pose estimation modules were adopted from existing work and could be replaced with more advanced methods. The selected algorithms will be briefly explained but with specific emphasis on necessary modification about the integration interface. The optimization module was the original work of this research and will be discussed in detail.

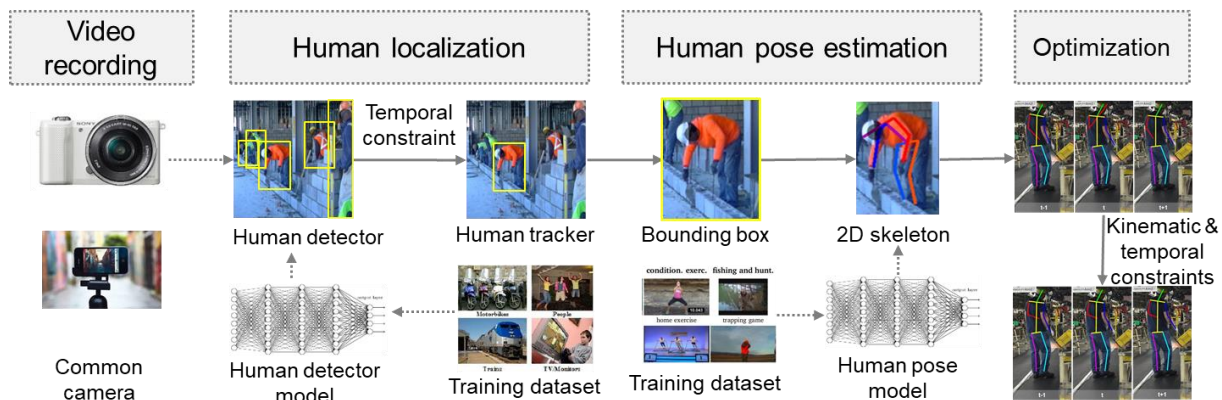


Figure 2.1 Framework of Video-based 2D Human Motion Capture

2.3.1 Human Localization

As the prior module to human pose estimation, a human localization module was developed to generate a bounding box within which all the body joints should present. In the computer vision community, this problem was addressed by object detection methods. Modern object detectors built with convolutional neural networks perform remarkably well in detecting a spectrum of objects, including humans. As the existing object detection methods detect all the humans in still images, the modification was included attempting to track the targeted worker throughout the frame sequence. A pipeline was developed as the human localization module (Figure 2.2) by

enforcing temporal constraints between the bounding box locations across frames and only retain the targeted worker.

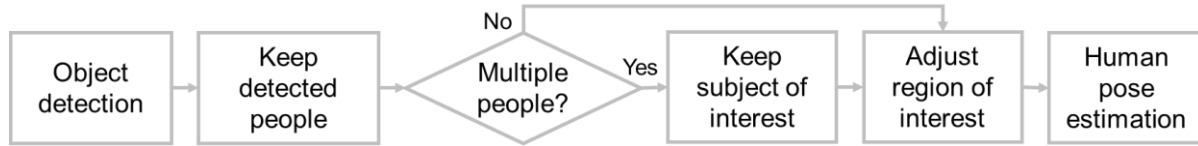


Figure 2.2 Pipeline of Human Localization

Working as the core algorithm of object detection, Faster R-CNN (Ren et al. 2015), which enables the detection of 20 different categories of objects, including humans, is selected. Faster R-CNN performs the task in two steps. First, some regions are proposed by a deep fully connected convolutional neural network. The regions, each with an “objectness score” (Ren et al. 2015), are image patches where the target is of probable presence with the score quantifying the confidence. In the second step, a Fast R-CNN (Girshick 2015) detector gives the eventual bounding box of the target from provided regions.

The algorithm originally reserves a detected bounding box for every human. To filter out irrelevant humans other than the worker of interest, a conditional statement is added. If multiple humans are detected, the conditional statement would be proved, and the worker of interest would be reserved. Specifically, to distinguish the worker of interest from the rest, a simple temporal constraint is applied, which favors the bounding box closer to the worker detected in the preceding frame while. As for the starting frame, this statement favors the worker with a bounding box closer to the image center.

The human localization module aims to generate the smallest bounding box of the target. In practice, specific body joints might be excluded from the boundary. The human pose estimation algorithm, however, needs a conservative estimate of the human bounding box where

every joint must be included. To address this practical issue, it is worth mentioning that in this study, the bounding box from the human detection algorithm is expanded by 50% about its width and height to include every joint.

2.3.2 Human Pose Estimation

Human pose estimation is a fundamental module in the proposed framework. The CNN-based algorithm (Newell et al., 2016) with state-of-the-art accuracy was selected as a representative of its kind. Other similar approaches could also work as alternatives for this module in the framework. The applied algorithm, Stacked Hourglass Networks (Newell et al., 2016), is comprised of multiple hourglass-like network modules (Figure 2.3). Each hourglass-like network has a symmetric architecture that processes the image in a fine-to-coarse (high resolution to low resolution) and then coarse-to-fine fashion. Going through multiple such modules (e.g., 8) enables the model to learn the image appearance of body joints both locally and globally. It is claimed that this network architecture plays an essential role in realizing optimal performance. The other major source that contributes to the performance lies in the strategy of regularizing the model in the midway by adding a loss function between two consecutive hourglass networks that work as an intermediate supervision process. The direct output is a set of heatmaps, each showing the confidence distribution of an individual joint's presence regarding each pixel location. The visualized output connects each joint's location to form a skeleton, and the joint location is the pixel with maximum confidence value in the corresponding heatmap.

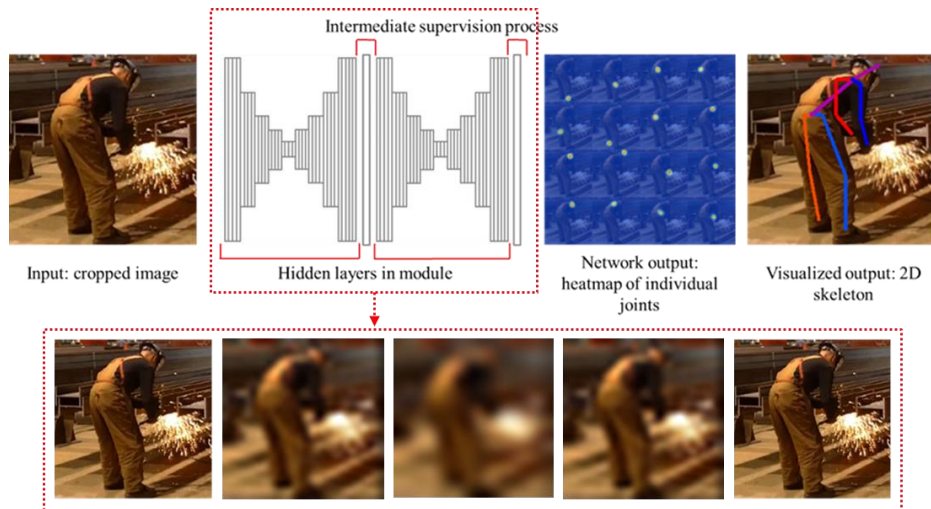


Figure 2.3 2D Human Pose Estimation

2.3.3 Optimization

Similar to Kobayashi et al. (2018), Liang et al. (2018), and Yang et al. (2005), this research addresses the long-duration occlusion handling issue by firstly separate the video to phases of before, during, and after occlusion. This is referred to as the “detection” phase. Subsequently, the “correction” phase is introduced to recover the occluded joints’ locations.

Detection Phase in Detection-and-Correction Scheme

Similar to Kobayashi et al. (2018), the detection phase is implemented by checking a series of rules. With a violation of any rule, the joint would be labeled as occluded/misdetected joint. To formulate this process in an algorithm, every joint in a frame is assigned with a “state.” Initialized as “reliable,” whenever considered to be occluded/misdetected the state would be updated to “unreliable” and would expect a correction in the later phase. Upon the completion of

the detection phase on all the joints throughout frames, the correction phase would be conducted and all the corrected joints' states would turn to "reliable" (Figure 2.4).

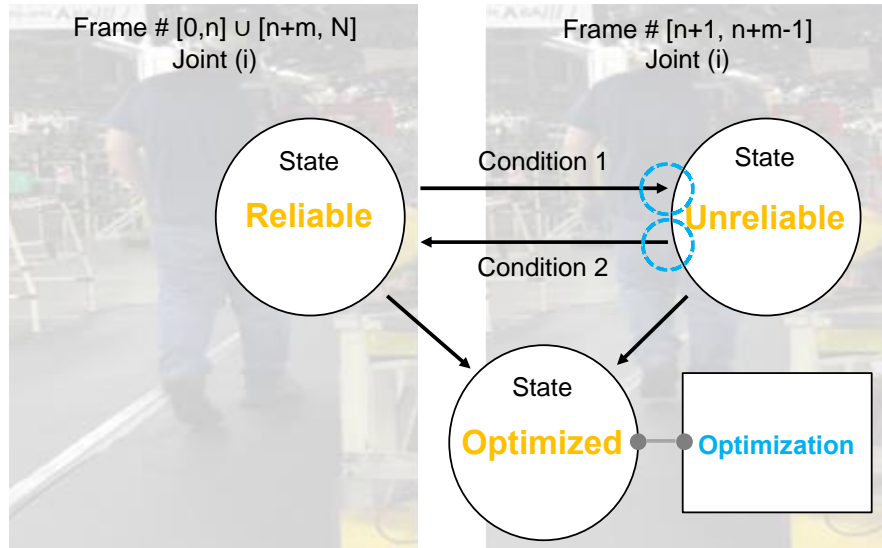


Figure 2.4 State Update Scheme for Optimization

The set of rules for occluded/misdetected joints detection are listed below, while sharing the first three with Kobayashi et al. (2018) and including several additional ones proposed originally on behalf of ergonomic risk assessment.

1) Confidence of joint location. With some preliminary analysis, high confidence of joint location estimation is strongly correlated with the fact if the joint is correctly detected. Different from Kobayashi et al. (2018), this rule is enforced more strictly that the high confidence should hold for a number of consecutive frames, instead of one. As being free of occlusion should not last longer than one frame, this modification is introduced to eliminate occasional success of joint location estimation.

2) Joints' translation between frames. The projected translation of a joint between consecutive frames should not be too large. Data-driven parameters are learned from the training dataset. Every joint has its specific range of movement so the parameters are joint-specific.

3) Bone length ratio maximum. Despite that bone length suffered from distortion when projected from 3D space to 2D images, the maximum bone length would not be unpredictable. In the field of ergonomics, anthropometric measurements were well-studied with accessible data. Such data could provide a quantity like maximum forearm length ratio regarding the subject's height/stature for the 95th percentile of the male population. An example of the data source is CAESAR anthropometric database.

4) Joint angle maximum. The maximum angle of some joints comes with a natural limit. For example, the elbow angle could not be less than $\sim 30^\circ$ even bent to its extreme. Such data could also be accessible in an anthropometric measurement dataset.

Correction Phase in Detection-and-Correction Scheme

The most relevant research (Kobayashi et al. 2018) only addresses the correction phase across three adjacent frames, assuming non-successive frames with the same occluded/misdetected joints. As this assumption would not hold in the target application, this research develops a novel method regarding the correction phase. Firstly, the two shoulder and two hip joints were considered as “root” joints, and would be corrected before optimizing other joints. They would then perform as foundations for localizing others.

According to the configuration of the occluded/misdetected joint and its adjacent ones, there are two configuration-conditioned scenarios: 1) end joint is occluded/misdetected (equivalently, with an “unreliable” state). This includes two scenarios if there is a connecting joint (e.g., elbow) between the root (e.g., shoulder) and the end joint (e.g., wrist), in terms of the connecting joint's state; and 2) end joint is well detected while the connecting joint was not. For the former, tree-structured optimization is deployed at the direction of the root, connecting, and

then end joint. However, as for the latter, this method would not be feasible if both the root and end joints' locations are known. In this case, inverse kinematics is introduced to estimate the connecting joint's location and solved by the Jacobian inverse technique.

As ergonomic evaluation focuses on the joint angle instead of its location, this research directly optimizes the joint angle, which is noticeably distinct from other works. Specifically, all the joints' location data was converted from cartesian coordinates to polar coordinates, expressed by bone length and joint angle following the kinematic model.

2.4 Field Testing

To test the feasibility of the proposed approach, field testing is conducted in a construction jobsite. The primary purpose of the testing is to examine whether the proposed method is applicable in a jobsite to handle the occlusion and generate robust human motion data for postural analysis in ergonomic risk assessment tools. To fulfill this objective, the joint angle is used for the validation metric to reflect the feasibility for automating assessment tools.

2.4.1 Testing Condition and Tasks

The testbed was sponsored by Power Construction, and 10 male construction workers participated in this study with written consent. Each construction worker performed a different task for several cycles. To have a balanced video length across subjects, roughly a 10 seconds video clip (~300 frames) for each subject was selected to validate the approach. All the subjects were asked to keep the personal protective equipment (PPE) and tools in order to reflect the usual appearance of body joints in the videos. Most tasks showed very typical challenges in terms of body joint localization, such as long-duration occlusion and potential interference from

adherent tools and materials. A sample of cropped video frames reflecting the jobs are shown in Figure 2.5.



Figure 2.5 Examples of Raw Video Frames for 2D Human Motion Capture

The ground truth was provided by manual annotation of all 16 body joints in every frame with a developed tool in MATLAB®. The human annotators were asked to click on the center of body joints with a visualized cursor, and the software recorded the pixel location of the click to obtain the optimal data quality.

2.4.2 Evaluation Metrics

This testing attempted to validate the human motion capture performance against the manually annotated body joints regarding both location and angle.

As for the evaluation metric for joints location, standard Percentage of Correct Keypoints with head size as acceptance threshold (PCKh) (Andriluka et al. 2014) is selected, which is widely accepted in the computer vision community. PCKh quantifies the percentage of body joints that is correctly located within a normalized distance of the ground truth location by a

fraction of the head size. For example, PCKh@0.5, which is used in this testing, indicates that half (0.5) of the head size is the allowed distance between the estimated location and the annotated one. Illustration about whether a joint is referred to as correctly detected is shown in Figure 2.6.



Figure 2.6 Illustration for Evaluation Metric: PCKh

Regarding the evaluation metric for joints angle, a typical frame-wise mean absolute error and its standard deviation were selected to reflect the feasibility to estimate joints angle.

2.4.3 Testing Results

Accuracy of Joint Location Estimation

Examples of the estimated 2D human skeleton are shown in Figure 2.8. As for the quantitative result, Table 2.2 shows the accuracy of body joints location estimation regarding individual body joint and task, expressed by the

evaluation metric PCKh@0.5. To summarize the accuracy statistics, body joints' locations generated from the proposed 2D human motion capture framework achieve an accuracy of 83.2% on average of the 10 tasks across all body joints. To

further analyze how occlusion affects the

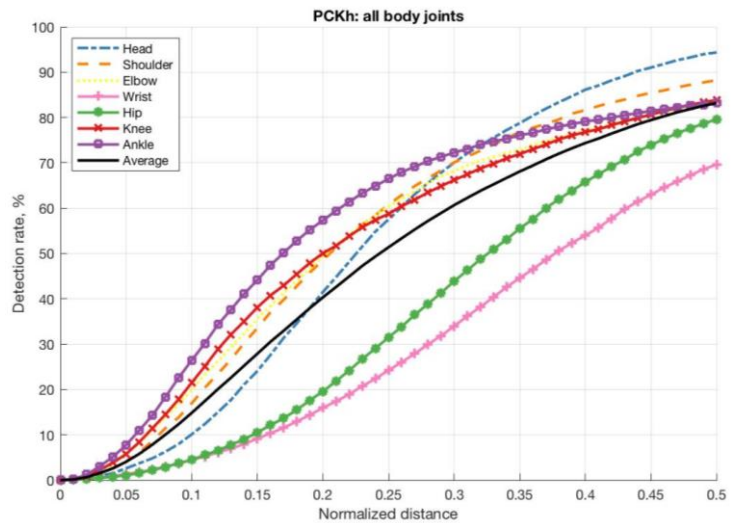


Figure 2.7 PCKh of Different Body Joints

accuracy, we excluded the occluded body joints. The accuracy of non-occluded body joints is 86.9% on average.

Table 2.1 Accuracy of 2D Joint Location Estimation

Task	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Carrying Lumber	96.0	82.2	79.8	57.7	91.1	91.2	92.9	84.4
Carrying Frame	99.8	99.5	70.3	57.1	86.8	73.9	78.1	80.8
Dumping Materials	94.7	89.3	87.4	66.7	81.9	71.3	81.2	81.8
Carrying Plywood	75.9	86.7	74.8	56.0	81.7	92.2	94.6	80.3
Stacking Lumber	95.0	69.5	95.5	85.6	83.6	64.1	81.9	82.2
Pushing Cart	100.0	97.2	92.0	85.1	86.2	95.8	96.4	93.3
Shoveling	100.0	97.5	100.0	95.7	95.1	95.7	100.0	97.7
Climbing Ladder	95.6	82.2	87.2	81.6	73.4	86.6	54.2	80.1
Hammering	100.0	90.7	92.9	82.8	85.1	99.4	98.4	92.7
Kneeling	98.2	98.2	69.9	53.7	34.7	71.3	59.7	69.4
All Joints	94.3	88.2	83.5	69.6	79.6	83.9	83.2	83.2
Visible Joints	94.2	96.0	94.2	78.0	78.1	85.9	85.6	86.9

To show the accuracy in terms of the normalized distance from the annotated joint location, the PCKh score is illustrated in Figure 2.7. The horizontal coordinate shows the PCKh normalized distance threshold (e.g., 0.5 indicates 0.5 x head size) while the vertical coordinate represents the proportion of joints that are predicted to lie within a corresponding threshold. Given one point along the curve, we can interpret the number of joints “correctly” identified (vertical coordinate) in terms of its normalized distance from the ground truth location (horizontal coordinate). In other words, the sooner the curve reaches the desired magnitude of the detection rate, the better its accuracy. Specifically, the resulting figure shows that ankle and head are best detected while the wrist and elbow had the lowest accuracy level.



Figure 2.8 Examples of Estimated 2D Human Pose

Accuracy of Joint Angle Estimation

Table 2.2 shows the accuracy of 2D joint angle estimation across tasks and body joints, expressed by the mean absolute error (MAE) and its standard deviation (STD.). To summarize the result, the average absolute difference between the vision-based approach and the manual annotation is 11.6° , with a standard deviation of 13.6° .

2.5 Discussion

The 2D joint location estimation accuracy is evaluated regarding all the joints first and then only visible joints that are free of occlusion. The accuracy of both scenarios is illustrated in Figure 2.9. In the way the data is visualized, the larger the “difference from visible joints” is, the less occluded body joints are corrected located. In the collected videos from the construction site, most of the jobs were captured from a side-view or diagonal-view. Head is defined as the top of one’s head so it is rarely occluded in the videos. All the rest body joints have left and right of its

kind, and are vulnerable to occlusion on one of either side. Considering the actual proportion of occluded body joints, the illustrated accuracy difference from all joints to visible/non-occluded body joints is less than or around 10% and is a small value. This small amount of difference suggested that the proposed method has a good ability to infer the occluded body joint locations.

Table 2.2 Accuracy of 2D Joint Angle Estimation

Task	Metric	Body Joint								Mean
		Left Shoulder	Right Shoulder	Left Elbow	Right Elbow	Back	Neck	Left Knee	Right Knee	
Carrying Lumber	MAE	8.0	9.2	12.4	14.5	3.3	9.3	7.3	7.9	9.0
	STD.	7.7	9.0	13.7	16.5	3.6	7.0	6.5	9.6	10.5
Carrying Frame	MAE	25.8	5.1	17.9	5.9	3.6	7.1	9.8	9.8	10.6
	STD.	31.8	3.5	21.7	4.7	3.0	4.4	11.6	7.8	16.3
Dumping Materials	MAE	8.0	5.9	4.7	9.3	5.3	6.8	20.2	9.5	8.7
	STD.	7.7	6.0	4.7	12.8	5.9	4.7	24.2	8.6	12.1
Carrying Plywood	MAE	5.7	32.8	5.9	44.8	3.8	15.4	5.1	10.6	15.5
	STD.	5.2	32.3	5.8	41.7	2.6	7.2	4.7	10.6	24.1
Stacking Lumber	MAE	22.2	30.8	12.0	11.1	17.9	19.7	14.7	15.5	18.0
	STD.	11.6	17.3	10.7	7.7	11.2	7.9	10.5	13.9	13.2
Cart Pushing	MAE	4.6	7.9	4.8	14.9	4.2	13.5	6.6	8.0	8.0
	STD.	3.4	6.3	5.4	10.8	3.4	6.1	5.1	8.1	7.5
Shoveling	MAE	18.5	17.4	6.2	5.6	21.1	8.3	4.2	3.5	10.6
	STD.	19.3	16.4	5.5	5.4	19.0	4.6	4.6	2.6	13.5
Ladder Climbing	MAE	12.6	5.8	14.3	5.9	4.2	5.2	19.0	16.1	10.4
	STD.	9.9	4.4	9.8	3.8	2.9	3.7	11.8	13.1	10.0
Hammering	MAE	23.6	16.5	12.2	5.6	19.7	7.2	7.8	4.9	12.2
	STD.	12.7	12.7	12.6	4.6	12.9	4.6	6.1	5.4	12.7
Kneeling	MAE	14.6	6.0	26.4	5.6	5.4	12.8	16.1	13.5	13.2
	STD.	11.7	4.5	13.9	13.9	5.2	7.1	14.8	11.0	15.7
All	MAE	14.4	13.8	11.7	14.0	8.8	10.5	11.1	10.0	11.6
	STD.	12.1	11.3	11.4	14.7	7.0	5.7	10.0	9.1	13.6

Performance comparison is conducted between the proposed method on the collected image set from a construction site and the state-of-the-art 2D human pose estimation method (Zhang et al. 2019) on the benchmark MPII Human Pose Dataset built by daily activity images (Figure 2.10). The average difference between these two scenarios is within 10% and is considered to be promising regarding the algorithm’s ability of generalization. Among all the

body joints concerned, the wrist has the largest discrepancy of almost 20%. This aligns well with the aforementioned hypothesis that construction jobs are expected to have huge bias and variation in hands' appearance due to continuous interaction with tools and materials. From the testing result, it is observed that for the three carrying tasks (i.e., carrying lumber, scaffold, and plywood) have their wrist localization accuracy consistently below 60%. The impact imposed by the contacting materials seems to be apparent. Similarly, the second to the largest such difference is related to the hip joint. Based on the observation on the visualized 2D skeleton on the test images, it is found that the detected hip joint positions are usually around the subjects' waist instead of the hips. It is reasonable if the manual annotation in MPII dataset regarding hip location has a bias from how that in collected dataset of construction frames are annotated. Also, the impact by being attributed to the fact the construction workers usually wear a tool belt on their waist, and the hanging tools may affect the appearance around the hips.

