On Natural Motion Processing using Inertial Motion Capture and Deep Learning

John H. Geissinger

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

Edward A. Fox, Chair

Alan T. Asbeck

Harpreet S. Dhillon

May 6, 2020

Blacksburg, Virginia

Keywords: inertial motion capture, deep learning, ergonomics, manual material handlers

On Natural Motion Processing using Inertial Motion Capture and Deep Learning

John H. Geissinger

(ABSTRACT)

Human motion collected in real-world environments without instruction from researchers - or natural motion - is an understudied area of the field of motion capture that could increase the efficacy of assistive devices such as exoskeletons, robotics, and prosthetics. With this goal in mind, a natural motion dataset is presented in this thesis alongside algorithms for analyzing human motion. The dataset contains more than 36 hours of inertial motion capture data collected while 16 participants went about their lives. The participants were not instructed on what actions to perform and interacted freely with real-world environments such as a home improvement store and a college campus. We apply our dataset in two experiments. The first is a study into how manual material handlers lift and bend at work, and what postures they tend to use and why. Workers rarely used symmetric squats and infrequently used symmetric stoops typically studied in lab settings. Instead, they used a variety of different postures that have not been well-characterized such as one-legged lifting and split-legged lifting. The second experiment is a study of how to infer human motion using limited information. We present methods for inferring human motion from sparse sensors using Transformers and Seq2Seq models. We found that Transformers perform better than Seq2Seq models in producing upper-body and full-body motion, but that each model can accurately infer human motion for a variety of postures like sitting, standing, kneeling, and bending, given sparse sensor data.

On Natural Motion Processing using Inertial Motion Capture and Deep Learning

John H. Geissinger

(GENERAL AUDIENCE ABSTRACT)

To better design technology that can assist people in their daily lives, it is necessary to better understand how people move and act in the real-world with little to no instruction from researchers. Personal assistants such as Alexa and Google Assistant have benefited from what researchers call natural language processing. Similarly, natural motion processing will be useful for everyday assistive devices like prosthetics and exoskeletons. Unscripted human motion collected in real-world environments - or natural motion - has been made possible with recent advancements in motion capture technology. In this thesis, we present data from 16 participants who wore a suit that captures accurate human motion. The dataset contains more than 36 hours of unscripted human motion data in real-world environments that is usable by other researchers to develop technology and advance our understanding of human motion. In addition, we perform two experiments in this thesis. The first is a study into how manual material handlers lift and bend at work, and what postures they tend to use and why. The second is a study into how we can determine what a person's body is doing with a limited amount of information from only a few sensors. This study could be useful for making commercial devices like smartphones, smartwatches, and smartglasses more valuable and useful.

Acknowledgments

First and foremost, I would like to thank Dr. Alan Asbeck. I have worked with him for nearly five years at the Assistive Robotics Lab. I have developed immensely under his guidance, as a mechanical engineer and later as a software engineer. He has always been enthusiastic, supportive, and encouraging, which has allowed me to work on a variety of challenging problems. Thank you to my committee chairman Dr. Edward Fox for valuable instruction in his course on Big Data Text Summarization and for allowing me to volunteer in the Digital Library Research Laboratory as a Master's student. His problem-based approach to teaching allowed me to explore Virginia Tech's computing clusters for the first time, which later proved instrumental to this thesis. I would also like to thank my committee member Dr. Harpreet Dhillon. His course on stochastic signals and systems was one of the most challenging and rewarding courses I have taken at Virginia Tech, and it greatly influenced my thinking about the topics covered in this thesis.

Thank you to everyone who participated in the collection of the dataset. The work presented in this thesis would not have been possible without Mehdi Alemi and Emily Chang, who collected the initial dataset at Lowe's Home Improvement. I also thank Taber Fisher for fruitful conversations about human motion modeling. I am grateful for everyone who made Virginia Tech and the Assistive Robotics Lab a collaborative and friendly environment, particularly Theo Long, Drew Giacalone, Joshua Hull, Tim Pote, David Weisbrodt, Taylor Pesek, Athulya Simon, Andrew Bocklund, Hubert Kim, Brandyn Greczek, Hani Awni, Liuqing Li, Ranger Turner, Hee Doo Yang, Evan Wood, and Maria-Fé Thielman.

Finally, I would like to dedicate this thesis to my parents, Brent Geissinger and Amy Nix, who have supported and encouraged me for years and years.

Contents

List of Figures

List of Tables

List of Abbreviations

GRU	Gated Recurrent Unit
IMU	Inertial Measurement Unit
LSTM	Long Short-Term Memory
MSD	Musculoskeletal Disorder
RNN	Recurrent Neural Network
Seq2Seq	Sequence-to-Sequence
SOS	Start-of-sequence

Chapter 1

Introduction

To aid the study of exoskeletons, robotics, and human-computer interaction, datasets must contain significant amounts of unscripted natural human motion instead of being overly biased towards scripted motion. Such natural motion should meet the following constraints: 1) the motion should not be scripted or directed by a researcher, 2) the motion should not be constrained to a highly limited environment that contains a limited number of objects the researchers want the participants to interact with, and 3) the motion should include subconscious habitual motion that people partake in, without realizing it. To capture natural motion, we used inertial motion capture to collect data as regular people continued with their lives.

This approach is important for several reasons. First, to the best of our knowledge, previous large human motion capture datasets contain motion recorded at the direction of researchers in settings constrained by the researchers. Also, since the motions recorded in previous datasets are scripted and limited in duration, subconscious habitual motion is not present. This limits the utility of these motion capture datasets for devices that will be used in day-to-day life where people perform unscripted motion in real-world environments.

To this end, we captured motion by placing an XSens MVN Link suit on participants on-site and then let the participants continue about their day with little to no interference in unconstrained real-world environments. We used XSens MVN Link because inertial motion capture is well suited to capturing natural motion. It is not constrained by limited motion capture volumes like optical motion capture, it does not suffer from line-of-sight limitations like cameras, and with recent advancements in Kalman filter design, it is capable of working near magnetic surfaces.

We apply our dataset in two experiments. The first is a study into how manual material handlers lift and bend at work, and what postures they tend to use and why. This experiment uses traditional methods for analyzing data such as manual data labeling to determine the lifting and bending postures the participants used. We present findings of interesting postures that the manual material handlers tend to perform like one-legged lifting and split-legged lifting that have not been studied frequently by researchers in comparison to stoop lifting and squat lifting.

The second experiment is a study of how to infer human motion using limited information. Inferring human motion with sparse information about the orientation and acceleration of different human body segments is essential for exoskeletons, robotics, stroke rehabilitation, and human-computer interaction. Our dataset was collected using an XSens MVN Link suit, which contains 17 inertial sensors. Can systems in the future use fewer sensors to infer accurate human motion? To that end, we conduct analyses on upper-body and full-body motion inference using deep learning, specifically with Seq2Seq models and Transformers.

In Chapter 2, we cover motion capture datasets that have been collected over the years, how motion capture has been used in relation to ergonomics, and the literature on human motion inference. Chapter 3 describes the process we followed for collecting our dataset using the inertial motion capture system XSens MVN. In Chapter 4, we cover the deep learning algorithms we used for human motion inference. We describe the experimental design in Chapter 5. In Chapter 6, we discuss the results of the experiments. Finally, Chapter 7 concludes the thesis with limitations, discussion, and future work.

Chapter 2

Review of Literature

2.1 Motion Capture Datasets

In recent years, the expanding need for accurate human motion analysis has resulted in multiple motion capture datasets [1, 18, 37, 56, 60, 88, 90]. In this section, we review motion capture datasets collected through various means.

Optical motion capture using markers is a widely accepted, popular methodology for ground-truth pose data. One of the earliest datasets was BioMotion, which researchers collected for analysis of gait patterns [88]. To collect as close to natural motion as possible, they recorded walking motion on 40 participants for 5 minutes before recording and did not let the participants know when they recorded the walking motion. Also, the researchers used a treadmill with speed set by the participants to make data collection as natural and comfortable as possible.

A later and widely used dataset is the CMU motion capture dataset [18]. This dataset was the most abundant motion capture dataset of its time in terms of timespan and number of motions. The dataset contains data on 144 participants and contains more than 9 hours of data. The data is collected using Vicon motion capture cameras, with 41 markers placed on the participants. The capture volume is approximately 3m x 8m. In terms of participants and size, this dataset is impressive. However, it is limited in that it is captured in a mostly artificial environment. This limitation influences how naturally the participants can act.

To make improvements on this dataset, CMU also produced the CMU-MMAC dataset [18]. In this dataset, optical motion capture was combined with inertial sensors, video, audio, and additional wearable sensors. This data was collected in a kitchen environment that the researchers built for 43 participants to simulate cooking five different recipes. This approach is closer to the ideal of natural motion capture. However, it is notable that the researchers segmented out data on “anomalous” situations such as fire and smoke, falling dishes, and some distractions that interrupted the cooking tasks. It is our view that this anomalous data contains real-world interruptions that are important for natural motion capture. Nevertheless, this dataset is significant and impressive because it shows the push for more natural motion capture.

The KIT motion capture project [56] introduced the largest optical motion capture dataset in addition to object interaction. The dataset contains 55 participants and approximately 11 hours of data [53]. KIT is essential because it focuses not only on human motion but also on object and environment interaction. Also, it attempts to aggregate motion data from other sources like the CMU motion capture database, to have one accessible format, the Master Motor Map (MMM), based on XML. This dataset is an important step in the direction of real-world natural motion akin to the CMU-MMAC dataset’s simulation of a kitchen environment. However, it is still constrained to small motion capture volumes due to its reliance on optical motion capture. It is also dependent on participants receiving instructions from researchers and researchers limiting and designing the environment.

The largest and most expansive motion capture dataset to date is the Archive of Motion Capture as Surface Shapes (AMASS). AMASS is the most recent step towards much bigger motion capture datasets for deep learning and computer vision use. It combines a number of previously mentioned motion capture datasets [18, 56, 60, 88]. Since many of the motion capture procedures differ, it uses surface shapes to combine different marker placements with the SMPL and MoSH algorithms [48, 49]. It contains 43.6 hours of motion capture data on 450 subjects, making it by far the largest publicly available motion capture dataset. However, it shares the same constraints as the other motion capture datasets. It is dependent on researchers instructing people to perform actions in areas designed and constrained by researchers.

Improved motion capture has been enabled by advancements in pose estimation using cameras. The Kinect from Microsoft, and OpenPose from CMU, are two popular examples [14, 81]. These systems do not require large marker sets on the human body and can be used in a wide range of locations. On the other hand, the 3D pose accuracy is generally lower when using these systems. OpenPose enables the collection of large datasets of pose estimates using images found on the internet or taken with smartphones. Researchers collected the CMU Panoptic dataset using OpenPose in a moderately sized 3D volume. It contains around 5.5 hours of data using more than 120 subjects. The novel contribution of the CMU Panoptic dataset is the 3D modeling of the interaction between multiple people while playing games. Although there are dependencies on the 2D pose detection method and issues with distinguishing left and right limbs, the Panoptic dataset is another dataset that has moved more in the direction of natural motion capture, specifically between multiple people [39].

A valuable dataset for recent advancements is the Human3.6M dataset [37]. The researchers took videos with four cameras to capture 3.6 million poses collected at 50 Hz with 11 participants, which comes in at approximately 20 hours of data. Importantly, the motivating factor in building this dataset was to simulate 3D postures seen in the open world. They purposefully collected images of people posing in real-world environments and had actors take on the same postures. This dataset is heavily benchmarked for deep learning purposes and has been used frequently to advance human motion prediction using deep learning

[13, 26, 29, 30, 38, 59, 63].

Optical motion capture systems are considered the standard in ground-truth motion capture. However, there are also motion capture solutions using inertial sensors that have several advantages [74, 96]. Critically, they are usable in outdoor, real-world settings and are not limited to indoor, small volume laboratory environments. This difference is essential because it allows for the collection of unscripted motion while people perform day-to-day activities. This advancement opens up the possibility of collecting accurate, natural motion in real-world settings like work environments and the outdoors. For example, [97] collected inertial motion capture data in environments like parks where the participants ride bikes and climb jungle gyms. Notably, [90] uses multi-view cameras and XSens inertial sensors to create a non-marker based motion capture dataset in a larger area than is possible with optical motion capture.

Other use cases for inertial motion capture systems include prosthesis and exoskeleton control, rehabilitation of athletes or medical patients after early discharge, or monitoring of ergonomics in straining work environments. However, despite the array of applications that inertial motion capture systems could benefit, the largest open datasets are from [34] and [90], which total 90 minutes and 50 minutes, respectively, according to [34].

2.2 Motion Capture and Ergonomics

Musculoskeletal disorders (MSDs) are highly prevalent work-related health problems. Laborers and freight, stock, and material movers had the highest number of MSDs, and 31% of the work-related injuries that required time away from work were MSDs in 2015 [11]. Awkward and extreme postures, a rapid work pace, and repetitive motions are among the most important risk factors related to MSDs [4, 41, 67, 71] and, among other causes, can be attributed to workers quickly and repeatedly loading their spines and lower backs in postures that require overexertion.

Manual material handlers (MMH) are workers at risk of MSDs due to the demand put on their bodies from manually moving freight and stock. As the job requires strenuous labor, the workers are exposed to the risk of injury in multiple parts of the body. Since 2.6 million people work in this occupation in America [11], it is important to study and measure postures and any risks they are exposed to throughout their day.

Several methods have previously been used to measure postures and ergonomic risks in workplaces. On-site observational studies and questionnaires can estimate risk profiles for workers [12, 17, 32, 40, 85]. Traditionally, this includes on-site observational studies and questionnaires. Systems such as Ovako Working Posture Analyzing System [40], Posture, Activity, Tools, and Handling [12], Rapid Entire Body Analysis [32], and Quick Exposure Check [17] consist of qualitative in-person heuristics that can be applied by observers to construct risk profiles for workers. Despite a wide variety of these methods [85], the process is subjective and, since the observer is watching the worker, the worker may act differently than they would normally.

Alternatively, posture recognition has been accomplished using cameras and computer vision techniques [64, 104] or inertial sensors [15, 25, 42, 44, 45, 55, 61, 77, 92, 93, 103, 106]. The first method, consisting of posture recognition using cameras and computer vision techniques, has been applied in both on-site studies and laboratory settings [64, 104]. For example, ergonomic posture assessment in outdoor construction environments was recently developed using 2D RGB cameras and machine learning models [104]. In another example, Pellegrini and Iocchi used stereo cameras and hidden Markov Models to recognize postures in a laboratory experiment [64]. Some scenarios lend themselves well to cameras, such as when there is a limited workspace, or there are no objects that could visually occlude the participant or worker. However, this is not always true, especially in warehouses or stores where the work area is large and visual occlusion is likely.

Researchers have conducted on-site studies of nurses [76] and construction workers [15, 44, 93, 106] using inertial sensors, which has been noted by [93] to help improve the quantitative analysis of worker habits. [76], for example, demonstrated that nurses did not extend to an extreme trunk or arm position during work, but were fatigued from a lack of rest during their shifts. Furthermore, Lee et al. calculated the percentage of time roofers spent with trunk flexion more significant than 60 and 90 degrees, which is an effective statistic for evaluating the risk of occupational injury [44].

Also, [23] studied the distribution of box height and reach distances during manual material handling in factories, as well as how frequently people twist during lifting. [50] presented quantitative data on how workers lifted and moved while unloading vans, including kinematics and box heights. [5] quantified how novice and expert manual material handlers differed in their strategies for lifting, and observed that the workers frequently supported the majority of their body weight on one foot. [47] used force sensors under each foot to estimate the distribution of weight between each foot, the loads lifted, and the lifting frequencies in nursing aides and warehouse workers, similarly finding that in most cases one foot carried most of the load during lifting (however, they did not collect kinematics, so the relative positioning of the feet was unknown). [20] studied the kinematics and dynamics of people lifting and lowering using two different prescribed foot motions to understand their relative injury risk.

2.3 Motion Capture and Human Motion Inference

2.3.1 Hybrid Approaches using Cameras and Sparse Inertial Sensors

There have been multiple previous approaches to full-body human motion inference based on sparse inertial sensors. Many have taken a hybrid approach in combining video data with inertial sensor data [54, 65, 66, 98]. In [65], the authors designed a method for fusing data from eight video cameras with sparse inertial sensor data from five inertial sensors. Their approach makes use of video data to obtain drift-free position data, which is not possible with inertial sensor data. By matching a person's surface vertices to a silhouette of their body, they can approximate positions of points on a person's body. Then from the inertial sensors, they obtain orientations of the limbs that they can use in addition to the position data. Later on, [66] proposed using particle-based optimization and five inertial sensors that constrain the number of valid poses, which they then weigh using visual information from four video cameras. In [98], the authors use inertial sensors and a single video camera to find 3D poses. Using convolutional neural networks to detect 2D poses [14], the authors find 3D postures with help from inertial sensor data.

