KINEMATIC ASSESSMENT FOR STROKE PATIENTS IN A STROKE GAME AND A DAILY ACTIVITY RECOGNITION AND ASSESSMENT SYSTEM

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

MENGXUAN MA

Dr. Marjorie Skubic, Dissertation Supervisor

DECEMBER 2022

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

KINEMATIC ASSESSMENT FOR STROKE PATIENTS IN A STROKE GAME AND A DAILY ACTIVITY RECOGNITION AND ASSESSMENT SYSTEM

presented by **Mengxuan Ma**,

a candidate for the degree of **Doctor of Philosophy**,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Marjorie Skubic

Professor Rachel Proffitt

Professor James Keller

Professor Ye Duan

ACKNOWLEDGEMENTS

I would like to express my appreciation to the people who helped me with this project. I would like to express my sincere gratitude to my supervisors Dr. Skubic and Dr. Proffitt for giving me such a fantastic topic. I am extremely grateful for their assistance and advice throughout my project and proposal. I would also like to thank my committee, Dr. Skubic, Dr. Proffitt, Dr. Keller and Dr. Duan. Thank you for their time and valuable feedback relating to my research work. I would like to thank my entire lab colleague. They provided lots of useful suggestions for my projects. My special thanks to Noah Marchal and students from the REU program for helping me collect and manage the stroke participants' data.

Finally, I would like to convey my heartfelt gratitude to my family. Without their support, it is not possible for me to study here and chase my dream. Also, I would like to thank Hancheng Wu and my son Ricky, who inspire me, support me and make my life colorful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

KINEMATIC ASSESSMENT FOR STROKE PATIENTS IN A STROKE GAME AND A DAILY ACTIVITY RECOGNITION AND ASSESSMENT SYSTEM

Mengxuan Ma

Dr. Marjorie Skubic, Dissertation Supervisor

ABSTRACT

Stroke is the leading cause of serious, long-term disabilities among which deficits in motor abilities in arms or legs are most common. Those who suffer a stroke can recover through effective rehabilitation which is delicately personalized. To achieve the best personalization, it is essential for clinicians to monitor patients' health status and recovery progress accurately and consistently. Traditionally, rehabilitation involves patients performing exercises in clinics where clinicians oversee the procedure and evaluate patients' recovery progress. Following the in-clinic visits, additional home practices are tailored and assigned to patients.

The in-clinic visits are important to evaluate recovery progress. The information collected can then help clinicians customize home practices for stroke patients. However, as the number of in-clinic sessions is limited by insurance policies, the recovery information collected in-clinic is often insufficient. Meanwhile, the home practice programs report low adherence rates based on historic data. Given that clinicians rely on patients to self-report adherence, the actual adherence rate could be even lower. Despite the limited feedback clinicians could receive, the measurement method is subjective as well. In practice, classic clinical scales are mostly used for assessing the qualities of movements and the recovery status of patients. However, these clinical scales are evaluated subjectively with only moderate inter-rater and intra-rater reliabilities.

Taken together, clinicians lack a method to get sufficient and accurate feedback from patients, which limits the extent to which clinicians can personalize treatment plans. This work aims to solve this problem. To help clinicians obtain abundant health information regarding patients' recovery in an objective approach, I've developed a novel kinematic assessment toolchain that consists of two parts.

The first part is a tool to evaluate stroke patients' motions collected in a rehabilitation game setting. This kinematic assessment tool utilizes body-tracking in a rehabilitation game. Specifically, a set of upper body assessment measures were proposed and calculated for assessing the movements using skeletal joint data. Statistical analysis was applied to evaluate the quality of upper body motions using the assessment outcomes.

Second, to classify and quantify home activities for stroke patients objectively and accurately, I've developed DARAS, a daily activity recognition and assessment system that evaluates daily motions in a home setting. DARAS consists of three main components: daily action logger, action recognition part, and assessment part. The logger is implemented with a Foresite system to record daily activities using depth and skeletal joint data. Daily activity data in a realistic environment were collected from sixteen post-stroke participants. The collection period for each participant lasts three months. An ensemble network for activity recognition and temporal localization was developed to detect and segment the clinically relevant actions from the recorded data. The ensemble network fuses the prediction outputs from customized 3D Convolutional-De-Convolutional, customized Region Convolutional 3D network and a proposed Region Hierarchical Co-occurrence network which learns rich spatial-temporal features from either depth data or joint data. The per-frame precision and the per-action precision were 0.819 and 0.838, respectively, on the validation set. For the recognized actions, the kinematic assessments were performed using the skeletal joint data, as well as the longitudinal assessments. The results showed that, compared with non-stroke participants, stroke participants had slower hand movements, were less active, and tended to perform fewer hand manipulation actions.

The assessment outcomes from the proposed toolchain help clinicians to provide more personalized rehabilitation plans that benefit patients.

CHAPTER 1 INTRODUCTION

1.1 Motivation

In the United States, stroke is a leading cause of serious long-term disability [1]. Each year, more than 795,000 individuals experience a stroke [1]. Stroke-related costs, both direct (cost of care, medicines) and indirect (lose of wages) in the United States summed up to nearly $53 billion between 2017 and 2018 [1].Thanks to advancement in acute neurological care, nearly 85% of them live [2]. However, stroke reduces mobility in more than half of stroke survivors with the age of 65 and older, [1] which seriously impacts daily life and bring difficulty to areas such as bathing, dressing, feeding, leisure, and paid/unpaid work [2]. Clinicians can address limitations in function and participation through evidence-based rehabilitation interventions. A set of classic clinical scales [3] are frequently used to assess upper extremity function and independence in daily living tasks. These scales are standard for rehabilitation clinicians; however, these quantitative measures still have a high subjective component, and scores can differ depending on the clinician who rates and scores the assessment [3]. In addition, high demands are placed on the clinician's time, and resources (insurance payments) are limited. Despite half of stroke survivors reporting some hemiparesis after 6 months, only 31% reported receiving outpatient rehabilitation [4]. Those that return to work and the community after a stroke strive to live daily life to the fullest. Once outpatient treatment programs are completed, few options exist to provide continuation of care for patients that still require much needed monitoring of progress.

To overcome resource limitations such as no insurance coverage and high-demanded clinician time, stoke patients are often prescribed with home programs [5] in the form of a list of exercises with pictures and instructions. As these activities and exercises are the key to successful rehabilitation, patients are required to complete them accurately and report progress back to the clinician. However, previous studies show that the adherence to self-guided exercise programs in the home setting is not only notoriously low [6-8], but also very difficult to quantify due to the reliance on patients' subjective feedback and the reliance on their accurate record maintaining of exercise sessions in an exercise diary. Factors such as fatigue, poor health, lack of motivation, and musculoskeletal issues, reported by previous research works [6-8], could prevent people with stroke from initiating and/or maintaining a structured exercise program. Studies also show that people may altogether avoid certain activities after a stroke due to the fear of failure or insecurity with having a disability [6]. As a result of decreased exercise participation and increased dependence in daily activities, burden of care could manifest on family members, caregivers and healthcare systems. Thus, indirect costs (e.g., caregiver time, lost wages) of stroke could account for up to one third of the total cost of stroke [2]. Fortunately, for those who participate in regular activity and exercise, the risk of a second stroke decreases significantly by 30-50% [9].

Virtual reality (VR) and video games have grown in popularity in healthcare [10-12] in the past decade and can address the limitations with standard home-based activity and exercise programs. VR and video games are fun, motivating, and can be tailored to meet the needs of a wide variety of individuals [11]. Specifically, physical rehabilitation has embraced novel VR applications in clinics, hospitals, nursing homes, and the community [11, 13, 14]. Robotic systems have long included game-based and VR-based user interfaces and most robotic devices provide some form of physical assistance to the patient and/or haptic feedback [15, 16]. With the release of the Nintendo Wii in 2008, many VR applications for healthcare moved away from bulky, expensive robotics and embraced the portable nature of movement and gesture recognition devices and systems. Microsoft released the Kinect sensor in 2010 to accompany its Xbox console system. Since then, there has been an exponential increase in the number of studies that report the use of the Kinect as the input device for a VR-based rehabilitation game or feedback application [17-22]. In recently years, 3Divi Inc. released the VicoVR sensor [23] and TVico system [24], and Foresite healthcare developed the Foresite depth camera-based system [25]. These devices provide the full body positional tracking with SDKs. In addition, kinematic assessments of body movements can be achieved with these systems using the recorded skeletal joint data. There are several advantages of performing kinematic assessments. First, the kinematics study can provide accurate objective information of the upper extremity motion. Second, the analysis of kinematics has also been considered more sensitive than using clinical scales [26].

Mystic Isle is a virtual reality Kinect-based video game which targets balance training and upper limb reaching exercises for people with orthopedic and neurological injury or impairments, including stroke [27]. To assess the quality of the movement of a player, I propose a set of assessment measures for the Mystic Isle. The assessment measures consist of not only common methods such as hand speed, extent of reach, and smoothness, but also new methods such as hand efficiency measures and density measures. The statistical analysis is also performed on the game data recorded for each assessment metric.

Motion data can only be captured when the patient is playing the game, but the movement information from game alone is still not sufficient, as clinicians have no information on how much and how well patients perform in their daily activities. A system able to observe and quantify the movement quality of these daily activities at home can provide valuable feedback to improve the personalization and the refinement of the rehabilitation plan. However, currently, there is no such system. In this work, I propose a system for passively tracking and assessing daily upper body movements at home in a kitchen setting. The system comprises three modules: (1) daily activity data logging and (2) activity recognition and (3) assessment. The daily activity data logging module is a depth sensor system that logs depth frames and skeletal joint data of a patient's daily activities; the depth sensor preserves privacy for in-home data logging. The activity recognition module utilizes a customized convolution-deconvolution neural network that learns the spatial features, preserves the temporal information and recognizes the actions from untrimmed videos. In the assessment module, motion evaluation metrics, such as hand speed, smoothness, and range of motion are applied to assess the quality of motion using skeletal joint data.