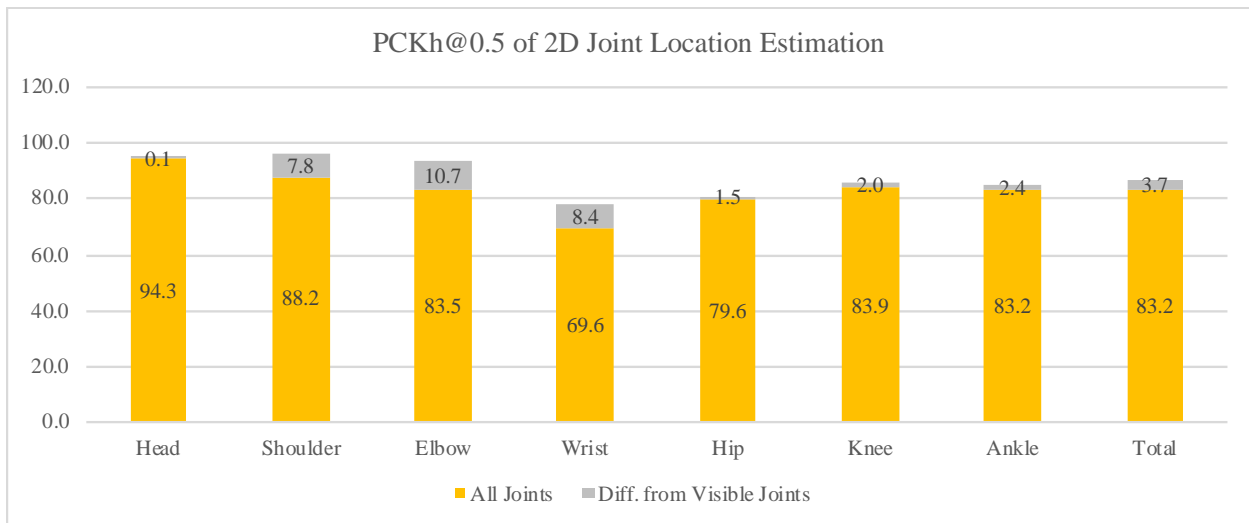


Figure 2.9 Accuracy for 2D Body Joint Location: All Joints VS. Visible Joints

Apart from the joint location, the angle estimation accuracy is also evaluated. Figure 2.11 shows the overall accuracy of 2D joint angle estimation error across the jobs regarding individual body joint. Considering the subjects have continuous movement, the angle estimation accuracy shows a promising performance for the on-site application. As the ground truth data is provided by manual annotation of joint location expressed in image pixel, the proposed method has a comparable estimation with human observation. Among the body joints, back has the highest accuracy of 8.8° difference. As the back angle is determined by mid-point of shoulders and mid-point of hips and that shoulder has a very high (88.2%) localization accuracy, the seemingly low localization accuracy of the hip does not have a large impact on the estimation of back angle.

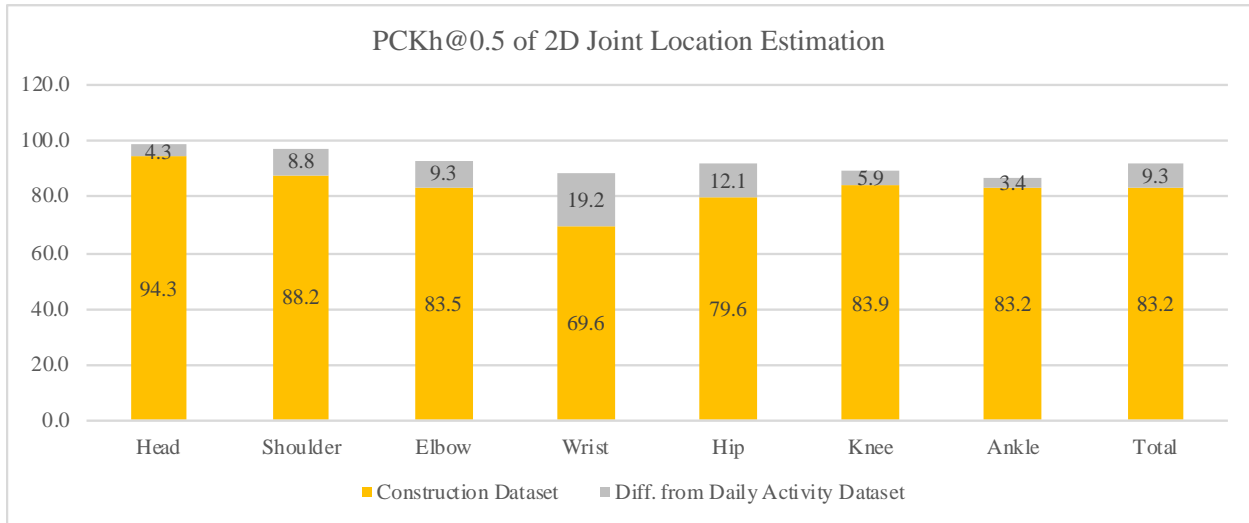


Figure 2.10 Accuracy of 2D Body Joint Location: Construction Dataset VS. Daily Activity Dataset

In the field test, a major failure case appears in knee angle estimation. As shown in Table 2.2, the two jobs involving kneeling postures (i.e., ladder climbing and kneeling) have large angle estimation error (i.e., 16-19 degrees) given little occlusion. This result correlates well with the joint location accuracy as shown in Table 2.1, that these two jobs have the lowest accuracy regarding the location of three joints (i.e., hip, knee, and ankle) that determine the knee angle. It suggests that the widely used 2D human pose datasets (e.g., MPII) collected mainly from daily

activities may involve insufficient kneeling or similar workplace postures. Supplementary data collection and model fine-tuning with domain-specific context may enhance the framework's capability in motion capture.

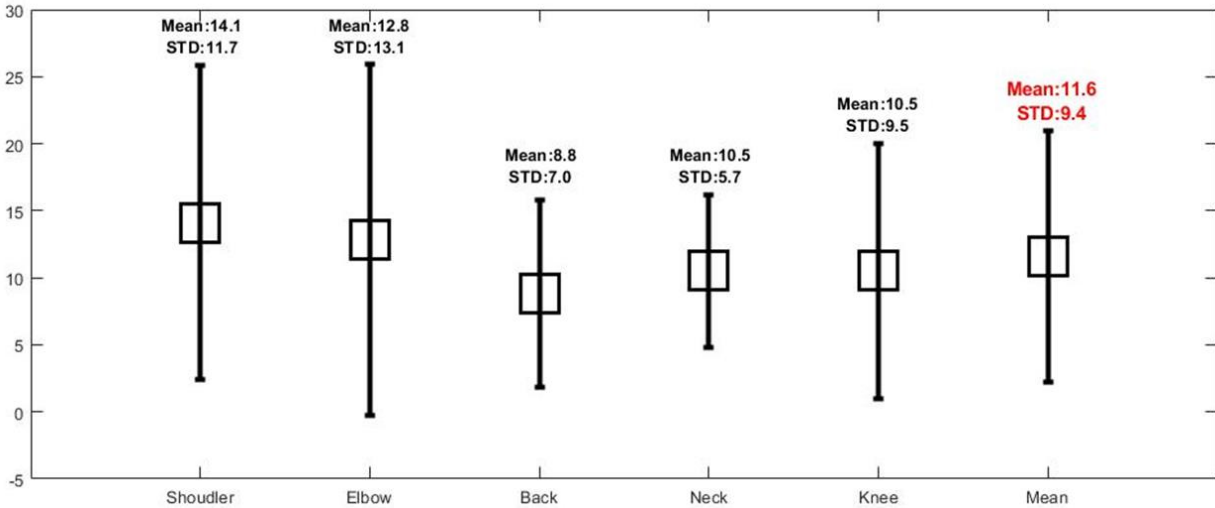


Figure 2.11 Accuracy of 2D Joint Angle Estimation Regarding Body Joints

2.6 Conclusion

In this chapter, this study proposed a video-based 2D human motion capture framework to estimate primary risk factors (i.e., posture and repetition) in ergonomic assessment tools. To address the knowledge gap that enforcing kinematic and temporal consistency across video frames under long-lasting occlusion is the key to enable accurate posture and repetition (frequency and duration) estimation, a convolutional neural network-based 2D human motion capture framework is proposed. The framework mainly consists of a frame-wise 2D human pose estimation module that applies the state-of-the-art algorithm and a novel optimization module that detects and correct occluded/misdetected joint location based on kinematic and temporal constraints. The pose estimation module is replaceable to incorporate continuous advancement in the computer vision community. The optimization module leverages the confidence level of pose

estimation and anthropometry data based on an extensive statistical analysis of human motion in the workplace. In addition, the joint location correction step in optimization is innovatively designed for ergonomic risk assessment that enforces a linear change of joint angle and projected bone length, instead of the joint location that most studies focus on for a general purpose.

To test the feasibility of the proposed approach, a field test was conducted with ten male construction workers who have different anthropometry. From the testing results, it was found that the proposed framework can provide robust body joint localization with 83.2% of accuracy in PCKh@0.5, and joint angle estimation with 11.6° of error compared with manual annotation. This result indicates the potential of the proposed framework to enforce temporal smoothness across frames regarding body joint angles and viable to estimate posture and repetition for ergonomic risk assessment.

From the test, issues are also identified as potential study in the future. By comparing with the performance of the state-of-the-art algorithm on benchmark dataset with images collected from daily activity, it is found that wrist localization has significantly lower accuracy of 19.2%. It may indicate the apparent bias and variation of hand appearance of construction jobs impose a noteworthy impact on the pose estimation's ability of generalization regarding wrist. As the human pose estimation performance largely depends on the training dataset, the low accuracy of wrist localization suggests the need for extensive training images collected from a construction site, especially including a large variation on hands appearance. It is also found that even the hip has low accuracy in localization, the issue may not be significant as the estimation accuracy of back angle, which directly relies on hip location, is very high (8.8° average error). By examining the estimated hip location visualized on images, it is found that the hip location is close to that of the waist. It indicates that the hip location has a large discrepancy in its vertical

coordinate but trivial in horizontal coordinate. For the purpose of ergonomic risk assessment, this issue may be ignored.

Despite that the framework needs to be further tested on a larger scale and to address remaining limitations such as training with the additional construction-focused dataset and enhancing occlusion handling ability, the proposed framework has great potential for on-site risk factor estimation in ergonomic risk assessment tools.

Chapter 3 Video-Based 3D Human Motion Capture for Posture and Repetition Estimation

3.1 Introduction

The preceding chapter introduced a 2D human motion capture framework to estimate posture and repetition data, which are primary risk factors for ergonomic risk assessment tools such as REBA, RULA, and OWAS. An estimated 2D joint angle on images is projected from 3D space and is sufficient for many assessment tools that either only need a rough posture category instead of numerical angle value, or a projected angle on the sagittal plane that can align with image plane given a good camera angle. There are, however, tools that require specific values of angles that cannot be approximated by being projected onto the image plane, such as the twisting angles of the trunk. NIOSH lifting equation (Waters et al. 1994) is one of the examples where such an angle (i.e., “asymmetric angle”) is needed to evaluate the ergonomic risk of lifting tasks. Some assessment tools further require the distinction between abduction/adduction and flexion/extension, regarding the limb’s pose that also cannot be addressed by 2D projected angles. 2D human motion capture does not have the capacity to handle such scenarios.

Apart from specific types of angles, tools that require distance-related measurements also find 2D human motion inapplicable. Such measurements include, but are not limited to, hands’ distance from the body or ground. Specifically, NIOSH lifting equation requires the hands’ horizontal distance from the ankles, and vertical distance from the ground, and vertical lifting

distance (Waters et al. 1994). Snook tables require hands' horizontal distance from the front of the body and vertical lifting distance (Snook and Ciriello 1991).

Beyond postural analysis that evaluates an overall risk level of jobs, biomechanical analysis is favored to estimate individual internal joints' exerted loads when joint location and external force exertion data are accessible. Among the spectrum of biomechanical models, some are static models (Chaffin and Baker 1970; Garg and Chaffin 1975; Martin and Chaffin 2007) that only require body joint location for static postures. The limitation of these models is the ignorance of inertial loads exerted on body parts due to dynamic postures. Dynamic biomechanical models (Marras and Sommerich 1991), can handle this scenario by utilizing joints' velocity and acceleration. As the velocity and acceleration require true-to-scale (e.g., in m/s), measurement of kinetics data, thus necessitates the motion data in 3D space.

The invasiveness and high cost of 3D human motion capture system are the major obstacles to collect such data in the workplace. To address these issues, some studies applied human motion modeling software, which provided a human model (Li et al. 2017, 2019). With relatively light manual effort required to manipulate the human model, it animates the expected motion under realistic movement constraints.

To eliminate the human modeling effort, automated motion capture approaches with economical devices were explored. They mainly focused on image sensors such as RGB-D sensor, stereovision camera, and multiple cameras. The performance evaluation comparing these methods were also conducted (Seo et al. 2017). Given the capability of the existing motion capture approaches, Seo (2016) investigated the expected motion data quality level to automate biomechanical analysis on lifting tasks. It was concluded that 10° of error in joint angle estimation is the maximum to obtain reliable biomechanical analysis result. Among the three

approaches under comparison, RGB-D sensor achieved the joint angle estimation accuracy of less than 10° error, while stereovision camera and multiple camera approach did not. Given RGB-D sensor was not applicable in an outdoor environment, which is a normal condition in construction, there is still a need to develop an economical method to capture quality human motion data for biomechanical analysis.

In summary, to evaluate the risk factor of posture that relies on more detailed motion data (e.g., twisting, flexion/extension, abduction/adduction) or biomechanical analysis that need true-to-scale 3D joint location and even kinetics data such as velocity and acceleration, a 3D human motion capture approach is needed. Specifically, the expected approach should achieve a joint angle estimation accuracy of within 10° error while reserving temporal smoothness for repetition estimation.

3.2 Literature Review

Among the major non-invasive imaging sensors applicable for outdoor deployment, stereovision camera and multiple camera approaches share a common limitation of relying on 3D reconstruction from multiple lenses. This process makes it sensitive to the variation of illumination condition and occlusions. With the remarkable progress of 3D human pose estimation from a monocular image made in computer vision domain recently, this study aims to apply a 3D human pose estimation method for still images, and integrate it with the proposed two-stage detection-and-correction scheme with needed modification to achieve 3D human motion capture on video sequences captured from construction sites.

Frame-wise human pose estimation

3D human pose estimation from the monocular/single 2D image is a severely ill-posed problem because a projected joint location can represent infinite possible locations in a 3D space (Sarafianos et al. 2016). It is also ill-conditioned as a small error in projected joint location estimation, which would result in a much greater error in a 3D space (Sarafianos et al. 2016). However, as training such a pose estimator requires a large amount of ground truth data, including 3D human motion and corresponding images. Compared with 2D human pose estimation that only requires manual annotation of the projected joint location on the images, 3D human pose estimation is much costlier. To achieve advanced performance, a noteworthy amount of research was conducted focusing on training an effective and generalizable pose estimator with a minimum amount of data or data collection effort.

Rogez and Schmid (2016) augmented the training dataset by developing an image-based synthesis engine. With identical human motion data, multiple synthetic images were generated with different appearances and backgrounds. Chen et al. (2016) augmented the training dataset by generating avatars with 3D human motion. By changing the appearance (e.g., clothing) of avatars and generating 2D images from diverse view angles, the model was more generalizable to different appearance and camera view. As the 3D human pose datasets do not include external occlusion (contrary to occlusion of one's other body parts), some studies also augment the dataset by synthesizing noises, e.g., image patch to partially occlude the subject (Sárándi et al. 2018).

Besides training dataset augmentation, another idea was to utilize dual-source training data. One is an image captured from lab condition with accurate 3D human motion data from a multiple camera system. Examples of such datasets are CMU Motion Capture Dataset and

Human3.6M Dataset (Ionescu et al. 2014). The other source is 2D images taken in the wild with the manual annotated 2D human pose, such as MPII Human Pose Dataset. Due to the much higher cost of collecting 3D pose, annotated 2D pose dataset comes on a much larger scale. Yasin et al. (2015) learned the mapping between 3D poses and its 2D projection in the training phase. In the testing phase, 2D pose would be estimated first with the model trained from the massive annotated dataset. The most similar 3D pose was then retrieved with the learned mapping between 2D and 3D poses, and it was refined to obtain the final pose. Zhou et al. (2017) proposed a framework that imposes weakly supervision on a 2D annotated pose by only enforcing the bone length symmetry of the left and right body parts, and 2D joints' projection loss.

Detection of occluded/misdetected joints

There were many constraints explored in the 3D human pose estimation area. Many of them were carefully reviewed and considered for application in this study, while some of them could be modified to better fit into the target application.

1) Confidence of joint location's 2D projection. 3D human pose estimation is usually regressed from the estimated 2D location from the image. It only comes with a confidence level for the location's 2D projection, rather than that of 3D space. Simo-Serra et al. (2012) used the confidence of joint location's 2D projection for evaluating the quality of the estimated 3D pose.

2) Bone length constancy. As 3D human motion capture estimates the joint's location in 3D space, bone length constancy was widely utilized to find the occluded/misdetected joint, and also to enforce a reasonable correction process to recover problematic initial estimation (Gupta et al. 2008; Ramakrishna et al. 2012). As it was found to be intractable to enforce length constancy

on each individual bone (Ramakrishna et al. 2012), studies generally only constrain a constant sum of bone length for all parts.

3) Bone length ratio. It was constrained according to anthropometry data. As the estimated 3D human pose came from a monocular image lacking scaling information, the individual's stature was not considered, and the bone length proportions were only counted as identical across individuals (Wang et al. 2014). This constraint could be modified to fit for ergonomic evaluation by incorporating the individual's stature and specific anthropometry data from authorized datasets.

4) Joint angle maximum. Radwan et al. (2013) identified the possible range of motion from the training dataset and enforced joint angles to be smaller than the maximum.

Many other ideas were also explored. For example, the appearance of symmetric body parts (e.g., left and right wrists) was constrained to be similar (Gupta et al. 2008). This was not considered feasible to apply in the construction site as workers frequently interact with different tools or materials and usually do not have a similar appearance on symmetric body joints.

Correction of occluded/misdetected joints

Following the detection of occluded/misdetected joints, correction of the problematic joint location is naturally different from that for 2D human motion, as the increase of the human pose's dimensionality.

In 2D space, the orientation of a limb can be parameterized by one angle (e.g., the angle between the vector and one of the axes). Thus, if the orientation of a limb is known at the start and end positions, while given the condition of linear joint angle interpolation, its orientations in

the intervening frames are straightforward to obtain. However, it becomes challenging in 3D space as it requires more than one angle to define a limb's orientation.

There are three systems to parameterize the orientation, in the area of computer graphics: 1) Euler angles; 2) quaternions; and 3) exponential map (Du et al. 2016). For human pose modeling, Euler angles are widely applied to represent a joint orientation, but not suitable to calculate the intervening states between the start and end states. The reasons include its gimbal lock and singularities issues. Quaternions have a long history of being applied for representing the rotation of joint (Du et al. 2016; Lu and Dai 2018; Urtasun and Fua 2004), and the unit quaternions are free of gimbal lock, thus in this study, the quaternions were selected and normalized to achieve this property. An exponential map is essentially a re-parameterization of quaternions, so it has similar properties with quaternions.

3.3 Method

3.3.1 Frame-wise 3D Human Pose Estimation

To maximize the pose estimator's ability to generalize with minimum training data or dataset augment effort, the major ideas could be summarized as: 1) synthesize training data, mainly 2D images with the different appearance or introducing occlusion; 2) utilize 2D training data with only manually annotated 2D poses. The second approach was adopted in this study due to the corresponding major update of the network architecture, while the first approach could be adopted upon this basis.

Among the studies utilizing the second approach, the weakly supervised method (Zhou et al. 2017) was selected as the frame-wise 3D human pose estimation module. This method trained the pose estimation network from both data sources of Human3.6M (2D image with 3D motion

data from lab setting) and MPII (2D image with 2D motion data “in the wild”). The network imposed strong supervision for 3D motion data source by directly enforcing the estimated 3D joints location to be the same as that of ground truth motion data. In contrast, weak supervision was applied for 2D motion data source by only enforcing bone length symmetry between left and corresponding right parts as no 3D ground truth motion data was available. It was also selected because the pose estimation network was built upon the 2D pose estimator (Newell et al. 2016) this study applied in Chapter 2. Because of its demonstrated feasibility to work for the target application, this study inherited the network architecture to expect a similar level of performance.

3.3.2 Optimization

Optimization module of the proposed 3D human motion capture framework was inherited from the 2D motion capture framework and also consists of two phases that form the detection-and-correction scheme.

Detection Phase in Detection-and-Correction Scheme

To identify the before, during, and after occlusion frame segment, most of the rules applied in the proposed 2D motion capture framework was adopted in the 3D framework. The major change was adding bone length consistency as an additional constraint. The inherited rules were modified accordingly as follows:

- 1) Confidence of joint location’s 2D projection. Low confidence of joint location’s 2D projection suggests a high probability of occlusion or misdetection. A misdetected joint naturally

expects re-evaluation of location in 3D space. An occluded joint may be estimated correctly regarding location, but deserves further evaluation based on other constraints.