Depth cameras and optical motion capture have been used alongside inertial sensors as well. In [31], the authors use a Kinect and six inertial sensors for full-body motion inference. They construct a database for querying inertial sensor orientations that constrain the Kinect's depth camera data. In [3], the authors propose a real-time system that makes use of six sparse inertial sensors and five reflectors for optical motion capture. They leverage mass and inertia estimates to perform inverse dynamics. Using inverse dynamics, they solve for motion that is satisfied by the position and orientation constraints of the inertial sensors and optical motion capture. Because of the inverse dynamics estimates and extra constraints from optical motion capture, they can model highly dynamic motion like jump kicking.

2.3.2 Approaches with Different Types of Wearable Sensors

Other than inertial sensors, human motion inference has been studied using magnetic sensors [6, 80], foot pressure sensors [105], accelerometers [82, 86], and inertial-ultrasonic motion sensors [46]. In pioneering work, [6] used magnetic sensors and inverse kinematics to model someone standing up. The magnetic sensors are placed on the back of the head, the waist, and both forearms. Similarly, in [80], the authors use eight magnetic sensors and inverse kinematics to model human motion. The sensors are placed on the pelvis, the head, both ankles, both hands, and both upper arms. In [105], the authors present FootSee, which is a system for extracting poses from a database based on foot pressure readings.

In [82], the authors use five accelerometers to match upper-body postures in a database to accelerometer data. In [86], the authors use only four accelerometers to extract full-body postures. With six inertial-ultrasonic motion sensors that provide both orientation and position, [46] models full-body motion in real-time. These are more powerful than regular inertial sensors because they also provide position data, but they are restrictive because they require a base station and thus constrict capture volume like optical motion capture.

2.3.3 Human Motion Inference with Sparse Inertial Sensors

Learning and optimization approaches have both been used for human motion inference using sparse inertial sensors. In [79], the authors use Gaussian Processes to learn mappings between data from four inertial sensors and optical motion capture data. The inertial sensors are placed on the wrists and ankles. In additional experiments, the authors were able to reduce the sensor count to estimate walking postures. However, Gaussian Processes have since been mostly replaced by neural networks because those are capable of scaling with training data and generalizing to unseen data.

Later on, [102] made use of densely connected neural networks to predict postures using five inertial sensors with several different configurations. [102] is similar to our study because they use an XSens MVN Link suit, allowing them to test multiple configurations. However, they only have around 2 hours of data collected in lab conditions. They also only predict a single posture at a time.

In [97], the authors present an offline optimization method by minimizing orientation, acceleration, and anthropometric errors and jointly optimizing postures over multiple frames using six inertial sensors. One interesting finding from this paper is the necessity of an acceleration term. They found that acceleration impacts posture approximation significantly. They also made use of an anthropometric term to ensure they generate human-like poses. Overall, they generate accurate human motion from a variety of postures. However, their method is computationally expensive and offline.

Recently, [34] presents an advancement on [97] that uses deep learning instead of offline optimization. They use a two-layer bidirectional LSTM to predict single postures using 20 past and 5 future frames. Their system uses six inertial sensors and runs in real-time. One interesting point is that they use the AMASS dataset to generate synthetic inertial sensor data, which allows them to generate 45 hours of training data. However, because they use synthetic data, they later have to fine-tune on real inertial sensor data that they collect manually to perform well on their test set.

Chapter 3

Data Collection Methodology

In this chapter, we discuss the use of an XSens MVN Link suit for data collection. We describe the procedure we followed for using this data collection system and our data collection process in real-world environments.

3.1 Data Collection System

To properly collect high-quality data, we made use of an XSens MVN Link suit (Xsens North America Inc., Culver City, CA, USA). The XSens MVN Link suit collects synchronized inertial sensor data and post-processes it to construct accurate human motion.

3.1.1 XSens Motion Capture using Inertial Sensors

The miniature motion trackers that come with the XSens MVN Link suit have 9 degrees of freedom (DoF) from 3D rate gyroscopes, 3D linear accelerometers, and 3D magnetometers. The motion trackers are placed on important segments of the body to measure the segment's underlying motion. For example, an inertial sensor placed on a footpad is shown in Figure 3.1.

The XSens MVN Link consists of 17 inertial motion trackers along with Velcro straps, gloves, and a headband with tracker pouches, as well as footpads with a tracker patch. To log

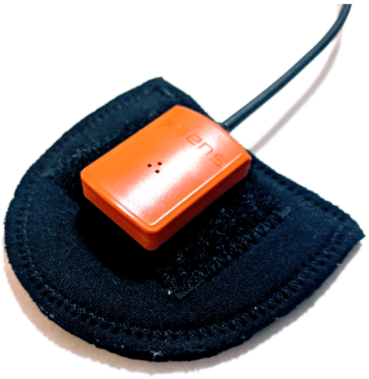


Figure 3.1: An XSens MVN Link inertial sensor (MTX2-4A7G6) placed on a footpad that is inserted into the participant's shoe during data collection. The inertial sensor has 9 degrees of freedom for measuring linear acceleration, angular velocities, and the Earth's magnetic field.

data while people go about their day-to-day activities, an on-body recording device called a bodypack is used for storing data. A battery that has a typical life of 9.5 hours is used for powering the bodypack and the sensors. A battery plugged into a bodypack is shown in Figure 3.2.

The sensors were velcroed to segments using a portion of the straps and then wrapped with the remainder of the straps to ensure minimal movement of the sensors. For example, Figure 3.3 shows the glove with a pouch for a sensor and the wrist with the constrained sensor. This wrapping procedure was followed for each sensor on the arms and legs. The feet and gloves are constrained by the shoelaces and webbing on the glove, respectively, so the wrapping procedure was not necessary for these sensors.

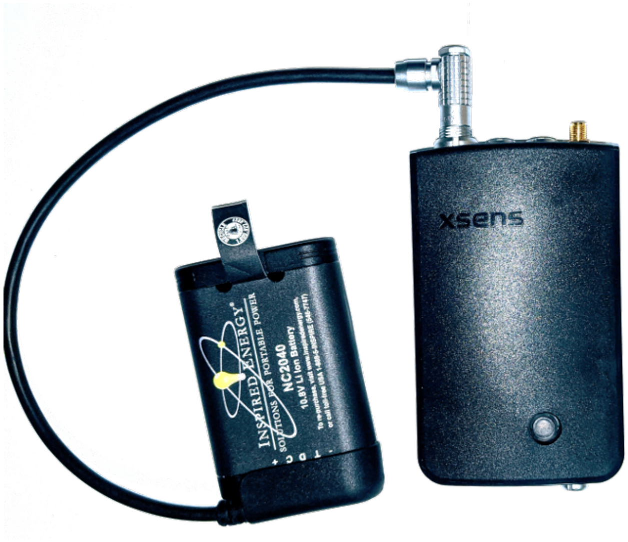


Figure 3.2: A bodypack connected to a battery. The bodypack also has two ports for connecting the sensors. The button on the lower right of the bodypack is for turning on/off the bodypack, and for starting a calibration or data collection session.

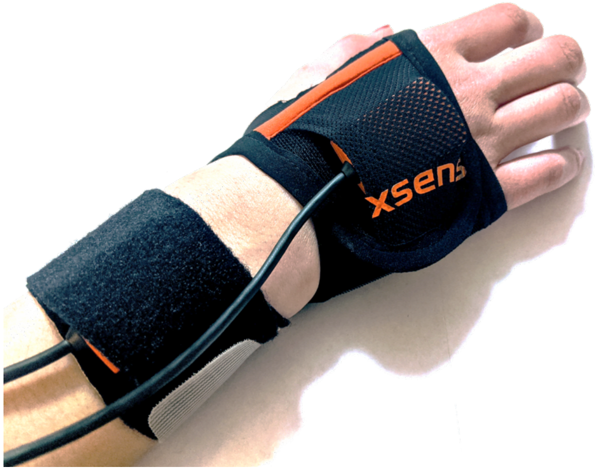


Figure 3.3: A pair of XSens inertial sensors attached to the hand and wrist. The glove contains a pouch with velcro for attaching an inertial sensor and webbing that constrains the sensor from rotating. The wrist's inertial sensor was attached to a portion of the strap and then wrapped with the remainder of the strap to constrain it further and avoid any additional rotation.

3.1.2 XSens Motion Capture Engine

Importantly, XSens uses a specialized Kalman filter design and advanced biomechanical model that compensates for magnetic disturbance, and they have found this significantly reduces errors due to ferromagnetic materials [72, 73]. Thus, materials like reinforced concrete that may have been present in surrounding environments would have had minimal impact on the inertial sensor readings. Also, the motion capture engine allows for data collection under semi-static and highly dynamic conditions. Since our participants were performing daily activities like lifting boxes, pushing carts, and working at computers, the engine meets our needs.

3.2 Data Collection Process

3.2.1 Study Design

Manual material handlers in a home improvement store, and students at Virginia Tech, were invited to participate. The study was approved by Virginia Tech's Institutional Review Board (IRB). We collected data from $N = 16$ participants. Each participant was asked beforehand if they would be willing to participate in the study and were asked at least 24-hours before signing a consent form. Measurements were taken of each participant following the guidelines given by XSens. Of the participants, 12 were Virginia Tech students, and 4 were employees of a local home improvement store. 13 were male, and 3 were female. The heights of the store employees and Virginia Tech participants were 175.1 cm ± 5.1 cm and 180.7 cm ± 8.1 cm, respectively. The age ranges for the store employees and Virginia Tech participants were 30-58 and 20-35, respectively. For an overview of our dataset, see Table 3.1.

Table 3.1: An overview of the natural motion dataset presented in this thesis. The number of hours we collected data for is greater than other datasets we know of apart from AMASS [53].

# Hours	# Frames	# Subjects	Age Range	M/F
36	31 million	16	20-58	13/3

We placed a total of 17 inertial sensors following the guidelines from XSens (Figure 3.4). Following the wrapping procedure shown in Figure 3.4, the inertial sensors were secured to prevent the extraneous movement of the sensors during data collection. The calibration process consisted of the participant going into a neutral pose (N-pose) for 10 seconds, walking forward and back in a straight line, and then returning to an N-pose.

3.2.2 Manual Material Handlers

The manual material handlers (N=4) worked at a local home improvement store while wearing the XSens inertial sensors. Researchers taught each participant how to turn on the system, hold the calibration pose, and turn off the system. Participants were accompanied for 30 minutes on the first day to answer questions and troubleshoot slippage. At the end of each data collection period, researchers revisited the store to download the data and reset the system. The workers wore the suit at the beginning of their shifts and took the suit off to complete their shift.

Workers performed a variety of tasks during the sessions, including opening boxes on pallets; sorting objects on pallets; moving items from pallets onto shelves, including both small, light objects and large or heavy objects; moving large appliances both in-store and into a truck; unloading large items from pallets such as vanities and counter-tops, and placing these in aisles or on shelves; moving bags of mulch and plants; and moving empty boxes into shopping carts or onto pallets.

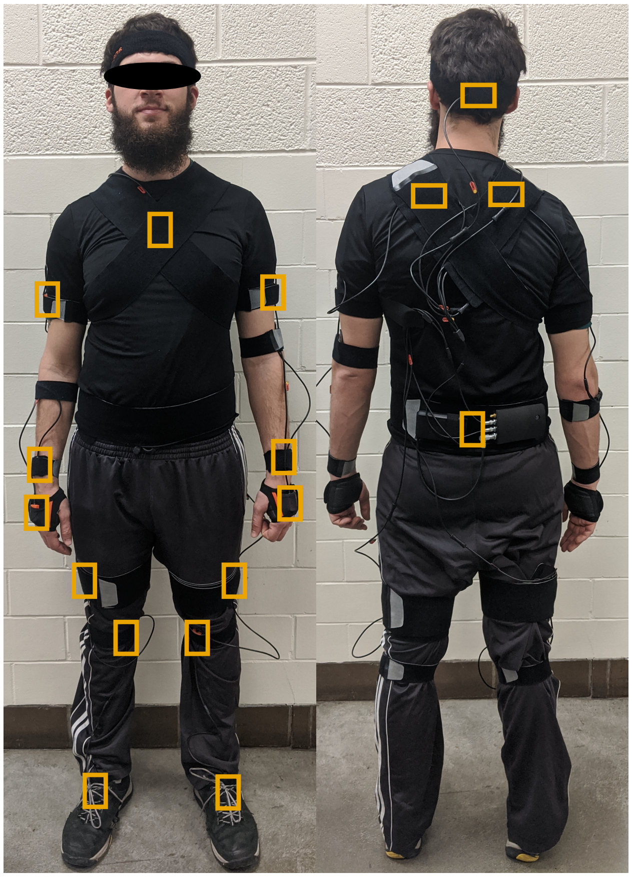


Figure 3.4: Inertial sensor placement used in this study following XSens guidelines. Since the inertial sensors were wrapped with straps to prevent movement, orange boxes were used to indicate the inertial sensor locations. Some additional straps were used to constrain cables on the forearms and the legs. Custom shirts were made for some workers since the standard XSens shirts did not fit.

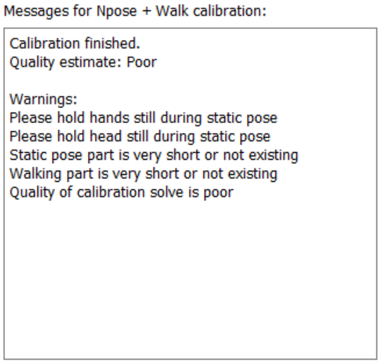
3.2.3 Virginia Tech Participants

For the Virginia Tech participants (N=12), data collection was done in the Assistive Robotics Lab and the surrounding buildings and stores around Virginia Tech's campus. The researchers were available for troubleshooting any slippage, calibrating the suit, guiding the participants through the calibration procedure, and turning off the system. The participants continued their day for up to two and a half hours. For most participants, a calibration file was collected at the start and end of the data collection period to avoid issues with slippage that may have occurred.

The participants from Virginia Tech conducted a myriad of tasks, including performing routine office work while sitting down and standing up, discussing and communicating with other people, walking around campus and going to class, sitting and talking to others in meetings, driving in cars to local stores, collecting data and orchestrating tests, cleaning and organizing the Assistive Robotics Lab, reading books while lounging in chairs, and other activities of daily life. Unlike the data collection sessions at the home improvement store, we maintained logs of general activities the participants performed and where they went on campus. The logs will be useful for other researchers to understand the context of the motion capture files.

3.2.4 Quality Assurance

To ensure data quality, we checked each file using XSens MVN Studio to verify that calibration was performed correctly. XSens provides a calibration quality estimate, and we flagged data if the quality estimate was "Poor." For example, see a flagged file in Figure 3.5 with a "Poor" quality estimate below. For each file, we cropped and removed any data using XSens MVN Studio, where there was evidence that sensors had slipped or fallen off.



```
Messages for Npose + Walk calibration:

Calibration finished.
Quality estimate: Poor

Warnings:
Please hold hands still during static pose
Please hold head still during static pose
Static pose part is very short or not existing
Walking part is very short or not existing
Quality of calibration solve is poor
```

Figure 3.5: A calibration quality estimate from XSens MVN Studio. The quality estimate was "Poor" so we flagged the data and did not use it in our dataset.

Chapter 4

Algorithm Design and Implementation

In this chapter, we discuss the algorithms that we have used for applying deep learning to human motion inference. We discuss Seq2Seq models and Transformers, which we used to model human motion as a sequence-to-sequence problem.

4.1 Seq2Seq Architectures

Sequence-to-sequence (Seq2Seq) models were introduced for the task of neural machine translation in [84]. Seq2Seq models consist of an encoder and decoder, typically containing one or more layers of long-short term memory (LSTM) layers or gated recurrent unit (GRU) layers ([16, 33]). The encoder takes in an input of arbitrary length, passes it through the recurrent layers, and produces an encoder hidden state as an output. This encoder hidden state is then fed into the decoder along with a start-of-sequence or <SOS> token. In [7] was the introduction of bidirectional encoders to Seq2Seq models. The bidirectional encoder will look at the input sequence in both directions and generate hidden states for both the forward and backward view. This approach can help the decoder generate sequences.

The decoder also contains recurrent layers that output both a hidden state and an output vector. The output vector can then be compared to a target output to compute the loss for the network. If a ground truth input vector is given to the decoder, then the decoder does not have to correct for its own mistakes. This methodology is called teacher forcing and causes exposure bias in the decoder [9, 63, 70, 101]. To improve the outputs during inference, the output vector can be fed back into the decoder to expose the decoder to its errors and adjust its predictions over time.

In all, we study two Seq2Seq models as baselines. The first has a simple encoder and decoder. The second has a bidirectional encoder and a decoder with attention. We make use of GRU layers in both the encoder and decoder. They simplify LSTMs by requiring fewer gates [16]. Both the encoder and decoder use a single GRU layer. We now describe attention and how it may help in modeling human motion.

Attention

Seq2Seq models were improved following their initial introduction using the attention mechanism in [7], commonly referred to as Bahdanau attention. Attention mechanisms allow for longer input sequences to be evaluated by the decoder by learning which of the encoder hidden states are most important to the sequence being generated by the decoder.