1.2 Advantages of the proposed system

The kinematic assessment toolchain proposed including an assessment tool for a VR-based rehabilitation game and a daily activity and recognition system (DARAS) has the potential contributions in both stroke rehabilitation area and action recognition area.

In stroke rehabilitation area, clinicians design a rehabilitation treatment based on the scale-based assessment outcomes performed during the in-clinic visits and patients self-reports of preserved home exercises. The number of in-clinic assessments is limited for a patient by insurance policies. Also, the scale-based assessment approach and self-reports are sometimes subjective. Thus, the treatment outcome can be low due to the insufficient and potential subjective assessment feedback.

(1) The assessment tools in both rehabilitation game and the DARAS provide kinematic assessment using the skeletal joint data. Thus, the tools objectively quantify and qualify upper extremity movement for post-stroke players.

(2) The game and the DARAS collects skeletal joint samples at least 10 frames per second. The proposed kinematic metrics in the tools were calculated for each trial in the game and each recognized clinically relevant daily action using the collected joint data. With this large amount of motion data collected from a stroke individual, the assessment outcomes are potentially sensitive to performance impairments that are unobservable. Also, it is possible to statistically provide a comprehensive assessment of a stroke individual.

(3) Clinicians are able to provide more personalized treatment for each stroke individuals by analyzing the assessment outcomes from the game and the daily motions of a patient in a natural realistic setting.

In the action recognition area, to the best of our knowledge, this is the first work to perform action recognition and temporal action localization on real daily motions in a realistic home environment of stroke population.

(4) The proposed DARAS has effectively collected daily activity data in actual home environments, running 24 hours/day for several months while preserving privacy through only depth sensing.

(5) A stroke population daily action dataset was generated by collecting daily activity data from sixteen stroke participants in realistic home environments. The data

collection time period of each participant was three months. This the first stroke-population dataset for action recognition and temporal localization.

(6) The proposed ensemble network is robust to accommodate different room layouts and light conditions. The algorithm accurately detects and segments the clinically relevant actions of stroke individuals.

1.3 Content organization

The proposal contents will be arranged as follow:

Chapter Two: Literature survey of the stroke rehabilitation strategies and status including the classical rehabilitation assessment and kinematic assessment with rehabilitation technologies. Literature review of daily activity monitoring and assessment.

Chapter Three: A detailed description of the kinematic assessment performed in Mystic Isle stroke rehabilitation game.

Chapter Four: A detailed description of the proposed daily activity recognition and assessment system. The implementation of the action logging module is first presented. Then, datasets collected from the system are described. Next, action recognition algorithms developed for the system is presented. Finally, the assessment of the daily activities is presented.

Chapter Five: In the discussion section, the main purpose of the study has been reviewed. The most important findings were explained. Limitations of the study has been discussed. Finally, the recommendations for future research were included.

Chapter Six: The project progress was presented.

Chapter Seven: The published papers of this work have been listed in this section.

Chapter Eight: In the appendix section, the detailed information of IRB protocol, recruited participants, and data collection was included. Then, the description of the collected data and the process of data processing and data labelling were included. The developed software, tools and source code of algorithms were summarized. Finally, the per-participant prediction accuracies were presented.

Chapter Nine: References.

CHAPTER 2 LITERATURE REVIEW

2.1 Stroke

The World Health Organization defines stroke as a clinical syndrome consisting of rapidly developing clinical signs of focal (or global in case of coma) disturbance of cerebral function lasting more than 24 hours or leading to death with no apparent cause other than a vascular origin [28]. A stroke occurs when the blood supply to part of your brain is interrupted or reduced, depriving brain tissue of oxygen and nutrients. A stroke may be caused by a blocked artery (ischemic stroke) or the leaking or bursting of a blood vessel (hemorrhagic stroke). Some people may experience only a temporary disruption of blood flow to the brain (transient ischemic attack, or TIA) that doesn't cause permanent damage.

Each year, approximately 795,000 people experience a new or recurrent stroke [1, 2]. 87% of strokes are classified as ischemic. Most patients survive from a first stroke, but they often have significant morbidity [1, 29]. Stroke is a leading cause of serious long-term disability in the United States [1]. Seventy to eighty five percent of first strokes are accompanied by hemiplegia [30]. Only 60% of people with hemiparesis who need inpatient rehabilitation achieve functional independence in activities of daily living (ADLs) after 6 months of a stroke [31].

2.2 Stroke rehabilitation

Rehabilitation is a complex set of processes usually involving several professional disciplines and aimed at improving quality of life for people facing daily living difficulties caused by chronic disease [32]. Stroke survivors can relearn lost skills through stroke rehabilitation after a stroke affected part of brain. Stroke rehabilitation helps stroke patients

Table 1. Common scale-based outcome measures

Body structure (impairments)	**Activities (limitations to activity-disability)**	**Participation (barriers to participation-handicap)**
Canadian Neurological Scale Clock Drawing Test Frenchay Aphasia Screening Test Fugl-Meyer Assessment General Health Questionnaire Geriatric Depression Scale	Action Research Arm Test Functional Independence Measure Motor Assessment Scale	Stroke Adapted Sickness Impact Profile Stroke Impact Scale Stroke Specific Quality of Life

regain independence and improve their quality of life. Researchers have found that people who participant in a focused stroke rehabilitation treatment perform better than most people who don't receive stroke rehabilitation [33].

2.3 Scale-based assessment

In attempting to discuss some of the commonly used measures available for use within the field of stroke rehabilitation, it is useful to have guidelines available for classifying these tools. The WHO International Classification of Functioning, Disability and Health (ICF: WHO, 2001, 2002) provides a multi-dimensional framework for health and disability suited to the classification of outcome instruments [34]. Outcomes may be measured at any levels -- Body functions/structure (impairment); Activities (refers to the whole person – formerly conceived as disability in the old ICIDH framework) and Participation (formerly referred to as handicap). Activity and participation are affected by environmental and personal factors (referred to as contextual factors within the ICF) [34]. Table 1 lists the common outcome measures under each level. These measures are validated and standardized; however, many are self-report and observational measures (scored by a clinician) can have lower inter-rater and intra-rater reliability [34].

2.4 Rehabilitation technology and objective kinematic assessment

Virtual reality (VR) and video games have grown in popularity in healthcare [10-12, 35] in the past decade. Specifically, physical rehabilitation has embraced novel VR applications in clinics, hospitals, nursing homes, and the community [11, 13, 14]. Robotic systems have long included game-based and VR -based user interfaces and most robotic devices provide some form of physical assistance to the patient and/or haptic feedback [15, 16]. With the release of the Nintendo Wii in 2008, many VR applications for healthcare moved away from bulky, expensive robotics and embraced the portable nature of movement and gesture recognition devices and systems. Microsoft released the Kinect sensor in 2010 to accompany its Xbox console system. Since then, there has been an exponential increase in the number of studies that report the use of the Kinect as the input device for a VR-based rehabilitation game or feedback application [17, 18, 36]. 3Divi Inc. released the VicoVR sensor [37] and TVico system [24]. The devices provide the full body positional tracking from Nuitrack SDK to Android and iOS smart devices. Kinematic assessments can be provided with the use of the technologies in rehabilitation. There are several advantages of performing kinematic assessments. First, they can provide accurate objective information of the upper extremity motion. Second, kinematic measures are more sensitive than traditional clinical measures [26].

The idea of using kinematic measures to assess the upper extremity movement was introduced by Burdet et al. who quantified the reaching movements [38]. Since the kinematic assessment can provide accurate and objective information and it is sensitive to the outcome of interest, many kinematic studies have been performed in laboratory settings with the aim of quantifying the upper extremity movements during the last twenty years [3]. The kinematic metrics which have been proposed [3] to evaluate the movement are listed in Figure 1.

Proffitt et.al have provided feasible [39] and successful [40] virtual reality-based game Mystic Isle to maximize adherence to home exercise and activity programs. Other developed rehabilitation technologies and devices, such as the Neofect Rapael® smartglove, report similar outcomes. Rehabilitation clinicians report that the data collected by these systems are clinically meaningful and relevant [41]. However, all of these technologies only assess patient performance when the patient is actively interacting.

2.5 Activity monitoring

Though VR-based rehabilitation technologies can provide objective assessments to individuals, unfortunately, we are only able to capture data on amount and quality of activity when the individual is in front of the motion capture device (e.g., Microsoft Kinect sensor). Stroke patients spend most of their time doing daily activities in their natural

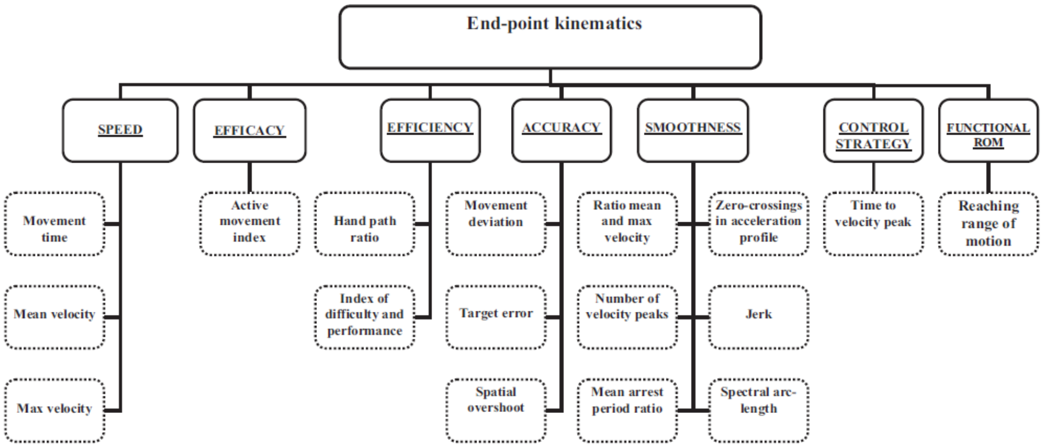


Figure 1 A set of kinematic metrics classified by the movement characteristics.

environment. A system able to monitor and quantify the movement quality of these activities at home could provide valuable feedback to improve the personalization and the refinement of the rehabilitation plan. There are many methods and sensors available to collect movement data and monitor activities. Each has its own unique strengths and drawbacks.