2) Bone length constancy. While it was found to be intractable to enforce length constancy on each individual bone when estimating the joints location in 3D space (Ramakrishna et al. 2012), it is not necessary to enforce this rule strictly to obtain a feasible level of accuracy to estimate posture and repetition. Thus this study only constrained the bone length to lie within $\pm 10\%$ of the median of the values throughout the frame sequence. Mean value was not used to prevent the impact of outliers.

3) Bone length ratio.

4) Joint angle maximum.

5) Joint location translation between frames.

The threshold values applied for 3) – 5) are either from anthropometric measurement datasets (e.g., CAESAR) or calculated from the training dataset.

Correction Phase in Detection-and-Correction Scheme

Given the correctly detected poses at the start and end frame of the occluded/misdetected segment, the intervening poses were recovered by spherical linear interpolation, abbreviated as Slerp, represented by a unit quaternion. Quaternion is expressed by a four-element vector as $q = [w, x, y, z]$. Before interpolation, the first step is to convert the 3D joint angle, or axis angle, to quaternion. The 3D joint angle is the angle value between the vector of two body parts' axes, v_1 and v_2 , in 3D space. Supposing the axis angle is α , then the conversion equation to quaternion can be expressed by:

$$w = \cos\left(\frac{\alpha}{2}\right)$$

$$[x, y, z] = \frac{v_1 \times v_2}{|v_1 \times v_2|}$$

Two unit quaternions q_0 and q_1 would be obtained at the start and end positions of the interpolation. To calculate the quaternions of in-between frames, the “distance” θ between the quaternions of the start and end positions was first derived by:

$$\theta = \arccos \frac{\text{dot}(q_0, q_1)}{|q_0||q_1|}$$

Supposing t is the time-frame index, and $t \in [0, 1]$, the quaternion of an intervening frame can be calculated as:

$$q(t) = \text{Slerp}(t, q_0, q_1) = \frac{q_0 \sin((1-t)\theta) + q_1 \sin(t\theta)}{\sin(\theta)}$$

Quaternions were then converted to the joint location of all intervening frames, given the parent joint (adjacent joint closer to the body center) location expressed by Cartesian coordinates $[x, y, z]$:

$$\text{joint location}_{\text{child}} = q \times \text{joint location}_{\text{parent}} \times \text{conjugate}(q)$$

3.4 Laboratory Testing

To test the feasibility of the proposed approach, laboratory testing was conducted in this study. To the best of my knowledge, there is no validated motion capture system that can provide ground truth for 3D human motion in a construction jobsite. As aforementioned in the introductory section of this chapter, Seo (2016) suggested the expected quality of 3D motion data captured in the lab condition is $\pm 10^\circ$ in joint angle estimation. This study aimed to test if the proposed approach could achieve the expected level of accuracy.

3.4.1 Testing Conditions

For this test, pushing task was selected for estimating the posture. The cart to be pushed was very heavy that it takes around 80 N initial force to move the cart. This heavy pushing naturally made the subject to involve dynamicity all major body joint angles, including shoulder, elbow, back, knee, and neck.

10 subjects participated in this test to provide a diversity of stature and the joint angles' configuration. The subjects' height ranges from 160 to 190 cm. As they pushed the same cart, subjects had to drive their body parts differently, by presenting with different joint angles, to perform the pushing with identical hand height. The range of motion (i.e., range of joint angle) for primary body joints are: 0° - 120° for shoulders, 0° - 120° for elbows, 0° - 90° for back, 0° - 60° for neck, and 0° - 90° for knees.

To demonstrate the advantage of 3D human motion capture regarding estimating joint angle at a non-sideview perspective, the smartphone camera was located at a diagonal view to the right side of the subject. To provide the ground truth for the estimated motion data, optical marker-based motion capture system (OptoTrakTM, Northern Digital, Inc., Waterloo, Canada) was set up for this test. The system provides the 3D location of each body-attached marker, and the joint angle was calculated from the vector connecting two adjacent markers. As the markers could only be attached at the joint's surface and the vision-based approach estimates the joint center location, the adjacent markers' locations were carefully adjusted to best align with the limb's rotational axis. The testing layout and an example frame are illustrated in Figure 3.1.

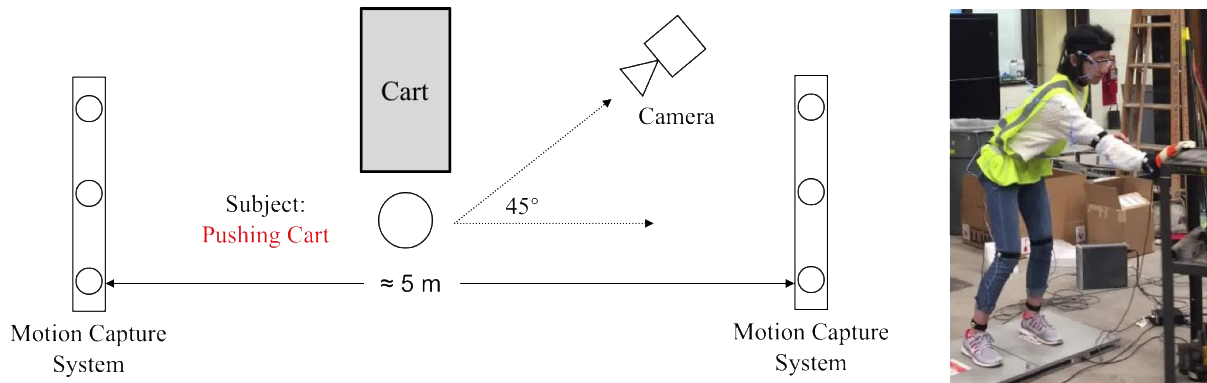


Figure 3.1 Testing Layout of 3D Human Motion Capture

3.4.2 Evaluation Metric

The proposed vision-based approach generates the 3D human skeleton in the camera coordinate system. The location of a body joint is expressed in the unit of the pixel. The skeleton constructed by the motion capture system's marker locations is expressed in the unit of millimeter. The discrepancy between the coordinate system of two approaches makes it challenging and unfair to compare the joints location estimation without any assumption. As ergonomic evaluation does not require absolute joint locations, this study only validates the joint angle estimation. As for some ergonomic evaluation tools require, relative joint locations (e.g., horizontal distance between hand and body) will be evaluated in Chapter 4.

To evaluate the joint angle estimation accuracy, frame-wise mean absolute error (MAE) and its standard deviation (STD.) was used, which is the same measurement with that of the relevant study (Seo 2016) suggesting the expected accuracy.

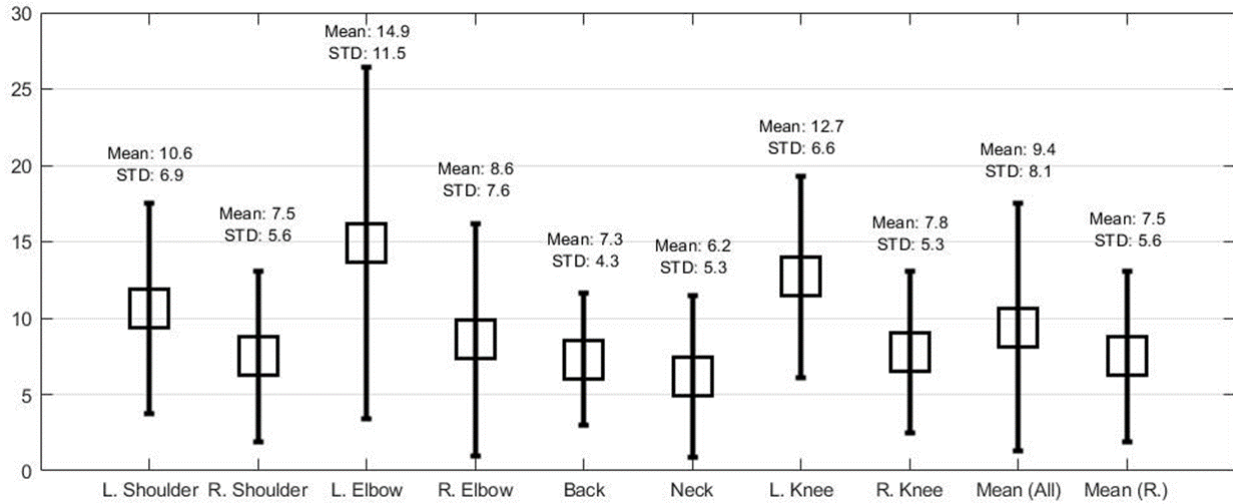


Figure 3.2 Accuracy of 3D Joint Angle Estimation

Table 3.1 Accuracy of 3D Joint Angle Estimation

Subject ID	Metric	Joint Angle Estimation Error (Unit: degree)									
		Left Shoulder	Right Shoulder	Left Elbow	Right Elbow	Back	Neck	Left Knee	Right Knee	Mean (all joints)	Mean (visible/right)
1	MAE	12.1	8.9	15.9	7.7	6.0	4.4	23.1	8.4	10.9	7.1
	STD.	7.7	8.0	9.9	7.0	3.4	2.6	7.8	5.6	9.0	5.3
2	MAE	7.1	5.7	10.4	10.9	7.3	7.6	10.5	6.5	8.1	7.6
	STD.	5.5	4.0	7.6	12.7	5.0	6.2	7.1	4.8	7.4	6.5
3	MAE	8.1	7.6	10.9	6.3	8.8	4.6	10.3	8.7	8.2	7.2
	STD.	5.8	6.3	8.5	4.9	3.9	5.8	6.0	5.2	5.8	5.2
4	MAE	8.8	8.1	18.2	11.3	6.2	5.5	10.0	6.8	9.7	7.6
	STD.	7.0	8.6	15.0	8.0	4.6	4.0	7.4	3.4	9.3	5.7
5	MAE	8.0	6.2	14.9	7.5	6.8	8.0	11.0	11.9	9.2	8.1
	STD.	5.4	3.7	12.7	7.1	3.7	5.7	6.9	7.5	7.7	5.5
6	MAE	8.8	5.3	17.4	9.3	13.8	8.6	18.9	9.0	11.0	9.2
	STD.	5.5	4.5	10.9	6.6	8.2	6.9	8.6	6.2	8.6	6.5
7	MAE	14.0	4.5	17.8	6.7	6.6	6.9	18.5	4.6	9.5	5.9
	STD.	7.7	3.2	14.9	5.7	3.8	5.7	5.9	2.8	8.9	4.2
8	MAE	12.2	12.6	16.2	12.2	1.9	4.5	6.5	6.5	9.9	7.5
	STD.	9.8	6.5	15.6	10.9	1.5	3.9	3.9	3.2	9.5	5.2
9	MAE	10.7	8.2	12.1	7.5	6.0	6.7	12.8	11.9	9.3	8.1
	STD.	6.7	7.0	9.8	7.7	4.7	7.7	8.3	10.7	8.0	7.6
10	MAE	15.7	7.6	15.0	6.5	9.6	4.8	5.7	3.8	8.3	6.5
	STD.	8.2	4.1	10.2	5.5	4.7	4.2	4.0	3.1	7.1	4.3
All	MAE	10.6	7.5	14.9	8.6	7.3	6.2	12.7	7.8	9.4	7.5
	STD.	6.9	5.6	11.5	7.6	4.3	5.3	6.6	5.3	8.1	5.6

3.4.3 Testing Result

To have a quantitative analysis of the proposed approach's accuracy on 3D joint angle estimation, the absolute error of estimated angles regarding each of the 10 subjects and 8 joints are listed in Table 3.1. Including all the subjects and body joints, the overall angle estimation error is 9.3° with a standard deviation of 8.0° . To exclude the impact of occluded body joints, the accuracy of visible joint angles is calculated additionally. With the exclusion, the overall angle error is reduced to 8.1° with a standard deviation of 7.6° . The body joints are specified as the left or right side of its type to show the discrepancy of accuracy between the visible side (right) and the partially occluded side (left).

3.5 Discussion

The mean absolute error of each body joint's angle estimation is summarized and visualized in Figure 3.2. Given the camera is placed at a diagonal-view from a subject's right, the right side of body parts are generally visible while the left side suffers a higher level of occlusion. This is shown consistently in the angle estimation error between within the left and right joints of a kind (e.g., left and right shoulders). For example, left shoulder commonly has a higher angle error than right shoulder, and this pattern is showing in the tests of almost every subject. In terms of body joint, back and neck have the lowest estimation error. This is reasonable as the related body parts are not on the limbs that are susceptible to occlusion and dynamic movements with variations. In contrast, the angles related to limbs have much larger estimation errors. Among the three types of such angles (i.e., shoulder, elbow and knee), the elbow has the largest error, followed by knee and subsequently the shoulder. Elbow angle directly relates to the three active joints: shoulder, elbow, and wrist. Thus its accuracy is vulnerable to any of these error sources.

Similarly, knee angle is a counterpart of the elbow in the lower limbs. It has a smaller error in this test, may because the cart pushing task has a more rapid and larger range of movements on upper limbs than lower limbs. Shoulder angle is defined by two vectors: arm vector calculated from elbow and shoulder, and back vector determined by both shoulders and hips. Apparently, the back vector is rather stable compared with limbs. Thus the error source of shoulder angle is less than that of the elbow and knee.

3.6 Conclusion

In this chapter, the proposed motion capture framework is enhanced with modification to capture 3D human motion data from a video to estimate primary risk factors (i.e., posture and repetition) in ergonomic assessment tools. Apart from the addressed knowledge gaps in the preceding chapter, this study focuses on applying a 3D human pose estimation algorithm that has a strong ability of generalization to handle the cluttered condition in construction sites. Additionally, necessary modifications are needed to address the 3D motion data's higher dimensionality and intrinsic constraints (i.e., bone length constancy) compared with the 2D framework. To address these issues, a weakly supervised approach was applied, which can utilize both lab-based 3D training data and in-the-wild 2D data. It demonstrated a strong ability of generalization to capture well frame-wise 3D human pose in the simulated lab environment which has a cluttered background and frequent interaction with and occlusion by the contacting cart. The optimization module leverages the bone length constancy constraint for the 3D human model. The higher dimensionality issue is addressed by developing a modified body joint trajectory interpolation process to enforce the linear incremental change of joint angle across frames, which is designed for ergonomic risk assessment.

To test the feasibility of the proposed approach, a lab test was conducted with ten subjects with different anthropometry. From the testing results, it was found that the proposed framework can provide robust body joint angle estimation with 9.4° of error compared with angles calculated from marker-based motion capture system (OptoTrak). This result indicates the potential of the proposed framework to enforce temporal smoothness across frames regarding body joint angles and viable to estimate posture and repetition for ergonomic risk assessment.

From the test, issues are also identified for potential study in the future. The accuracy discrepancy between the visible and occluded joint is large, which indicates a limited ability to infer the occluded joint's location in 3D space. It is a reasonable issue given the extra dimension to infer and the limitation on the training dataset's scale. Further research is expected to address it. Additionally, body joints with a rapid and large range of motion are vulnerable to angle estimation error. It suggests a need for further study on the joint-conditioned body part movement with solid analysis to modify the proposed framework.

Despite that the framework needs to be further tested on a larger scale dataset, regarding joint localization accuracy, and ideally a real construction site, and to address remaining limitations such as training with additional construction-focused dataset and enhancing occlusion handling ability, the proposed framework has great potential for on-site risk factor estimation in ergonomic risk assessment tools.

Chapter 4 Applications of Video-Based Human Motion Capture on Ergonomic Postural Analysis

4.1 Introduction

The preceding chapters demonstrated the feasibility of the proposed method in estimating frame-wise body joint angles for posture estimation in ergonomic risk assessment. However, it remains unclear how the methods perform in estimating the risk factors required by various tools (e.g., posture's occurrence, frequency, and duration) as well as how the performance would affect the final ergonomic risk level calculation. To answer this question, I select 3 of the most popular postural analysis tools for ergonomic risk assessment: 1) REBA (Rapid Entire Body Assessment); 2) Snook's Tables and 3) NIOSH Lifting Equation.

REBA is selected because it focuses on estimating an overall full-body risk level (1-15) with major posture input of joint angle values, and the frequency and duration of the postures' occurrence. It is referred to as an angle-based postural analysis tool. This study showed the feasibility of the proposed 2D motion capture in estimating these input data directly for calculating the risk level, by comparing its performance with 27 experienced human observers.

Some tools rely on distance-based measurements rather than joint angles, such as Snook's Tables requiring horizontal and vertical distance between the hands and the front of the body or the ground. The horizontal or vertical distance is different from the 3D distance, which is rather a distance either projected to the ground plan (horizontal distance) or projected to the normal vector of the ground plane (vertical distance). For example, "hand distance" is, in fact, a

horizontal distance that is projected to the ground level from the 3D distance between the mid-point of the hands and the front of the body.

Lastly, NIOSH Lifting Equation requires a set of distance-based measurements similar to that of Snook's Tables. Additionally, it also requires an angle-based measurement, asymmetric angle representing the trunk's twisting angle. As such, it is a mixture of distance- and angle-based measurements.

As the distance-based and asymmetric angle require 3D human motion data to estimate, the proposed 3D motion capture method is applied to Snook's Tables and NIOSH Lifting Equation. To provide a baseline of risk factors estimation, a tapeline measurement of distance-based variables and observed asymmetric angle are used to evaluate the accuracy of proposed video-based motion capture for these representative ergonomic risk assessment tools. As the measurements do not require full-body joint angle that requires a complex motion capture system to provide ground truth data, this testing was able to be conducted on a real construction job-site with construction workers' material handling tasks.

4.2 REBA (Rapid Entire Body Assessment)

4.1.1 REBA

Among the postural analysis tools for ergonomic evaluation, REBA has been widely applied since it focuses on full-body assessment and provides a dense 15-level scoring for the job. Many tools do not provide such a dense overall scoring mechanism because the input posture data needs to be detailed enough to generate a wide spectrum of combinations. Tools usually only require qualitative posture data input, (e.g., whether back-bending happens) as it is time-consuming to collect quantitative data, (e.g., whether the back-bending angle lies between 20° -

60°). The proposed approach was aimed to provide such detailed data to facilitate tools like REBA for a detailed understanding of a job's ergonomic risk level.

4.1.2 Method

The proposed 2D human motion capture method provides frame-wise joints' location in image pixel. The joint angle can be calculated from the vector formed by the adjacent joints' locations. This approach has the advantage of being widely adapted to various joint-angle based postural analysis tools. On the other hand, its performance largely correlates to the camera's view angle to mitigate the distortion effect brought to the joint angle projected from 3D space onto the 2D image.

If the postural analysis tool is pre-defined and only requires categorical posture data as input, such as REBA, an additive module could be suggested to mitigate the distortion effect. The proposed module is a supervised machine learning algorithm called K-Nearest Neighbor. This module takes in a vector of numerical values as "feature," such as all the joint angles of the human model in a video frame as used in this study. The training data is Human3.6M that includes 2D images of subjects performing various postures and provides 2D and 3D joints location. The 2D joints location data was converted to 2D joint angles as feature vectors, and the 3D joints location data was converted to REBA posture code as labels. The testing data is similarly converted to feature vectors and labels. By retrieving the "closest" set of training data to testing data in the high dimensional feature space and the mode of their labels, the training data is assigned with the same label.

This study demonstrated the performance of proposed 2D human motion capture alone and with the additive machine learning module, for estimating posture and repetition data in

REBA. The proposed 2D motion capture method alone was referred to as “vision-2D,” and it, along with the machine learning module was referred to as “vision-2D + ML” for the rest of this chapter.

4.1.3 Testing Condition

In this test, one male subject was selected to perform a sequence of common and ergonomically awkward postures in a lab condition. Two 72-second videos were taken of a subject interacting with items on a table and the floor, involving typical postures for ergonomic evaluation such as back-bending, arm-reaching, etc. The subject performed essentially the same movements in both videos, with one video taken closer to the side-view and the other closer to the diagonal view.

Different camera views aimed to help analyze how they impact the evaluation result, for both the vision-based method and manual observation. The sample snapshots of both videos are shown in Figure 4.1. The subject was wearing an IMU-based motion capture system (Perception Neuron®, Noitom, Miami, USA) that extracted the frame-wise 3D skeleton and derived the frequency and repetition for all postures to work as the baseline.

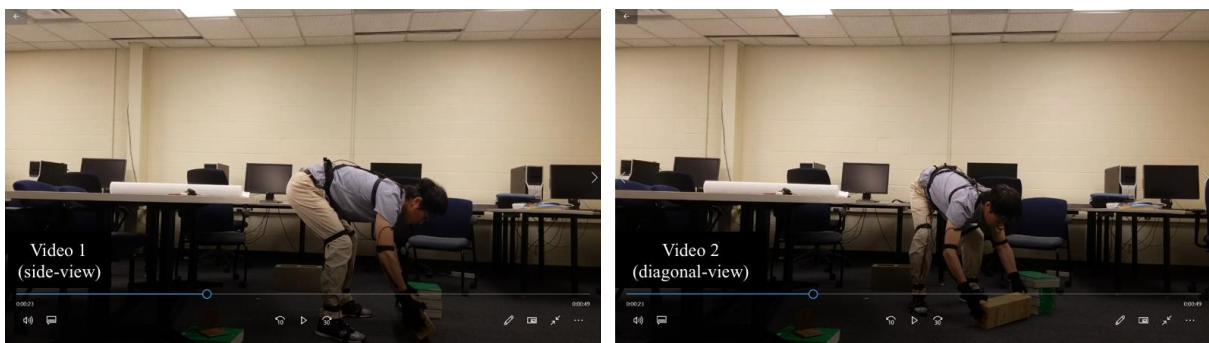


Figure 4.1 Snapshots of Test Videos for Performance Evaluation on REBA

The proposed “vision-2D” and “vision-2D + ML” were used to process the testing videos, and this study analyzed the subject’s posture for each frame of the video and calculated

the repetition expressed by frequency. As a human could not estimate frame-wise posture, the testing protocol selected frequency and duration of each posture as the target measurement. According to REBA, the selected postures are listed in Table 4.1. Given the test videos, the proposed method and the participating specialists were supposed to estimate the frequency and duration of all the 12 postures listed in the table.