The two inputs to an attention layer are the annotations from the encoder and the previous hidden state of the GRU layer. The hidden state is just the output of the GRU layer at the previous time step. The annotations h_j and output s_{i-1} are passed into the attention layer where the layer computes an alignment model [7]:

$$e_{i,j} = a(s_{i-1}, h_j) \tag{4.1}$$

The alignment model is a multi-layer perceptron that contains learnable parameters (W_a,

W_b, and v_a):

$$a(s_{i-1}, h_j) = v_a^T \tanh(W_a s_{i-1} + W_b h_j) \tag{4.2}$$

These learnable parameters are optimized during training through backpropagation so that the model pays better attention to important features. Note that here, we are casting both s_{i-1} and h_j to a new vector space so that they can be added together. This fact means that the annotations from the encoder and the output from the GRU can have different representations that are converted through the multi-layer perceptron.

In the literature, this mode of attention is denoted *additive* attention. There are other forms of attention, but we cover additive attention for simplicity [52].

Once the annotations and the GRU output are passed through the alignment model, we can compute the weights of the annotations:

$$\alpha_{i,j} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})} \tag{4.3}$$

Note that these weights are simply the outputs of the alignment model passed through a softmax layer. The softmax layer normalizes the alignments to probabilities. This detail is important: it converts the alignment model outputs into probabilities that can be used to weight the annotations for a specific output from the GRU. What this means is that the decoder can weight certain features (the annotations) from the encoder as relevant or not relevant to a particular posture produced by the GRU. In other words, our model gains the ability to denote the importance of features to a posture.

To use these weights, we find the expectation of the annotations. Mathematically, we find the expected *context vector*:

$$\mathbb{E}_{p(s_t|a)}[\widehat{\mathbf{z}}] = \sum_{i=1}^{N} \alpha_{t,i}\mathbf{h}_i \tag{4.4}$$

Here we are taking the expectation over the annotations given the weights that come from our alignment model. We denote this as the expectation of the *context vector*. This context vector can help generate better outputs because we now have some idea of the importance of the features generated by our encoder.

We finally use this context vector alongside the GRU's hidden state and our previous posture from our decoder to compute the next posture. Mathematically, we pass the context vector $\hat{z}_t$, the previous posture y_{t-1}, and the previous output from our GRU s_t through a multi-layer perceptron:

$$p(y_t|\mathbf{h}, \mathbf{y}_{t-1}) = act(W_o(W_z\widehat{\mathbf{z}}_t + \mathbf{W}_y\mathbf{y}_{t-1} + W_s\mathbf{s}_t)) \tag{4.5}$$

The new output from the decoder is the next posture. It should be noted that the decoder is more than just a GRU. From here, we concisely summarize what is contained in the decoder. The decoder takes in annotations and the previous hidden state. These inputs are passed through the attention layer to generate a context vector $\hat{z}_t$. The GRU is used to generate s_t from the previous hidden state s_{t-1} and the previous word y_{t-1}. Finally, the multi-layer perceptron makes use of $\hat{z}_t$, s_t, and y_{t-1} by casting them to the same representation with learnable parameters, and uses them to generate the output posture.

4.2 Transformer Architectures

In addition to seq2seq architectures, we use Transformers to study whether they can model motion more effectively [95]. Transformers have emerged as the model architecture of choice in natural language processing because they are capable of modeling natural language effectively across multiple tasks [21, 68, 69]. Transformers, similar to seq2seq models, have an encoder and decoder, and work similarly to seq2seq models. Transformers are of interest to human motion modeling because of the multi-head attention layer present in both the encoder and decoder. Instead of attending to the entire sequence with one attention function as the seq2seq model does, the model instead is capable of attending to multiple subspaces of the data at different positions. Since the human body is made up of multiple parts, it is interesting whether attending to these multiple parts with multi-head attention may allow for improvements.

Transformers are challenging to discuss in detail. Our use of the Transformer is almost identical to the original implementation, so we point the reader to the original paper [95] and two excellent tutorials [2, 75].

We make use of the Transformer architecture in two ways. The first is through the use of a bidirectional Transformer encoder similar to [21]. This model is not autoregressive like the seq2seq model or the full Transformer, meaning it produces all postures as output at once without predicting one posture at a time. We refer to this architecture as the "Transformer Encoder" in later sections.

The other Transformer architecture we study includes both the encoder and decoder and follows the description in [95]. The model is autoregressive. It will predict one pose at a time, and the decoder can attend to the entirety of the encoder's outputs as it makes predictions. During training, the entire target sequence is passed into the decoder with a vector of zeros as a start-of-sequence token. The target sequence is masked, so it will only attend to previous target postures. Since the ground-truth target posture is fed into the model during training, the Transformer does not make use of teacher forcing like the seq2seq architecture. During inference, the model is passed the start-of-sequence vector, and the first posture is predicted. This posture and the start-of-sequence vector are then passed again to the model. This procedure is repeated until the last posture is predicted.

Chapter 5

Experimental Design

We describe the experimental design in this chapter for both our study of manual material handlers and our application of the dataset to human motion inference.

5.1 Quantification of Postures for Low-Height Object Manipulation

In this section, we describe our study of worker lifting and bending behavior at the local home improvement store. These results are for the workers (N=4) and motivated our exploration of machine learning and natural motion that we describe later on.

5.1.1 Categorization of Postures

To categorize lifting and bending behavior broadly, we enumerated all possible postures that could be used during low-height object manipulation, including both those seen and not seen in this study, in Figure 5.1. The descriptions of these postures are provided in Table 5.1. In the categorization, some possible postures are physically challenging or more awkward than others (indicated by '-' symbols). Postures are awkward due to joint rotations near their ends of motion, involving compound rotations, or requiring muscle force to maintain [41]. Excluding these, there are only nine possible posture classes that can be used in low-height object manipulation. We include all of the variables of knee flexion, foot position, number of feet on the ground, and number of heels off the ground in the decision tree because the combination of these variables results in some postures being awkward while others are not. Instead of using ankle angle or hip angle, using the relative foot position, the number of feet on the ground, and the number of heels off the ground minimizes the use of thresholds (e.g., for angular ranges) in determining postures.

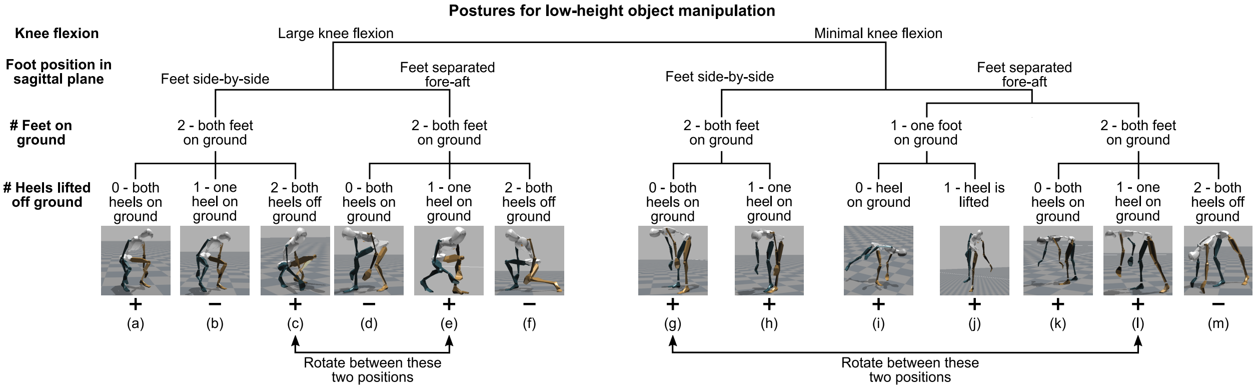


Figure 5.1: A decision tree categorization for postures used in low-height object manipulation. A '+' indicates that the posture is feasible, while a '-' indicates an awkward posture that would be unlikely ever to be used. Table 5.1 provides a description of each posture.

Several of these classes have been discussed frequently in the literature. Traditionally, the deep squat has been studied, where the heels remain on the ground (Figure 5.1a) [94]. A squat where the heels come off the ground ("catcher's squat") is also commonly studied (Figure 5.1c) [22]. A form of squatting, which we refer to as fore-aft squatting (Figure 5.1e), is similar to "straddling" in the literature [19, 24, 43], except in our categorization an object does not need to be between the legs. Symmetric stooping, where the knees remain straight (Figure 5.1g), has been frequently studied [83, 91]. One-legged stooping or the "golfer's lift," shown in Figure 5.1i, keeps one leg planted on the ground while the other leg is lifted off the ground, acting as a counterbalance. The person often rests their free hand on a nearby table or countertop to help them keep balance.

Table 5.1: Description and feasibility of different postures in our decision tree categorization of postures used during low-height object manipulation. The row highlighting is for clarity and does not denote groupings.

Letter	Description	Feasible?
(a)	Squat, symmetric, heels down (“Deep squat”), requires high flexibility in knees and ankles	Y
(b)	Squat, symmetric, one toe, awkward if feet are close together and side-by-side in the sagittal plane	N
(c)	Squat symmetric, on toes (“Catcher’s Squat”), requires good balance and less flexibility than deep squat	Y
(d)	Squat, legs split fore-aft, two heels, difficult to do when reaching down to the floor; possible for shallow bends	N
(e)	Squat, legs split fore-aft (“Fore-aft squat” in this thesis), one heel	Y
(f)	Squat, legs split fore-aft with no heels, possible but awkward	N
(g)	Stoop, symmetric, commonly used and investigated thoroughly in literature	Y
(h)	Stoop, two legs, one heel; useful only when reaching to the side in sagittal plane, uncommon	Y
(i)	Stoop, one-legged (“Golfer’s lift”), may require external support for balance	Y
(j)	Stoop, one-legged, one heel; useful primarily when reaching extremely far forward with external support.	Y
(k)	Stoop, two legs split fore-aft, two heels; uncommon as it requires more ankle flexion than only one heel on the ground	Y
(l)	Stoop, two legs split fore-aft, one heel; commonly seen in the dataset	Y
(m)	Stoop, two legs split fore-aft, both heels off the ground; uncommon as it may not be useful without external support	N

Another posture, which we refer to as the split-legged stoop, was hardly mentioned in the literature. It is similar to symmetric stooping, but the legs are split fore-aft while the rear leg remains on the ground. There are two types of split-legged lifting: when the rear foot is entirely on the ground (Figure 5.1k), and when the rear foot has its heel raised (Figure 5.1l). Split-legged lifting with a heel raised is similar to a one-legged lift in that most of the person’s weight is on the leading foot.

Within the classification, several postures are easy to transition between, as indicated by arrows in Figure 5.1. It is relatively more straightforward to pivot about a forefoot (front of the foot) on the ground, as compared to a foot that is flat on the ground, due to the point-like contact with the forefoot. This maneuver allows for smooth transitions between several of the poses.

Asymmetric motion, where the person reaches to the side during a bend or lift, usually involves longitudinal rotation about the axis of the spine [57, 94]. This could be used with any of the postures in this categorization.

5.1.2 Labeling of Bends and Classification

We identified bending postures in a semi-automated manner. We first algorithmically identified bent postures using kinematic features. This process was for pre-labeling events, after which we manually confirmed and classified events. If the pitch (bend angle) of the T8 segment of the mid-back was greater than 40°, we determined that the person was bending and marked the event for post-processing. If the T8 pitch was greater than 25° and either knee had flexion angles greater than 50°, which identifies squats, we also marked the event. These were conservative thresholds: we performed a sensitivity analysis to check what percentage of the bends of each type had a maximum T8 pitch angle of 5% past these bounds (26.25° for asymmetric squatting and 42° for other bend types). For stoop lifting, 99% of bends satisfied this criterion. With one-legged lifting and split-legged lifting with a heel raised, 100% of bends satisfied this criterion. In fore-aft squatting and split-legged lifting, 97% and 96% of bends satisfied this criterion, respectively.

Once the bend postures were pre-labeled, we then post-processed the marked events by (1) combining events less than a second apart, (2) removing events less than a half-second in length, and (3) expanding events by a half-second at the start and end of the bend to capture all of the motion.

We then manually reviewed and classified the posture at each of the marked events according to the decision tree in Figure 5.1, using XSens MVN Studio [78]. This was done in two rounds with a single reviewer in the first round and five reviewers in the second round; in cases where the reviewers disagreed on their classification, two third-round reviewers also reviewed the posture (297 out of 920 pre-labeled events). The posture was then classified according to the majority opinion. Of the 920 pre-labeled events, 666 were bending postures, and 254 were outliers. If any of the first and second round reviewers labeled the events as outliers, they were treated as outliers. Outliers included postures such as slight back bends, sitting down on stools or chairs, and operating hand-crank lifts that were incorrectly labeled as bending postures. We then categorized the bends into five categories: fore-aft squatting (Figure 5.1e and 5.1f; symmetric squatting was not observed), stoop lifting (Figure 5.1g), split-legged lifting with a heel raised (Figure 5.1l), split-legged lifting with no heels raised (Figure 5.1k), and one-legged lifting (Figure 5.1j and 5.1i). We present examples of these lifts for clarity in Figure 5.2. Fore-aft squatting includes both heels raised (Figure 5.1e) and one heel raised (Figure 5.1f) because both were infrequent. One-legged, heel raised type (Figure 5.1j) was also classified as one-legged lifting (Figure 5.1i) for our subsequent analysis since it was very infrequent. These were the only postures from our classification structure observed in the dataset. An additional sixth "Mixed"-type category was designated for events that were a mixture of different postures. For example, a stoop posture that goes into a one-legged posture while the participant kept their back bent, was considered a mixed event.

For bends where only the heel was lifted off the ground, the heel lifting was visible, and the toe was notably bent (Figure 5.2c). When the entire foot was lifted off the ground, the toe remained straight (Figure 5.2e). Categorizing these lifts was minimally affected by noise in the XSens IMUs. The motion occurred over a short time window and coincided with the motion of other joints. Additionally, XSens uses a specialized Kalman filter design that compensates for magnetic disturbance, and they have found this significantly reduces errors due to ferromagnetic materials [72, 72]. Thus, materials like reinforced concrete in the floor would have had minimal impact on the IMU readings.

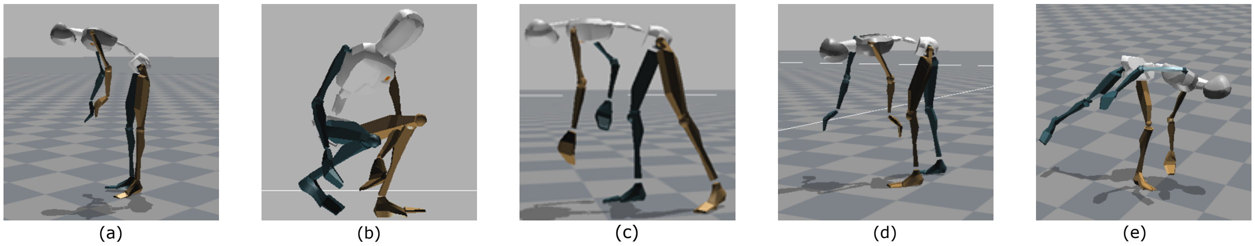


Figure 5.2: Examples of postures that were considered in the classification process. (a) Stooping, (b) Fore-aft squatting (symmetric squatting was not observed), (c) Split-legged with heel raised, (d) Split-legged, (e) One-legged ("Golfer's lift").

5.1.3 Metrics for Evaluating Lifting

We determined several metrics to quantify the bending/lifting profile of each participant. We used the T8 pitch angle to quantify the bend angle of each participant. The bend duration was the time from the beginning to the end of a bend (as described above). We counted the number of bends within a sliding one-minute window to quantify the bending frequency. To understand one-legged lifting further, we used the foot height difference, measured between the foot segment origin, located at the inter-malleolar point (approximately halfway between the ankle and heel), according to XSens. To quantify asymmetric lifting, we determined the average yaw of both the feet and the shoulders. We found the difference between these averages throughout each bend.

5.1.4 Hand Position Analysis

We investigated the range of hand motion during the different postures, to gain insight into why certain postures were used. The position of the dominant hand during the lift was determined relative to the average position of the participant's feet. The hand segment height was measured compared to the foot placed on the ground for one-legged lifting, and the average of both feet for the other lift types. To find the dominant hand in use during the bend, we projected each hand's position onto the sagittal axis and summed this over the entire bend, effectively finding the hand that was furthest away from the body for the most prolonged time during the bend. When the dominant hand's position was at the lowest point in each lift (i.e., along the vertical axis), that point was recorded for analysis. The effective operational radius of the hand position was computed as $R = \sqrt{x^2 + y^2 + z^2}$, where x, y, and z were the distances from the hand at the lowest point in the bend to the average foot position. The X-direction was to the subject's right, the Y-direction extended forward, and Z was measured vertically. For visualization purposes, a 3-D Normal distribution with a standard deviation of 1 cm was centered on each point, and these were added together to form a smoothed distribution of the hand positions for each lifting type.

5.1.5 Statistical Analysis

We analyzed the X, Y, and Z hand positions and radius R during lifting with a mixed model analysis of variance (ANOVA) in JMP Pro 14 (SAS, Cary, NC) using a minimum level of significance of 0.05 and considering each variable independently. For measures that were significant using ANOVA, we then used a Tukey-Kramer honest significant difference post hoc test.