2.5.1 Radio frequency identification

Radio Frequency Identification (RFID) is a technology that transmits and receives unique serial information using radio frequency waves [42]. The key elements of RFID systems consist of RFID readers, tags. RFID readers are silicon-based radio transceivers, which interrogate and communicate with RFID tags by electromagnetic waves. RFID tags have a tiny on-board memory up to several kilobytes, storing their unique identification as well as some additional information [42]. Initially, RFID tags were placed on objects and the reader was attached to the arm of interest to recognize the activities related to arms [43, 44]. RFID was also considered as a solution to make aware of the location and movement direction of a person. Wang et al. [45] designed a system in which passive tags were embedded into the clothes and a small RFID reader is also worn on the user's body to extend the detection coverage as the user moves. The sitting, standing and walking motions can be recognized by analyzing the radio signal strength information. However, such approaches have two disadvantages. Firstly, it is uncomfortable to attach wires and sensors on the body [45, 46]. Secondly, the performance may be affected when multiple persons wearing the sensors are close to each other due to the limitation of the number of tags which can be read by a RFID reader [45]. Yao et al. presented an unobtrusive system that interpreted what a person is doing by deciphering signal fluctuations using radio-frequency identification (RFID) technology [46]. However, Passive RFID tags must be deployed in an environment (e.g., on the wall in a room) forming a tag array.

2.5.2 Smartphone

Recently, many researchers started using mobile phones for activity recognition [47-49], since these ubiquitous devices are equipped with various sensors including accelerometers, gyroscope, magnetic field sensors, and so on. Cheng et al. [48] presented a smartphone-based remote gait and mobility monitoring system for patients with Parkinson's disease. Coni et al. [49] designed an activity monitoring system using a smartphone and then investigated the association between mean and extreme values of physical activity and gait characteristics derived from daily living activities and well-established clinical tools. Mekruksavanich et al. [50] developed a CNN-based model to recognize the global-body activities via data collected from an accelerometer and gyroscope of a smart phone. The highest accuracy was 93.54%.

Though mobile phones have become more powerful in terms of available resources, there are still some limitations on using battery, CPU and memory usage for activity recognition. In addition, the recognition results are sensitive to the orientation changes of the accelerometer and gyroscope sensors. Another challenge in activity recognition using mobile phones is that motion sensors are sensitive to body position. In most studies, the position of the mobile phones is kept fixed, because any changes in position may result in a loss of recognition performance [51, 52].

2.5.3 Inertial

Inertial sensors are sensors built to measure metrics related to inertia. They come in various forms that provide different range of measurements. While MEMS inertial sensors can measure in the range of only a few square millimeters, ring laser gyroscopes can measure accurately in the range of 50 cm in diameter [53, 54].

With the advance of hardware technologies, inertial sensors have become portable and affordable, leading to their wide adoption as wearable sensors for human activity detection and classification. Examples are accelerometers and gyroscopes that are widely used in cellphones and wearable devices for either gaming or healthcare purposes.

Inertial sensors can detect various ambulatory-type activities [55]. For example, previous studies propose fall detection systems using accelerometers [56-58]. Song et al. [59] have developed a gait monitoring system, based on inertial sensors, to estimate the user gait parameters such as walking speed, stride time and stride length. Other accelerometer-based systems are proposed to distinguish global body motion activity types such as walking, running, standing, lying and stairs [47, 54, 60-63]. The arm movements including reaching, grabbing and wrist rotation of stroke individuals were recognized using a single Inertial Measurement Unit. The test Correlation Coefficient score was 0.58 on the functional dataset [64].

However, inertial sensors also present some limitations. First, although they work fine for global body motion activity types, they fail to provide competitive performance when used alone for local interaction type activities such as eating, food preparation and house cleaning. This is due to that they only capture two domains of features (time domain and frequency domain based), which is not sufficient for local interaction type activity types [55]. To enrich the features extracted from these activities. Researchers fuse the information from inertial sensors with that from other sensors. Given that camera-based sensors can provide contextual information, many researchers fuse the inertial data with camera-based data [65]. Nam et al. [66] and Hafeez et al. [67]fused accelerometer features with camera-based features to increase the accuracy of classification for ambulatory activities. Doherty et al. [68] used the context provided by cameras to identify the specific class of activity once an accelerometer has identified the level of activity being undertaken. Meng et al. [69] collected acceleration, angular velocity, surface electromyography data synchronously from 5 upper-limb-worn sensor modules to evaluate the upper-limb Brunnstrom Recovery Stage (BRS) via three typical ADLs (tooth brushing, face washing and drinking). Second, several studies have revealed that using a single accelerometer might not be sufficient for activity classification [55, 70]. Researchers address the problem by combining the data from multiple accelerometer and gyroscope sensors [71-74]. Third, another disadvantage of using inertial sensor is that they need to be worn all the time. As a result, the sensors must be lightweight and small enough to provide the highest degree of comfort and convenience [75]. Fourth, inertial sensors may drift away from their ideal waring positions during movements, which causes the loss of contact with skin. Thus, the accuracy and quality of the data may not be stable. Last, inertial sensors are usually difficult to use in case of long-term monitoring due to their low battery capacities [75, 76].

2.5.4 Videos

Activity recognition has been widely studied within the field of computer vision. With the development of computing ability and the improvement of sensor techniques, various data modalities including RGB data, depth data and skeleton data have been introduced. Depth sensors have the ability to sense the 3D visual world and capture low-level visual information, thus the depth data of a person can be extracted more easily and accurately [77]. Skeletal joints encode 3D joint positions of a person. Since the movements of the human skeleton can distinguish many actions, it is promising to exploit skeleton data for action recognition [77].

To recognize an action from a given video, features are extracted and encoded to represent the input video. The encoded features are processed by a classifier to output the class of the action [78]. During the last two decades, a large number of novel approaches for activity recognition have been proposed for both feature extraction and classification. As to extracting features, the three-dimensional corners, blobs, and/or junctions representations can be extracted from a video using spatiotemporal interest points (STIPs) method [79, 80]. A large set of gradient-based descriptors have appeared for action recognition, such as histogram of oriented gradients (HOG) [81, 82], cuboid descriptor [83] and scale-invariant feature transform (SIFT) [84]. Dense trajectory (DT) [78, 85] features were introduced as a form of descriptors that track the path of motion and improved dense trajectory (iDT) [86] feature has achieved great performance among hand crafted features. In recent years, there has been a surge of algorithms relying on Convolutional Neural Networks which can also be used as feature extractor [78]. As to classification, support vector machines (SVMs) and Hidden Markov Model (HMM) have been widely used to provide classification decisions. A fuzzy rule based system was designed to learn activities of daily living [87].

Before CNN-based algorithms took the field of action recognition and detection by storm, iFV-encoded iDT features with HOG, HOF, and MBH descriptors using a linear SVM classifier were top performing hand-crafted features achieving an accuracy of 57.2% and 85.9% on HMDB51 and UCF101 datasets, respectively [78, 88, 89]. The classification accuracy was increased to 94.6% by using deep-learned convolutional features.

Since most realistic action-related videos are untrimmed with sparse segments of interest, action recognition itself is insufficient for analyzing actions in real-life videos as it requires as inputs trimmed action segments. Recently, researchers have started investigating temporal action localization on THUMOS '14, MPII Cooking Activities and MPII Cooking 2 Activities, as well as the Activity Net datasets [78], as action recognition and localization from untrimmed videos have been demanded in many scenarios such as action monitoring in home environment.

The objectives of temporal activity localization are to localize the temporal boundary of actions and to classify the action categories simultaneously. Existing works have investigated the temporal action localization task in full and limited supervision settings based on the level of action annotations, including supervised learning and weakly supervised learning [90].

In supervised learning temporal action localization, the temporal boundaries and action category labels of action instances are needed for each untrimmed video of training set. As for inference, the goal is to predict the temporal boundaries and the action labels of action instance [90]. The prediction is based on temporal proposal generation methods which can be categorized to anchor-based approach and anchor-free approach [90]. The anchor-based approaches generate dense multi-scaled temporal proposals, and the extract proposal feature with the same length of each proposal, such as 3D RoI pooling in R-C3D algorithm [91]. TSA-Net [92] employ a multi-tower network and achieve a higher performance compared with 3D RoI pooling methods.

Weakly-supervised temporal action localization usually requires only the video-level labels of actions during training. During testing, both temporal boundaries and action categories are predicted. The most common approach of weakly-supervised temporal localization is to use attention mechanism to focus on discriminative snippets and combine salient snippet-level features into a video-level feature [90]. The attention scores are utilized to localize the action boundaries and eliminate background frames. Attention signals are predicted with class-specific attention and class-agnostic attention methods [90]. Temporal localization with class-specific attention includes UNet [93], Action graphs [94], BaSNet [95], and the temporal localization with class-agnostic attention includes STPN [96] and BG-modeling [97].

Through a rich set of approaches for action recognition and temporal action localization have proposed, most of the existing approaches were developed on RGB datasets. The RGB videos can't preserve the privacy of a patient. Thus, it is needed to build a system to recognize and monitor daily activities using depth videos. The approaches for the depth videos were trained and tested on simulated datasets with predesigned actions, which is different from the real-life action videos. In addition, no studies have investigated the action recognition and localization on stroke population. Our proposed system temporal localizes clinically relevant actions of stroke individuals using the recorded realistic in-home depth videos.