Table 4.1 Selected Postures for Performance Evaluation on REBA

Joint	Posture		
	Safe	Cautious	Hazardous
Back	$< 20^\circ$	$\geq 20^\circ, < 60^\circ$	$\geq 60^\circ$
Neck	$< 10^\circ$	$\geq 10^\circ, < 20^\circ$	$\geq 20^\circ$
Shoulder	$< 45^\circ$	$\geq 45^\circ, < 90^\circ$	$\geq 90^\circ$
Knee	$< 30^\circ$	$\geq 30^\circ, < 60^\circ$	$\geq 60^\circ$

The experience of the 27 participating professionals ranged from 3 months to 27 years. These specialists were also identified by their status: 1) ergonomic Interns (lowest level of proficiency); 2) AEP - Associate Ergonomics Professionals (middle level of proficiency); and 3) CPE - Certified Professional Ergonomist (highest level of proficiency). The detailed information of each specialist is listed in Table 4.2.

The specialists were asked to fill in a spreadsheet with the observed posture and repetition data. Following is the description of how they recorded observations into the data collection form (Table 4.2). The data in the form is for explanation purposes only. While watching the video, the specialist identified the first instance of the back-bending angle between 20 and 60 degrees and recorded its duration as 15 seconds. The specialist then identified a second instance of the back-bending angle between 20 and 60 degrees and recorded its duration

as 12 seconds. The total frequency would be the total number of occurrences regarding each posture, and the duration would be the sum of duration for all instances.

Table 4.2 Information on Participating Ergonomists for Performance Evaluation

ID	Status	Experience*
S1	Intern	7 months
S5	AEP	1 year 6 months
S7	CPE	17 years
S8	CPE	11 years
S9	AEP	2 years
S10	CPE	15 years
S11	CPE	10 years
S12	CPE	6 years
S13	AEP	5 months
S14	CPE	20 years
S15	CPE	17 years
S16	CPE	18 years
S17	CPE	6 years
S18	CPE	13 years
S19	AEP	2 years 6 months
S20	CPE	27 years
S21	CPE	6 years 6 months
S22	CPE	10 years
S23	AEP	3 years
S24	Intern	3 months
S25	Intern	3 months
S26	Intern	3 months
S27	CPE	3 years
S28	AEP	2 years 6 months
S29	CPE	9 years
S30	AEP	10 years
S31	CPE	6 years
Average		7 years

<i>Total evaluation time:</i>															
Back	Duration (Unit: sec)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
< 20°	27	18	23	14	19										
≥ 20°, < 60°	15	12	11	17	10										
≥ 60°	22	21	20	5	4										

Figure 4.2 Example of Data Collection Form for Specialists

The specialists were allowed to playback the videos with no constraints. After collecting the estimated frequency and duration from both the proposed methods and specialists, the estimation accuracy was expressed by frequency error and duration error calculated as follows:

1) Frequency error (# of miscounts) = |baseline frequency – estimated frequency|

2) Duration error (mean absolute deviation in seconds) = |baseline duration – estimated duration|

Additionally, the evaluation time was also collected from specialists or recorded from the proposed methods.

4.1.4 Testing Result

This study showed the accuracy comparison of frequency and duration estimation by proposed “vision-2D” and “vision-2D + ML” and specialists’ observation. Table 4.3 and Table 4.4 showed the result comparison on “video1” (side-view), regarding frequency and duration, respectively. Similarly, Table 4.5 and Table 4.6 showed the result comparison on “video2” (diagonal-view). Each table was sorted showing the specialist that had the least errors at the top. The last row in each table showed the average errors by the body joint. The column headings represent each body part and its posture: N_C is neck-cautious ($10 \leq \text{Flexion} < 20$), N_H is neck_hazardous (> 20 Flexion), SL_C is left shoulder-cautious ($45 \leq \text{Flexion} < 90$), etc. Detailed meaning of each column heading can be found below:

N_C: neck_cautious; N_H: neck_hazardous;
SL_C: shoulder-left_cautious; SL_H: shoulder-left_hazardous;
SR_C: shoulder-right_cautious; SR_H: arm-right_hazardous;
B_C: back_cautious; B_H: back_hazardous;
KL_C: knee-left_cautious; KL_H: knee-left_hazardous;
KR_C: knee-right_cautious; KR_H: knee-right_hazardous.

Table 4.3 Performance Comparison of Frequency Estimation on REBA: Video1 (side-view)

Subject	Frequency Error ((Ground Truth - Estimation), Unit: # Occurrence)												Average Error
	N C	N H	SL C	SL H	SR C	SR H	B C	B H	KL C	KL H	KR C	KR H	
S1	2	6	4	0	2	0	2	0	0	1	1	1	1.58
S5	1	4	0	0	0	0	1	0	0	1	1	1	0.75
S7	2	2	1	0	1	0	1	0	2	1	3	1	1.17
S8	2	4	1	0	1	1	3	1	0	1	1	1	1.33
S9	0	2	1	1	1	1	3	0	0	1	1	1	1.00
S10	3	1	0	1	0	1	0	1	0	1	1	1	0.83
S11	1	2	4	1	4	1	5	0	0	1	1	1	1.75
S12	2	3	2	0	0	2	4	1	1	1	2	1	1.58
S13	1	4	1	0	1	0	1	0	0	1	1	1	0.92
S14	3	8	1	0	0	0	1	1	2	1	3	1	1.75
S15	3	2	0	1	3	2	1	0	0	1	1	1	1.25
S16	1	4	2	0	1	0	4	0	0	1	1	1	1.25
S17	2	5	1	2	0	0	2	2	1	1	3	1	1.67
S18	2	8	3	0	1	2	3	2	0	1	1	1	2.00
S19	4	2	1	1	1	0	1	0	0	1	1	1	1.08
S20	4	2	5	1	4	1	3	1	4	1	5	1	2.67
S21	0	5	3	0	2	0	2	1	2	1	3	1	1.67
S22	3	5	3	5	10	2	2	1	1	1	2	1	3.00
S23	3	10	1	1	1	1	2	1	1	1	2	1	2.08
S24	3	4	2	0	2	0	3	2	0	1	1	1	1.58
S25	1	5	2	2	3	3	2	1	0	1	1	1	1.83
S26	2	3	3	0	1	1	2	1	0	1	1	1	1.33
S27	0	2	2	2	8	1	0	0	1	1	1	1	1.58
S28	2	5	2	4	2	5	2	2	0	1	1	1	2.25
S29	0	2	0	0	1	1	0	0	4	3	5	3	1.58
S30	0	3	2	4	0	5	0	1	0	1	1	1	1.50
S31	0	5	3	3	4	5	0	0	0	1	1	1	1.92
Vision-2D	0	0	0	2	1	1	0	2	4	1	2	1	1.17
Vision-2D + ML	0	1	0	0	0	0	0	0	0	0	2	0	0.25
Vision-2D	0	0	0	2	1	1	0	2	4	1	2	1	1.17
Vision-2D + ML	0	1	0	0	0	0	0	0	0	0	2	0	0.25
Specialists	1.62	3.76	1.72	1.07	1.90	1.24	1.72	0.72	0.79	1.03	1.72	1.03	1.53

Table 4.4 Performance Comparison of Duration Estimation on REBA: Video1 (side-view)

Subject	Duration Error ((Ground Truth - Estimation), Unit: Sec)												Average Error
	N C	N H	SL C	SL H	SR C	SR H	B C	B H	KL C	KL H	KR C	KR H	
S1	1.47	2.27	1.57	1.13	4.00	0.83	6.50	3.47	7.47	8.37	9.03	8.37	4.54
S5	10.53	12.73	0.43	0.87	2.00	0.83	6.50	1.53	8.47	8.37	10.03	8.37	5.89
S7	9.53	11.27	3.57	1.87	0.00	0.17	6.50	0.47	7.47	8.37	9.03	8.37	5.55
S8	8.53	5.27	3.43	0.13	1.00	0.83	5.50	1.47	8.47	8.37	10.03	8.37	5.12
S9	2.53	7.73	2.57	0.13	1.00	0.83	6.50	0.47	12.47	8.37	14.03	8.37	5.42
S10	12.53	4.73	3.43	1.13	0.00	0.83	10.50	3.47	6.47	8.37	8.03	8.37	5.66
S11	9.53	17.73	1.57	2.13	12.00	3.83	3.50	2.53	3.53	8.37	1.97	8.37	6.26
S12	0.53	13.27	5.43	2.13	0.00	2.83	9.50	1.47	10.47	8.37	12.03	8.37	6.20
S13	11.53	4.73	1.43	1.87	1.00	5.17	7.50	1.47	9.47	8.37	11.03	8.37	5.99
S14	0.53	15.27	1.43	2.87	2.00	3.17	8.50	1.53	6.47	8.37	8.03	8.37	5.54
S15	12.53	9.73	5.57	3.13	6.00	7.83	5.50	0.47	4.47	8.37	6.03	8.37	6.50
S16	9.53	10.73	1.57	2.87	7.00	2.17	8.50	3.53	11.47	8.37	13.03	8.37	7.26
S17	7.53	3.77	2.07	0.13	1.00	1.17	6.00	0.53	5.97	4.37	10.53	8.37	4.29
S18	2.53	2.27	2.43	4.13	1.00	5.83	9.50	6.47	9.47	8.37	11.03	8.37	5.95
S19	1.47	9.73	8.57	0.13	10.00	2.17	3.50	1.47	13.47	8.37	15.03	8.37	6.86
S20	6.53	2.27	8.43	2.87	5.00	1.17	1.50	0.47	3.53	8.37	1.97	8.37	4.21
S21	11.53	1.27	5.93	2.37	1.00	2.67	1.00	0.47	10.47	8.37	8.03	8.37	5.12
S22	5.53	5.27	11.57	11.87	22.00	0.83	3.50	3.47	5.47	8.37	7.03	8.37	7.77
S23	10.53	7.27	9.57	1.87	19.00	9.17	4.50	2.53	9.47	8.37	6.03	8.37	8.06
S24	6.53	1.73	0.43	0.13	4.00	0.17	0.50	3.47	8.47	8.37	10.03	8.37	4.35
S25	0.47	2.27	2.43	2.13	3.00	1.83	4.50	5.47	6.47	8.37	8.03	8.37	4.44
S26	10.53	13.73	8.43	5.87	5.00	1.83	3.50	1.47	8.47	8.37	5.03	8.37	6.72
S27	5.47	3.73	20.57	0.13	44.00	0.83	11.50	2.53	8.47	8.37	8.03	8.37	10.17
S28	8.53	5.27	2.57	3.87	8.00	0.83	6.50	5.47	10.47	8.37	11.03	8.37	6.61
S29	4.53	2.27	6.57	11.13	14.00	0.17	1.50	1.47	3.53	3.63	1.97	3.63	4.53
S30	13.53	1.27	3.43	7.87	1.00	11.83	7.50	7.53	5.47	8.37	7.03	8.37	6.93
S31	2.53	10.73	15.57	11.13	16.00	10.83	0.50	0.47	11.47	8.37	13.03	8.37	9.08
Vision-2D	9.23	21.00	6.57	0.53	1.67	4.13	1.20	6.00	3.53	8.37	1.97	8.37	6.05
Vision-2D + ML	11.27	4.07	2.40	3.63	4.20	0.20	5.40	4.23	7.97	3.27	8.80	3.27	4.89
Vision-2D	9.23	21.00	6.57	0.53	1.67	4.13	1.20	6.00	3.53	8.37	1.97	8.37	6.05
Vision-2D + ML	11.27	4.07	2.40	3.63	4.20	0.20	5.40	4.23	7.97	3.27	8.80	3.27	4.89
Specialists	7.16	7.36	5.16	3.11	6.75	2.93	5.42	2.60	7.89	7.89	8.52	8.03	6.07

Table 4.5 Performance Comparison of Frequency Estimation on REBA: Video2 (diagonal-view)

Subject	Frequency Error ((Ground Truth - Estimation), Unit: Sec)												Average Error
	N C	N H	SL C	SL H	SR C	SR H	B C	B H	KL C	KL H	KR C	KR H	
S1	1	4	1	0	1	0	2	0	2	3	1	3	1.50
S5	1	1	1	1	2	1	1	1	2	3	1	3	1.50
S7	2	5	1	1	0	3	2	0	2	3	1	3	1.92
S8	1	3	0	1	0	2	2	2	2	3	1	3	1.67
S9	1	0	1	0	2	0	1	0	2	3	1	3	1.17
S10	3	1	2	1	0	3	3	0	2	3	1	3	1.83
S11	0	1	1	1	4	0	4	0	1	3	2	3	1.67
S12	2	7	3	0	0	2	3	0	2	3	1	3	2.17
S13	2	0	1	1	0	1	1	1	2	3	1	3	1.33
S14	1	8	1	1	1	2	2	0	1	3	0	3	1.92
S15	3	0	1	1	2	2	1	0	2	3	1	3	1.58
S16	2	0	1	0	1	3	3	1	2	3	1	3	1.67
S17	1	3	2	1	0	1	3	0	2	2	1	3	1.58
S18	1	3	2	1	1	4	4	3	2	3	1	3	2.33
S19	0	3	1	0	1	0	2	0	3	3	2	3	1.50
S20	2	3	5	1	4	3	1	1	1	3	2	3	2.42
S21	4	2	4	1	0	3	1	1	2	3	0	3	2.00
S22	3	1	8	6	11	3	1	0	1	3	0	3	3.33
S23	3	6	3	1	1	1	3	0	2	3	0	3	2.17
S24	0	1	2	1	0	3	2	0	2	3	1	3	1.50
S25	1	5	3	1	0	3	2	0	2	3	1	3	2.00
S26	3	0	5	1	1	2	2	0	2	3	0	3	1.83
S27	0	2	5	1	11	1	2	1	2	3	1	3	2.67
S28	2	4	1	0	0	2	3	3	3	3	1	3	2.08
S29	1	3	1	4	2	3	0	0	1	0	2	0	1.42
S30	4	3	2	3	2	6	4	0	1	3	0	3	2.58
S31	0	1	0	4	0	6	0	0	2	3	1	3	1.67
Vision-2D	1	7	3	1	1	2	2	2	1	3	2	3	2.33
Vision-2D + ML	1	3	0	0	1	2	1	0	3	2	1	2	1.33
Vision-2D	1	7	3	1	1	2	2	2	1	3	2	3	2.33
Vision-2D + ML	1	3	0	0	1	2	1	0	3	2	1	2	1.33
Specialists	1.59	2.76	2.10	1.21	1.69	2.21	2.00	0.55	1.86	2.83	0.97	2.86	1.89

Table 4.6 Performance Comparison of Duration Estimation on REBA: Video2 (diagonal-view)

Subject	Duration Error ((Ground Truth - Estimation), Unit: Sec)												Average Error
	N C	N H	SL C	SL H	SR C	SR H	B C	B H	KL C	KL H	KR C	KR H	
S1	11.57	8.80	3.94	0.92	1.15	1.79	5.20	4.91	3.40	5.13	2.80	5.13	4.56
S5	0.43	16.20	3.94	1.92	3.15	2.29	5.20	4.91	1.40	5.13	0.80	5.13	4.21
S7	1.43	10.20	0.06	0.08	1.15	0.79	0.80	0.09	4.60	5.13	5.20	5.13	2.89
S8	2.43	11.80	2.94	0.08	0.15	2.79	2.20	0.91	3.40	5.13	2.80	5.13	3.31
S9	8.57	2.20	2.94	4.92	0.15	3.79	9.80	3.91	7.40	5.13	4.80	5.13	4.90
S10	2.43	4.20	2.06	6.92	3.85	6.79	1.20	6.91	1.40	5.13	0.80	5.13	3.90
S11	2.57	17.20	8.94	0.08	6.85	1.79	8.80	3.91	3.40	5.13	2.80	5.13	5.55
S12	0.43	1.20	6.94	3.92	2.15	6.79	4.20	2.91	4.60	5.13	5.20	5.13	4.05
S13	2.57	1.20	5.94	0.92	4.15	0.79	4.20	3.91	2.40	5.13	1.80	5.13	3.18
S14	16.57	14.80	2.94	0.92	0.85	2.21	2.20	2.91	4.60	5.13	5.20	5.13	5.29
S15	1.43	11.20	1.94	5.92	1.85	2.79	2.20	3.09	0.40	5.13	0.20	5.13	3.44
S16	7.57	4.20	4.94	3.08	2.85	1.21	10.20	6.09	5.40	5.13	4.80	5.13	5.05
S17	6.57	6.80	3.94	0.42	3.15	2.79	6.80	6.91	5.40	5.13	5.20	2.87	4.67
S18	13.57	14.80	5.94	3.92	1.15	7.79	2.20	8.91	2.40	5.13	1.80	5.13	6.06
S19	4.43	2.20	5.06	6.08	5.85	4.21	1.20	1.09	6.40	5.13	5.80	5.13	4.38
S20	4.43	6.20	10.94	7.08	8.15	1.21	5.20	7.91	11.60	5.13	12.20	5.13	7.10
S21	6.07	7.80	1.44	3.08	6.85	3.71	5.70	1.41	2.60	5.13	3.20	5.13	4.34
S22	5.57	1.80	12.06	2.92	18.85	7.79	2.80	6.91	2.60	5.13	3.20	5.13	6.23
S23	3.57	20.80	8.06	9.08	9.85	7.21	7.80	0.09	5.40	5.13	4.80	5.13	7.24
S24	2.43	3.80	2.94	0.08	4.85	3.21	0.20	5.91	5.40	5.13	3.80	5.13	3.57
S25	4.57	4.80	3.94	5.08	3.15	6.79	2.20	8.91	2.40	5.13	1.80	5.13	4.49
S26	2.57	11.20	2.94	1.08	0.85	0.21	2.80	2.91	4.40	5.13	2.80	5.13	3.50
S27	2.57	0.20	19.06	3.08	27.85	2.79	1.80	11.09	3.60	5.13	2.80	5.13	7.09
S28	3.57	6.80	17.06	14.92	20.85	15.79	6.80	7.91	2.40	5.13	1.80	5.13	9.01
S29	3.57	8.20	3.06	2.92	4.85	3.79	4.80	2.09	11.60	8.87	12.20	8.87	6.24
S30	0.57	2.80	7.94	14.92	0.15	14.79	0.80	1.09	4.40	5.13	3.80	5.13	5.13
S31	7.57	2.20	22.06	13.92	24.85	15.79	4.80	4.09	6.40	5.13	5.80	5.13	9.81
Vision-2D	0.70	17.40	2.97	11.52	5.08	7.09	2.17	6.61	11.60	5.13	3.63	5.13	6.59
Vision-2D + ML	0.16	1.47	0.11	0.15	0.05	0.16	0.10	0.21	0.10	0.03	1.43	0.03	0.33
Vision-2D	0.70	17.40	2.97	11.52	5.08	7.09	2.17	6.61	11.60	5.13	3.63	5.13	6.59
Vision-2D + ML	0.16	1.47	0.11	0.15	0.05	0.16	0.10	0.21	0.10	0.03	1.43	0.03	0.33
Specialists	4.50	7.67	6.10	4.48	6.02	4.79	3.94	4.43	4.52	5.08	3.91	5.01	5.04

The summarized frequency and duration estimation accuracy with comparison was shown in Figure 4.3. Without the machine learning (“ML”) module, the proposed 2D human motion capture method (“Vision-2D”) presented a comparable estimation accuracy on frequency and duration for REBA postures. With the machine learning module, the overall estimation error drops significantly from specialists’ observation. Also, similar to the human observation’s correctness regarding the camera’s view angle, vision-based approaches showed a higher estimation accuracy on the side-view video than the diagonal-view video.

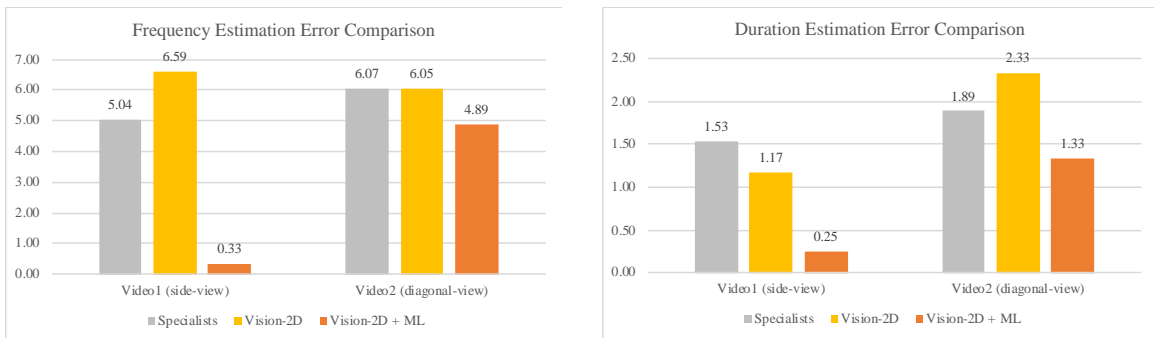


Figure 4.3 Performance Comparison of Frequency and Duration on REBA

4.2 Snook’s Tables

4.2.1 Summary of Snook’s Tables

Snook’s Tables (Snook and Ciriello 1991), sometimes referred as “Snook’s Lifting Recommendation” or “Liberty Mutual Manual Materials Handling Tables,” were developed to evaluate manual materials handling tasks including lifting, lowering, pushing, pulling and carrying. It consists of a series of tables regarding each of the 5 tasks and different genders of the participating subjects. Snook’s Tables for lifting tasks require several input variables as listed below:

1) Hand distance. The hand distance is defined as the horizontal distance between the mid-point of the hands to the front of the body. Snook's Tables does not require numerical value, but only categorical data that which value it is closer to 7, 10, 15 inches.

2) Lifting distance. Lifting distance is the vertical distance between the origin and destination of the evaluated lift, and also comes with categorical data. The three key values taken as the mean of the three categories are 10, 20, 30 inches.

3) Hand location at the origin. It is the vertical distance between the mid-point of the hands and the ground when the lifting task starts. There are three categories: below knuckle height; between knuckle and shoulder height; above shoulder height.

4) Frequency. The frequency is quantified as one lift happening in every specific period. A value can be selected from 15 sec., 30 sec., 1 hour, 5 hours, and 8 hours.