5.2 Human Motion Inference with Deep Learning

We study two scenarios for potential application to evaluate the efficacy of human motion inference or approximation. Based on a limited number of sensors, we evaluate upper-body motion inference and full-body motion inference.

We study how to infer the entire upper-body (15 segments, including both arms, the back, the neck, and the head) using sparse sensors by evaluating three different configurations. The first two configurations use both forearm orientation and acceleration values. The first configuration uses the head segment, while the second configuration uses the T8 (sternum) segment. Each orientation and acceleration is normalized relative to the pelvis (see Section 5.2.3). The third configuration uses both forearms normalized relative to the sternum. This requires one less sensor and is a more challenging configuration.

These configurations require a sensor on both wrists but could have the potential of approximating and inferring the entire upper-body. This could be useful for monitoring upper-body posture while working in office environments, approximating postures for virtual reality applications, emerging human-computer interaction applications, or in studying the entire upper-body during rehabilitation.

The final task we study is the inference of full-body motion (23 segments) using sparse segment orientation and acceleration. We study two configurations. Both configurations use orientation and acceleration of both forearms and both lower legs. However, like with the previous task, we use the head's orientation and acceleration in the first configuration and the T8's (sternum) orientation and acceleration in the second. In the third configuration, similar to the last, we normalize both forearms and both lower legs relative to the sternum.

The potential use cases for this task are numerous. In addition to the tasks listed above, several endpoint applications are possible: monitoring the lifting postures of manual laborers; facilitating feedback in stroke rehabilitation and physical therapy; and developing exoskeletons and prosthetics.

We present the different configurations in Table 5.2 as reference.

Table 5.2: Letters denoting the different configurations. W, L, H, S, and P stand for wrists, lower legs, head, sternum, and pelvis, respectively. The last letter of each sequence is the normalization segment (either pelvis or sternum), and the other letters are what is entered into the model. Note the only differences between the upper-body and full-body inputs is information about the lower legs. Using this additional information the full-body models are responsible for predicting the orientation of the 8 segments of the lower-body in addition to the upper-body.

	Config. 1	Config. 2	Config. 3	Config. 4
Upper-Body	W,H,P	W,S,P	W,S	W,P
Full-Body	W,L,H,P	W,L,S,P	W,L,S	W,L,P

Our motivation for choosing configuration 1 for full-body motion inference was to follow the configurations used in [34, 97, 98] that made use of inertial sensors on the wrists, lower legs, head, and pelvis. We then expanded on this configuration to see if better accuracy was possible using configurations 2 to 4. The upper-body motion inference configurations only require the removal of the legs so that they are similar to the full-body configurations.

5.2.1 Inputs and Outputs

We frame this problem as a sequence-to-sequence problem, meaning the inputs and outputs are both sequences over time. For each task, we input a sequence of orientation and acceleration values for different segments and then predict the orientations of additional segments over the same sequence. The orientation and acceleration values are from each segment following sensor-to-segment calibration. This procedure is done using an N-pose where the segment orientations can be approximated [78]. Each segment that is used has an inertial sensor, meaning any end applications could also make use of an N-pose to find the rotation between the sensors' and segments' reference frames. Also, no segment lengths are passed as input to make sure the model generalizes between participants.

To construct sequences, we take a sequence from the dataset, which was recorded at 240 Hz, and downsample it to 40 Hz. Downsampling is common practice [63, 87], and some datasets used for human motion prediction have lower sampling rates of 50 Hz [37]. For every task, we use an initial sequence length of 30, and, after downsampling, we have five frames of data as both input and output.

For our validation set, we stride over the dataset to test on more data. Starting at index 0, we read s frames of data where s is the sequence length. Then we move forward by the stride length l and read another s frames of data. We use a stride of 15 frames in our validation set, whereas in our training set, we use a stride of 30. This effectively tests on similar sequences from different starting points and “doubles” the amount of data we can test.

5.2.2 Representing Rotations

Possible rotational representations include Euler angles, rotation matrices, exponential mapping, and quaternions. Each rotation representation has benefits, but for Euler angles and rotation matrices, the downsides are too large to ignore. We refer the reader to [28] for additional information on each representation.

Euler angles are a useful representation in R^3 for visualizing rotations in space. They provide an interpretation that engineers and mathematicians can easily visualize. However, Euler angles have multiple problems that make them unsuitable for our application: singularities and non-uniqueness. Singularities come from gimbal lock where a loss of a degree of freedom can occur when two axes align, effectively “locking” one of the degrees of freedom. Non-uniqueness, which also happens in other representations like quaternions, occurs because multiple angles on a unit circle can represent the same rotation.

Rotation matrices are another representation that could be used for our application. They avoid gimbal lock and non-uniqueness, but they present a challenge because the set of rotation matrices form the group SO(3). The columns of a rotation matrix form the basis for a reference frame. Thus, constraining rotation matrices to remain in SO(3) requires the columns to be of unit length and orthogonal. To predict rotation matrices, one must satisfy these constraints by normalizing each column and ensuring orthogonality. In addition to this issue, rotation matrices consist of 9 values, the most of any rotation representation, so they require the most computation. Although these issues are present, they have been used in prior work [34]. However, the authors make no mention of constraining the rotation matrices to SO(3), so it is unclear how they did so.

Another possible representation that avoids the singularities present in Euler angles is the exponential map. Exponential maps, although represented with three dimensions, can avoid singularities through the use of mathematical substitutions and Taylor series (see [28]). The exponential map has been used in prior work in human motion prediction and modeling [26, 38, 59, 87]. Thus, this representation is a possible choice in representing rotations and could potentially be used in our application.

Quaternions are a 4-dimensional number system that forms a set in R^4. By constraining them to unit length, we lose a degree of freedom, and we can refer to this new set of quaternions as S^3. Quaternions have been found to work well in human motion prediction [63], most likely because quaternions have the same local geometry and topology as rotation matrices [28]. However, the constraint that quaternions must have unit length adds extra computational load. This requires a manual normalization layer in the network to enforce unit length so that the predicted rotations are a part of S^3.

In deciding between exponential maps and quaternions, we decided to use quaternions for several reasons. First, the results from [63] show a benefit in using quaternions that outweighs the normalization layer that must be added to the neural networks. Second, XSens provides segment orientations in quaternion format, which makes working with them more accessible. Third, we found during testing that quaternions performed better than exponential maps for our use case.

One unaddressed issue with quaternions is that q and $-q$ represent the same rotation. We found that the XSens dataset alternates between these representations, resulting in discontinuities. These discontinuities are similar to the findings in [63] that used the Human3.6M dataset for human motion prediction. We made use of their process for enforcing continuity. The function finds where discontinuities in the time series arise and then chooses the representation that maximizes the dot product.

5.2.3 Normalization

To properly learn and generalize, we normalized the orientation and acceleration data that comes as input into the model. This normalization procedure is nearly the same as in other papers such as [34, 102]. First, we ensured that the orientation was invariant to the direction the person was facing by normalizing the orientation of each segment relative to the root segment. With root orientation $\mathbf{R}_{GP}$, we normalized each segment orientation:

$$\mathbf{R}_{PB} = \mathbf{R}_{GP}^{-1} \cdot \mathbf{R}_{GB} \tag{5.1}$$

Next, we also normalize the acceleration data relative to the acceleration of the root segment and put it in the same frame of reference:

$$\overline{\mathbf{a}}_B = \mathbf{R}_{GP}^{-1} \cdot (\mathbf{a}_B - \mathbf{a}_P) \tag{5.2}$$

For configurations 1, 2, and 4, the root segment is the pelvis. Configuration 3, on the other hand, uses the sternum as the root segment.

After this normalization procedure, we also zero the mean and divide by the standard deviation of each feature in the training set. The mean and standard deviation of the training set is used to also normalize the validation set, under the assumption that both sets come from the same underlying distribution.

5.2.4 Training and Evaluation

We experimented with different model architectures using PyTorch [62]. We evaluate four different neural networks for each scenario and configuration. In total, we train 16 different models to compare and contrast the configurations and algorithms. The algorithms, described in Chapter 4, are a vanilla Seq2Seq model, a Seq2Seq model with a bidirectional encoder and attention, a Transformer Encoder, and a full Transformer.

For both Seq2Seq models, we used a hidden size of 512. For the Transformer Encoder, we used a feedforward size of 200, 2 encoder layers, and e heads in the multi-head attention layer where e is the dimensionality of the input. We found this worked the best for training. For the Transformer, we used two different configurations for upper-body and full-body motion inference. Unlike other models, like Seq2Seq architectures, we found different hyperparameter settings worked best for the two different output settings. We found a feedforward dimension of 2048, 2 layer encoder and decoders, and four heads in the multi-head attention layers worked the best for upper-body motion inference. Also, we were able to get better results with full-body motion inference when using a feedforward dimension of 512, 6 layer encoders and decoders, and four heads in the multi-head attention layers.

We train our models on a single V100 GPU. We used the AdamW optimizer [51] with a learning rate of 0.001. For the Transformer models, we multiplied this learning rate by 0.1 every two epochs. We use a batch size of 32 for each model. We found that the models learned quickly using AdamW. We experimented with SGD and found it required around 20 epochs to reach accuracy that AdamW reached in only 3-5 epochs. The models can be trained in under 4 hours using a V100 GPU and AdamW.

For each scenario, we used the same training/validation split. We placed Virginia Tech participants 1, 2, 3, 4, 6, 7, 8, 9, 11, and 12 along with workers 1 and 4 in the training set. In the validation set, we placed participant 5 along with workers 2 and 3. We split the data this way to have representative samples from the Virginia Tech participants and manual material handlers in both the training and validation set. In all, we used 850,114 and 290,888 sequences for training and validation, respectively.

The loss function used for training is mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{5.3}$$

where $\hat{y}$ is the predicted posture, and y is the ground truth posture.

We provide the best test results for each model using the mean angle difference between the ground truth quaternions q and the predicted quaternions $\hat{q}$ for each segment in degrees:

$$\bar{\theta} = \frac{360}{\pi n}\sum_{i=1}^{n} \arccos(|\langle \hat{q}_i, q_i \rangle|) \tag{5.4}$$

where i is indexing the individual body segments in the output, n is the number of segments, and $\langle \cdot, \cdot \rangle$ is the inner product between two quaternions. We found that MAE was more effective as a loss function compared to mean angle difference. There are multiple options for computing the closeness of two quaternions that are less computationally expensive [35]. However, we found that this metric was useful to report differences in degrees.

Forward kinematics was performed after training to validate the models qualitatively. This procedure requires building the kinematic chain starting at the pelvis. The forward kinematics software uses the segment orientations and then multiplies by a single participant's segment lengths taken from an XSens MVNX file. We used the following equation to perform forward kinematics given the underlying rotation of the segment [78]:

$$\mathbf{p}_G^{landmark} = \mathbf{p}_G^{origin} + \mathbf{R}_{GB} \cdot \mathbf{x}^{landmark} \tag{5.5}$$

where G refers to the global reference frame, and B refers to the segment's reference frame. As an example, the origin is initially the pelvis. Then using the orientation and the set length of the right upper leg, the position of the right upper leg is determined. This is continued throughout the kinematic chain.

Although normalization is performed to improve generalization, we multiply by the orientation of the pelvis to view the posture as it would be viewed without normalization for qualitative evaluation. To do so, we use the following equation on the predicted poses:

$$\mathbf{R}_{GB} = \mathbf{R}_{GP} \cdot \mathbf{R}_{PB} \tag{5.6}$$

We present a qualitative evaluation for the full-body scenario as a means of showing what the models are capable of approximating. We provide a set of predictions from a seq2seq model, a Transformer encoder, and a full Transformer alongside a ground truth reference posture. All of the postures come from the validation set.

Chapter 6

Experimental Results

We describe the experimental results in this chapter for both our study of manual material handlers and our application of the dataset to human motion inference.

6.1 Quantification of Postures for Low-Height Object Manipulation

6.1.1 Overall Bending Postures, Frequencies, and Durations

We generated plots to understand the overall bending/lifting profile of each participant (Figure 6.1). Figure 6.1a shows a cumulative distribution function (CDF) of each worker's T8 pitch angle during the entire dataset. Each participant spent the majority of their time walking or standing with a straight back, and less than 10% of their time at pitch angles greater than 60 degrees. Most lifts were quite brief, between 1.5 and 3.5 seconds in duration (Figure 6.1b). Although a small portion of time was spent lifting, each worker's bending frequency was high (Figure 6.1c), with up to 8 lifts per minute.

A histogram of foot height distance (Figure 6.1d) reveals that workers would frequently lift their back heel off the ground by a centimeter or two, which leads to a height change of the back foot without the lift being one-legged. One-legged lifting corresponded to a height

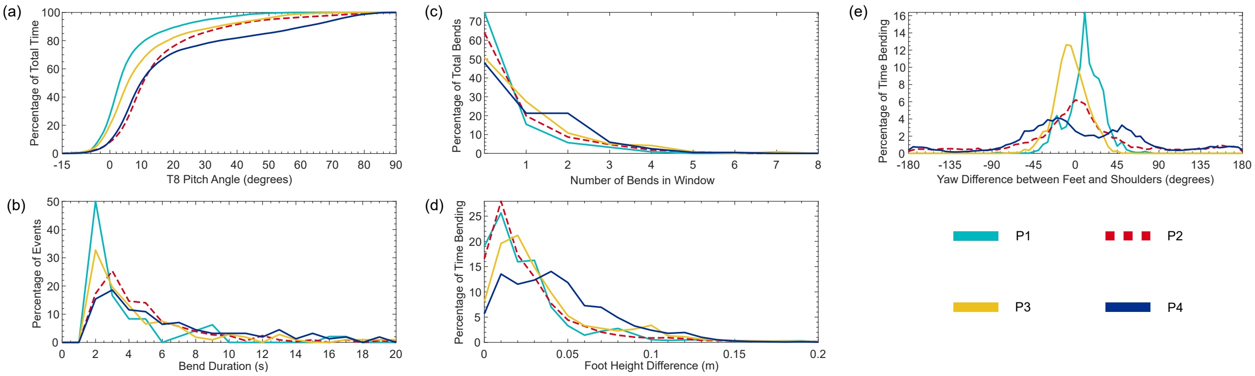


Figure 6.1: Generated plots from participant data. (a) A Cumulative Distribution Function (CDF) of T8 pitch over the entire work period for each participant. (b) A Probability Density Function (PDF) of bend duration for each participant. (c) A PDF of bends per one-minute window for each participant. Participants bent or lifted once per one-minute window approximately 20% of the time. (d) A PDF of foot height difference during each lift for each participant. The maximum frequency at 1 centimeter of height difference is likely due to noise in the XSens IMUs. Overall, 12.3% of the lifts were split-legged lifting with a heel off the ground, and 22.1% of the lifts had a foot entirely off the ground as in one-legged lifting. During the one-legged lifts, typically, the heel was raised for a long time, then the foot left the ground entirely for only a short time. (e) PDF of foot and shoulder yaw difference during lifting, illustrating lifting posture asymmetry.

change of greater than roughly 6 cm. Asymmetric lifting was observed in each of the lifters' profiles (Figure 6.1e).

6.1.2 Lifting Types

Table 6.1 shows a summary of the different bending postures observed. Only participant 4 performed fore-aft squatting, but each participant used the other bend postures frequently. Notably, mixed lifts had an average duration of 12.75 seconds. The participants were moving objects and sorting bins during the mixed lifts, which accounts for this length of duration.

Table 6.1: Overall lifting data for the participants in the study. Total lift count refers to the total number of lifts that we manually labeled for that day. Duration for each lift type ("Dur.") refers to the average minutes per hour spent in that posture. N is the number of lifts of that type that were classified during labeling. Average duration (Avg. Dur.) refers to the average time spent in each posture overall. Finally, the average percentage (Avg %) of the lift types observed in the data is reported. The posture indicators (a - e) correspond to Figure 5.2.

	Length	Total #	Stoop (a)		Fore-aft Squat (b)		Split-legged, heel raised (c)		Split-legged (d)		One-legged (e)		Mixed	
P1	(HH:MM)	Lifts	Dur. (min/hour)	*N*	Dur. (min/hour)	*N*	Dur. (min/hour)	*N*	Dur. (min/hour)	*N*	Dur. (min/hour)	*N*	Dur. (min/hour)	*N*
Day 1	1:58	47	0.61	18	0	0	0.07	4	0.32	6	0.27	14	0.49	5
P2														
Day 1	2:40	145	1.14	29	0	0	0.32	13	1.98	52	0.58	20	3.33	31
Day 2	4:19	118	0.83	41	0	0	0.25	16	0.42	21	0.37	28	0.37	12
Day 3	1:20	29	0.39	9	0	0	0.15	3	0.24	6	0.31	10	0.10	1
Day 4	0:43	31	1.41	7	0	0	0.80	9	0.74	8	0.14	3	0.74	4
P3														
Day 1	1:58	106	1.50	30	0	0	0.35	13	0.79	24	0.58	22	1.32	17
P4														
Day 1	1:54	98	3.34	23	2.11	13	0.59	18	0.77	17	0.24	10	3.60	17
Day 2	1:22	92	0.97	7	6.20	22	0.38	7	1.78	22	1.72	23	2.09	11
All			Avg. Dur. (sec)	Avg %	Avg. Dur. (sec)	Avg %	Avg. Dur. (sec)	Avg %	Avg. Dur. (sec)	Avg %	Avg. Dur. (sec)	Avg %	Avg. Dur. (sec)	Avg %
			7.46	27.3	20.82	4.6	3.31	12.3	5.16	20.3	3.24	22.1	12.75	13.3

6.1.3 Hand Position and Range of Motion During Lifting

Though our sample size was small (666 bends across the four participants), some significant differences could be seen between lifting types. The average y-distances (distance in front of the person) and effective reach radius R were significantly different: quantitative data on the hand position for each lifting type and the Tukey-Kramer honest significant difference test gave results shown in Table 6.2 ($p < 0.05$ for all pairwise differences). Also, the average z-height of fore-aft squat lifting was significantly lower than that of the other lift types ($p < 0.001$). Note that since only one subject completed fore-aft squat lifting, the results for this lifting type (in all of the directions) are not necessarily generalizable. There was no statistically significant difference between any of the lift types in the x-direction (frontal or horizontal axis). These results can be qualitatively observed in Figure 6.2, which shows contours of the participants' hand position during the different lifting postures. The contours occupy different regions based on the lift type.