2.6 Activity assessment

Walking and gait measurement are vital metrics for health and rehabilitation assessments. By analyzing daily walking activity researchers can get a good idea for how a person's physical health changes over time. This has been accomplished using inertial sensors to classify actions (stairs, ramp, level ground) [59] and analyze real life walking movement data to predict falls [98]. It's quite common for rehabilitation research to correlate sensory measurements with existing clinical rating scales such as "Get up and go". However, the assessment on walking-related motion is only focused on lower body, the quality of upper-body movement is important for patients with stroke. Some researchers [99, 100] sought to analyze the data using metrics otherwise immeasurable by standard in clinic tests e.g., movement intensity/smoothness. One downside to these approaches is that it paints a vague picture of categories of actions that have been assessed. An alternative to quantifying the relevance of measured movement is to assess the person's daily motions by the category of actions. However, no system has been introduced for evaluating the quality of daily actions. My research work fills the gap by developing a daily activity recognition and assessment system to perform kinematic assessments on different action categories of a post-stroke individual.

CHAPTER 3 ASSESSMENTS IN MYSTIC ISLE GAME

3.1 Mystic Isle

Mystic Isle is a platform for rehabilitation that allows a user to interact with a virtual environment by using their bodies as shown in Figure 2. The Mystic Isle created in Unity 3D allows the tracked user to interact with virtual environments and objects in a 3-D world. Using Mystic Isle, specific movements, distances, and locations of objects can be tailored to the abilities and requirements of the user. The system uses the Microsoft Kinect V2 camera to track participant movements. The Kinect V2 tracks 20 discrete points/joints on the body of the user. Both gross motor (stepping, jumping, squatting) and fine motor (waving the hand, turning the palm facing up, open/close hand) movements can be tracked. The Kinect V2 tracks the user in 3-dimensional space and then inputs the data in real time to the associated software, Mystic Isle. The Kinect V2 tracks and records the x, y, and z coordinates (and confidence) of each discrete joint at either 15 or 30 frames per second. The kinematic measures can then be employed to assess the movement quality using the joint samples.

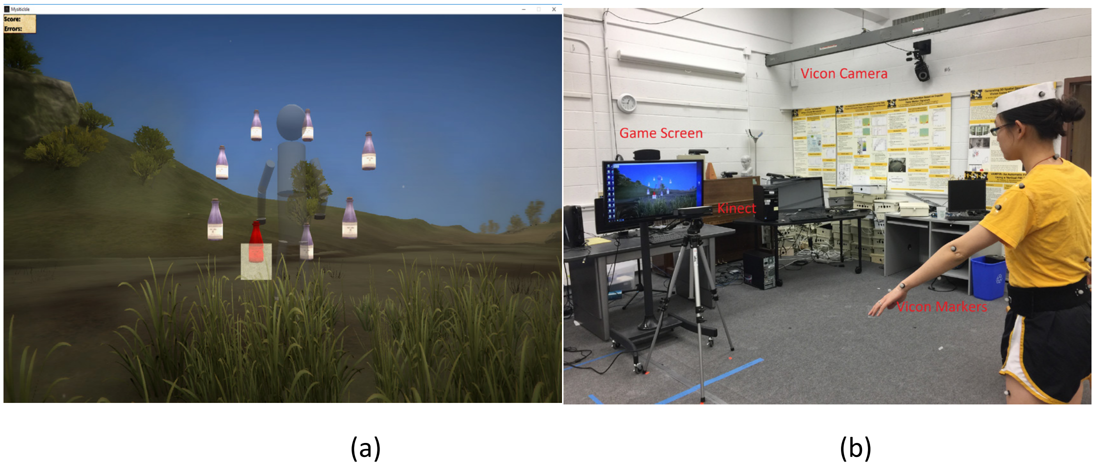


Figure 2 (a) A virtual avatar collecting targets in a Kinect-based rehabilitation game, Mystic Isle. (b) A participant playing the game with Vicon markers on the body. Joint data of game trials were recorded by a Kinect and the Vicon system for validation.

3.2 Kinematic measures

3.2.1 Extent of reach

Extent of reach was calculated for each trial. Extent of reach was defined as the distance from the hand joint to the shoulder center, where shoulder center is the middle of the left and right shoulder joints. Suppose the hand joint and the shoulder center are represented by $j_{hand} = \{h_x, h_y, h_z\}$ and $j_{shoulderC} = \{s_x, s_y, s_z\}$, then the extent of reach for each frame is calculated by

$$Extent\ of\ Reach = \sqrt{(h_x - s_x)^2 + (h_y - s_y)^2 + (h_z - s_z)^2} \tag{1}$$

3.2.2 Speed

We also calculated maximum and mean velocities for each trial. Suppose the hand joint of the i^{th} frame is represented by $j_{hand} = \{x_i, y_i, z_i\}$, the velocity of this frame is calculated by

$$Hand\ Velocity = \frac{\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2}}{t_{i+1} - t_i} \tag{2}$$

where t, time, is measured and stored automatically with each frame by the Kinect V2 SDK.

3.2.3 Smoothness

A healthy player can perform a smooth reaching movement in the game. However, because of impaired brain function after suffering stroke, patients may perform uncoordinated movements that include various acceleration and deceleration periods or brisk movements [3, 101]. To assess the smoothness of movements, we evaluated the ratio between maximum and mean velocities and normalized jerk metrics on hand joint data from the game. Both can reflect the alterations of hand acceleration.

3.2.3.1 Ratio between max and mean velocities

In a game trial, the mean velocity is average velocity over all hand joint samples. The max velocity is the maximum reported among the samples. The ratio calculated for a constant-velocity movement is 1. Thus, for healthy players, the ratio is expected to be close to 1 since their movements tend to be steady. A larger value indicates jerky or less smooth movements.

3.2.3.2 Normalized jerk

Jerk is the third derivative of position. Reference [102] presented that the time-integrated squared jerk decreases as the smoothness of the movement increases. The integrated squared jerk has dimensions of (squared length)/(5th power of duration) [103]. To make the measurement dimensionless, normalized Jerk is computed using (3):

$$\text{Normalized Jerk} = \sqrt{\frac{1}{2} \times \frac{d^5}{l^2} \int_{t_i}^{t_f} J^2(t)dt} \tag{3}$$

where d denotes the overall movement duration, and l denotes the overall movement length, and J denotes the jerk function, the third derivative of position. The normalized jerk was evaluated on two different paths: paths between two adjacent targets (target paths) and valid paths (defined in Section 3.3.1). The interquartile range approach was applied to detect outliers for the outcomes calculated on target paths. For a path, the normalized jerk was calculated using equation (3). Specifically, the 3-dimensional samples of a hand joint of a path were input. Then the length was computed, which is the

Euclidean distance from the start to the end positions. The duration, which is the time difference between the start and end positions, was computed. Then, compute the jerk of each sample position, which is the third derivative of the distances. Compute the integral of jerk over time. Finally, the normalized jerk of the path was computed based on the equation (3).

3.2.4 Efficiency

An efficient movement is defined as the movement to a target without extraneous or abnormal trajectories [3, 104]. The reaching movement of a healthy player is highly stereotypic and has a well-executed trajectory [3, 105]. The trajectory of a reaching movement for someone with stroke tends to be more curved. We used two metrics to measure the hand movement efficiency.

3.2.4.1 Hand path ratio

The hand path ratio has been frequently used to measure movement efficiency. It is the ratio of the real reaching path to the shortest straight line between two targets as in (4):

$$Path\ Ratio = \frac{the\ length\ of\ the\ real\ path}{The\ length\ of\ the\ shortest\ path} \tag{4}$$

3.2.4.2 Average sway distance (ASD)

To quantitatively describe the difference between a real/sample reaching and an ideal reaching, we propose a metric called sway distance. During a hand movement between two targets, the sway distance of a sample point is the distance from the sample point to the shortest path between the two targets. To evaluate how close the real path is to the shortest path, the average sway distance is computed as in (5):

$$ASD = \frac{\sum distance(sample\ position,ideal\ position)}{number\ of\ samples} \qquad (5)$$

3.3 Density and path analysis

Not all the samples collected can be used to calculate the smoothness and efficiency assessment metrics proposed in Section 3.2.4 since some data samples are not related to movements. In this part, we first introduce how these redundant data samples are generated in the game; we then explain how we extract the useful data samples by performing a clustering algorithm; we last present the pseudo algorithm for the efficiency assessment.

3.3.1 Define and extract valid paths

In the Mystic Isle game, players were asked to perform an upper limb motion to reach a virtual target. Most players began with their hands in a rest position by their side or in their lap. Once the hand position matches the target position in 3D space, the player gains a point and the target disappears. Players can then return to a resting position or move directly to the next target. Some players took longer than others to achieve the correct target position in 3D space. At the end of each game trial, Mystic Isle exported the 3D positions of joints and sample indexes corresponding to scoring events.

To assess the smoothness and efficiency of the hand movement paths, it is necessary to understand the structure of hand joint samples. Figure 3 shows the visualization of the trajectories of a hand with target positions in lateral-vertical plane. Samples can be divided into three groups: a) curved paths between either targets or dense clusters; b) dense clusters surrounding the targets; c) a dense cluster not surrounding a target (natural rest position). Since the game records the samples in either 15 or 30 frames per second, the more the hand appears in a location, the denser the samples are in this location.

Hand efficiency metrics evaluate the paths of the hand movement from one position to another position. Thus, the data samples generated from the resting position should be excluded. This is to say, as shown in Figure 3, the data samples surrounding the rest position or the circles around the targets (group b and c according to our analysis) should not be considered when analyzing the path efficiency. We only consider a valid path as a curved trajectory excluding clusters around targets or the hand rest cluster.