5) Object weight. This is the only variable that needs numerical value in lb. With these 4 variables, a population percentage (%) output would be calculated based on the research. The population percentage indicates the percentage of the population from a given gender could safely perform the evaluated task. Thus, the higher the risk, the less the population percentage would be.

4.2.2 Method

According to the required input data for Liberty Mutual Tables, distance-based measurements are required with true-to-scale values (e.g., in inch). However, the proposed 3D motion capture method estimated the 3D canonical human model from the monocular camera only generated the location expressed in the camera's coordinate system and does not directly provide data with a unit of inch or centimeter. To address this issue, this study introduced a human model scaling

step to scale the canonical human model to individualized human model by scaling with the subject's height, as shown in Figure 4.4 (left).

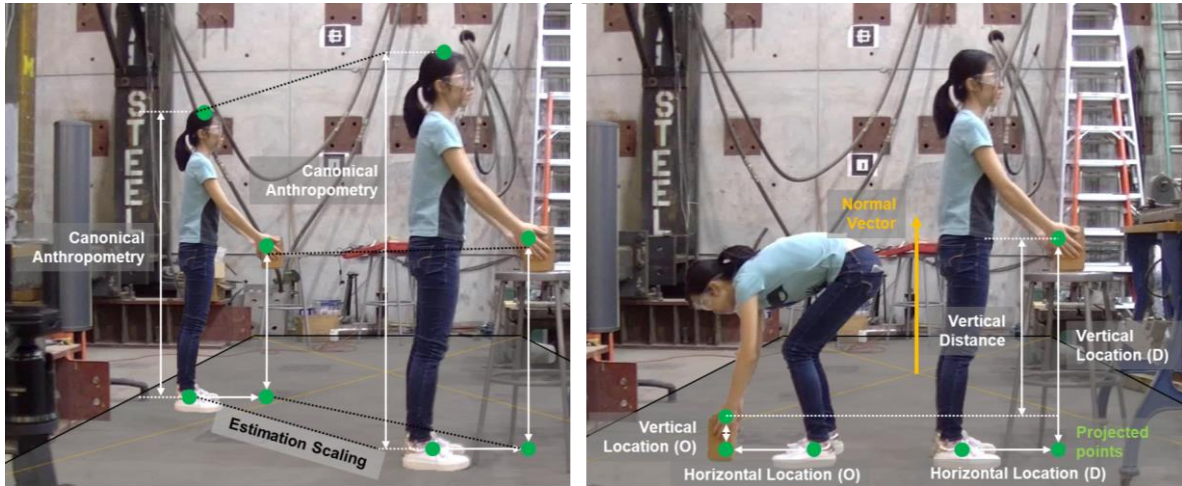


Figure 4.4 3D Human Model Scaling (left) and World Coordinate Rectification (right)

Another challenge is that the required distance should be projected to the horizontal or vertical direction while the projection direction remains unknown. To address this issue, either the camera should be assumed to be leveled throughout the video recording or the camera's real-time pose is known to estimate the horizontal or vertical direction. To enable a free-moving camera during video recording, the second scenario was explored. The proposed solution was to extract the real-time gravity direction from the smartphone while using the built-in camera to record the video. With the gravity direction provided, it was converted to the camera's coordinate system that aligns with the one used by the 3D human motion capture. Then this converted direction, called normal vector's direction, was used to project the hands' location on and calculate the vertical location and vertical travel distance. As for the horizontal location, the ground plane was extracted from the orthogonal plane of the normal vector. The horizontal

location was calculated by projecting the hands' location onto the ground plane. Figure 4.4 (right) illustrated this step with a lifting task.

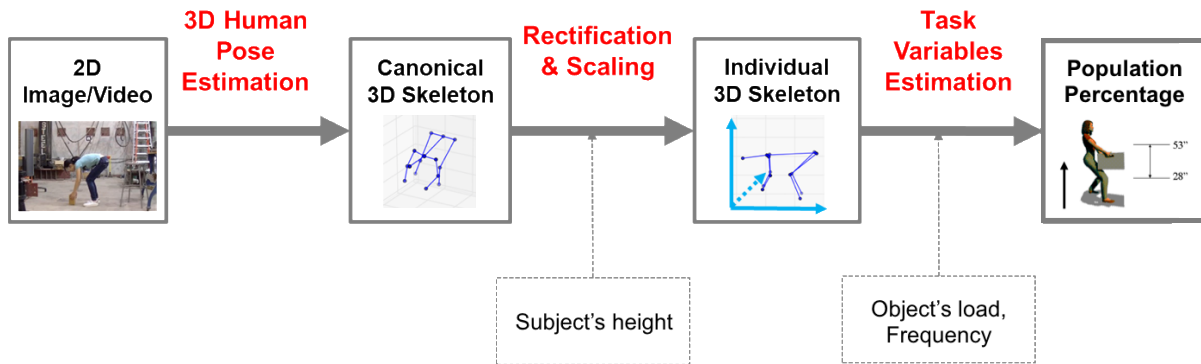


Figure 4.5 Pipeline of Assessing Lifting Tasks with Snook's Tables



Figure 4.6 Examples of Recorded Lifting Tasks

The overall procedure was summarized in Figure 4.5. With 2D image/video captured from a monocular camera, a canonical 3D skeleton is estimated by the 3D human motion capture framework introduced in Chapter 3. With user input of the subject's height, the skeleton is then scaled to an individualized skeleton with joint location expressed in the unit of the height information. If the gravity direction of the moving camera was recorded, the 3D skeleton is

rectified to align with the world coordinate of 3D space. Except for the object's load and frequency that needs to be manually input, major task variables and final risk level expressed as population percentage could be estimated.

Table 4.7 Performance Comparison of Task Variables Estimation for Snook's Tables

	Hand Distance (in.)	Lifting Distance (in.)	Population Percentage (%)
Est.	15	30	59
Meas.	15	30	59
Diff.	0	0	0
Est.	15	30	59
Meas.	15	30	59
Diff.	0	0	0
Est.	15	20	62
Meas.	15	20	62
Diff.	0	0	0
Est.	15	30	59
Meas.	15	30	59
Diff.	0	0	0
Est.	15	20	62
Meas.	15	30	59
Diff.	0	1	1
Est.	10	20	81
Meas.	15	30	59
Diff.	1	1	1
Est.	15	10	72
Meas.	15	10	72
Diff.	0	0	0
Est.	15	30	59
Meas.	15	30	59
Diff.	0	0	0
Est.	15	30	59
Meas.	15	30	59
Diff.	0	0	0
Est.	15	30	59
Meas.	15	30	59
Diff.	0	0	0
# Misdetection	1	2	2
% Misdetection	10%	20%	20%

4.2.1 Testing Condition

The testbed was sponsored by Power Construction, and 10 male construction workers participated in this study with written consent. Each construction worker performed a different lifting task for several cycles. One cycle from each subject was selected to validate the approach. All the subjects were asked to keep the personal protective equipment (PPE) and tools to reflect the usual appearance of body joints in the videos. Additionally, the tasks covered a wide range of different variables' values, such as hands' height at origin can be as low as a ground level to around the waist level, along with various camera angle. All the subjects were asked to lift real objects, and occlusion was naturally introduced to reflect the real site's condition. A sample of cropped video frames reflecting the jobs are shown in Figure 4.6. The ground truth data was provided by how a manual observer would evaluate a lifting task, including using a tape measure to measure distance related variables and observation for an asymmetric angle. The variables that cannot be automatically estimated by the proposed method were assumed to be known, including the subject's height, object's weight, and frequency.

4.3.2 Testing Result

Table 4.7 shows the accuracy of using the proposed video-based 3D motion capture to estimate the major task variables (i.e., hand distance, lifting distance) and the final risk level of population percentage. The task variable, hand location at the origin, is also estimated and measured. As this task variable is estimated to be correct for all the 10 lifting tasks, it is not included in the table for comparison. For the three measurements under comparison, estimation by the proposed method is referred to as "Est.," Ground truth data by a tape measure measurement is referred to as "Meas.,". As the task variables are categorical, the difference between estimation and measurement is

represented by a digit “0” for correct estimation (no difference) and “1” for incorrect estimation (with a difference). In summary, the second to the last row in the table shows the number of incorrect estimations, named by “# misdetection.” It is followed by the percentage of tasks with incorrectness for the variable concerned. From the test result, most of the variables are correctly categorized. Specifically, 90% of tasks have a correct estimation of hand distance, and 80% of tasks have a correct estimation of lifting distance. As an overall accuracy, 80% of tasks are estimated correctly on the risk level or population percentage.

4.3 NIOSH Lifting Equation

4.3.1 Summary of NIOSH Lifting Equation

NIOSH Lifting Equation was developed and widely applied to evaluate lifting tasks in manual materials handling. To use this tool, the observer needs to identify the origin and destination of the lifting. Usually, the time frames when the handled object leaves and arrives at the designated location, respectively. At both time frames, several input data need to be collected (Waters et al. 1994):

- 1) Horizontal hand location. The horizontal distance from the mid-point of the hands to the mid-point of the ankles.

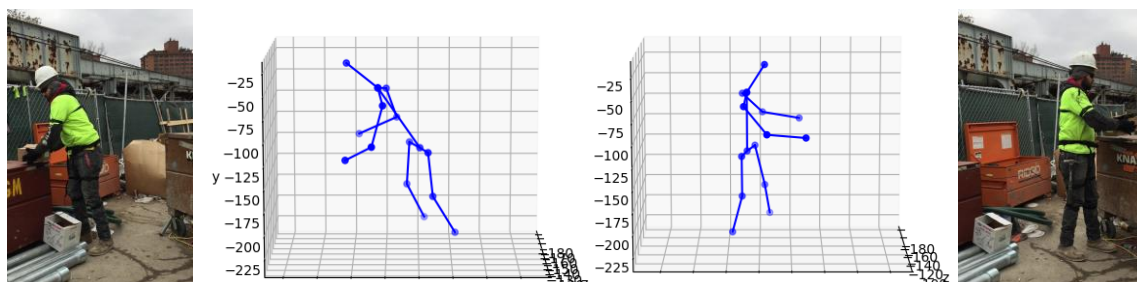


Figure 4.7 Example of Lifting Task and Captured Motion Data

2) Vertical hand location. The vertical distance from the mid-point of the hands to the ground.

3) Vertical travel distance. The vertical displacement of mid-point of the hands between the origin and destination.

4) Asymmetric angle. The twisting angle of the trunk, or the angle of symmetry between the mid-plane of the body and the direction of lift (Berlin and Adams 2017).

5) Frequency and duration of lifting.

6) Coupling. The goodness of how the subject handles the object (e.g., with or without handles).

Each of the 6 variables' value was converted to a decimal between 0-1 as a multiplier (HM, VM, DM, AM, FM, CM), by retrieving the data from the given table (Waters et al. 1994). The higher the risk each variable correlates with, the smaller the multiplier. The product of all 6 multipliers is also a decimal between 0-1.

It was assumed that the maximum weight load could be lifting for the majority of healthy people under the best possible lifting circumstance, for up to 8-hour shift, is 23 kg or 50 lb. A load constant "LC" is multiplied to the product of all multipliers to obtain a principal product of NIOSH Lifting Equation, Recommended Weight Limit (RWL):

$$\text{Recommended Weight Limit (RWL)} = LC \times HM \times VM \times DM \times AM \times FM \times CM$$

This output variable suggests the highest weight could be lifted under given workplace condition. To further suggest the injury risk of lifting an actual weight, other than the suggested highest weight, Lifting Index (LI) can be calculated from the lifted object's weight and RWL by:

$$\text{Lifting Index (LI)} = \frac{\text{Load Weight (L)}}{\text{Recommended Weight Limit (RWL)}}$$

If Lifting Index is larger than 1.0, the job is considered to have high risk and needs to be modified. Otherwise, the job has a nominal risk.

4.3.2 Method

The estimation for task variables of NIOSH Lifting Equation is mostly similar to that for Snook's Tables. The pipeline (Figure 4.8) is then inherited with modification on the required task variables and the lifting index as the evaluated risk level.

The major difference of NIOSH Lifting Equation from Snook's Tables regarding the task variables is the additional angle-based measurement, asymmetric angle. This angle value is quantified by the trunk's twisting angle at both the origin and destination of the lift. Since the asymmetric angle is challenging to estimate through manual observation, it is specifically evaluated in this study to demonstrate the feasibility of the proposed approach in such a tool.

The testing condition is similar to that of Section 4.2 and is not described again.

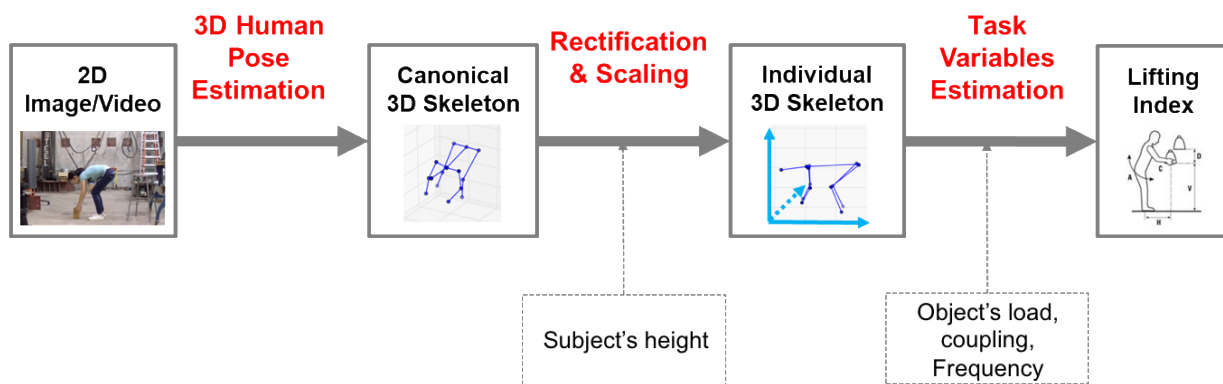


Figure 4.8 Pipeline of Assessing Lifting Tasks with NIOSH Lifting Equation

Table 4.8 Performance Comparison of Task Variables Estimation on NIOSH Lifting Equation

	Hand Location (in.)				Vertical Travel Distance (in.)	Asymmetric Angle (°)		Lifting Index
	Origin		Destination			Origin	Destination	
	Horizontal	Vertical	Horizontal	Vertical				
Est.	17.3	8.5	19.0	43.5	35.0	3.8	2.0	1.43
Meas.	15.0	3.0	15.0	38.0	35.0	0.0	0.0	1.48
Diff. (%)	2.3	5.5	4.0	5.5	0.0	3.8	2.0	3.4%
Est.	3.7	5.1	12.9	47.9	42.7	23.4	85.9	1.40
Meas.	4.2	3.0	14.0	48.0	45.0	45.0	90.0	1.40
Diff. (%)	0.5	2.1	1.1	0.1	2.3	21.6	4.1	0.0%
Est.	6.2	22.2	17.4	40.5	18.3	25.2	0.5	1.21
Meas.	8.3	24.0	15.0	42.0	18.0	0.0	0.0	1.04
Diff. (%)	2.3	5.5	4.0	5.5	0.3	3.8	2.0	16.3%
Est.	9.2	3.5	21.8	33.3	29.8	3.8	5.7	1.52
Meas.	9.0	3.0	20.0	34.0	31.0	0.0	0.0	1.38
Diff. (%)	2.3	5.5	4.0	5.5	1.2	3.8	2.0	10.1%
Est.	24.1	23.7	21.8	40.3	16.6	3.1	75.6	1.97
Meas.	24.0	33.5	24.0	43.0	9.5	0.0	90.0	2.20
Diff. (%)	2.3	5.5	4.0	5.5	7.1	3.8	2.0	10.5%
Est.	12.1	21.7	11.5	39.5	17.8	4.8	5.6	0.84
Meas.	14.0	25.0	14.0	36.0	11.0	0.0	0.0	0.88
Diff. (%)	2.3	5.5	4.0	5.5	6.8	3.8	2.0	4.5%
Est.	14.1	23.9	17.1	38.3	14.4	5.8	2.3	1.14
Meas.	19.0	29.0	22.0	43.0	14.0	0.0	0.0	1.51
Diff. (%)	2.3	5.5	4.0	5.5	0.4	3.8	2.0	24.5%
Est.	10.3	1.8	12.7	39.5	37.7	1.7	5.2	0.94
Meas.	13.0	3.0	13.0	33.5	30.5	0.0	0.0	1.09
Diff. (%)	2.3	5.5	4.0	5.5	7.2	3.8	2.0	13.8%
Est.	14.3	10.1	24.7	42.0	32.0	6.6	5.2	1.85
Meas.	16.5	3.0	22.0	33.3	30.3	0.0	0.0	1.51
Diff. (%)	2.3	5.5	4.0	5.5	1.7	3.8	2.0	22.5%
Est.	10.2	1.0	25.2	42.0	41.0	12.5	17.2	1.98
Meas.	8.3	3.0	24.5	42.9	39.9	0.0	30.0	2.04
Diff. (%)	2.3	5.5	4.0	5.5	1.1	3.8	2.0	2.9%
Mean Diff.	2.1	5.2	3.7	5.0	2.8	5.6	2.2	10.9%

4.3.3 Testing Result

The performance of task variables estimation for NIOSH Lifting Equation is shown in Table 4.8. The proposed video-based 3D motion capture approach with rectification and scaling modules, achieve an overall accuracy of lifting index estimation with an error of 10.9%. From the estimation accuracy of individual task variables, distance-based measurements have a mean absolute error of around or within 5 inches for each variable. The asymmetric angle estimation has an accuracy of fewer than 10 degrees and at maximum 21.6 degrees for a specific task.

4.3.4 Discussion

NIOSH Lifting Equation partitions the value of task variables too much smaller intervals (i.e., one inch), compared with a rough categorical range in Snook's Tables. The result of NIOSH Lifting Equation should give more detailed information about the performance of the proposed motion capture approach in ergonomic risk factors quantification for on-site lifting jobs.

As the overall accuracy in NIOSH Lifting Equation estimation is similar to that in Snook's Tables, it is promising to claim the robustness of the 3D human motion capture method for on-site application. As for the individual task variables, it seems to yield conflicting conclusion with that in the preceding section that vertical travel distance is most vulnerable to error, as it has a mean difference of only 2.8 inches. However, the problematic task which has an excessive error in the lifting distance estimation in Snook's Tables also has the maximum estimation difference in NIOSH Lifting Equation. It is averaged out as the task variable is provided in numerical value, instead of a categorical one.

4.4 Conclusion

To apply the proposed 2D and 3D human motion capture approach in risk factors estimation for ergonomic risk assessment, this study applied the approach with a specific modification regarding three types of assessment tools and used a representative tool from each type to demonstrate the potential.

For an angle-based postural analysis tool (e.g., REBA), 2D human motion capture is sufficient to estimate the joint angle given a proper camera view angle, such as a side view. The joint angle is calculated directly from the detected 2D joint location regarding image pixel. To mitigate the distortion brought by imperfect camera view angle, a supervised classification module is developed for assessment tools that only require categorical value for risk factors. The lab-based test yielded a comparable performance of raw 2D human motion capture with average professionals' observation. Further, with the additional classification module, the accuracy of frequency and duration estimation of the video-based approach is much higher than the professionals' observation. It suggests a great potential to apply the proposed 2D human motion capture framework in estimating posture and repetition for angle-based postural analysis.

Rather than angle-based, an ergonomic risk assessment tool that relies on distance-based measurements (e.g., vertical hand location from the ground) such as Snook's Tables, can apply the developed 3D human motion capture approach to estimate the task variables. As the raw 3D motion data does not reflect true-to-scale location or distance measurement, a human model scaling and rectification module is developed given the subject's height and camera's pose. With the true-to-scale 3D human skeleton, the required distance-based measurements were evaluated by a tapeline measurement on 10 workers' lifting tasks in a construction site. With categorical task variables for Snook's Tables, 80% of the tasks have a correct estimation of all the three

distance-based measurements. The result demonstrated a feasible application of the proposed 3D motion capture approach to distance-based postural analysis on a real job-site.

NIOSH Lifting Equation is a widely used tool that needs both distance- and angle-based (i.e., asymmetric angle) measurements. The proposed 3D human motion capture approach is applied and validated by a tapeline measurement for distance-based variables and human observation for the angle variable. The field test was conducted with 10 lifting tasks each performed by a construction worker. The result shows an accuracy of within 5.2 inches distance estimation error, within 10 degrees angle estimation error and 10.9% lifting index estimation deviation. It demonstrates the proposed approach is robust enough to not only properly estimate the risk factors' rough range (for categorical input) and also accurate value (for numerical input), given a realistic job-site condition.

Chapter 5 Vision-Based Hand Push Force Estimation from 3D Motion Capture

5.1 Introduction

Apart from posture and repetition, force is considered among the most impactful ergonomic risk factors to evaluate jobs. Many observation-based methods require external force exerted on the body part. Among them, several posture-based methods ask for a categorical value of force. For example, OWAS asks if the exerted force lies in one of the following three ranges: 0 - 10 kg, 10-20 kg, and >20 kg. REBA requires similar data with slightly different ranges: 0-11 lb., 11-22 lb., and >22 lb. Biomechanics-based methods, which impose more emphasis on the force, needs a numerical value of external force. For example, NIOSH Lifting Equation takes the object's weight for evaluating lifting tasks. Similarly, Snook's Tables require the lifted/lowered object's weight for lifting/lowering tasks while initiated/sustained force for pushing/pulling/carrying tasks.

Some methods only require the object's weight to calculate the risk level are straightforward (e.g., NIOSH Lifting Equation). However, some methods like Snook's Tables that try to capture more complex interaction with the environment (e.g., pushing tasks) are challenging to estimate the force exertion. In practice, force sensors are applied to measure the exposure level directly. For example, a force gauge was attached between the object (e.g., the cart) and the contacting body part (e.g., hand) to measure how much hand push force was exerted

when pushing a cart. Electromyography (EMG) sensors can be attached to body parts (e.g., back) to measure muscle force exerted on the specific body segment.