Table 6.2: Measures of hand position throughout different lifting postures. Averages that have different superscript letters in each column are significantly different (Tukey-Kramer HSD pairwise comparisons). The letters are arbitrary designators of different groupings.

	Avg. (cm)				Std. Dev. (cm)				Min. (cm)				Max. (cm)			
Posture	x	y	z	R	x	y	z	R	x	y	z	R	x	y	z	R
Fore-aft Squat	9	20^a	10^c	31^e	13	13	13	11	-16	-4	-10	11	34	50	35	59
Stoop	1	28^a	53^d	64^f	15	15	14	14	-31	-2	27	40	30	55	86	98
Split-legged	1	30^a	57^d	70^{fg}	21	14	14	15	-42	3	32	42	47	60	89	104
Split-legged, heel raised	5	39^{ab}	55^d	76^{gh}	27	15	16	19	-38	14	31	46	48	67	84	109
One-legged	6	45^b	53^d	79^h	29	20	16	17	-46	8	24	51	58	86	86	117

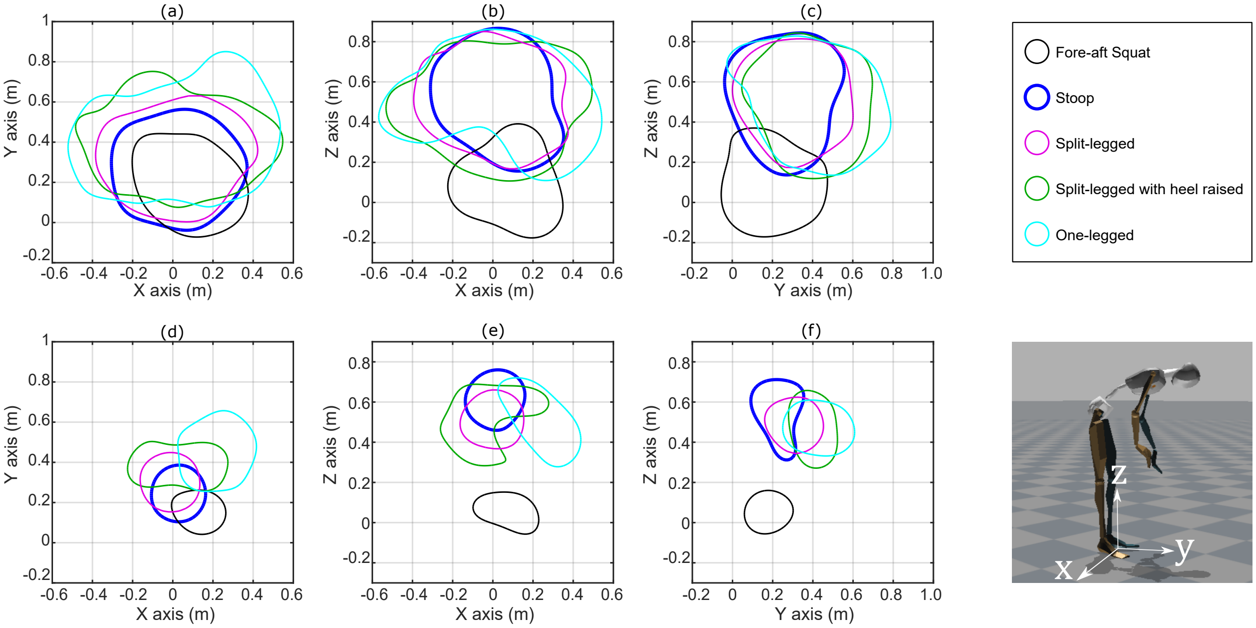


Figure 6.2: In-plane contour plots for the position of the hand-in-use during different bending motions. The outer contours along the top row are at levels capturing 75% of the distribution's volume. The inner contours along the bottom row are at levels capturing 25% of the distribution's volume. (a) Outer contour of hand position in xy (transverse) plane; (b) Outer contour of hand position in xz (frontal) plane; (c) Outer contour of hand position in yz (sagittal) plane; (d) Inner contour of hand position in xy (transverse) plane; (e) Inner contour of hand position in xz (frontal) plane; (f) Inner contour of hand position in yz (sagittal) plane. The contour plots have negative positions along the z-axis for fore-aft squatting. This is due to the hand being level with the floor in some lifts, combined with the standard deviation of the Normal distribution used for smoothing, which slightly expanded the volume of hand positions.

6.2 Human Motion Inference with Deep Learning

To understand the performance of our models, we study them using quantitative and qualitative evaluations. The quantitative evaluation is in two stages. The first covers the performance on the validation set using mean angle difference metric described in Section 5.2.4. We then evaluate each model using histograms of this metric to better understand the performance. For qualitative evaluation, we view various representative postures to determine how well the model generalizes and where it fails. Each evaluation is performed on our validation set which contains data from participant 5 and workers 2 and 3. Participant 5 went to a drawing/sketching club meeting. Workers 2 and 3 worked at a home improvement store. We chose these three participants as a representative group since they come from both the college campus participants and the home improvement store employees.

6.2.1 Quantitative Evaluation

We present model performance for upper-body inference with four configurations (Table 6.3). The Transformer performs marginally better than the other models under all configurations.

Table 6.3: The performance of the various models with differing configurations for upper-body motion inference. The values reported are the mean angle difference in degrees for the validation set. Config. 1 and Config. 2 use accelerations and orientations normalized relative to the pelvis. Config. 1 and 2 both make use of the forearms, but they differ in using the head and sternum, respectively. Config. 3 and Config. 4 use both forearms normalized to the sternum and pelvis, respectively. Config. 1 and 2 use 4 sensors while Config. 3 and 4 use 3 sensors. Each configuration uses the entire upper-body as output (15 segment orientations with the pelvis included). See Table 5.2 for the configurations.

Model	Config. 1	Config. 2	Config. 3	Config. 4
Seq2Seq	12.80	11.81	13.55	15.97
Seq2Seq (BiRNN, Attn)	12.49	11.63	13.39	15.71
Transformer Enc.	12.59	13.67	13.50	15.90
Transformer	**12.36**	**11.59**	**13.34**	**15.66**

We also present model performance for full-body inference under four configurations (Table 6.4). The Transformer model performs the best under all configurations except for four, most significantly under configuration two.

Table 6.4: The performance of the various models with differing configurations for full-body motion inference. The values reported are the mean angle difference in degrees for the validation set. Config. 1 and Config. 2 use accelerations and orientations normalized relative to the pelvis. Config. 1 and 2 differ in using the head and sternum, respectively. Config. 3 and Config. 4 use both forearms and both lower legs normalized to the sternum and pelvis, respectively. Config. 1 and 2 use 6 sensors while Config. 3 and 4 use 5 sensors. Each configuration uses the full-body (23 segment orientations) as output. See Table 5.2 for the configurations.

Model	Config. 1	Config. 2	Config. 3	Config. 4
Seq2Seq	12.87	12.20	12.91	14.73
Seq2Seq (BiRNN, Attn)	12.50	11.94	12.64	**14.46**
Transformer Enc.	12.65	11.93	14.10	14.88
Transformer	**12.39**	**11.61**	**12.50**	19.69

Overall, configuration two performs better than every other configuration, using the mean angle difference for both full-body and upper-body motion inference. However, it is surprising that configuration three is competitive in full-body motion inference. It uses information about only five segments but still maintains similar (although slightly reduced) accuracy compared to the configurations using six segments.

Compressing the performance of different models to a single number can be misleading, so we also report histograms of sequence angular error in degrees for upper-body motion inference in Figure 6.3.

It is noticeable that for each model, except the Transformer Encoder, the end of the distribution's tail drops off faster when using configuration two in comparison to other configurations. Configuration four has the worst performance for each model and the longest tail in each case. It is of note that configuration three performs better than configuration four. This encourages the conclusion that the sternum is more informative as a point of reference for

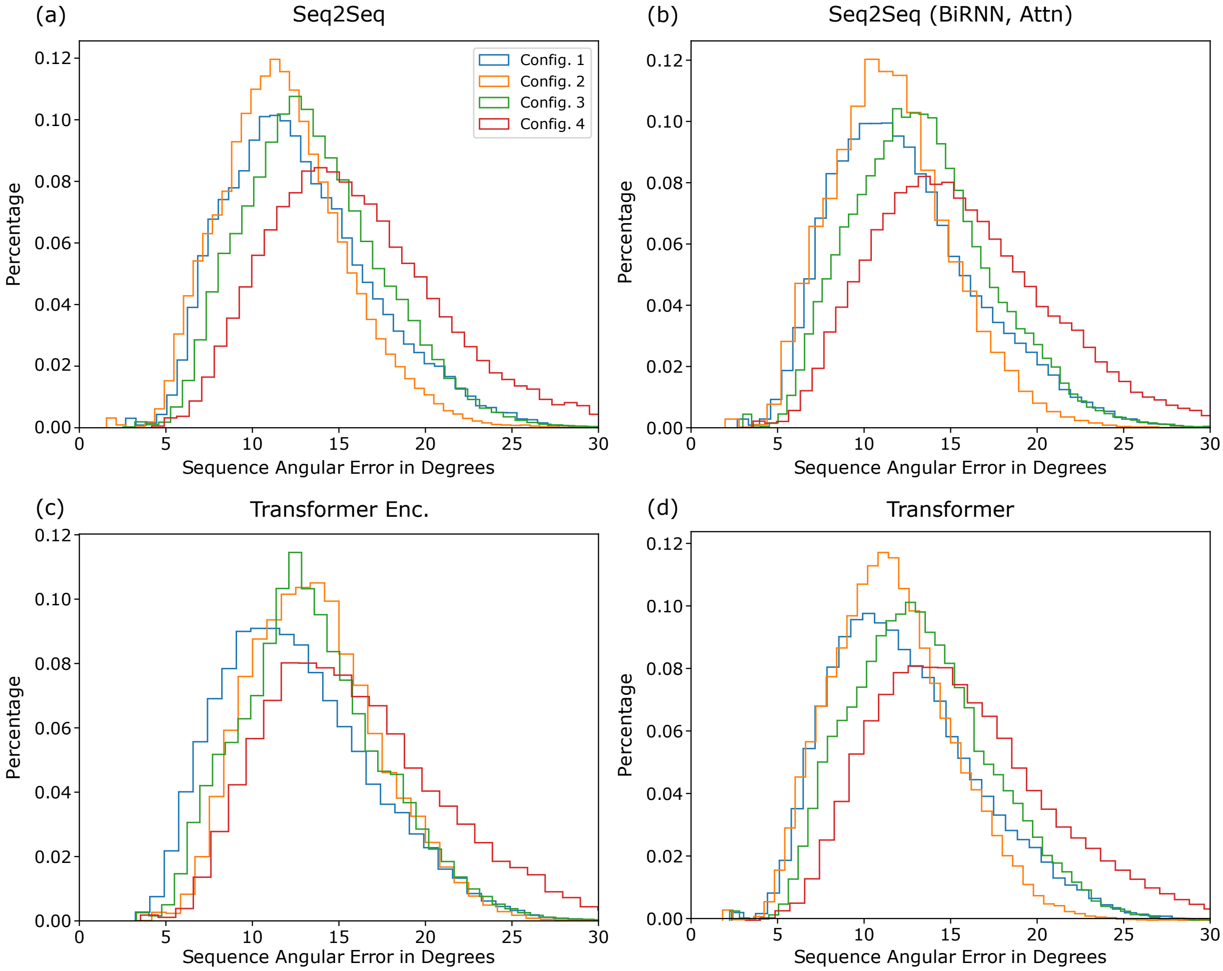


Figure 6.3: Histograms showing the performance of the different models on the validation set under different upper-body configurations. The letters correspond to different models: (a) Seq2Seq model, (b) Seq2Seq model with bidirectional encoder and attention, (c) Transformer Encoder, and (d) Transformer.

upper-body motion inference than the pelvis.

We also report histograms of sequence angular error in degrees for upper-body motion inference in Figure 6.4.

For configuration two, the end of the tail noticeably drops off faster in comparison to other configurations. The center of the distributions for configuration two is also lower, in agreement with Table 6.4. It is also clear that configuration four performs the worst for each

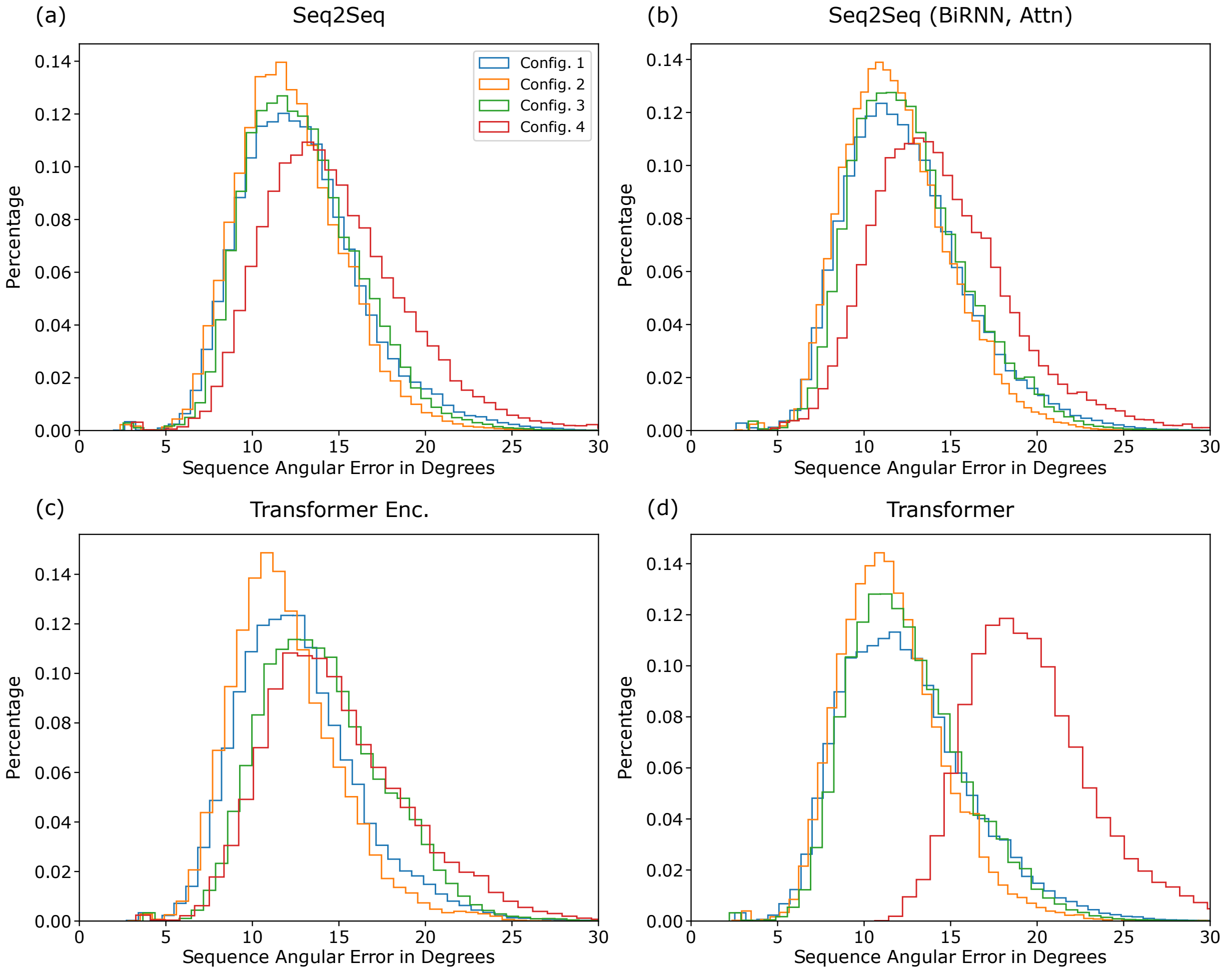


Figure 6.4: Histograms showing the performance of the different models on the validation set under different full-body configurations. The letters correspond to different models: (a) Seq2Seq model, (b) Seq2Seq model with bidirectional encoder and attention, (c) Transformer Encoder, and (d) Transformer.

model. This is very noticeable when using the Transformer (Figure 6.4d). Under each model, configuration one and three are quite similar. This is true despite configuration one using both the head and the pelvis and configuration three using only the sternum with no reference to the head or the pelvis. Configuration three and four use the same number of sensors, but configuration three performs noticeably better.