To get the valid paths, we need to locate the clusters first. Since the sample structure is related to the density, we applied the Ordering Points To Identify the Clustering Structure (OPTICS) clustering algorithm to perform the path segmentation. OPTICS is a density-based clustering algorithm proposed by Ankerst et al [106]. It not only produces a clustering result explicitly, but also gives a specific order of the samples in the data set.

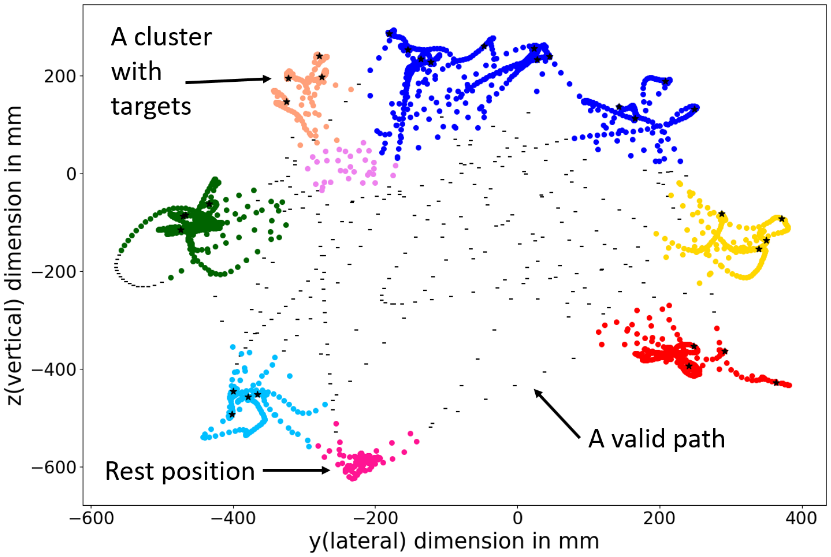


Figure 3 Trajectory in lateral-vertical plane of one hand from a game trial. Target locations are marked using black stars. There are dense clusters around most targets. Hand rest position is located at the cluster without any target.

```
1  PathAssessment(HandJointPts,ListoftargetIdx):
2   order = OPTICS(HandJointPts,ε,MinPts);
3   PtsClusterLabel
       = ExtractCluster(ClusterOrderedPts,ε');
    // Find the start points of paths.
4   FOR i FROM 1 TO ListoftargetIdx.size-1 DO
5       pathStart = ListoftargetIdx(i);
6       If pathStart.Label == InClusterLabel
7          pathStart += 1;
8           UNTIL pathStart.Label != InClusterLabel
9       pathStartList(i) = pathStart;
    // Find the end p
10   FOR i FROM 2 TO ListoftargetInx.size DO
11      pathEnd = ListoftargetIdx(i);
12      If pathEnd.Label == InClusterLabel
13         pathEnd -= 1;
          UNTIL pathEnd.Label != InClusterLabel
14      pathEndList(i) = pathEnd;
15  Paths
     = GetPath(HandJointPts,pathStartList,pathEndList);
16  HandPathRatio,AverageSway =Assessment(Paths);
17  Return: HandPathRatio,AverageSway
```

Figure 4. Algorithm of efficiency assessment of valid paths of a hand.

The order represents its density-based clustering structure and contains the information about the clustering level (such as the number of clusters and the size of each cluster) [106].

Figure 4 shows the algorithm of performing movement efficiency assessment of one hand. The algorithm requires joint samples of a movement and a list of target sample indexes as inputs. The first step is to extract the structure of the joint samples by obtaining the order of the samples using OPTICS algorithm as in line 2. MinPts and ε are parameters controlling the reachability distance of sample points. After having the order of data samples, a reasonable clustering decision can be made by adjusting the parameter ε' ≤ ε using ExtractCluster method whose output is a list of clustering labels corresponding to each sample (line 3). The next step is to find the indexes of start points and end points (line 10 - 14) of valid paths. With the indexes, we can easily segment the valid paths out. Ideally, target indexes should be the start and end indexes of paths if they are not included in any clusters. Thus, the algorithm checks the label of each target in the finding path process (line 4-14). For the start points, if the label of the target is not an in-cluster label, the index of the target is then the start index. Otherwise, the algorithm finds the first point out of the cluster from the target by increasing the index (line 4-9). The path end index is located similarly. The algorithm assigns the index of a target point as the end index if the target label is not an in-cluster label. Otherwise, the algorithm finds the first point entering that cluster by decreasing the index from the target index (line 10-14). With the indexes of the start and end points, the samples of each valid path are extracted (line 15), and the hand path ratio and average sway distance assessment are performed on each valid path (line 16-17).

The overall assessment value of a game trial is the average over the values from all the valid paths.

3.3.2 OPTICS

In this part, we detail how we utilize the OPTICS algorithm and tune its parameters to accommodate our path segmentation need. The OPTICS algorithm defines two types of distances for each point in input dataset: core-distance and reachability-distance [106].

Core distance: Suppose p is a point in a dataset D, and let ε be a distance value, then Nε(p) represents the set of neighbor points whose distance to point p is no larger than ε. Let MinPts be a natural number and let MinPts-distance(p) be the distance from p to its MinPts's closest neighbor. Then, the core-distance of p is defined as $core - distance_{\varepsilon,MinPts}(p) =$

$$\begin{cases} Undefined, if\ Count\big(\mathrm{N\varepsilon(p)}\big) < MinPts \\ \quad MinPts - distance(p),\ otherwise \end{cases} \tag{6}$$

What this equation means is that: we draw a circle with radius ε around point p and expect there are at least a number of MinPts points falling inside this circle. If this expectation of density holds true, point p is defined as a core point and the core distance is defined as the distance from p to its MinPts's closest neighbor, otherwise the core distance is undefined.

Reachability distance: Suppose q is a point in the database D, Then, the reachability distance of q with respect to p is defined as $reachability - distance_{\varepsilon,MinPts}(q,p) =$

$$\begin{cases} Undefined, if\ Count\big(\mathrm{N\varepsilon(p)}\big) < MinPts \\ \max\big(core - distance(p), distance(q,p)\big), otherwise. \end{cases} \tag{7}$$

What this equation means is that: if point p meets our expectation of density (with radius ε and number MinPts), then the reachability distance from point q to point p is the maximum between distance (q, p) and core-distance(p).

```
OPTICS (HandJointPts,ε,MinPts):
Initialize OrderList;
FOR i FROM 1 TO HandJointPts.size DO
  Object := HandJointPts(i);
  IF NOT Object.Processed THEN
    neighbors := HandJointPts.neighbors(Object, □);
    Object.Processed := TRUE;
    Object.reachability_distance := UNDEFINED;
    Object.setCoreDistance(neighbors,ε, MinPts);
    OrderList.insert(Object);
    IF Object.core_distance <> UNDEFINED THEN
      OrderSeeds.update(neighbors, Object);
      WHILE NOT OrderSeeds.empty() DO
        currentObject := OrderSeeds.next();
        neighbors : = HandJointPts.neighbors(currentObject, ε);
        currentObject.Processed := TRUE;
        currentObject.setCoreDistance(neighbors, ε, MinPts);
        OrderList.insert(currentObject);
        IF currentObject.core_distance<>UNDEFINED THEN
          OrderSeeds.update(neighbors, currentObject);
Return OrderList;
END; // OPTICS
```

Figure 5. Pseudo code of the OPTICS algorithm.

The procedure of OPTICS algorithm is shown in Figure 5. The algorithm aims to get three outputs: the cluster-ordering (the order in which the input data samples are processed), the core-distance for each sample and the smallest reachability distance for each sample. With joint data, distance value ε and threshold MinPts as inputs, OPTICS keeps an input queue and a higher-priority queue called OrderSeeds (line 12, 14 and 20). OPTICS will always try to consume and process a data point off the top of the priority queue first. If the high-priority queue is empty, OPTICS will then try to process the next unprocessed data point in the input queue (line 3 and 13). The algorithm ends when all data points have been processed (line 3 and 21). For a point being processed, its ε distance-bounded neighbors are found firstly and its core-distance are computed based on the parameter MinPts. If the point proves to be a core point, its neighbors are inserted into the higher-priority queue (if not in the queue) and their reachability-distances are updated by OrderSeeds.update method (line 12, 20). Points in the higher-priority queue are always sorted based on the reachability-distance (line 12). Finally, the cluster-ordering, core-distances and reachability-distance (line 12). Finally, the cluster-ordering, core-distances and reachability distances are outputted. Figure 6 illustrates the reachability-ordering graph of a 2D data set. The deep "valleys" indicates the density areas.

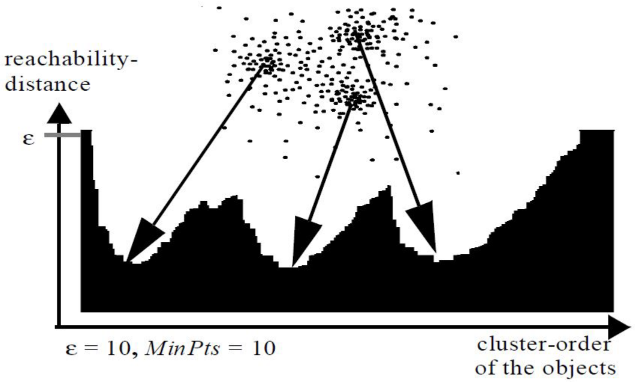


Figure 6. Reachability-ordering graph of a 2D data set. The shape of the order graph indicates the structure of the data set. The deep "valley" represents the dense area of

Now with the output of ordered data points and the parameter ε and MinPts, ExtractCluster method is able to extract a specific density-based clustering result with respect to a tuning parameter called clustering distance ε' (ε'≤ ε). The method loops through the points in cluster-ordering and assigns them cluster-memberships depending on their reachability-distance and the core-distance. The method contains three steps for each point. First, if the reachability-distance of a point is larger than the clustering-distance ε', the point is not density-reachable with respect to ε' and MinPts from any other points located before the current point in the cluster-ordering. Second, it re-evaluates the core point status. If the core-distance of a point is smaller than ε', then the point is identified as a core point and a new cluster starts; otherwise, the point is identified as not in any cluster. Last, if the reachability-distance of the current point is smaller than ε', the method assigns this point to the current cluster due to it is density-reachable with respect to ε' and MinPts from a preceding core point in cluster-ordering.