In addition to observation-based methods that only take advantage of a static external load for a complete task, biomechanical analysis requires vectorized external force data for every time-frame. Specifically, biomechanical models were developed to analyze force exerted on internal body joints (e.g., elbow and shoulder) from known force exertion on external body joints (e.g., hand) and the whole-body posture (all body joints' locations). Biomechanical analysis requires pair-wise whole-body postures and force exertion. Force data collection becomes rather challenging as it needs to be time frame-specific.

Force sensors for continuous data collection include force transducers and pressure sensors. A force transducer usually measures tri-axial forces and moments that come with 3 or 6 degrees of freedom. A pressure sensor presents as a capacitive thin-film sensing grid and could be mounted on a various contact surface, even wrapped around a grip handle (Welcome et al. 2004). A pressure sensor generates pressure distribution in a mesh-grid with numerical values for each grid and can provide the center of pressure. However, it does not give the direction of force or moments and could be inapplicable for biomechanical analysis.

Regardless of the pros and cons of various force sensors against one another, shared drawbacks include being invasive and cost-prohibitive for on-site application. Yu et al. (2019a) applied smart insoles to collect ground reaction force exerted on both feet. It significantly reduced the level of invasiveness from the aforementioned body attached force sensors, but still could not avoid the interference from sensors setup for in-field deployment. As Wang et al. (2015) argued, there lacks a non-invasive approach to measure force exertion for ergonomic evaluation.

5.2 Literature Review

Non-invasive Force Estimation

To estimate the exerted force using a non-invasive approach, some studies beyond the field of ergonomics explored the potential of applying vision-based methods. Gaddam et al. (2016) demonstrated the potential of estimating ground reaction force from the spine's frame-wise location collected by RGB-D sensor (Microsoft Kinect). Pham et al. (2015) used the same sensor to estimate hand contact force by capturing the hand pose and forming a physics-based reconstruction module followed by an optimization module with an artificial neural network. To apply ubiquitous visual sensing device, Sartison et al. (2018) developed a machine learning-based approach to estimate finger grip force from an RGB frame sequence with visual markers on fingers. These studies showed a potential of estimating force from visual and motion data, but only focused on a single body part. Such potential to be generalized to whole-body and various tasks remain unknown.

Pham et al. (2018) demonstrated the feasibility of estimating hand contact force of multiple tasks from whole-body motion data captured from IMU sensors. It inherited the two-stage scheme from their previous work that estimate required force to perform the task by applying physics-based reconstruction module to motion data, and recurrent neural network (RNN) to estimate the actual exerted force. Jahanbanifar and Akhavian (2019) also showed the potential of using an artificial neural network to estimate hand push force directly from the wrist's motion data collected by accelerometers.

These studies suggested the feasibility of estimating force on a single body part from visual data and force involving whole-body movement from reliable motion data. However, it still remains a question if the force (e.g., hand push force) could be estimated from motion data

extracted from visual frame sequence. Motion data captured from a commercial system comes in a comprehensive human model. For example, Pham et al. (2018) applied a human model with 22 joints connected by rigid body part model. A full number of joints can describe the human motion accurately, and rigid body part model enables the modeling part's rotation. However, motion data captured from images/videos come in a much simpler human model (e.g., 16 joints from the majority work in the computer vision community). Modeling human with fewer joints may suffer from inaccurate motion data (e.g., spine is modeled by a straight-line segment instead of a series of short connected line segments). In addition, vision-based human modeling that uses a line to represent each limb thus cannot model the body part's rotation.

As force reconstruction from comprehensive human modeling assumes motion data with full degrees of freedom, the essential challenge breaks down to the question if a simplified human model extracted from images can be converted to a compatible one with whole-body physics-based equations to reconstruct the required force.

Kinematic Model for Force Estimation

Multibody system (MBS) modeling serves as the most popular way to represent the biomechanical behavior of the human body (Raison 2009). This study modeled the human body with the tree-like structure, instead of the constrained structure as it only considered joint force, excluding muscle force. To represent the kinematics for force reconstruction, generalized coordinate was used as it can unambiguously describe the MBS configuration, including position and orientation. There is no agreed way to define the generalized coordinate and bi-directional conversion between Cartesian coordinate (or natural coordinate defined by the captured motion data) and generalized coordinate.

The generalized coordinate of a rigid part with 6 degrees of freedom (DOF) in 3D space consists of 6 elements, 3 representing position and 3 for orientation. The 3 position components are usually expressed in a Cartesian coordinate. The 3 orientation components are commonly expressed by Euler angles, quaternions, or exponential map for a robotic system (Liu and Sumit 2011). The parameters may have intuitive meaning in the specific applications but do not apply in human motion modeling. For example, Euler angles express a joint angle (between two connected links) by three angle parameters that the joint angle can be achieved by rotating the link about three orthogonal axes at these three angles successively. This is achievable by a programmable robotic system but remains challenging to parameterize in modeling human motion explicitly.

In biomechanics, segment coordinate systems were well developed to model human motion incorporating the intuitive meaning of parameters (Doriot and Chèze 2004; Dumas et al. 2007; Dumas and Chèze 2007). The most widely accepted one is Joint Coordinate System (JCS) recommended by the International Society of Biomechanics (ISB) (Merico et al. 2002). It defines the origin of the coordinate system at the proximal point of a body part and aligns one axis with the rotation axis of the part. As JCS still relies on the rigid volumetric body part, it cannot be fully adopted in this study that a body part is represented with a non-volumetric line. It lacks a kinematic model that incorporates the advantages of JCS with intuitive biomechanical parameters and works with simplified vision-based human model.

5.3 Method

Inspired by Pham et al. (2018), this study proposed a two-stage force estimation framework with a prior additive module to convert vision-based human motion data to an innovative kinematic

model (Figure 5.1). This conversion module facilitates force reconstruction with physics-based equations and the following neural network-based force optimization.

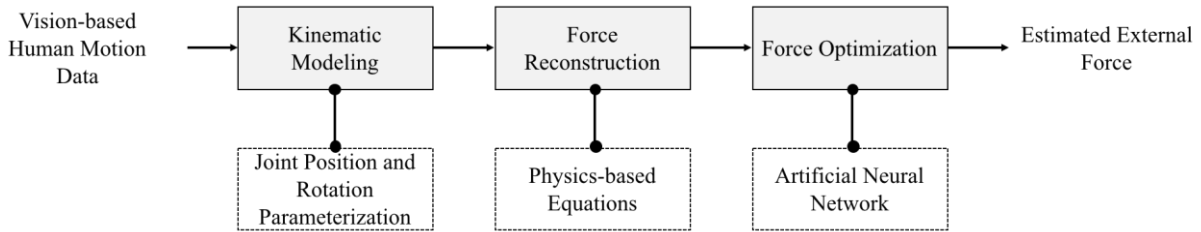


Figure 5.1 Framework of Force Estimation from Vision-based Human Motion Capture

5.3.1 Kinematic Model for Force Estimation

Among generalized coordinates that can represent a rigid body’s configuration unambiguously, there are absolute coordinates, relative coordinates, and natural coordinates to select from. A joint angle in biomechanical analysis can be broken down to three postural angles (i.e., flexion/extension, abduction/adduction, and axial rotation [NIOSH 2014]). These angles are defined as relative angles from the parent or proximal body part. To incorporate this intuitive meaning of angles, the relative coordinate approach was selected and used the three postural angles to represent the rotation of a body part, following Zatsiorsky and Zaciorskij (2002). As for the other 3 position components in a coordinate, the body part’s proximal point was defined as the origin of the coordinate system, following most of such models developed in the biomechanics field.

It needs three postural angles to represent a rigid volumetric body part, with 3 DOF. However, for the simplified vision-based human model, it represents link-shape parts by non-volumetric lines and cannot express axial rotation, would have a smaller number of DOF. An upper limb was taken as an example to show how the simplified model was defined while assigning the center of the pelvis as the root joint of the whole body. Following the way

Zatsiorsky and Zaciorskij (2002) representing the rotation matrix of a rigid body with the product of three rotation matrix about the postural angles, the rotation matrix can be calculated by

$$R_\phi = \begin{bmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix}, R_\theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix}, R_\psi = \begin{bmatrix} \cos \psi & 0 & \sin \psi \\ 0 & 1 & 0 \\ -\sin \psi & 0 & \cos \psi \end{bmatrix}$$

$$R = R_\phi R_\theta R_\psi = \begin{bmatrix} \cos \phi \cos \psi - \sin \psi \sin \theta \sin \phi & \sin \phi \sin \psi - \sin \psi \sin \theta \cos \phi & -\sin \psi \cos \theta \\ -\cos \theta \sin \phi & \cos \theta \cos \phi & \sin \theta \\ \sin \psi \cos \phi + \cos \psi \sin \theta \cos \phi & \sin \psi \sin \phi - \cos \psi \cos \theta \cos \phi & \cos \psi \cos \theta \end{bmatrix}$$

where ϕ is flexion/extension, θ is abduction/adduction, and ψ is axial rotation.

The advantage of this coordinate system is not only that it is parameterized by the three intuitive postural angles. Additionally, it is also very flexible for simplified human motion data. As the loss of DOF could only be axial rotation, the kinematics model could be easily modified by removing some of the R_ψ from the rotation matrix R . In the converted kinematic model for this study, the hip has 3 DOF and its rotation matrix $R_{hip} = R_\phi R_\theta R_\psi$, the shoulder has 3 DOF and $R_{shoulder} = R_\phi R_\theta R_\psi$, and elbow has only 1 DOF (flexion/extension) and $R_{elbow} = R_\phi$.

With the defined number of postural angles for each joint, the generalized coordinate of the whole-body system can be represented by concatenating the postural angles. With the translational terms, the transformation matrix could be formulated by combining rotation and translation matrix. With the transformation matrix, the Jacobian matrix J could be represented by

$$V = \begin{bmatrix} v \\ \omega \end{bmatrix} = \begin{bmatrix} J_v \\ J_\omega \end{bmatrix} \dot{q} = J \dot{q}$$

where v is the linear velocity of a joint and ω is the angular velocity of a joint, and \dot{q} is the generalized velocity. Up to this step, the relationship between the Cartesian coordinate of body joints and the generalized coordinate q is formulated and can proceed to force reconstruction using physics-based equations.

5.3.2 Physics-based Force Reconstruction

According to Liu and Sumit (2011) that directly applying Newton's second law on a complex articulated rigid body system, therefore, Lagrange's equations derived from D'Alembert's principle were used to describe the dynamics of motion.

The primary equation this section applied is equations of motion in vector form:

$$M(q)\ddot{q} + C(q, \dot{q}) = Q$$

where $M(q)$ is the mass matrix, \ddot{q} is the generalized acceleration, $C(q, \dot{q})$ is the Coriolis and centrifugal term, and Q is the vector of generalized forces.

Using the transformation from the Cartesian coordinate to the generalized coordinate, equations of motion could be formulated as:

$$(J^T M_c J)\ddot{q} + (J^T M_c \dot{J} + J^T [\tilde{\omega}] M_c J)\dot{q} = J_v^T f + J_w^T \tau$$

This equation is essentially identical with the prior one, and thus the mass matrix, Coriolis and centrifugal term, as well as the generalized forces can be represented as:

$$\begin{aligned} M(q) &= J^T M_c J \\ C(q, \dot{q}) &= (J^T M_c \dot{J} + J^T [\tilde{\omega}] M_c J)\dot{q} \\ Q &= J_v^T f + J_w^T \tau \end{aligned}$$

Detailed information about representing M_c , J_v , J_w etc. could refer to Liu and Sumit (2011).

5.3.3 Force Optimization

Physics-based equations could only model the "physically plausible" distribution of force (Pham et al. 2018) while there remains a gap to estimate the actual force due to its issue of indeterminacy. It is also intuitive to sense this issue that one can usually impose a larger force even if the same body movements were allowed. To address this issue, this study followed the idea from Pham et al. (2018) to apply an artificial neural network, to learn the latent mapping

between motion data with reconstructed force to actual forces. Among different types of neural networks, recurrent neural network (RNN) was selected as it effectively incorporates temporal information behind time-series data like the studied dynamic forces.

It was explained that to apply RNN, motion data and reconstructed forces were encoded to feature vectors for every time-frame. The encoding method could not be adopted due to the aforementioned issue that the existing literature modeled the human kinematics in a more complex manner, and the simplified vision-based human model lacks specific motion data representation to replicate the validated approach. Therefore, a compatible feature extraction method and corresponding RNN structure were devised to encode the motion data that can feed the network along with the reconstructed force.

The feature vector for every image frame incorporated the reconstructed external forces and torques exerted on the contacting hand, and both grounded feet. Motion data was also included in the feature vectors. This study focused on push force as it is a prevalent task with high exposure to ergonomic risk while its estimation relies on instrumentation (Seo 2016). Consequently, the motion data extracted as the feature was the wrist's location of the primary hand (e.g., right hand) performing the pushing task.

The RNN structure consisted of four layers: long short-term memory (LSTM) layer, fully connected layer, dropout layer, and another fully connected layer Figure 5.2. LSTM is a type of RNN, and a common such unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is where “memory” is stored, and the three gates are “regulators” that control the flow of information depends on its feasibility to be reserved. LSTM was found to work well on processing time-series data while learning the latent temporal pattern.

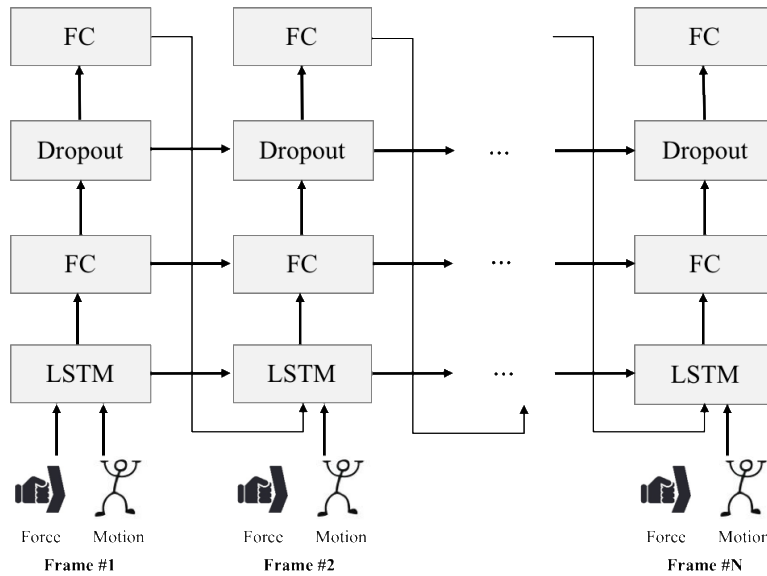


Figure 5.2 Recurrent Neural Network Structure for Force Optimization

5.4 Lab Testing

To test the feasibility of the proposed approach, lab testing was conducted. With the subject simulating a task of pushing a cart, vision-based force estimation was validated against the actual force exerted on the subject's hand collected by a 6 DOF force transducer.

5.4.1 Testing Condition

The same testing condition with that in Chapter 3 was applied in this study, as shown in Figure 5.3. In addition to the motion data collected from the marker-based system (OptoTrak™, Northern Digital, Inc., Waterloo, Canada), this study also recorded hand push force and ground reaction forces.

To collect the hand push force including both magnitude and direction, 6 DOF force transducer (model Mini45, ATI Industrial Automation, Inc., Apex, USA) was selected which provided tri-axis force and tri-axis torque. The force transducer was attached to the subject's right hand with specific instructions to impose the pushing force directly on it. As the subjects wore a glove outside the force transducer, it was torn before the pushing to eliminate the impact of any contact, e.g., support from the glove. Ground reaction forces exerted on both feet were recorded with two 6 DOF force plates (model AccuGait Optimized, Advanced Mechanical Technology, Inc., Watertown, USA).

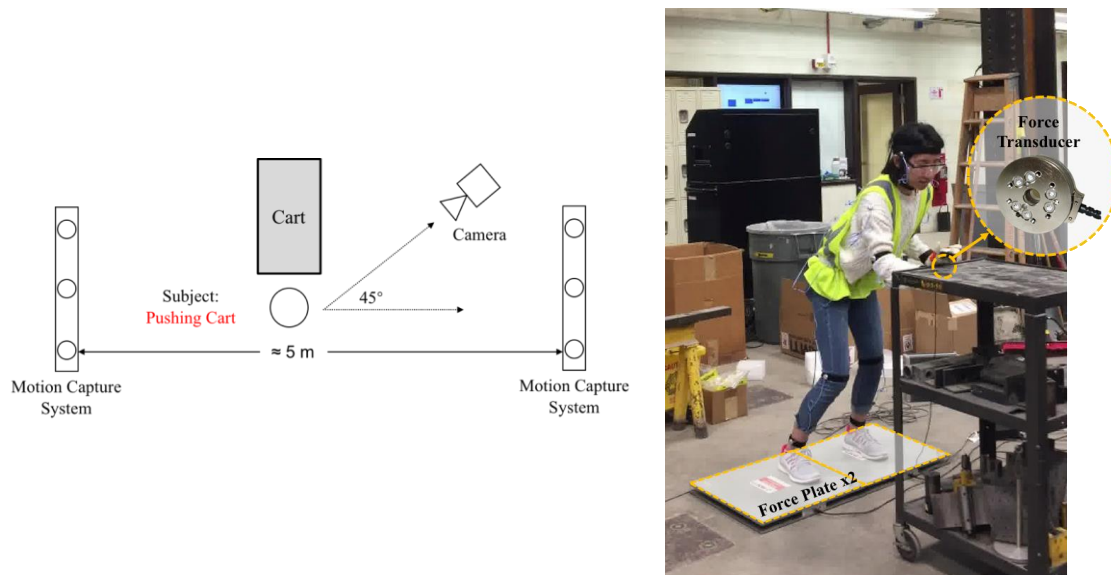


Figure 5.3 Test Setup for Force Estimation

5 subjects participated in this test to provide a diversity of stature and ways of pushing. The subjects' height ranges from 160 to 190 cm. As they pushed the same cart, subjects had to drive their body parts differently by presenting with diverse body motion and different hand push force regarding both magnitude and direction. Collectively 14 cycles of pushing tasks were

performed, with completely free of hand contact with the cart between each cycle to eliminate the impact of inertia accumulated from the prior cycles.

5.4.2 Measure of Accuracy

Every cycle of an evaluated pushing task starts from the subject's natural standing pose, through the forceful exertion during pushing, until returning to a next standing pose. Identifying the time frame a forceful exertion happens and its peak magnitude are critical information for ergonomic risk assessment. The peak force can be used to analyze the severity of the exertion, while the time of an exertion's occurrence helps identify the combined impact from co-occurred forceful exertion and awkward posture.

To facilitate the analysis, the testing videos are trimmed into individual cycles of pushing tasks. The peak force and its time frame are identified and compared. Given the frame rate of the videos as 30 fps, the time of peak force is calculated from the frame index divided by the frame rate. In the testing tasks, tri-axial force is collected from both the proposed approach and the force transducer. As the only the force component orthogonal to the contacting surface has a noteworthy magnitude of value, the force estimation focuses on this component that is most critical to evaluate a pushing task.

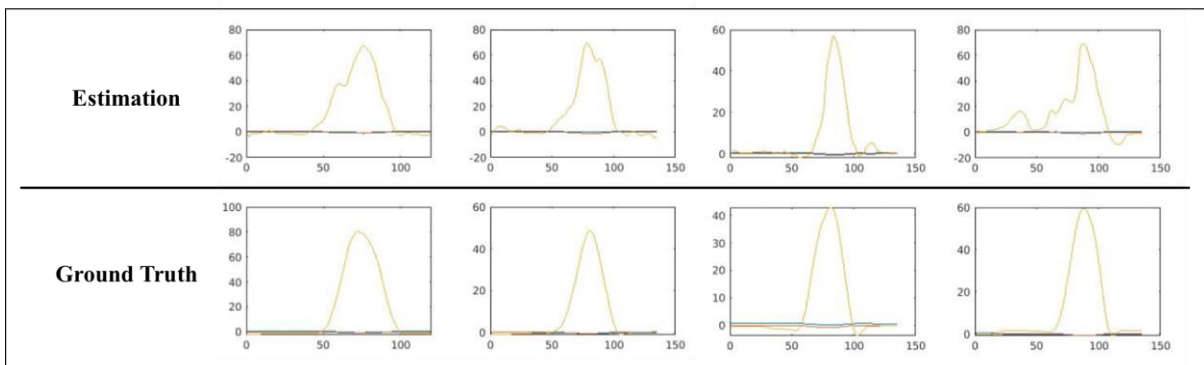


Figure 5.4 Examples of Force Profile: Estimation VS. Ground Truth

5.4.3 Testing Result

Figure 5.4 shows several examples of the estimated force profile from the proposed video-based approach against the ground truth data collected from the force transducer. Table 5.1 shows the comparison of force estimation between the proposed method and the force transducer, regarding the time frame of peak force and its magnitude. In summary, within a cycle with an average duration of 3.8 seconds, the mean absolute error in time estimation of peak force occurrence is 0.3 seconds, taking 7.9% of the total duration. The peak force magnitude estimation has a mean error of 9.5 N, taking 23.6% of the average magnitude.