6.2.2 Qualitative Evaluation

Though quantitative evaluation is useful for viewing concise metrics about the performance of the models, qualitative evaluation is necessary to build intuition for how the models make predictions and where they fail. Our goal in this section is to visualize the postures that the models predict to build intuition about what they get wrong. Qualitative evaluation is performed in most every study of human motion inference such as in [34, 65, 105]. In this section, we look at postures generated by the Seq2Seq model with a bidirectional encoder and attention, the Transformer Encoder model, and the Transformer model. We use configuration two from the full-body set to investigate how the models differ. Configuration two was chosen because it gave the best results out of the four configurations. The upper-body set has very similar output, so we only present the full-body set here to avoid repetition.

In each of the figures in this section, we label the different rows with letters (a, b, c, d), which represent different postures that the model must infer using sparse segment orientation and acceleration. Our goal here was to choose various representative postures such as standing, sitting, and bending to understand where the models fail and succeed. The first column is the reference (ground truth) posture, and the remaining columns are predictions from each model. We provide descriptions of the postures in the captions.

The first set of postures comes from participant 5 in our validation set (Figure 6.5).

Of note is that each model correctly predicts the participant is sitting down in each posture. In posture (a), the participant is sitting and typing on a laptop in their lap. Each model has subtle inaccuracies in inferring this motion, but most seem to infer realistic motion. In posture (b), the participant has their elbows on the table, but the Transformer incorrectly infers the participant is lifting their right elbow, and the Transformer Encoder is inferring inaccurate feet orientation. In posture (c), the participant is reaching down into their bag.

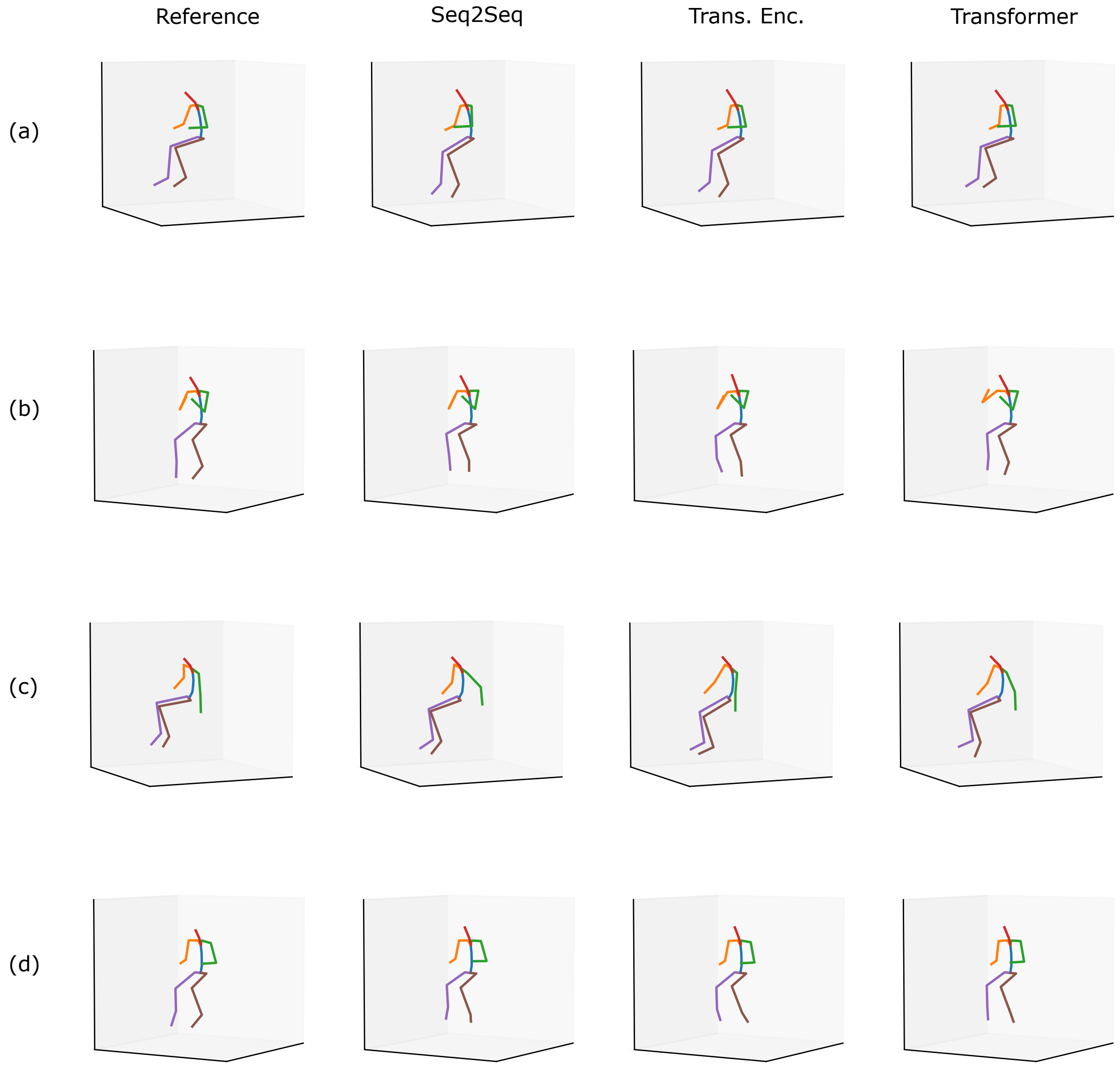


Figure 6.5: A set of sitting postures from the validation set where the participant went to a drawing/sketching club meeting. The rows are of different postures such as (a) sitting at a desk and typing on a computer, (b) sitting at a desk with elbows on the table, (c) sitting at a desk reaching down into their bag, and (d) sitting at a desk and sketching.

The Seq2Seq model infers that the person is reaching further back than they are in reality. In posture (d), the participant is sketching instead of typing, and the Transformer Encoder and Transformer infers inaccurate feet orientation.

Another set of postures comes from worker three, who was working at a home improvement store (Figure 6.6).

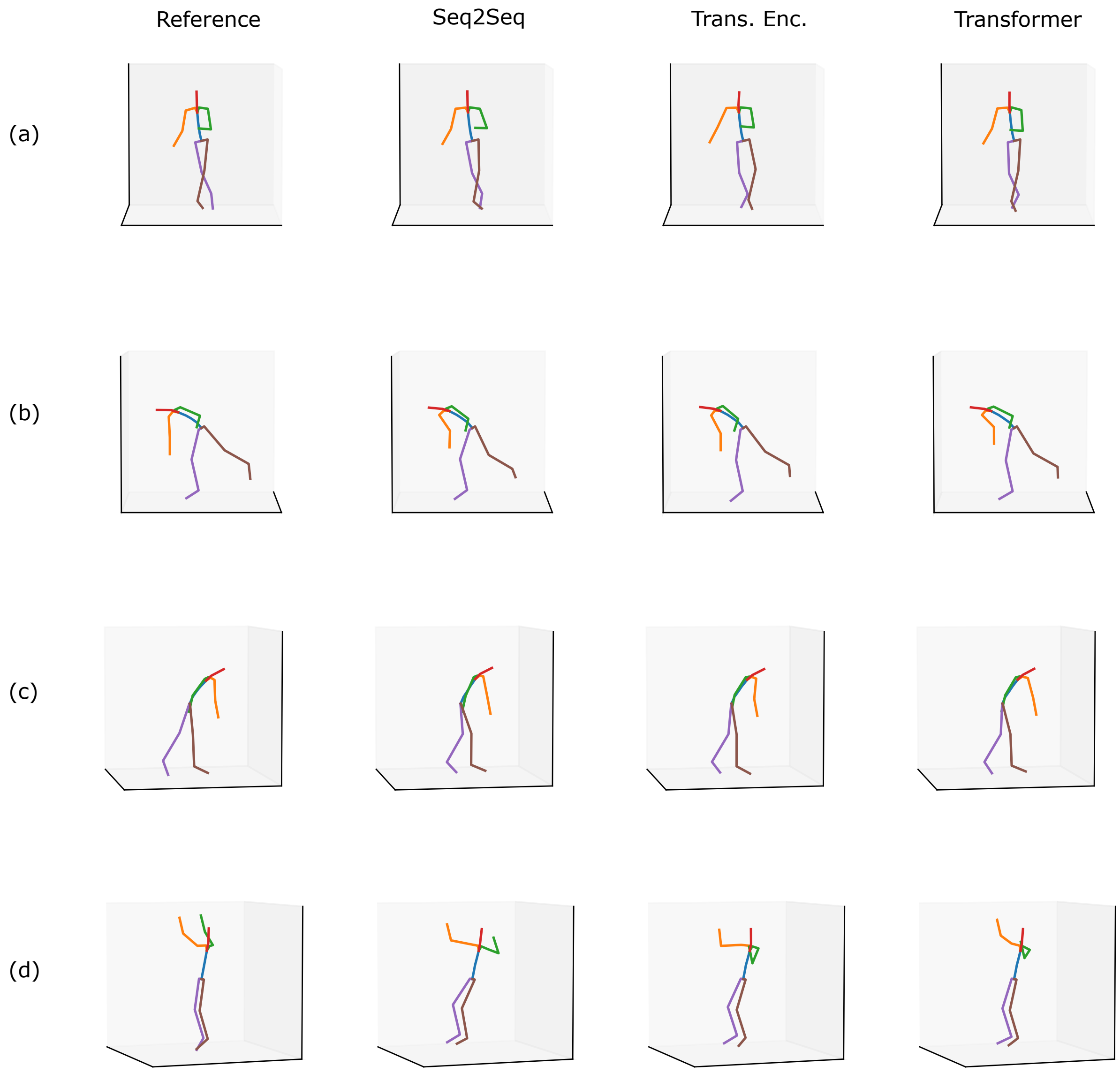


Figure 6.6: A set of postures from the validation set where the participant was working at a home improvement store. The rows are of different postures such as (a) standing with legs crossed talking with someone, (b) reaching for something in a one-legged ("golfer's") lift, (c) reaching for something with legs split fore-aft, and (d) reaching up for something overhead.

Similar to Figure 6.5, the models can all determine that the participant is standing up instead of sitting down. In posture (a), the participant has their legs crossed while leaning on something. Each model except for the Transformer Encoder accurately models the participant crossing their legs. In postures (b) and (c), the participant is doing a one-legged lift and split-legged lift with a heel raised, respectively. Each model can infer this correctly, though there is some inaccuracy in the right arm orientation for posture (b). In posture (d), the Seq2Seq model and Transformer Encoder fail to infer accurate upper arm motion, whereas the Transformer is reasonably accurate in inferring the right arm's orientation. Each model also fails to infer that the participant is on their toes.

The third set of postures is from worker two who was also working at a home improvement store (Figure 6.7).

This set of postures is interesting because it demonstrates the range of motions the models can perform inference for, and where the models fail. In posture (a), the participant is walking with their hand on a cart behind them. Each model has accurate inference about the posture. In posture (b), the participant is kneeling on the ground while reaching for something. Each model has varying predictions for how extreme the knees' flexion is, but each seems to correctly infer that the participant is kneeling. In posture (c), the participant is putting on a vest or jacket. This posture is particularly challenging because of the unusual way the arms move during the activity. The Transformer Encoder, in particular, incorrectly predicts the participant is sitting down. Also, each model inaccurately infers the upper arm orientation. Each model has varying inaccuracies. Finally, in posture (d), the participant is lifting something with both hands. There are varying degrees of inaccuracy between each model. Each model seems to be conservative about how far apart the arms are.

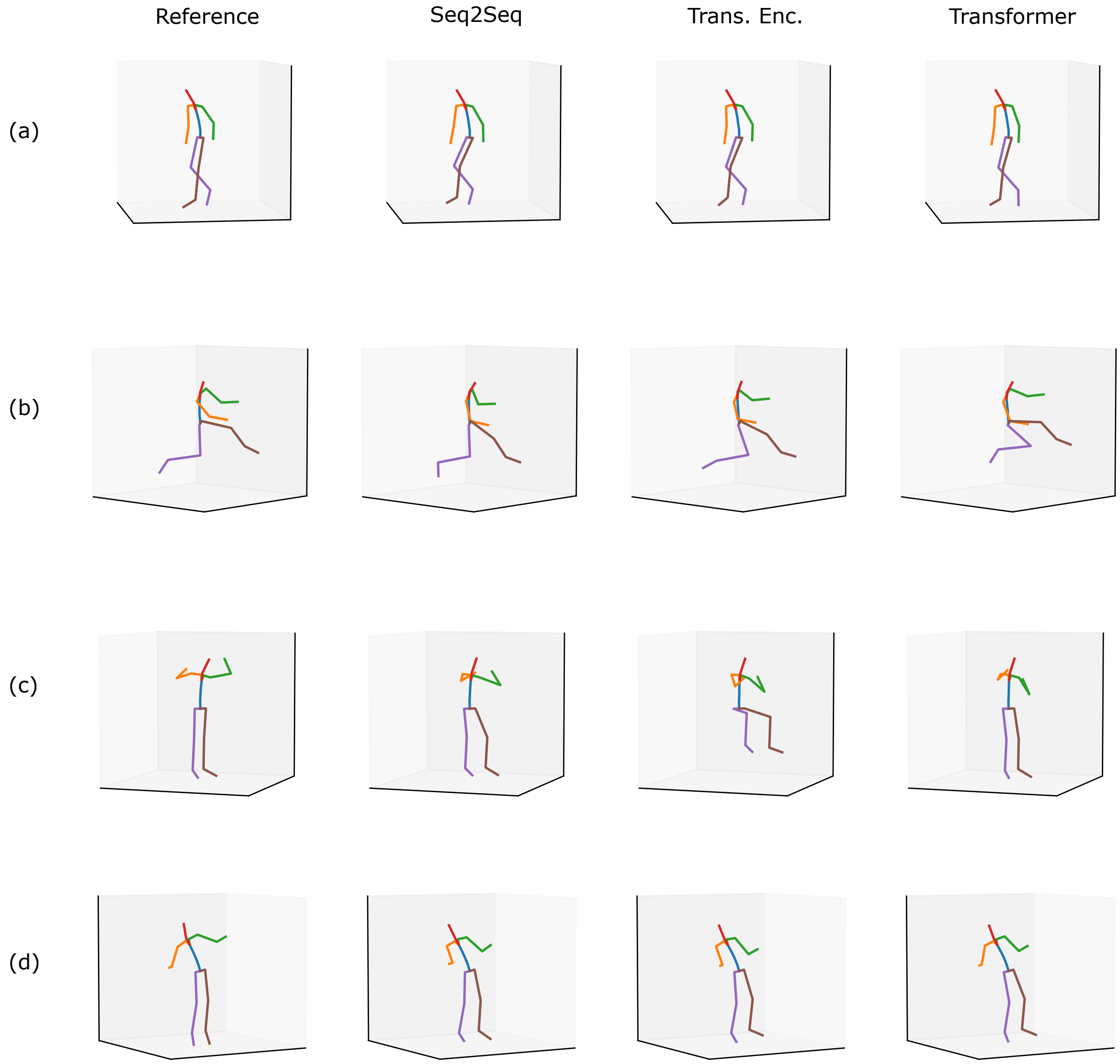


Figure 6.7: A set of postures from the validation set where the participant was working at a home improvement store. The rows are of different postures such as (a) walking with their hand on a cart, (b) kneeling on the ground while reaching for something, (c) putting on a vest, and (d) lifting something with both hands.

6.2.3 Failure Cases

We also present a typical set of failure cases (see Figure 6.8). Each of these postures has greater than 20^o of angular error when using the Transformer. These inaccurate predictions are few in number as shown in Figure 6.4. However, for endpoint applications such as ergonomic monitoring and stroke rehabilitation, it is good to know about typical failure cases to be aware of.

In (a), the Transformer predicts that the person is sitting down while they are instead operating machinery (possibly a stand-up forklift). The accelerations from the forklift possibly cause this error to occur. In (b), the participant correctly predicts the participant is reaching across their chest, but incorrectly predicts the amount of knee flexion present in the posture. In (c), the participant is reaching across their chest with one hand. The Transformer model output looks decent, but this still resulted in high angular error. Finally, in (d), the participant is performing an action with high elbow flexion. The Transformer correctly predicts this but has a large amount of error in predicting the leg orientation.