3.4 Validation

The current Mystic Isle game involves multi-planar, full body movements. Designed for individuals with diverse abilities, games can be played in a sitting or standing position, depending on the therapy treatment plan. In standing, the player is able to move around in the 3-dimensional space, akin to real-world rehabilitation. Few studies have evaluated the tracking and measurement capabilities of the Microsoft Kinect V2 for full-body, multiplanar movements in both sitting and standing. The purpose of this study was to determine the spatial accuracy and measurement validity of the Microsoft Kinect V2 sensor in a Natural User Interfaces rehabilitation game in comparison to a gold-standard marker-based motion capture system (Vicon™) [40].

3.4.1 Materials and methods

3.4.1.1 Participants

Participants were recruited via convenience sample at the University of Missouri-Columbia campus. Participants were included if they: 1) were over the age of 18, 2) could understand conversational English, and 3) had no medical conditions which prevented them from playing video games. All potential participants were screened and consented before beginning the study. The Health Sciences Institutional Review Board at the University of Missouri approved this study.

3.4.1.2 Vicon^{TM}

The Vicon system is a marker-based motion capture system that uses infrared cameras to track the 3-dimensional locations of reflective markers placed on the body. It can be used to measure or give real-time feedback on the movements of the whole body. The Vicon system has been used as an assessment tool for posture analysis, and in balance and reaching studies [23]. It is a gold standard tool for biomechanical kinematic assessment

Table 2. The mapping of joints from the Kinect V2 and the joints from Vicon Plug-in gait model

Kinect joints	Vicon markers	Kinect joints	Vicon markers
Hand_Left	LFIN	Hand_Right	RFIN
Elbow_Left	LELB	Elbow_Right	RELB
Shoulder_Left	LSHO	Shoulder_Right	RSHO
Hip_Left *	LASI or (LASI+LPSI)	Hip_Right *	RASI or (RASI+RPSI)
Spine_Mid	CLAV	Spine_Base *	(RASI+LASI) or (RASI+LASI+RPSI+LPSI)

* Multiple mappings to Vicon markers have been tested for these Kinect joints. The Lip joint is mapped to either the ASI marker or the middle position of the ASI and PSI markers. The Spine_Base joint is mapped to either the middle position of left and right the ASI marker or

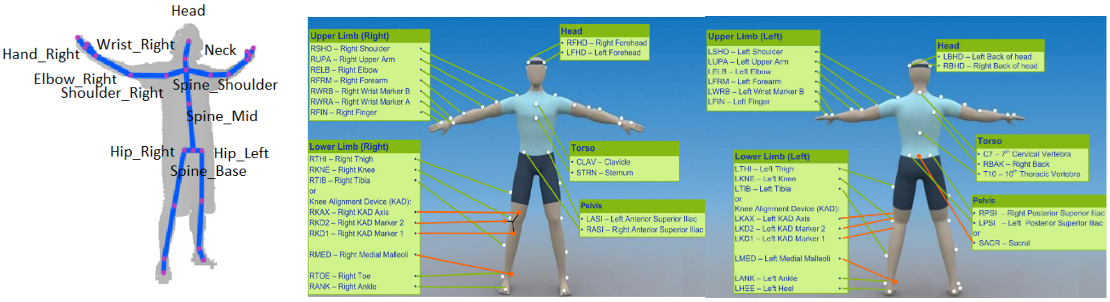
[23]. The sample rate of the Vicon system is 100Hz. For this study, the system included 7 individual cameras placed in a space with a ceiling height of 13 feet.

3.4.1.3 Mapping of the joints

The Kinect V2 provides a skeleton model [107] (Figure 7 (a)) of a game player by recording the x, y, and z coordinates of each discrete joint. The full-body Plug-in Gait model template [108] (Figure 7 (b)) is commonly used in a Vicon system to build the skeleton model. The joint locations in these two models are not the same. In order to validate the results using joint data from these two skeleton models for this study, we mapped the joints between the two systems (Table 2). For hand, elbow, shoulder and chest joints, the mapping was direct; for the hip and spine base joints, we took the average of several joint locations in the Plug-in gait model to optimize the matching.

3.4.1.4 Data collection

The sampling rate of the Kinect V2 is either 15 or 30 frames per second (f/s), depending on computer performance. In this study, 15 participants' Kinect V2 data were collected at a rate of 15 f/s on a lower performance laptop computer. The remaining 15



(a) (b)

Figure 7. The joint locations of the Kinect V2 skeleton model and a Vicon Plug-in gait model. (a) The joint labels and positions of Kinect V2 skeleton model. (b) the marker placement of a Vicon Plug-in gait model.

participants' Kinect V2 data were collected at a rate of 30 f/s on a higher performance desktop computer. In order to investigate how the sample rate influences the accuracy of the measurement outcomes in Mystic Isle, we analyzed the errors of extent metrics and speed metrics using the data collected under different frame rates separately. The average difference between the two frame rates in hand extension metrics were 0.70 ± 0.55 centimeters. The average difference between the two frame rates of hand speed metrics were 1.08 ± 1.09 centimeters/second. This variation of errors is tolerable and nearly negligible. Therefore, we will combine samples together for all analyses.

The layout of the data collection room and the coordinate systems of the two systems are displayed in make the coordinate space of the Kinect V2 overlay with the Vicon coordinate space, the z dimension of the Kinect V2 was aligned with the x dimension of the Vicon system shown in Figure 8. The distance between the origin points of the two systems was 2 meters. The display screen of the game was placed right behind the Kinect V2 and was not occluded by the Kinect V2.

Participants stood 1.8-2.4 meters (6-8 feet) from the Kinect V2 and close to the origin point of the Vicon system.

1. Sitting close: Two rings of eight objects were presented to each participant. The locations of the objects were within arm's length and no torso movement was required. The subject was seated.
2. Sitting far: Two rings of eight objects were presented to each participant. The locations of the objects required the participant to lean with their torso to be successful. The subject was seated.

3. Standing close: Two rings of eight objects were presented to each participant. The locations of the objects were within arm's length and no torso movement was required. The subject was standing and did not take a step.
4. Standing far: Two rings of eight objects were presented to each participant. The locations of the objects required the participant to lean with their torso to be successful. The subject was standing and did not take a step.
5. Standing step: Two rings of eight objects were presented to each participant. The locations of the objects required the participant to take a step in order to reach the virtual object.
6. Sorting game: Two rings of eight brightly colored objects were presented to each participant. Four color areas appeared in the virtual environment. The participant was then instructed to select an object and "drag" it into the matching-colored area. This game used the same calibration for the "standing close" game.

3.4.1.5 Data analysis

Data pre-processing and statistical analysis were performed in R2017a MATLAB. The Kinect V2 coordinates were transformed, data from both systems were filtered and synchronized, and the Vicon data were down sampled. These steps are described in detail below.

Coordinate Transformation

As shown in Figure 8, the coordinates of the Kinect V2 and the Vicon system are different. In order to visualize the similarities in different dimensions and compute the correlation of the data from the two systems, it was necessary to perform coordinate transformation. We transformed the Kinect V2 coordinates to be the same as the Vicon's,

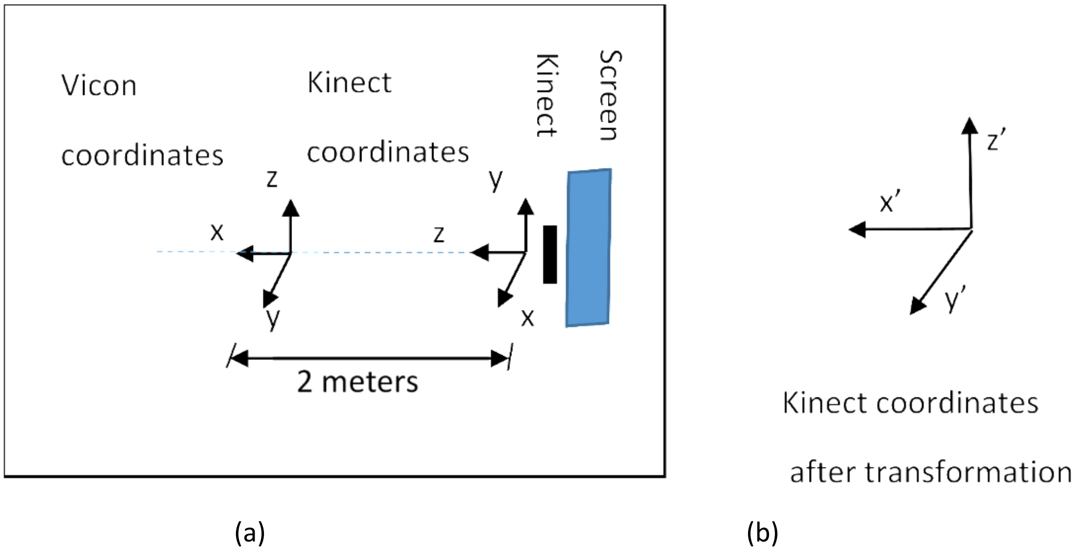


Figure 8. (a) The settings of the Vicon system and the Kinect V2. The origin of the Vicon system is set in the center of the room. The z dimension of the Kinect V2 coordinate is lined with x dimension of the Vicon's. (b) The transformed coordinates of the Kinect.

which means x, y and z dimension of the Kinect V2 Data have been transformed to y, z and x dimension, respectively.