Table 5.1 Performance Comparison of Force Estimation

Cycle ID	Time Frame of Peak Force (Sec.)			Magnitude of Peak Force (N)		
	Estimation	Ground Truth	Diff.	Estimation	Ground Truth	Diff.
1	2.2	2.1	0.1	29.4	33.9	4.6
2	1.9	1.6	0.3	33.0	46.1	13.1
3	3.3	4.0	0.7	47.7	33.9	13.9
4	4.3	4.5	0.2	34.7	44.6	9.9
5	4.5	4.8	0.3	27.5	34.8	7.3
6	4.6	5.4	0.8	12.7	18.3	5.6
7	5.0	5.1	0.1	29.3	27.0	2.3
8	5.7	5.6	0.1	14.0	24.9	10.9
9	2.5	2.4	0.1	67.5	80.2	12.7
10	2.6	2.7	0.1	69.3	48.7	20.6
11	2.8	2.7	0.1	57.0	42.9	14.1
12	2.9	2.9	0.0	69.1	59.6	9.6
13	5.1	4.9	0.2	34.8	39.8	4.9
14	5.1	3.9	1.2	33.5	30.4	3.1
Mean	3.8	3.8	0.3	40.0	40.3	9.5
STD.	1.3	1.3	0.4	19.1	15.7	5.2

Bland-Altman plot is a graphical method to show the agreement between two methods, especially if one of them is a reference or “golden standard” method. In this test, the force data collected from the force transducer is considered as the reference, and the agreement between it

and the force estimation by the proposed video-based method is explored. Bland-Altman plot is used to display the mean difference between the two methods plotted against the average of the two, as shown in Figure 5.5. In the left figure, the assay measurement is the time frame of the peak force, and in the right figure, the measurement is the magnitude of the peak force. For the time frame of the peak force, the middle horizontal line is close to zero which suggests little evidence of systematic bias of the proposed method compared to the reference one. The higher and lower horizontal lines show the upper and lower bounds of “limit of agreement” and displays the range of difference of future measurements with a 95% confidence level. It can be interpreted that for 95% of the time, the future estimation of the time frame of the peak force by the video-based approach should have an error of less than 0.92 seconds. Similarly, from the right figure, it is suggested that the peak force magnitude estimation of the video-based approach has little evidence of systematic bias and the future estimation should have less than 21N of error compared to the force transducer. Although these two numbers are significantly larger than the average error calculated from the test, less than a one-second error in peak force’s occurrence time is considered to be highly accurate. The error of peak force magnitude estimation of 21N may seem large and will be further analyzed regarding its impact on the accuracy of biomechanical analysis.

5.5 Conclusion

In this chapter, this study proposed a continuous tri-axial hand push force estimation framework from the 3D human motion data captured by the video-based approach. To address the knowledge gap of estimating the hand push force with a simplified human motion model of reduced DOF, compared with that from a body-attached motion capture system with full DOF, a

kinematic model is developed that can estimate the force data yet with intuitive biomechanical parameters. Subsequently, physics-based plausible force is estimated by equations of motion. It is followed by a recurrent neural network to estimate the actual force exertion by addressing the indeterminacy issue of force with motion data only.

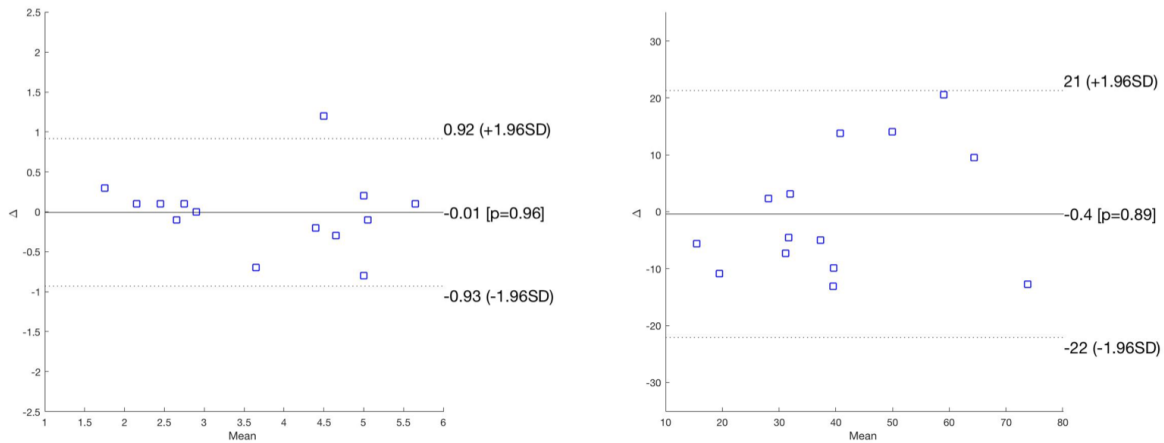


Figure 5.5 Bland-Altman Plot of Estimation and Groundtruth Agreement on Force Estimation

Lab testing was conducted to demonstrate the feasibility of estimating hand push force with the video-based 3D motion capture. The ground truth data for the validation is provided by a 6 DOF force transducer attached to the subjects' hand. The test yields an error of 0.3 (7.9%) seconds in time frame estimation of peak force's occurrence and a peak force estimation error of 9.5N (23.6%). With the Bland-Altman plot, it is interpreted that the peak force occurrence time estimation does not present apparent systematic bias from the data collected by the force transducer and expect an error of within 0.92 seconds for future estimation with 95% confidence level. The peak force magnitude quantification does not embed a systematic bias and can expect an accuracy level of within 21N difference in a future test. The peak force occurrence time and magnitude are critical measurements for biomechanical analysis to evaluate the ergonomic risk

level of individual body joint. The test result shows a great potential to estimate hand push force with an ordinary camera to quantify the force exertion for ergonomic risk assessment.

Chapter 6 Application of 3D Motion Capture and Force Estimation in Biomechanical Analysis

6.1 Introduction

To obtain a quantitative measurement of risk level, regarding individual body joints, biomechanical analysis is used to estimate the internal load exerted on each body joint. The required input data for such analysis primarily focuses on human motion data expressed by whole-body joint location or angle, along with external force with its magnitude and direction. As being costly to collect such data, biomechanical analysis is commonly applied with the direct measurement as the data collection approach. The major type of sensors involved include, but are not limited to, motion data capture system (e.g., optical marker-based, accelerometer-based) and force data collection devices (e.g., force gauge, load cell, pressure mat).

Many studies in construction have demonstrated how biomechanical analysis can help improve the occupation. Seo et al. (2014) developed an application to leverage a commonly used human motion data format (.bvh), especially applied by an RGB-D sensor (Microsoft Kinect), converted into a compatible input format for the two most prevalent biomechanical analysis software, and evaluated a simulated masonry work's joint moments on L5/S1 (an intervertebral disc between the fifth lumbar and first sacral vertebra), knees and elbows. Golabchi et al. (2015) used a human motion modeling tool (3ds Max) to collect motion data for biomechanical simulation in workplace design, aiming to prevent the designed job's excessive exposure to ergonomic risks. Yu et al. (2019) developed computer vision-based 3D human motion capture

method with smart insoles to collect posture and force data simultaneously, and conduct a biomechanical analysis of simulated tasks (i.e., brick lifting, rebar tying, and plastering). These studies also suggested an emphasis on the data collection process, especially human motion and external force data.

Seo (2016) suggested that, to conduct biomechanical analysis with a reliable accuracy of 10% error, the motion data should achieve an accuracy of joint angle estimation to be less than 10° of error. Despite that the proposed 3D human motion capture approach in a preceding chapter (Chapter 3) is demonstrated to achieve this level of accuracy regarding joint angle estimation, it remains unknown how the additive error in estimated force data could simultaneously impact the biomechanical analysis result. This study aims to explore this question by applying the video-based motion and force data collection approach in biomechanical analysis.

6.2 Computerized Biomechanical Analysis Tools

Among the spectrum of biomechanical models, some are static models (Chaffin and Baker 1970; Garg and Chaffin 1975; Martin and Chaffin 2007) that only require body joints locations for static postures. The limitation of these models is the ignorance of inertial loads exerted on body parts due to dynamic postures. Dynamic biomechanical models (Marras and Sommerich 1991), can handle this scenario by utilizing joints' velocity and acceleration. As the velocity and acceleration require true-to-scale measurement (e.g., in m/s), thus necessitates the motion data in 3D space.

3D SSPPTM (Three-Dimensional Static Strength Prediction ProgramTM) is a widely used software for static biomechanical analysis, with its GUI shown in Figure 6.1. It utilizes frame-

wise static human posture without considering the velocity, acceleration, and their impact. Its major strength is the ability to not only quantify the exerted internal loads on the individual joint, but also a quantitative risk level of a joint by comparing the forceful exertion with the relevant human capacity (NIOSH 1981).

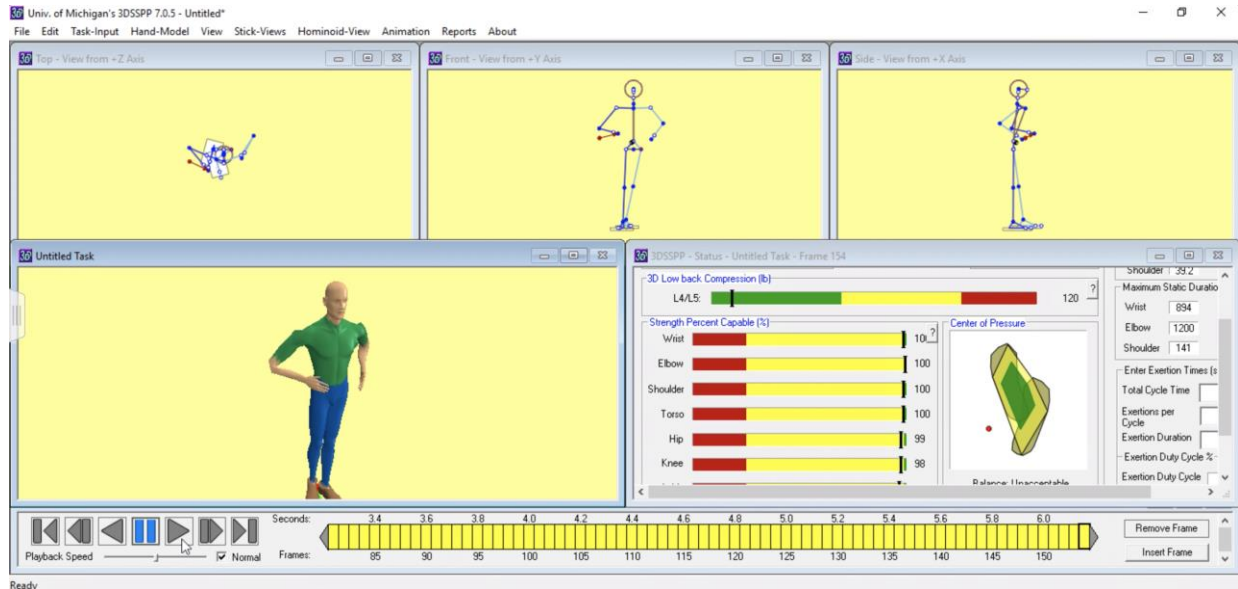


Figure 6.1 Screenshot of Biomechanical Analysis Tool 3D SSPP™

Apart from whole-body motion data expressed by joint location, external contacting force data exerted on hands is the other key input. As 3D SSPP™ models the human in 3D space, the force's magnitude and direction are both needed and can be visualized by a red arrow contacting the animated subject's hand. The direction of force is expressed by a tri-axial variable with three components aligned with the three orthogonal axes.

The proposed posture and force data collection approach is developed for dynamic biomechanical analysis that requires a frame-wise data stream to estimate velocity and acceleration. Due to the motion data format's accuracy issue of OpenSim, this study selects the

static biomechanical analysis tool 3D SSPP™ to validate the impact of input risk factors data on the output analysis result.

6.3 Method

To utilize 3D SSPP™ to conduct an automated biomechanical analysis with available 3D human motion data and tri-axial forces exerted on the hands, there are mainly three steps to perform: 1) Identify the conversion rule of 3D coordinate system from motion and force data to that used in 3D SSPP™. 2) Convert motion and force data to the required input data format. 3) Generate a batch file to run the biomechanical analysis process on a frame sequence automatically.

To minimize the effort of identifying how to convert the motion and force data to the coordinate system used by 3D SSPP™, this study deliberately defined a 3D coordinate system that all the devices' coordinate systems share its axes directions (Figure 6.2). The conversion between the coordinate system becomes effort-saving to only focus on the correct indexing of an axis's label and its positive direction.

3D SSPP™ has three acceptable formats of motion data for biomechanical analysis. The three formats are LOC, SLOC, and LLOC file, with a descending complexity of the required types of motion data. For example, both LOC and SLOC require a grip center apart from wrist location, that the 3D motion data from the proposed video-based approach does not provide. Therefore, the LLOC file is selected, but it does not allow the assignment of elbow and knee locations, which could only be generated by the embedded posture prediction feature.

3D SSPP™ is a static biomechanical analysis tool and does not automatically run the biomechanical analysis process on the video frames successively. Thus, a batch file needs to be

generated to program the automatic triggering of the analysis process while exporting the summary report frame by frame.

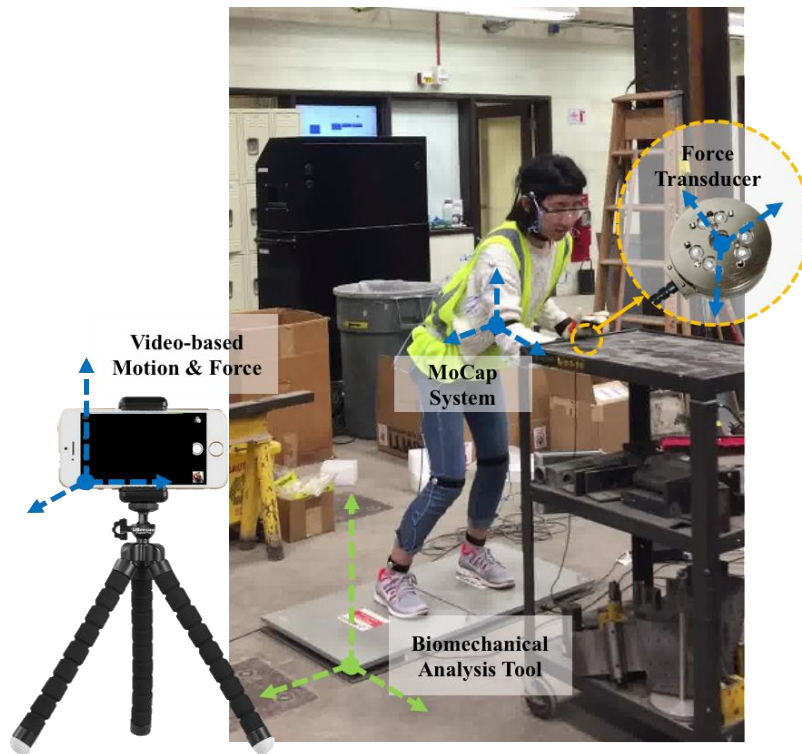


Figure 6.2 Coordinate Systems Involved for Biomechanical Analysis

6.4 Lab Testing

The setting is inherited from that for the preceding Chapter 5. Four cycles of pushing tasks of a male subject are used for the test. This subject has an overall joint angle estimation accuracy almost identical to the average value across the 10 participating subjects and is believed to be representative for a preliminary study.

A primary output of biomechanical analysis in SSPP™ is 3D low back compression (unit: lb.). As the elbow and knee locations are predicted and highly unreliable, the estimated load on joints from limbs may not be reflecting the performance. Thus, the low back compression is selected as the estimation variable to evaluate the performance.

Following the force estimation, the measures of accuracy include time frame difference of peak compression's occurrence, the corresponding peak magnitude, and mean absolute error of frame-wise compression's magnitude. Between the two assay methods, the reference measurement is 3D low back compression estimation from 3D SSPP with hand push force collected by the force transducer and the 3D motion data collected by the marker-based motion capture system (OptoTrak). The other assay method is the proposed one, and the measurement is 3D low back compression estimation generated from video-based hand push force and 3D human motion data capture.

6.5 Testing Result

Table 6.1 shows the low back compression estimation between the biomechanical analysis result generated by video-based force estimation and that by force data collected from the force transducer. The test results yield an average of 0.2 seconds (6.4%) in peak compression occurrence time on the low back, and a 121.2lb (12.6%) mean absolute difference on peak compression's magnitude.

The accuracy of peak compression occurrence time and magnitude are both around 12.5%, but the standard deviation has a much smaller scale of around 5%. This suggests a systematic bias in the estimation but with very high precision. Given the 3D human motion data from the video-based approach has a joint angle estimation error of slightly less than 10 degrees, the biomechanical analysis result is not too sensitive to the combined errors from both force and motion data. Based on the findings, a future research thrust could be suggested to explore the root cause of the systematic bias of the low back compression, which might further improve the biomechanical analysis accuracy.

Table 6.1 Accuracy of Biomechanical Analysis on Low Back Compression Estimation

Cycle ID	Time Frame of Peak Compression (Sec.)				Magnitude of Peak Compression (lb.)			
	Estimation	Ground Truth	Diff.	% Diff.	Estimation	Ground Truth	Diff.	% Diff.
1	2.1	1.8	0.3	14.5%	926.7	1064.2	137.5	12.9%
2	2.6	2.8	0.2	6.0%	877.3	939.0	61.7	6.6%
3	2.2	2.8	0.6	20.2%	715.6	871.7	156.1	17.9%
4	2.5	2.8	0.2	8.4%	827.4	956.8	129.4	13.5%
Mean	2.4	2.5	0.3	12.3%	836.8	957.9	121.2	12.6%
STD.	0.2	0.5	0.2	6.4%	90.4	79.8	41.2	4.7%

6.6 Conclusion

Biomechanical analysis is a highly comprehensive ergonomic risk assessment approach. It utilizes whole-body 3D human motion data and external force data and can quantify the internal force and torque exerted on the individual body joint and its risk level. This study demonstrates the feasibility of using both hand push force and 3D human motion data extracted from an ordinary video to conduct biomechanical analysis. To validate the performance of such a non-invasive, rapid, and economical approach, laboratory-based testing is conducted. A commonly used biomechanical analysis software 3D SSPP™ is applied to perform the analysis from provided input of force and motion data. As the primary measurement generated from the software, 3D low back compression is used to represent biomechanical analysis result for validation purpose. In the validation, a marker-based motion capture system (OptoTrak™) provides ground truth motion data, and a hand-attached 6 DOF force transducer provides tri-axial ground truth force data.

The biomechanical analysis result suggests that the proposed approach can estimate the peak compression occurrence time with 0.3 seconds' error and peak compression with 12.6% mean error. The standard deviation of the three measurements are significantly smaller (i.e., ~5%) and suggests a rather high precision in the estimation.

This study demonstrated a great potential of video-based biomechanical analysis that evaluates internal load and the risk level of individual body joints with 3D human motion capture and force estimation with an ordinary camera.

Chapter 7 Conclusion and Recommendation

7.1 Summary of Research

This research effort started with the following overarching research goals: 1) to provide a rapid, non-invasive and accurate human motion capture approach that estimates the risk factors of posture and repetition for ergonomic risk assessment of on-site construction jobs; 2) to provide a non-invasive and non-contact hand force estimation approach to quantify forceful exertion which is critical to ergonomic risk assessment; 3) to apply the collected motion and force data for comprehensive ergonomic risk assessment including postural analysis that focuses on posture data and biomechanical analysis that requires pair-wise posture and force data.

Considering these goals, the research had these three more specific research objectives:

1) to develop and validate a video-based human motion capture framework to quantify ergonomic risk factors of posture and its repetition by extracting continuous 2D/3D human model with enforced kinematic and temporal consistency while handling long-lasting occlusion in construction sites; 2) to propose a hand push/pull force estimation framework with simplified human model of reduced DOF via the proposed video-based 3D human motion capture; and 3) to apply the collected risk factors such as posture, repetition, and force to ergonomic risk assessment tools.

To achieve these research objectives, five inter-related studies were conducted. A summary of these studies' results and implications are as follows.

1. Video-based 2D Human Motion Capture for Posture and Repetition Estimation:

This study introduces a proposed framework to capture continuous 2D human motion from a video and demonstrates its robustness in quantifying the ergonomic risk factors of posture and repetition. 2D body joint location and angle estimation accuracy are validated by manual annotation on images for on-site construction jobs, with 83.2% joint location is correctly estimated and 11.6° joint angle error across body joint and 10 tasks.

This result supports the potential of the developed motion capture framework with the ability to enforce kinematic and temporal consistency to estimate the key risk factors of posture and repetition (frequency and duration) for ergonomic risk assessment while handling long-lasting occlusion in a construction site.

2. Video-based 3D Human Motion Capture for Posture and Repetition Estimation:

This study adopts the framework from the prior chapter with modification to incorporate 3D human motion capture. 3D body joint angle estimation accuracy is validated by marker-based motion capture system for lab-based simulated tasks. The results showed that the collected motion data with less than 10° of error in body angles. The result shows the ability to capture detailed information about human motion compared to the 2D approach (e.g., measuring trunk twisting angle). Additionally, based on the previous study's suggestion that such an accuracy level of motion data could yield a reliable biomechanical analysis result with less than 10% error. The proposed approach demonstrated great potential for reliable biomechanical analysis with video-based 3D motion capture.

3. Applications of Video-based Human Motion Capture on Ergonomic Postural

Analysis: This study shows the feasibility of the motion data capture approach in three types of postural analysis tools requiring quantified risk factors of angle-based, distance-based and a

mixture of both. Three tools from each category are selected. The estimation of frequency and duration for an angle-based postural analysis tool (REBA) achieved comparable accuracy with the average performance of 27 ergonomists' observation with significantly less evaluation time. The estimation of risk level for a distance-based analysis tool (Snook's Tables) from categorical risk factor estimation yields 80% of 10 on-site jobs correctly evaluated compared to analysis result from the tapeline-based measurement. The estimation of risk level for an angle- and distance-based analysis tool (NIOSH Lifting Equation) achieved a mean absolute error of 10.9% in the estimated score compared with tapeline and observation-based measurement for 10 on-site lifting jobs. The results collectively exhibit the immense potential to apply the proposed motion data capture approach for on-site postural analysis to evaluate the ergonomic risks of construction jobs.

4. Video-based Hand Push Force Estimation: This study introduces the proposed hand push force estimation framework from captured whole-body human motion data via a video recording. Lab testing was conducted to demonstrate the feasibility of estimating hand push force with video-based 3D motion capture. The estimated force has a peak force estimation error of 9.5N (23.6%) and its occurrence time estimation error of 0.3 seconds, validated by a 6 DOF force transducer attached to the hand. The result shows a possibility to estimate hand push force with an ordinary camera to quantify the force exertion for ergonomic risk assessment. Further, biomechanical analysis is conducted using the 3D motion and force captured from video. The result, 3D low back compression, is compared with that from the sensor-based approach. The result shows a potential to perform biomechanical analysis with the proposed video-based 3D motion and force estimation approach.

7.2 Future Research

While this work has extended the potential of a video-based comprehensive data collection approach for on-site ergonomic risk assessment, many methodological and technical challenges remain which warrants further attention in future research efforts. A few such questions are listed below.