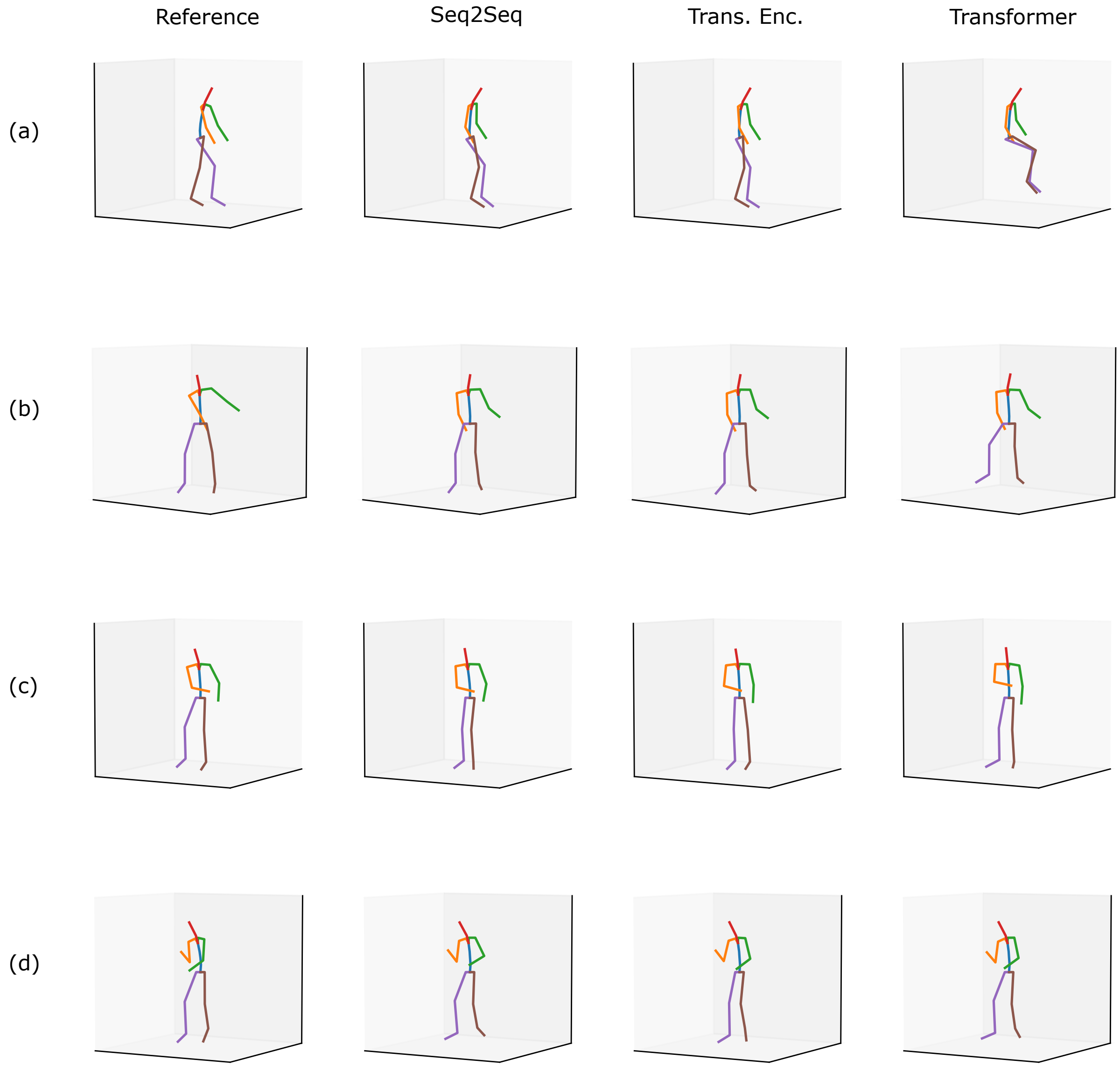


Figure 6.8: A set of postures from the validation set where the participant was working at a home improvement store. The Transformer had low accuracy ($> 20^o$) for each of these postures. The rows are of different postures such as (a) operating machinery (possible a stand-up forklift), (b) reaching for an object or box across their chest, (c) reaching for an object with one hand across their chest, and (d) performing some action with high elbow flexion.

Chapter 7

Discussion

7.1 Limitations

There are many limitations to our study of worker postures. Some of our limitations are similar to those of other studies concerning posture monitoring [76]. The number of days during which we collected data was limited, and in the literature it is recommended that four to twelve days be collected to determine habitual postures [10, 89, 100]. Though our dataset may contain habitual postures, these postures cannot be determined. Even after monitoring workers for multiple days, work conditions or tasks may change that cause a change in lifting habits.

Also, our study of workers cannot be generalized to material handlers overall. Since the study was limited to four (N=4) material handlers in a single retail store, future work could include a more extensive study in multiple stores with a more significant number of workers. Only a single participant performed squat lifting, so the lifting type's statistical results should be treated as a within-subject result. More experiments should be done to determine if this holds more broadly.

Although our natural motion dataset contains around 36 hours of human motion in unconstrained real-world environments, the number of participants (N=16) is still limited compared to other motion capture datasets [53]. It is also unknown how many unique postures are in the dataset. In future work, this dataset can be expanded upon and labeled.

Finally, the XSens suit only records kinematics. Recording dynamics using force sensors would be much more informative as to why workers choose specific postures and to get a clearer picture of when participants are interacting with their environment. The profile of lifting motions and other activities is affected by object weight, and our study was not able to characterize this. Also, synchronized on-body cameras were not used, which could have been used to capture additional context. Future studies could incorporate force sensors and cameras to gain information and context about human behavior.

7.2 Conclusions

In this paper, we have introduced, as far as we know, the largest inertial motion capture dataset. The dataset contains 36 hours of unscripted natural human motion in real-world, unconstrained environments. Unlike other large-scale motion capture datasets, we were able to capture data in several interesting environments like a home improvement store, all across a college campus, a person's car, and a shopping store.

Though we only had data from four manual material handlers and the data collection days were limited, many postures that were used frequently by the participants have not been frequently investigated: fore-aft squatting, split-legged lifting, poses with one heel off the floor, holding onto external objects for balance, and one-legged lifts. Our posture classification scheme for low-height object manipulation allows for studying different postures in more detail. In general, this study provides evidence for a much broader set of activities during manual material handling than previously discussed. These should be studied additionally, to understand the biomechanics and MSD risk for these motions. Much work has been done in developing advanced approaches for lifting types like stooping and squatting [8, 36, 99].

For example, lumbar lordosis during split-legged lifting and one-legged lifting could be a topic of interest [36]. In contrast, asymmetric lifting was used frequently in our study, but it has been researched a great deal by other groups [19, 24, 43, 58, 94].

There are many reasons why specific postures may be chosen over others. Different postures vary in their energy use, flexibility required, the possible range of motion, comfort in the posture, stability, and ease of transitioning to/from the posture. Certain hand positions seem to be easier or more comfortable in specific postures: the workers used one-legged lifts or split-legged lifts with a heel raised to reach objects that were extremely far forward. The average bend duration for a posture is possibly related to stability in that posture: we note that the average stoop lift is nearly twice as long as the average one-legged lift or split-legged lift with a heel raised. More data and participants are needed to make claims or determine why the fore-aft squat is used. We hypothesize that while workers were trained to lift using squats, most did not do so, possibly due to the higher energy requirements of squatting [27] or the fact that many objects lifted were light.

In contrast, the one-legged or split-legged postures may have lower energy use than stoop lifts, since the rear leg acts as a counterbalance. In this posture, the pelvis may be more forward, or the back may be more in lordosis than a stooped posture, which may reduce muscular effort. Agility is needed to do stocking quickly and efficiently if a work area is large, and stooping and split-legged lifting provide this. Further studies are needed to reach definite conclusions about these matters. Understanding why workers select each lift type can lead to ergonomic interventions, physical therapy, or training programs (e.g., stretching to increase flexibility) that will facilitate workers using bending postures with lower injury risks. Other reasons for choosing postures should be analyzed in future on-site studies including the impact of Body Mass Index, weight, body type, musculature, and amount of weight training.

Our study into human motion inference has shown that a wide range of human motion can be inferred with limited segment information using Seq2Seq models and Transformers. Other approaches, such as [34], also make use of neural networks for human motion inference using sparse sensors. They predict SMPL [49] model parameters using 20 past frames and 5 future frames at test time using a bidirectional LSTM. They synthesize their IMU data using AMASS [53] and then fine-tune on a real IMU dataset that is 90 minutes in length. An important point is that their system is also real-time while ours is not. Overall, their real-time system and use of SMPL model parameters is an impressive and novel contribution to the field.

In contrast, we predict segment orientations instead of SMPL model parameters. We also frame the problem as a sequence-to-sequence problem and use Seq2Seq models and Transformers instead of only RNNs that predict single poses. We do not make use of future frames; we map a sequence of sparse segment orientation and acceleration data to a sequence of full-body or upper-body orientations. Also, our dataset for training and validation is, to the best of our knowledge, the largest real inertial motion capture dataset and is not synthetic. With all of this said, one of our novel contributions, to the best of our knowledge, is the application of the sequence-to-sequence paradigm (Seq2Seq models and Transformers) to the largest inertial motion capture dataset, which also contains unscripted, natural motion in real-world environments.

In comparison to [34], we also found that our histograms have shorter tails, which is worth noting as this may lead to fewer failure cases in endpoint applications. However, this may be due to different underlying motions in each of our datasets and that they use joint angles instead of segment orientations to measure angle differences. Though their system is real-time, both of our approaches use neural networks, so there are not infeasible technical challenges to making our system real-time as well.

Importantly, we found that using the sternum is more useful than the head for both upper-body and full-body motion inference. This is a novel contribution of our work, to the best of our knowledge, because other research has not shown a direct benefit to using the sternum. As a point of reference, the sternum is also more useful than the pelvis. This is most likely because the head and pelvis provide less information about the segments along the back. This has important implications for future applications. An IMU may not move much if attached to a person's glasses or hat instead of the front or back of their shirt but may lower the accuracy in inferring their motion based on our results. In terms of comfort, people may find different configurations intrusive such as wearing a hat or a clip on the front of their shirt's neck. Wearing a clip on the back of their shirt's neck may be less noticeable and more acceptable.

Several postures are ambiguous and hard to model with sparse segment information. For example, putting on a jacket like in Figure 6.7c contains forearm orientation and acceleration that is rare as the person's forearms are behind them awkwardly. Another example includes reaching up high for something, such as in Figure 6.6d. The models tend to predict postures that are more common and less rare in these cases. Surprisingly, the models do not tend to fail catastrophically from the samples that we have seen. In other words, when viewing failure cases, we did not see cases where the limbs were tangled together or unrealistic joint angles were predicted. The models are inaccurate in some cases but do not seem to predict unrealistic postures.

In this thesis, we have presented methods for understanding natural human motion using deep learning and traditional techniques. We plan to release our dataset, our source code and our trained models for human motion inference for use by other researchers.

Bibliography

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015.

[2] Jay Alammar. The Illustrated Transformer, 2018. http://jalammar.github.io/illustrated-transformer, last accessed on 04/25/20.

[3] Sheldon Andrews, Ivan Huerta, Taku Komura, Leonid Sigal, and Kenny Mitchell. Real-time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European conference on visual media production (CVMP 2016)*, pages 1–10, 2016.

[4] MF Antwi-Afari, H Li, DJ Edwards, EA Pärn, J Seo, and AYL Wong. Biomechanical analysis of risk factors for work-related musculoskeletal disorders during repetitive lifting task in construction workers. *Automation in Construction*, 83:41–47, 2017.

[5] Marie Authier, Monique Lortie, and Micheline Gagnon. Manual handling techniques: comparing novices and experts. *International Journal of Industrial Ergonomics*, 17(5): 419–429, 1996.

[6] Norman I Badler, Michael J Hollick, and John P Granieri. Real-time control of a virtual human using minimal sensors. *Presence: Teleoperators & Virtual Environments*, 2(1): 82–86, 1993.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[8] Babak Bazrgari, Aboulfazl Shirazi-Adl, and Navid Arjmand. Analysis of squat and stoop dynamic liftings: muscle forces and internal spinal loads. *European Spine Journal*, 16(5):687–699, 2007.

[9] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[10] Jaime E Berlin, Kristi L Storti, and Jennifer S Brach. Using activity monitors to measure physical activity in free-living conditions. *Physical therapy*, 86(8):1137–1145, 2006.

[11] BLS. Nonfatal Occupational Injuries and Illnesses Requiring Days Away From W, 2015. Technical report, United States Bureau of Labor Statistics, 2016. www.bls.gov/iif/oshcdnew.htm, last accessed on 04/25/20.

[12] Bryan Buchholz, Victor Paquet, Laura Punnett, Diane Lee, and Susan Moir. Path: a work sampling-based approach to ergonomic job analysis for construction and other non-repetitive work. *Applied ergonomics*, 27(3):177–187, 1996.

[13] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6158–6166, 2017.

[14] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[15] Jiayu Chen, Jun Qiu, and Changbum Ahn. Construction worker's awkward posture

recognition through supervised motion tensor decomposition. *Automation in Construction*, 77:67–81, 2017.

[16] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[17] Geoffrey David, Valerie Woods, Guangyan Li, and Peter Buckle. The development of the Quick Exposure Check (QEC) for assessing exposure to risk factors for work-related musculoskeletal disorders. *Applied ergonomics*, 39(1):57–69, 2008.

[18] Fernando De la Torre, Jessica Hodgins, Javier Montano, Sergio Valcarcel, R Forcada, and J Macey. Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database. *Robotics Institute, Carnegie Mellon University*, 5, 2009.

[19] Michiel P de Looze, Patricia Dolan, Idsart Kingma, and Chris TM Baten. Does an asymmetric straddle-legged lifting movement reduce the low-back load? *Human Movement Science*, 17(2):243–259, 1998.

[20] Alain Delisle, Micheline Gagnon, and Pierre Desjardins. Kinematic analysis of footstep strategies in asymmetrical lifting and lowering tasks. *International Journal of Industrial Ergonomics*, 23(5-6):451–460, 1999.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Emily Dooley, James Carr, Eric Carson, and Shawn Russell. The effects of knee support on the sagittal lower-body joint kinematics and kinetics of deep squats. *Journal of biomechanics*, 82:164–170, 2019.

[23] Colin G Drury, Chau-Hing Law, and Christopher S Pawenski. A survey of industrial box handling. *Human Factors*, 24(5):553–565, 1982.

[24] Gert S Faber, Idsart Kingma, Anja JM Bakker, and Jaap H Van Dieën. Low-back loading in lifting two loads beside the body compared to lifting one load in front of the body. *Journal of biomechanics*, 42(1):35–41, 2009.

[25] Yi-Cho Fang and Ren-Jye Dzeng. Accelerometer-based fall-portent detection algorithm for construction tiling operation. *Automation in Construction*, 84:214–230, 2017.

[26] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[27] Arun Garg and Gary D Herrin. Stoop or squat: a biomechanical and metabolic evaluation. *AIIE transactions*, 11(4):293–302, 1979.

[28] F Sebastian Grassia. Practical parameterization of rotations using the exponential map. *Journal of graphics tools*, 3(3):29–48, 1998.

[29] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018.

[30] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 432–450, 2018.

[31] Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE international conference on computer vision*, pages 1105–1112, 2013.

[32] Sue Hignett and Lynn McAtamney. Rapid Entire Body Assessment (REBA). *Applied ergonomics*, 31(2):201–205, 2000.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[34] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37 (6):1–15, 2018.

[35] Du Q Huynh. Metrics for 3D rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009.

[36] Seonhong Hwang, Youngeun Kim, and Youngho Kim. Lower extremity joint kinetics and lumbar curvature during squat and stoop lifting. *BMC musculoskeletal disorders*, 10(1):15, 2009.

[37] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2014.

[38] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[39] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system

for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[40] I Kant, JHV Notermans, and PJA Borm. Observations of working postures in garages using the Ovako Working Posture Analysing System (OVVAS) and consequent workload reduction recommendations. *Ergonomics*, 33(2):209–220, 1990.

[41] Waldemar Karwowski. *International Encyclopedia of Ergonomics and Human Factors, 3 Volume Set*. Crc Press, 2006.

[42] Sunwook Kim and Maury A Nussbaum. Performance evaluation of a wearable inertial motion capture system for capturing physical exposures during manual material handling tasks. *Ergonomics*, 56(2):314–326, 2013.

[43] Idsart Kingma, Gert S Faber, Anja JM Bakker, and Jaap H Van Dieen. Can low back loading during lifting be reduced by placing one leg beside the object to be lifted? *Physical therapy*, 86(8):1091–1105, 2006.

[44] Wonil Lee, Ken-Yu Lin, Edmund Seto, and Giovanni C Migliaccio. Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction. *Automation in Construction*, 83:341–353, 2017.

[45] Wonil Lee, Edmund Seto, Ken-Yu Lin, and Giovanni C Migliaccio. An evaluation of wearable sensors and their placements for analyzing construction worker's trunk posture in laboratory conditions. *Applied ergonomics*, 65:424–436, 2017.

[46] Huajun Liu, Xiaolin Wei, Jinxiang Chai, Inwoo Ha, and Taehyun Rhee. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*, pages 133–140, 2011.

[47] Ann-Sofie Ljungberg, Åsa Kilbom, and Gōran M Hāgg. Occupational lifting by nursing aides and warehouse workers. *Ergonomics*, 32(1):59–78, 1989.

[48] Matthew Loper, Naureen Mahmood, and Michael J Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014.