Filtering

Noise, such as spike noise, quantization noise and white noise, can be introduced by digital devices when collecting data [109]. In addition, for the Vicon system, marker occlusion is possible, and gaps are filled in, introducing noise. To reduce noise, Butterworth filters were applied to both Kinect V2 and the Vicon data. A sixth-order Butterworth filter with 4Hz cut-off frequency was selected for Vicon data, while a 6th order Butterworth filter with 3Hz cut-off frequency was chosen for filtering the Kinect V2 Data. The combination of filter parameters were selected with the largest average Pearson's r correlation coefficient of the joints, which is also applied to our previous study [110].

Synchronization

Mystic Isle and the Vicon system started recording data at different times and through different input streams. In order to synchronize the data, the participants clapped three times at the beginning of each trial. The end of the clapping motion was considered to be the start point of a trial and the time stamp of the last game event of Mystic Isle was the end of the data trial. The data from two systems were cut based on the start and stop points.

Down Sampling

The sampling rate of the Vicon system (100 Hz) is different from the sampling rate of the Kinect V2 (15Hz or 30Hz). The velocity metric is affected by different sample rates. Thus, the Vicon data was down sampled close to either 15Hz or 30Hz to match Kinect V2 data's.

3.4.1.6 Outcomes

Spatiotemporal Accuracy

The signals representing the location of joints captured by the Kinect and the Vicon systems are spatial temporal signals. When analyzing the similarity of the spatiotemporal signals from the two systems, the mean of each signal was subtracted from the signal to minimize the bias.

Signal to noise ratio (SNR) compares the level of the ground truth signal with the level of noise. We applied SNR to compare the level of the signals from the Vicon system with the level of the signal difference between the two systems. The formula of SNR is

$$SNR = 10log_{10}(\frac{Vicon\ data}{Kinect\ data-Vicon\ data}) \quad (8)$$

We averaged the SNR results for each joint in different types of games. SNR is typically computed in decibels (dB). A SNR with 0 dB means the signal and the noise have the same level. A SNR below 0 dB indicates that the noise is larger than the desired signal; a 10 dB SNR indicates that the signal is 10 times larger than the noise [111].

Measurement Validity

Extent of reach was calculated for each trial. Extent of reach was defined as the distance from the hand joint to the shoulder center, where the shoulder center is the middle of the left and right shoulder joints. Suppose the hand joint and the shoulder center are represented by $j_{hand} = [h_x, h_y, h_z\}$ and $j_{shoulderC} = [s_x, s_y, s_z\}$, then the extent of reach for each frame is calculated by

$$Extent\ of\ Reach = \sqrt{(h_x - s_x)^2 + (h_y - s_y)^2 + (h_z - s_z)^2} \tag{9}$$

We also calculated maximum and mean velocities for each trial. Suppose the hand joint of the ith frame is represented by $j_{hand} = \{x_i, y_i, z_i\}$, the velocity of this frame is calculated by

$$Hand\ Velocity = \frac{\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2}}{t_{i+1} - t_i} \tag{10}$$

where t, time, is measured and stored automatically with each frame by the Kinect V2 SDK.

3.4.1.7 Statistical analysis

For spatiotemporal accuracy, we calculated the mean Euclidean 3-D distance, Pearson’s r correlation coefficient and SNR of each joint to determine the strength of association. For measurement validity, we calculated the difference and percentage error between the two systems for each participant per each game trial. These values were then averaged for different types of games. We also calculated the standard error of the difference, Pearson's r correlation coefficient and intra-class correlation (ICC) with 95 percentage confidence internal.

3.4.2 Results

3.4.2.1 Participants

Thirty subjects participated in this study, including 24 females and 6 males, with an average age of 24.2 years ± 6.6. Only two participants were left-handed.

3.4.2.2 Spatiotemporal accuracy

Upper Body

The average correlation coefficient of the arm joints was high; most of the correlation values were above 0.9 (Table 3). In addition, the SNR values of the arm joints (Table 4) were above 5, indicating a signal at least 5 times greater than noise. The hand joints had the greatest correlation between the two systems and very high SNR values. The chest (Spine Mid) "joint" had lower correlation between the two systems along with lower SNR values, ranging from 3 to 10. The mean 3D distance differences of joints were less than 10 centimeters (Table 5). The distance differences of chest (Spine Mid) "joint" were smaller than the joints on the arms. In addition, the distance differences were larger in the "standing step" game where the participants were required to take a step to reach an object.

Lower Body

When comparing the two systems, the lower body joints (Table 3 and Table 4) demonstrated less stability overall showing lower correlation values (0.5 to 0.9) than upper body and large variation in SNR values. However, lower body joints had smaller 3D distance differences than the values of upper body joints (Table 5). The differences were larger when the players performed a step motion in the game trial “standing step”.

Table 3. Correlation coefficients of spatiotemporal signals from the Vicon and the Kinect V2 for each of the six trials.

	Trial	**Sitting Close**			**Sitting Far**			**Standing Close**		
Upper-body joints	**Correlation coefficients**	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**
	Left Hand	**0.94**	**0.98**	**0.98**	**0.93**	**0.94**	**0.94**	**0.96**	**0.97**	**0.98**
	Right Hand	**0.90**	**0.95**	**0.97**	**0.94**	**0.98**	**0.98**	**0.98**	**0.97**	**0.98**
	Left Elbow	**0.94**	**0.97**	**0.97**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**
	Right Elbow	**0.94**	**0.95**	**0.97**	**0.93**	**0.95**	**0.96**	**0.97**	**0.96**	**0.97**
	Left Shoulder	0.88	**0.91**	**0.91**	**0.94**	**0.97**	**0.91**	**0.91**	**0.93**	**0.93**
	Right Shoulder	**0.90**	**0.90**	**0.91**	**0.92**	**0.96**	**0.92**	**0.92**	**0.94**	0.86
	Spine Mid-CLAV	0.83	0.75	0.85	0.87	0.89	0.84	0.88	0.79	0.78
	Trial	**Standing Far**			**Standing Step**			**Game**		
	Correlation coefficients	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**
	Left Hand	**0.96**	**0.96**	**0.97**	**0.95**	**0.92**	**0.95**	**0.96**	**0.98**	**0.98**
	Right Hand	**0.96**	**0.97**	**0.97**	**0.92**	**0.95**	**0.94**	**0.92**	**0.96**	**0.98**
	Left Elbow	**0.96**	**0.95**	**0.95**	**0.93**	**0.91**	**0.92**	**0.97**	**0.97**	**0.94**
	Right Elbow	**0.96**	**0.96**	**0.92**	**0.92**	**0.93**	**0.92**	**0.92**	**0.93**	**0.94**
	Left Shoulder	**0.93**	**0.97**	**0.84**	**0.90**	**0.93**	0.85	**0.96**	**0.98**	**0.93**
	Right Shoulder	**0.92**	**0.97**	0.88	**0.90**	**0.92**	0.87	**0.95**	**0.97**	**0.92**
	Spine Mid-CLAV	0.89	0.86	0.79	**0.90**	0.89	0.81	0.68	**0.95**	**0.95**
Lower-body joints	**Trial**	**Sitting Close**			**Sitting Far**			**Standing Close**		
	Correlation coefficients	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**
	Left Hip-LASI	0.66	0.61	0.60	0.82	0.76	0.57	0.89	0.88	0.55
	Right Hip-RASI	0.71	0.61	0.63	0.79	0.77	0.64	0.89	0.83	0.60
	Left Hip-LASI+LPSI	0.67	0.60	0.63	0.84	0.80	0.59	**0.92**	0.86	0.53
	Right Hip-RASI+RPSI	0.69	0.54	0.67	0.77	0.77	0.65	**0.93**	0.73	0.55
	Spine base-RASI+LASI	0.74	0.61	0.62	0.87	0.80	0.43	**0.96**	0.88	0.49
	Spine base-L/RASI+L/RPSI	0.73	0.58	0.61	0.84	0.81	0.58	**0.95**	0.76	0.63
	Trial	**Standing Far**			**Standing Step**			**Game**		
	Correlation coefficients	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**	**r_x**	**r_y**	**r_z**
	Left Hip-LASI	**0.90**	0.88	0.69	**0.91**	0.89	0.70	**0.90**	0.87	0.87
	Right Hip-RASI	0.89	0.87	0.70	**0.92**	0.87	0.64	0.88	0.82	0.71
	Left Hip-LASI+LPSI	**0.90**	0.89	0.57	0.87	0.89	0.55	**0.90**	**0.92**	0.87
	Right Hip-RASI+RPSI	**0.91**	0.85	0.69	0.87	0.86	0.63	**0.91**	0.78	0.77
	Spine base-RASI+LASI	**0.94**	0.88	0.65	**0.91**	0.89	0.66	**0.96**	0.89	0.91
	Spine base-L/RASI+L/RPSI	**0.93**	0.86	0.70	**0.92**	0.87	0.62	**0.95**	0.79	0.79

*Notes: Rx,Ry and Rz represent the Pearson’s *r* correlation coefficient in x, y and z dimensions. Values > 0.90 are bolded. CLAV, L/RASI and L/RPSI are the joint labels from Vicon plug-in-gait model (Table 1). CLAV represents the clavicle position. L/RASI represents the left and right anterior superior lilac, and L/RPSI represents the left and right posterior superior lilac [108].

Table 4 Signal-to-noise ratios of spatiotemporal signals from the Vicon and the Kinect V2 for each of the six trials.