1. How can the video-based 2D human motion capture framework achieve a higher accuracy level in construction sites similar to that of the state-of-the-art performance on image dataset collected from daily activities?

2. How can the video-based 3D human motion capture framework better infer an occluded body joint location?

3. How can we increase the level of automation in ergonomic risk assessment through the quantification of extended types of risk factors, such as the hand posture including grip?

4. How to estimate force exertion of tasks other than pushing and pulling with the proposed force estimation framework, such as carrying? As for the on-site deployment, especially the motion capture approach, there are several questions and corresponding recommendation for future study to work on:
1. How to improve the processing speed of motion capture approach? The most time-consuming module, a human pose estimation algorithm, applied from the computer vision domain, is designed to handle a high level of diversity in daily activities with complex neural network structure. Jobs from a construction site do not have such diversity in postures and appearance and may not require a similarly complex network structure to handle. An engineering effort could be devoted to simplify the network structure by identifying the most impactful layers and the minimum number of nodes. With a simplified network, the processing speed can be significantly reduced for an on-site application.

2. *How to utilize the on-site ambient cameras for motion capture (e.g., surveillance camera)?* This research uses videos taken from an observer's view, which might incur extra cost for data collection, compared with utilizing ambient cameras under operation on the jobsites. The major engineering effort required is replacing the human pose estimation module with one where the pose estimation model is trained from images collected by ambient cameras whose view angle is significantly different from an observer.

Bibliography

- [1] Bernard B, ed. “Musculoskeletal Disorders and Workplace Factors: A Critical Review of the Epidemiological Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back.” *Cincinnati: DHHS (NIOSH) Publication*, 1997; No. 97 141.
- [2] Bureau of Labor Statistics (BLS) (2018).
- [3] Achilles, F., Ichim, A., Coskun, H., Tombari, F., Noachtar, S., and Navab, N. (2016). “Patient MoCap: Human Pose Estimation Under Blanket Occlusion for Hospital Monitoring Applications.” 491–499.
- [4] Akkas, O., Lee, C. H., Hu, Y. H., Harris Adamson, C., Rempel, D., and Radwin, R. G. (2017). “Measuring exertion time, duty cycle and hand activity level for industrial tasks using computer vision.” *Ergonomics*, Taylor & Francis, 60(12), 1730–1738.
- [5] Alwasel, A. (2017). “Use of Kinematics to Minimize Construction Workers ’ Risk of Musculoskeletal Injury.”
- [6] Alwasel, A., Elrayes, K., Abdel-Rahman, E. M., and Haas, C. (2011). “Sensing Construction Work-Related Musculoskeletal Disorders (WMSDs).” *28th International Symposium on Automation and Robotics in Construction (ISARC 2011)*, 164–169.
- [7] Andriluka, M., Pishchulin, L., Gehler, P. V., and Schiele, B. (2014). “2D Human Pose Estimation - New Benchmark and State of the Art Analysis.” *Cvpr*, 3686–3693.
- [8] Bao, S., Howard, N., Spielholz, P., and Silverstein, B. (2006). “Quantifying repetitive

- hand activity for epidemiological research on musculoskeletal disorders – Part II: comparison of different methods of measuring force level and repetitiveness.” *Ergonomics*, 49(4), 381–392.
- [9] Baradel, F., Wolf, C., and Mille, J. (2017). “Pose-conditioned Spatio-Temporal Attention for Human Action Recognition.”
- [10] Beheshti, M. H., Javan, Z., and Yarahmadi, G. (2016). “Ergonomic Evaluation of Musculoskeletal Disorders in Construction Workers Using Posture, Activity, Tools, Handling (PATH) Method.” *International Journal of Occupational Hygiene*, 8(2), 110–115.
- [11] Berlin, C., and Adams, C. (2017). *Production Ergonomics: Designing Work Systems to Support Optimal Human Performance*. Ubiquity Press Ltd.
- [12] Borg, G. (1990). “Psychophysical scaling with applications in physical work and the perception of exertion.” *Scandinavian Journal of Work, Environment and Health*, 16(SUPPL. 1), 55–58.
- [13] Boschman, J. S., Van Der Molen, H. F., Sluiter, J. K., and Frings-Dresen, M. H. (2012). “Musculoskeletal disorders among construction workers: A one-year follow-up study.” *BMC Musculoskeletal Disorders*.
- [14] Chaffin, D. B., and Baker, W. H. (1970). “A biomechanical model for analysis of symmetric sagittal plane lifting.” *AIIE Transactions*, 2(1), 16–27.
- [15] Chen, C. H., Hu, Y. H., Yen, T. Y., and Radwin, R. G. (2013). “Automated video exposure assessment of repetitive hand activity level for a load transfer task.” *Human Factors*, 55(2), 298–308.
- [16] Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., and

- Chen, B. (2016). “Synthesizing Training Images for Boosting Human 3D Pose Estimation 2. Synthetic Images for Boosting 2D Pose Estimation 2.1. Domain adaptation on 2D Pose Estimation.” *In 3D Vision (3DV), 2016 Fourth International Conference on*, 479–488.
- [17] Cheraghi, M., Farahani, M. S., Ali, S., and Najarkola, M. (2018). “Ergonomic Risk Factors Evaluation of Work-related Musculoskeletal Disorders by PATH and MMH in a Construction Industry Data gathering Methods.” *Irian Journal of Health, Safety & Environment*, 6(1), 1175–1189.
- [18] Cho, Y. K., Kim, K., Ma, S., and Ueda, J. (2018). “A Robotic Wearable Exoskeleton for Construction Worker’s Safety and Health.” *Construction Research Congress 2018*, (April), 19–28.
- [19] Dabral, R., Mundhada, A., Kusupati, U., Afaque, S., Sharma, A., and Jain, A. (2018). “Learning 3D human pose from structure and motion.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11213 LNCS, 679–696.
- [20] Dai, F., and Ning, X. (2013). “Identification of Construction Cumulative Trauma Disorders: A Machine Vision Approach.” 661–668.
- [21] David, G. C. (2005). “Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders.” *Occupational medicine (Oxford, England)*, 55(3), 190–199.
- [22] Diego-Mas, J. A., and Alcaide-Marzal, J. (2014). “Using Kinect™ sensor in observational methods for assessing postures at work.” *Applied Ergonomics*, Elsevier Ltd, 45(4), 976–985.

- [23] Diraco, G., Leone, A., and Siciliano, P. (2013). “Human posture recognition with a time-of-flight 3D sensor for in-home applications.” *Expert Systems with Applications*, Elsevier Ltd, 40(2), 744–751.
- [24] Doriot, N., and Chèze, L. (2004). “A Three-Dimensional Kinematic and Dynamic Study of the Lower Limb during the Stance Phase of Gait Using an Homogeneous Matrix Approach.” *IEEE Transactions on Biomedical Engineering*, 51(1), 21–27.
- [25] Du, H., Manns, M., Herrmann, E., and Fischer, K. (2016). “Joint Angle Data Representation for Data Driven Human Motion Synthesis.” *Procedia CIRP*, Elsevier B.V., 41, 746–751.
- [26] Dumas, R., and Chèze, L. (2007). “3D inverse dynamics in non-orthonormal segment coordinate system.” *Medical and Biological Engineering and Computing*, 45(3), 315–322.
- [27] Dumas, R., Chèze, L., and Verriest, J. P. (2007). “Adjustments to McConville et al. and Young et al. body segment inertial parameters.” *Journal of Biomechanics*, 40(3), 543–553.
- [28] Dzung, R. J., Hsueh, H. H., and Ho, C. W. (2017). “Automated Posture Assessment for construction workers.” *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings*, 1027–1031.
- [29] Everett, J. G. (1999). “Overexertion Injuries in Construction.” *Journal of Construction Engineering and Management*, 125(2), 109–114.
- [30] Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., and Cucchiara, R. (2018). “Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World.”

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11208 LNCS, 450–466.

- [31] Fransson-Hall, C., Gloria, R., Kilbom, Å., Winkel, J., Karlqvist, L., and Wiktorin, C. (1995). “A portable ergonomic observation method (PEO) for computerized on-line recording of postures and manual handling.” *Applied Ergonomics*, 26(2), 93–100.
- [32] Gaddam, S. P. R., Chippa, M. K., Sastry, S., Ange, A., Berki, V., and Davis, B. L. (2015). “Estimating forces during exercise activity using non-invasive kinect camera.” *Proceedings - 2015 International Conference on Computational Science and Computational Intelligence, CSCI 2015*, 825–828.
- [33] Garg, A., and Chaffin, D. B. (1975). “A biomechanical computerized simulation of human strength.” *AIIE Transactions*, 7(1), 01-15.
- [34] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., and Tran, D. (2018). “Detect-and-Track: Efficient Pose Estimation in Videos.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 350–359.
- [35] Girshick, R. (2015). “Fast R-CNN.” *IEEE International Conference on Computer Vision (ICCV 2015)*, 1440–1448.
- [36] Gkioxari, G., Toshev, A., and Jaitly, N. (2016). “Chained predictions using convolutional neural networks.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9908 LNCS, 728–743.
- [37] Golabchi, A., Han, S., Seo, J., Han, S., Lee, S., and Al-Hussein, M. (2015). “An Automated Biomechanical Simulation Approach to Ergonomic Job Analysis for Workplace Design.” *Journal of Construction Engineering and Management*, 141(8), 04015020.

- [38] Gong, J., Caldas, C. H., and Gordon, C. (2011). “Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models.” *Advanced Engineering Informatics*, 25(4), 771–782.
- [39] Greene, R. L., Hu, Y. H., Difrancio, N., Wang, X., Lu, M. L., Bao, S., Lin, J. H., and Radwin, R. G. (2019). “Predicting Sagittal Plane Lifting Postures From Image Bounding Box Dimensions.” *Human Factors*, 61(1), 64–77.
- [40] Gupta, A., Mittal, A., and Davis, L. S. (2008). “Constraint integration for efficient multiview pose estimation with self-occlusions.” *IEEE transactions on pattern analysis and machine intelligence*, 30(3), 493–506.
- [41] Han, S., and Lee, S. (2013). “A vision-based motion capture and recognition framework for behavior-based safety management.” *Automation in Construction*, 35, 131–141.
- [42] Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15.
- [43] Jacobs, D. A., and Ferris, D. P. (2015). “Estimation of ground reaction forces and ankle moment with multiple, low-cost sensors.” *Journal of NeuroEngineering and Rehabilitation*, Journal of NeuroEngineering and Rehabilitation, 12(1), 1–12.
- [44] Jahanbanifar, S., and Akhavian, R. (2019). “Evaluation of wearable sensors to quantify construction workers muscle force: An ergonomic analysis.” *Proceedings - Winter Simulation Conference*, IEEE, 2018–Decem, 3921–3929.
- [45] Jain, A., Tompson, J., Andriluka, M., Taylor, G. W., and Bregler, C. (2014). “Learning Human Pose Estimation Features with Convolutional Networks.” 1–11.
- [46] Jebelli, H., and Lee, S. (2019). “Advances in Informatics and Computing in Civil and

- Construction Engineering.” (March).
- [47] Kobayashi, S., Kaseda, H., and Miyamoto, R. (2018). “Robust Localization of Body Parts Based on Interframe Failure Correction.” *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, IEEE, 3903–3908.
- [48] Kuorinka, I., Jonsson, B., Kilbom, A., Vinterberg, H., Biering-Sørensen, F., Andersson, G., and Jørgensen, K. (1987). “Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms.” *Applied Ergonomics*, 18(3), 233–237.
- [49] Li, X., Han, S., Gül, M., Al-Hussein, M., and 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.
- [50] Li, X., Han, S. H., Gül, M., and Al-Hussein, M. (2019). “Automated post-3D visualization ergonomic analysis system for rapid workplace design in modular construction.” *Automation in Construction*, 98(November 2018), 160–174.
- [51] Liang, W., Zhu, Y., and Zhu, S.-C. (2018). “Tracking Occluded Objects and Recovering Incomplete Trajectories by Reasoning about Containment Relations and Human Actions.” 7106–7113.
- [52] Liu, C. K., and Sumit, J. (2011). “A Quick Tutorial on Multibody Dynamics.” 1–25.
- [53] Liu, M., Han, S., and Lee, S. (2017). “Potential of Convolutional Neural Network-Based 2D Human Pose Estimation for On-Site Activity Analysis of Construction Workers.” *computing in civil engineering*, 3, 141–149.
- [54] Liu, M., Hong, D., Han, S., and Lee, S. (2016). “Silhouette-Based On-Site Human Action Recognition in Single-View Video.” *Construction Research Congress 2016: Old and*

New Construction Technologies Converge in Historic San Juan - Proceedings of the 2016 Construction Research Congress, CRC 2016.

- [55] Lu, Y., and Dai, S. (2018). “Motion Transition Based on Bézier Quaternion Curve.” *Lecture Notes in Electrical Engineering*, 458, 663–671.
- [56] Mahyuddin, A. I., Mhradi, S., Dirgantara, T., and Maulido, P. N. (2011). “Gait Parameters Determination by 2D Optical Motion Analyzer System.” *Applied Mechanics and Materials*, 83, 123–129.
- [57] Marras, W. S., and Sommerich, C. M. (1991). “A three-dimensional motion model of loads on the lumbar spine: I. Model structure.” *Human Factors*, 33(2), 123–137.
- [58] Martin, J. B., and Chaffin, D. B. (2007). “Biomechanical Computerized Simulation of Human Strength in Sagittal-Plane Activities.” *A I I E Transactions*, 4(1), 19–28.
- [59] Merico, A., Levedianos, G., Gallo, P., Zanco, P., Piccione, F., and Tonin, P. (2002). “ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine.” *Journal of Biomechanics*, 8(2), 179–179.
- [60] Nath, N. D., Akhavian, R., and Behzadan, A. H. (2017). “Ergonomic analysis of construction worker’s body postures using wearable mobile sensors.” *Applied Ergonomics*, Elsevier Ltd, 62, 107–117.
- [61] Newell, A., Yang, K., and Deng, J. (2016). “Stacked hourglass networks for human pose estimation.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9912 LNCS, 483–499.
- [62] NIOSH. (2014). “Observation-Based Posture Assessment: Review of Current Practice and Recommendations for Improvement.”

- [63] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., and Murphy, K. (2017). “Towards Accurate Multi-person Pose Estimation in the Wild.” *Computer Vision and Pattern Recognition, 2017. Proceedings of the 2017 IEEE Computer Society Conference on*, 4903–4911.
- [64] Pham, T. H., Caron, S., and Kheddar, A. (2018). “Multicontact Interaction Force Sensing From Whole-Body Motion Capture.” *IEEE Transactions on Industrial Informatics*, 14(6), 2343–2352.
- [65] Pham, T. H., Kheddar, A., Qammaz, A., and Argyros, A. A. (2015). “Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07–12–June, 2810–2819.
- [66] Radwan, I., Dhall, A., and Goecke, R. (2013). “Monocular image 3D human pose estimation under self-occlusion.” *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 1888–1895.
- [67] Raison, M. (2009). “On the Quantification of Joint and Muscle Efforts in the Human Body During Motion.” *Engineer*.
- [68] Ramakrishna, V., Kanade, T., and Sheikh, Y. (2012). “Reconstructing 3D human pose from 2D image landmarks.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7575 LNCS(PART 4), 573–586.
- [69] Ray, S. J., and Teizer, J. (2012). “Real-time construction worker posture analysis for ergonomics training.” *Advanced Engineering Informatics*, Elsevier Ltd, 26(2), 439–455.
- [70] Ren, S., He, K., Girshick, R., and Sun, J. (2015). “Faster R-CNN: Towards Real-Time

- Object Detection with Region Proposal Networks.” *Advances in neural information processing systems*, 91–99.
- [71] Rogez, G., and Schmid, C. (2016). “MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild.”
- [72] Russell, S. J., Winnemuller, L., Camp, J. E., and Johnson, P. W. (2007). “Comparing the results of five lifting analysis tools.” *Applied Ergonomics*, 38(1), 91–97.
- [73] Saha, P., Basu, B., and Devashish Sen, D. (2017). “Ergonomic evaluation of physiological stress of building construction workers associated with manual material handling tasks.” *Progress in Health Sciences*, 7(1), 54–62.
- [74] Sarafianos, N., Boteanu, B., Ionescu, B., and Kakadiaris, I. A. (2016). “3D Human pose estimation: A review of the literature and analysis of covariates.” *Computer Vision and Image Understanding*, 152(October 2017), 1–20.
- [75] Sárándi, I., Linder, T., Arras, K. O., and Leibe, B. (2018). “How Robust is 3D Human Pose Estimation to Occlusion?” 1–5.
- [76] Sartison, A., Mironov, D., Youcef-Toumi, K., and Tsetserukou, D. (2018). “Finger Grip Force Estimation from Video using Two Stream Approach.”
- [77] Seo, J. (2016). “Evaluation of construction workers’ physical demands through computer vision-based kinematic data collection and analysis.” *ProQuest Dissertations and Theses*, 144.
- [78] Seo, J. O., Alwasel, A., Lee, S. H., Abdel-Rahman, E. M., and Haas, C. (2017). “A comparative study of in-field motion capture approaches for body kinematics measurement in construction.” *Robotica*, 1–19.
- [79] Seo, J., Starbuck, R., Han, S., Lee, S., and Armstrong, T. J. (2014). “Dynamic

- Biomechanical Analysis for Construction Tasks Using Motion Data from Vision-Based Motion Capture Approaches.” *INTERNATIONAL CONFERENCE ON COMPUTING IN CIVIL AND BUILDING ENGINEERING*, 1005–1012.
- [80] Seo, J., Yin, K., and Lee, S. (2016). “Automated Postural Ergonomic Assessment Using a Computer Vision-Based Posture Classification.” *Construction Research Congress 2016*, 809–818.
- [81] Simo-Serra, E., Ramisa, A., Alenya, G., Torras, C., and Moreno-Noguer, F. (2012). “Single image 3D human pose estimation from noisy observations.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2673–2680.
- [82] Snook, S. H., and Ciriello, V. M. (1991). “The design of manual handling tasks: Revised tables of maximum acceptable weights and forces.” *Ergonomics*, 34(9), 1197–1213.
- [83] Starbuck, R., Seo, J., Han, S., and Lee, S. (2014). “A Stereo Vision-based Approach to Marker-less Motion Capture for On-Site Kinematic Modeling of Construction Worker Tasks.” *Computing in Civil and Building Engineering*, 1094–1101.
- [84] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). “Efficient object localization using Convolutional Networks.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07–12–June, 648–656.
- [85] Toshev, A., and Szegedy, C. (2014). “DeepPose: Human Pose Estimation via Deep Neural Networks.” *IEEE Conference on Computer Vision and Pattern Recognition*.
- [86] Urtasun, R., and Fua, P. (2004). “3D Tracking for Gait Characterization and Recognition.” *In Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 17–22.

- [87] Wang, C., Wang, Y., Lin, Z., Yuille, A. L., and Gao, W. (2014). “Robust estimation of 3D human poses from a single image.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (013), 2369–2376.
- [88] Wang, D., Dai, F., and Ning, X. (2015). “Risk Assessment of Work-Related Musculoskeletal Disorders in Construction: State-of-the-Art Review.” *Journal of Construction Engineering and Management*, 141(6), 04015008.
- [89] Waters, T. R., Putz-Anderson, V., and Garg, A. (1994). “Applications Manual for the Revised NIOSH Lifting Equation.pdf.”
- [90] Welcome, D., Rakheja, S., Dong, R., Wu, J. Z., and Schopper, A. W. (2004). “An investigation on the relationship between grip, push and contact forces applied to a tool handle.” *International Journal of Industrial Ergonomics*, 34(6), 507–518.
- [91] Xiao, B., and Zhu, Z. (2018). “Two-Dimensional Visual Tracking in Construction Scenarios: A Comparative Study.” *Journal of Computing in Civil Engineering*, 32(3), 04018006.
- [92] Yan, X., Li, H., Li, A. R., and Zhang, H. (2017a). “Wearable IMU-based real-time motion warning system for construction workers’ musculoskeletal disorders prevention.” *Automation in Construction*, Elsevier B.V., 74, 2–11.
- [93] Yan, X., Li, H., Wang, C., Seo, J. O., Zhang, H., and Wang, H. (2017b). “Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion.” *Advanced Engineering Informatics*, Elsevier, 34(June), 152–163.
- [94] Yang, T., Li, S. . Z., Pan, Q., and Li, J. (2005). “Real-Time Multiple Objects Tracking with Occlusion Handling in Dynamic Scenes.” (60172037), 970–975.

- [95] Yasin, H., Iqbal, U., Krüger, B., Weber, A., and Gall, J. (2015). “A Dual-Source Approach for 3D Pose Estimation from a Single Image.”
- [96] Yu, Y., Li, H., Umer, W., Dong, C., Yang, X., Skitmore, M., and Wong, A. Y. L. (2019a). “Automatic Biomechanical Workload Estimation for Construction Workers by Computer Vision and Smart Insoles.” *Journal of Computing in Civil Engineering*, 33(3), 04019010.
- [97] Yu, Y., Yang, X., Li, H., Luo, X., Guo, H., and Fang, Q. (2019b). “Joint-Level Vision-Based Ergonomic Assessment Tool for Construction Workers.” *Journal of Construction Engineering and Management*, 145(5), 04019025.
- [98] Zeiler, M. D., and Fergus, R. (2014). “Visualizing and understanding convolutional networks.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1), 818–833.
- [99] Zhang, D., and Shah, M. (2015). “Human pose estimation in videos.” *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter, 2012–2020.
- [100] Zhang, H., Ouyang, H., Liu, S., Qi, X., Shen, X., Yang, R., and Jia, J. (2019). “Human Pose Estimation with Spatial Contextual Information.”
- [101] Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. (2017). “Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach.” *Proceedings of the IEEE International Conference on Computer Vision*, 2017–Oct, 398–407.
- [102] Ringleberg J, Voskamp P. “Integrating Ergonomic Principles into C-Standards for Machinery Design. TUTB Proposals for Guidelines.” *Brussels: European Trade Union Technical Bureau for Health and Safety*, 1996.