[49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[50] Monique Lortie and Geneviève Baril-Gingras. Box handling in the loading and unloading of vans. *International journal of occupational safety and ergonomics*, 4(1):3–18, 1998.

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[52] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[53] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. *arXiv preprint arXiv:1904.03278*, 2019.

[54] Charles Malleson, Andrew Gilbert, Matthew Trumble, John Collomosse, Adrian Hilton, and Marco Volino. Real-time full-body motion capture from video and IMUs. In *2017 International Conference on 3D Vision (3DV)*, pages 449–457. IEEE, 2017.

[55] Zahra Sedighi Maman, Mohammad Ali Alamdar Yazdi, Lora A Cavuoto, and Fadel M Megahed. A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. *Applied ergonomics*, 65:515–529, 2017.

[56] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 329–336. IEEE, 2015.

[57] William S Marras, Steven A Lavender, Sue E Leurgans, Fadi A Fathallah, Sue A Ferguson, W Gary Allread, and Sudhakar L Rajulu. Biomechanical risk factors for occupationally related low back disorders. *Ergonomics*, 38(2):377–410, 1995.

[58] William S Marras, Sue A Ferguson, Deborah Burr, Kermit G Davis, and Purnendu Gupta. Spine loading in patients with low back pain during asymmetric lifting exertions. *The Spine Journal*, 4(1):64–75, 2004.

[59] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[60] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation Mocap Database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[61] MN Nyan, Francis EH Tay, and E Murugasu. A wearable system for pre-impact fall detection. *Journal of biomechanics*, 41(16):3475–3481, 2008.

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[63] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018.

[64] Stefano Pellegrini and Luca Iocchi. Human posture tracking and classification through stereo vision and 3D model matching. *EURASIP Journal on Image and Video Processing*, 2008:1–12, 2007.

[65] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3D full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 663–670. IEEE, 2010.

[66] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling. In *2011 International Conference on Computer Vision*, pages 1243–1250. IEEE, 2011.

[67] Laura Punnett and David H Wegman. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. *Journal of electromyography and kinesiology*, 14(1):13–23, 2004.

[68] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training, 2018. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, last accessed on 04/25/20.

[69] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1 (8):9, 2019.

[70] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[71] Tânia Ribeiro, Florentino Serranheira, and Helena Loureiro. Work related musculoskeletal disorders in primary health care nurses. *Applied Nursing Research*, 33:72–77, 2017.

[72] Daniel Roetenberg, Henk Luinge, and Peter Veltink. Inertial and magnetic sensing of human movement near ferromagnetic materials. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 268–269. IEEE, 2003.

[73] Daniel Roetenberg, Henk J Luinge, Chris TM Baten, and Peter H Veltink. Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation. *IEEE Transactions on neural systems and rehabilitation engineering*, 13(3):395–405, 2005.

[74] Daniel Roetenberg, Henk Luinge, and Per Slycke. XSens MVN: Full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep*, 1, 2009.

[75] Alexander M Rush. The annotated transformer. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, 2018.

[76] Mark C Schall Jr, Nathan B Fethke, and Howard Chen. Working postures and physical activity among registered nurses. *Applied ergonomics*, 54:243–250, 2016.

[77] Mark C Schall Jr, Nathan B Fethke, Howard Chen, Sakiko Oyama, and David I Douphrate. Accuracy and repeatability of an inertial measurement unit system for field-based occupational studies. *Ergonomics*, 59(4):591–602, 2016.

[78] Martin Schepers, Matteo Giuberti, and Giovanni Bellusci. XSens MVN: Consistent tracking of human motion using inertial sensing. *Xsens Technologies*, pages 1–8, 2018.

[79] Loren Arthur Schwarz, Diana Mateus, and Nassir Navab. Discriminative human full-body pose estimation from wearable inertial sensor data. In *3D physiological human workshop*, pages 159–172. Springer, 2009.

[80] Sudhanshu K Semwal, Ron Hightower, and Sharon Stansfield. Mapping algorithms for real-time control of an avatar using eight sensors. *Presence*, 7(1):1–21, 1998.

[81] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[82] Ronit Slyper and Jessica K Hodgins. Action capture with accelerometers. In *Symposium on Computer Animation*, pages 193–199, 2008.

[83] Leon Straker. Evidence to support using squat, semi-squat and stoop techniques to lift low-lying objects. *International Journal of Industrial Ergonomics*, 31(3):149–160, 2003.

[84] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[85] Esa-Pekka Takala, Irmeli Pehkonen, Mikael Forsman, Gert-Åke Hansson, Svend Erik Mathiassen, W Patrick Neumann, Gisela Sjøgaard, Kaj Bo Veiersted, Rolf H Westgaard, and Jørgen Winkel. Systematic evaluation of observational methods assessing biomechanical exposures at work. *Scandinavian Journal of Work, Environment & Health*, pages 3–24, 2010.

[86] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruc-

tion using sparse accelerometer data. *ACM Transactions on Graphics (ToG)*, 30(3): 1–12, 2011.

[87] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352, 2007.

[88] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2–2, 2002.

[89] Stewart G Trost, Russell R Pate, Patty S Freedson, James F Sallis, and Wendell C Taylor. Using objective physical activity measures with youth: how many days of monitoring are needed? *Medicine & Science in Sports & Exercise*, 32(2):426, 2000.

[90] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total Capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC*, volume 2, page 3, 2017.

[91] Waleed Umer, Heng Li, Grace Pui Yuk Szeto, and Arnold Yu Lok Wong. Identification of biomechanical risk factors for the development of lower-back disorders during manual rebar tying. *Journal of Construction Engineering and Management*, 143(1):04016080, 2017.

[92] Enrique Valero, Aparajithan Sivanathan, Frédéric Bosché, and Mohamed Abdel-Wahab. Musculoskeletal disorders in construction: A review and a novel system for activity tracking with body area network. *Applied ergonomics*, 54:120–130, 2016.

[93] Enrique Valero, Aparajithan Sivanathan, Frédéric Bosché, and Mohamed Abdel-Wahab. Analysis of construction trade worker body motions using a wearable and wireless motion sensor network. *Automation in Construction*, 83:48–55, 2017.

[94] Jaap H van Dieën and Idsart Kingma. Total trunk muscle force and spinal compression are lower in asymmetric moments as compared to pure extension moments. *Journal of biomechanics*, 32(7):681–687, 1999.

[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[96] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)*, 26(3):35, 2007.

[97] Timo von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs. *Computer Graphics Forum*, 36(2):349–360, 2017.

[98] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[99] James C Walsh, John F Quinlan, Robert Stapleton, David P FitzPatrick, and Damian McCormack. Three-dimensional motion analysis of the lumbar spine during "free squat" weight lift training. *The American Journal of Sports Medicine*, 35(6):927–932, 2007.

[100] Gregory J Welk, James McClain, and Barbara E Ainsworth. Protocols for evaluating equivalency of accelerometry-based activity monitors. *Medicine & Science in Sports & Exercise*, 44(1S):S39–S49, 2012.

[101] Sam Wiseman and Alexander M Rush. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*, 2016.

[102] Frank J Wouda, Matteo Giuberti, Giovanni Bellusci, and Peter H Veltink. Estimation of full-body poses using only five inertial sensors: an eager or lazy learning approach? *Sensors*, 16(12):2138, 2016.

[103] Xuzhong Yan, Heng Li, Angus R Li, and Hong Zhang. Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. *Automation in Construction*, 74:2–11, 2017.

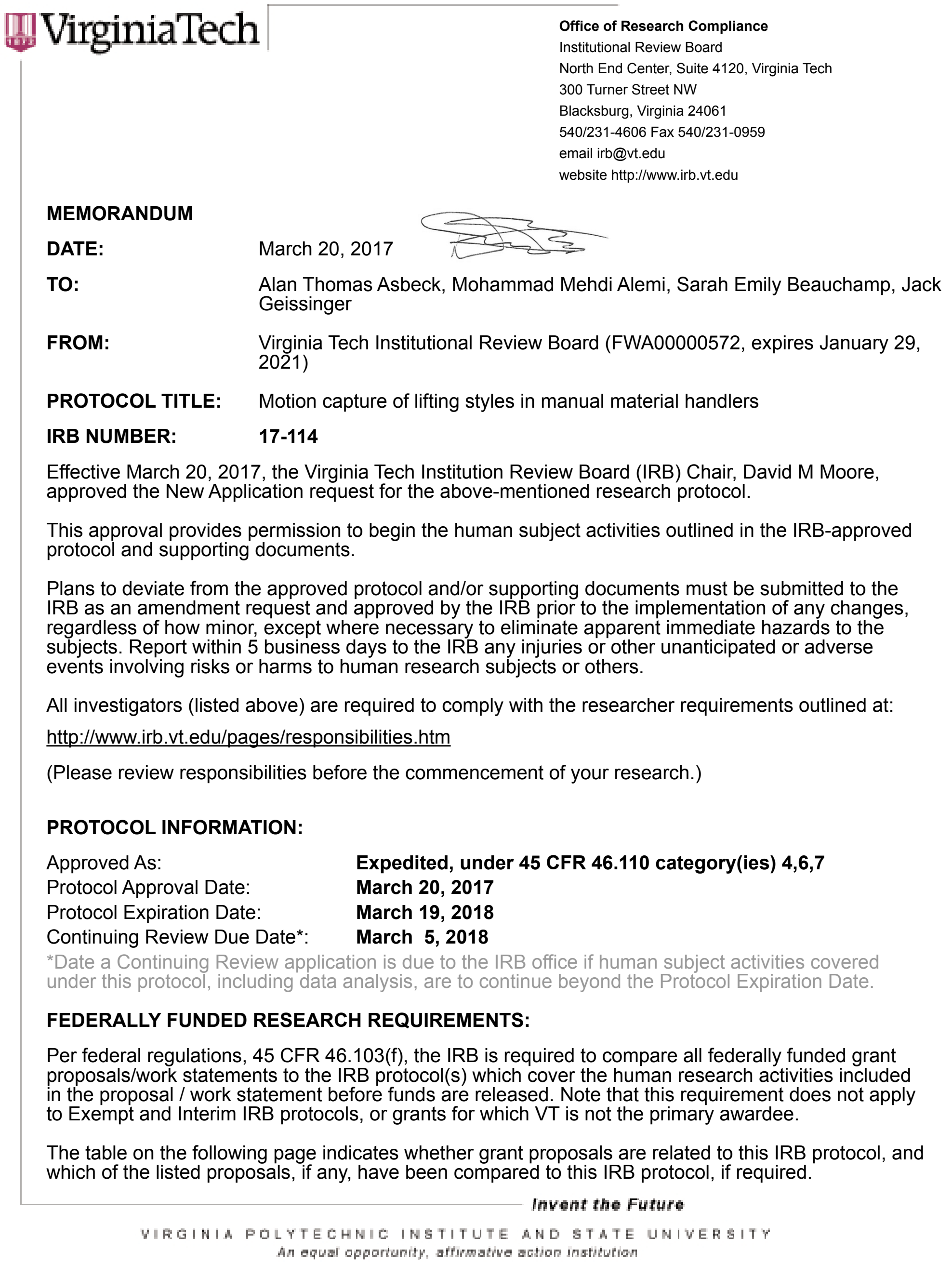
[104] Xuzhong Yan, Heng Li, Chen Wang, JoonOh Seo, Hong Zhang, and Hongwei Wang. Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. *Advanced Engineering Informatics*, 34:152–163, 2017.

[105] KangKang Yin and Dinesh K Pai. Footsee: an interactive animation system. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 329–338. Eurographics Association, 2003.

[106] Junqi Zhao and Esther Obonyo. Towards a data-driven approach to injury prevention in construction. In *Workshop of the European Group for Intelligent Computing in Engineering*, pages 385–411. Springer, 2018.

Appendices

Appendix A



Virginia Tech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: March 20, 2017

TO: Alan Thomas Asbeck, Mohammad Mehdi Alemi, Sarah Emily Beauchamp, Jack Geissinger

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)

PROTOCOL TITLE: Motion capture of lifting styles in manual material handlers

IRB NUMBER: **17-114**

Effective March 20, 2017, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the New Application request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 4,6,7**
Protocol Approval Date:	**March 20, 2017**
Protocol Expiration Date:	**March 19, 2018**
Continuing Review Due Date*:	**March 5, 2018**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

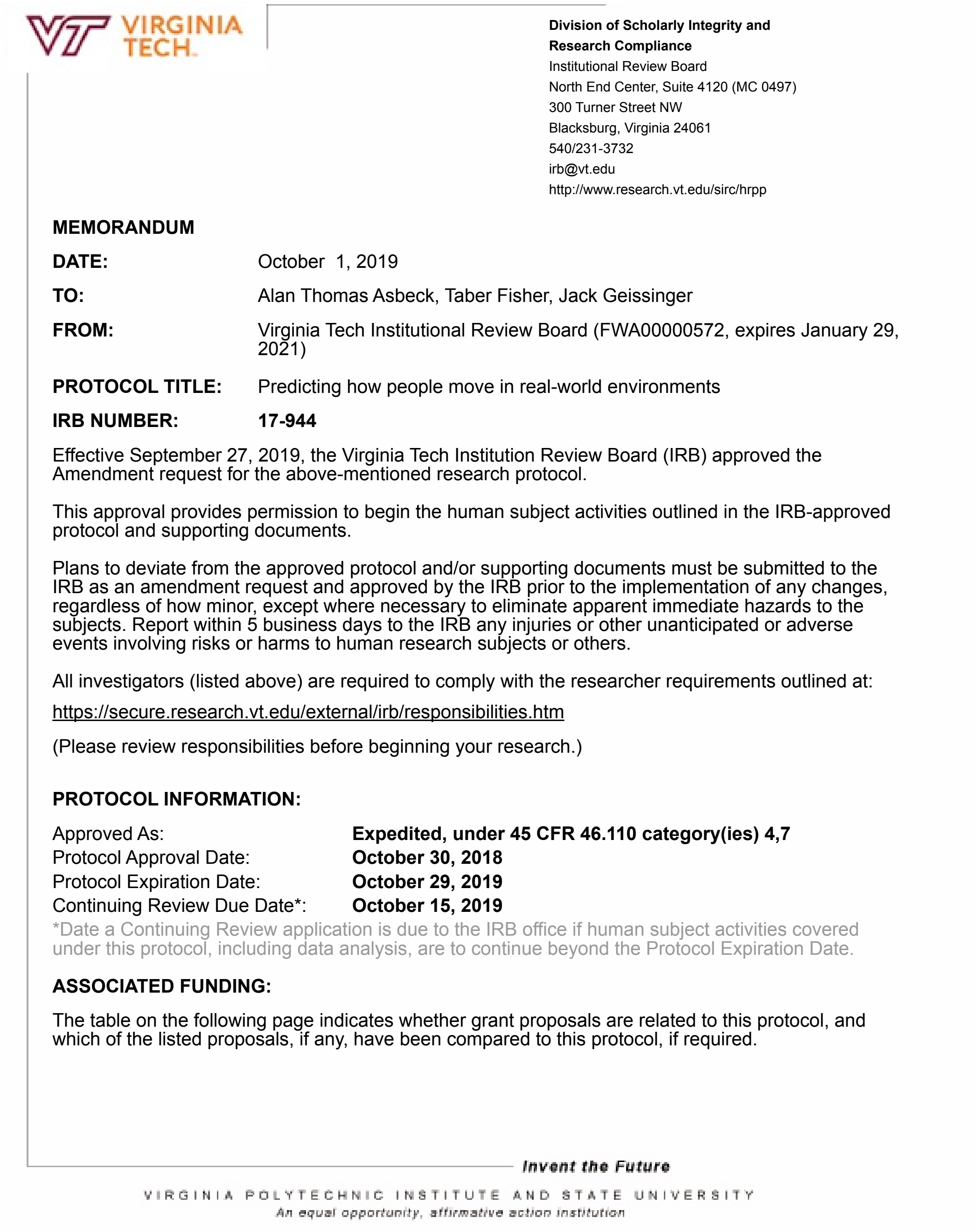
Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution

Figure A.1: The initial approval letter for the data collection study conducted at the home improvement store.



Division of Scholarly Integrity and
Research Compliance
Institutional Review Board
North End Center, Suite 4120 (MC 0497)
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732
irb@vt.edu
http://www.research.vt.edu/sirc/hrpp

MEMORANDUM

DATE:	October 1, 2019
TO:	Alan Thomas Asbeck, Taber Fisher, Jack Geissinger
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)
PROTOCOL TITLE:	Predicting how people move in real-world environments
IRB NUMBER:	**17-944**

Effective September 27, 2019, the Virginia Tech Institution Review Board (IRB) approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

https://secure.research.vt.edu/external/irb/responsibilities.htm

(Please review responsibilities before beginning your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 4,7**
Protocol Approval Date:	**October 30, 2018**
Protocol Expiration Date:	**October 29, 2019**
Continuing Review Due Date*:	**October 15, 2019**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

ASSOCIATED FUNDING:

The table on the following page indicates whether grant proposals are related to this protocol, and which of the listed proposals, if any, have been compared to this protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution

Figure A.2: The approval letter for the data collection study conducted at the college campus.