	Trial	**Sitting Close**			**Sitting Far**			**Standing Close**		
Upper-body joints	**SNR**	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**
	Left Hand	**12.4**	**15.2**	**16.5**	**11.6**	**12.8**	**15.0**	**14.7**	**13.3**	**17.4**
	Right Hand	**9.8**	**12.1**	**14.6**	**12.6**	**15.6**	**15.9**	**15.5**	**14.3**	**19.0**
	Left Elbow	**11.3**	**12.7**	**14.6**	**12.7**	**13.2**	**13.2**	**12.9**	**11.0**	**12.5**
	Right Elbow	**10.3**	**9.9**	**13.4**	**11.3**	**11.7**	**13.0**	**12.9**	**11.7**	**13.7**
	Left Shoulder	6.5	7.5	7.4	**10.3**	**12.1**	7.5	**9.7**	**10.1**	5.3
	Right Shoulder	7.3	7.4	6.8	**9.6**	**11.1**	**8.2**	**9.2**	**9.5**	5.5
	Spine Mid-CLAV	4.4	4.1	5.0	7.5	7.9	4.9	**9.0**	5.5	3.6
	Trial	**Standing Far**			**Standing Step**			**Game**		
	SNR	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**
	Left Hand	**13.7**	**12.6**	**14.5**	**12.4**	**10.4**	**12.9**	**12.8**	**16.0**	**16.1**
	Right Hand	**13.9**	**13.7**	**15.4**	**10.8**	**11.5**	**11.1**	**11.9**	**14.6**	**15.7**
	Left Elbow	**12.2**	**11.1**	**10.5**	**11.1**	**10.3**	**10.0**	**12.1**	**14.4**	**10.0**
	Right Elbow	**12.4**	**11.6**	**10.7**	**9.9**	**10.9**	**8.7**	**9.7**	**11.8**	**9.8**
	Left Shoulder	**10.3**	**12.2**	6.1	**9.4**	**10.8**	6.4	**9.4**	**12.3**	**8.9**
	Right Shoulder	**10.1**	**11.9**	7.6	**8.8**	**10.3**	6.5	**8.7**	**12.8**	**8.2**
	Spine Mid-CLAV	**9.6**	6.6	5.4	**9.1**	7.9	4.8	3.1	**8.5**	**10.6**
Lower-body joints	**Trial**	**Sitting Close**			**Sitting Far**			**Standing Close**		
	SNR	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**
	Left Hip-LASI	-0.7	-3.1	-3.8	5.5	2.8	-1.0	7.5	**8.2**	-5.4
	Right Hip-RASI	1.2	-3.5	-7.4	5.1	2.8	-1.2	**8.2**	6.0	-8.9
	Left Hip-LASI+LPSI	-1.3	-5.3	-4.4	5.8	2.4	-0.4	**8.3**	7.9	-5.4
	Right Hip-RASI+RPSI	0.2	-5.5	-4.3	4.8	2.0	0.2	**9.0**	4.9	-8.0
	Spine base-RASI+LASI	1.5	-3.3	-8.9	6.9	3.3	-6.2	**12.2**	**8.1**	-7.3
	Spine base-L/RASI+L/RPSI	0.8	-4.9	-5.6	6.4	2.5	-1.9	**11.7**	5.4	-8.2
	Trial	**Standing Far**			**Standing Step**			**Game**		
	SNR	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**	**SNR_x**	**SNR_y**	**SNR_z**
	Left Hip-LASI	**8.5**	**8.2**	-3.9	**9.0**	**8.4**	-1.1	7.0	7.9	-2.3
	Right Hip-RASI	**8.3**	7.0	-5.4	**9.8**	7.6	-3.7	6.2	5.2	-7.7
	Left Hip-LASI+LPSI	7.7	**8.3**	-5.3	7.3	**8.3**	-1.7	7.1	**8.5**	-3.8
	Right Hip-RASI+RPSI	**8.1**	6.7	-6.5	7.6	7.2	-4.8	**8.3**	3.9	-9.9
	Spine base-RASI+LASI	11.5	**8.4**	-8.1	**10.1**	**8.5**	-3.4	**10.7**	**8.4**	-3.5
	Spine base-L/RASI+L/RPSI	**9.7**	7.0	-8.9	**9.5**	7.5	-5.8	**9.2**	4.3	-9.2

*Note: SNRx, SNRy and SNRz represent the signal to noise ratio in x, y and z dimensions. Values > +8 are bolded. CLAV, L/RASI and L/RPSI are the joint labels from Vicon plug-in-gait model (Table 2). CLAV represents the clavicle position. L/RASI represents the left and right anterior superior lilac, and L/RPSI represents the left and right posterior superior lilac [108].

Table 5. Spatiotemporal accuracy of joint signals from the Kinect V2 against the Vicon markers for each of the six trials. The accuracy is evaluated by the mean 3D Euclidean distance in centimeter and corresponding standard deviation.

Game Type	**Sitting Close**	**Sitting Far**	**Standing Close**	**Standing Far**	**Standing Step**	**Game**
Joint Name	Diff3d	Diff3d	Diff3d	Diff3d	Diff3d	Diff3d
			Upper Body			
Left Hand	4.07(4.11)	4.34(4.05)	3.80(3.69)	5.02(4.52)	8.70(7.21)	6.17(5.34)
Right Hand	4.35(4.36)	5.42(4.81)	3.80(3.37)	5.69(5.53)	9.12(9.57)	8.39(7.02)
Left Elbow	2.80(2.12)	3.07(2.29)	2.92(2.15)	4.60(2.80)	7.19(5.15)	5.69(3.92)
Right Elbow	3.02(2.43)	4.03(2.80)	3.20(2.02)	4.86(3.31)	7.90(6.04)	7.41(5.44)
Left Shoulder	1.87(1.11)	2.39(1.25)	1.97(1.08)	3.47(1.93)	6.19(4.65)	4.71(3.33)
Right Shoulder	1.92(1.09)	2.63(1.43)	2.07(1.12)	3.33(1.90)	6.29(4.62)	4.40(3.18)
Spine Mid-CLAV	1.36(0.65)	2.19(1.03)	1.87(0.85)	2.97(1.51)	6.64(5.09)	5.01(2.76)
			Lower Body			
Left Hip-LASI	3.28(1.34)	2.95(0.82)	1.45(0.74)	2.67(1.44)	6.10(4.53)	3.82(2.66)
Right Hip-RASI	1.23(0.53)	1.95(0.77)	1.45(0.74)	2.65(1.46)	6.21(4.38)	4.21(2.94)
Left Hip-LASI+LPSI	2.11(0.91)	2.08(0.73)	1.64(0.82)	2.88(1.51)	5.65(4.09)	4.28(2.62)
Right Hip-RASI+RPSI	1.17(0.53)	1.66(0.73)	1.62(0.80)	2.86(1.43)	5.72(3.87)	4.68(3.05)
Spine Base-RASI +LASI	1.93(0.53)	2.05(0.68)	1.21(0.64)	2.38(1.34)	5.29(3.82)	3.35(2.32)
Spine Base-L/RASI +L/RASI	1.44(0.56)	1.61(0.64)	1.45(0.73)	2.66(1.28)	5.29(3.69)	4.24(2.66)

*Note: CLAV, L/RASI and L/RPSI are the joint labels from Vicon plug-in-gait model (Table 2). CLAV represents the clavicle position. L/RASI represents the left and right anterior superior lilac, and L/RPSI represents the left and right posterior superior lilac [108].

3.4.2.3 Measurement validity

Extent of Reach

Overall, the average difference values of maximum extent of reach were less than 3 cm across all six trials and the percentage error was less than five percent (Table 6). More errors were introduced in measurements of the right hand as compared to the left hand. The Pearson's r correlation coefficient of extent of hand in x, y and z dimension are high. Most were greater than 0.8. Only one trial had the lowest value 0.7. Extent of reach in 3D had lower Pearson's r correlation coefficient correlation values compared to extent of reach in each dimension. But the values were not less than 0.6. The intra-class correlation values of extent of reach around the sagital and frontal axes were very high (>0.96) and larger than movements around the vertical axis for most of the trials. The intra-class correlation values of extent of reach in 3D were relatively low in standing-type trials.

Maximum and Mean Velocity

Maximum velocity had larger errors than mean velocity over all the trials (Table 6). The largest average error of maximum velocity was about 10cm/s from the "game" trial. For mean velocity, the largest amount of error was less than 4 cm/s. When considering percentage error, the average percentage error of mean velocity was about 10% and the average percentage error of maximum velocity was less than 5%. The errors from the "game" trial were greater than other trials and mean velocity errors were larger in sitting versus standing trials. The Pearson's r correlation coefficient values of maximum and mean velocities were not less than 0.9 and the intra-class correlation values were not less than 0.97.

Table 6. Accuracy of clinical measures from the Kinect V2 against the Vicon for each of the six trials. Accuracies were validated using mean difference, standard error, mean percentage error, Pearson's r correlation coefficient and intra-class correlation with corresponding 95% confidence internal.

	Diff		**SE**		**Percentage error**		**Pearson's r**		**ICC(3,1) 95% Confidence internal**	
Metrics	L	R	L	R	L	R	L	R	L	R
Sitting close										
Max Ext_X	0.8	2.4	0.2	0.3	1.9	5.4	1.0	1.0	0.99(0.99;1.00)	0.97(0.95;0.99)
Max Ext_Y	1.8	1.9	0.4	0.5	3.2	3.2	1.0	1.0	0.98(0.91;0.99)	0.98(0.97;0.99)
Max Ext_Z	1.4	3.0	0.4	0.4	2.6	5.4	0.9	0.8	0.93(0.84;0.96)	0.86(0.71;0.93)
Max Ext_3D	1.7	2.3	0.3	0.3	2.4	3.3	0.9	0.8	0.85(0.92;0.95)	0.85(0.37;0.95)
Max Speed	6.7	5.0	1.4	0.9	3.1	2.5	1.0	1.0	0.99(0.97;1.00)	1.00(0.99;1.00)
Mean Speed	2.0	2.0	0.3	0.3	12.0	10.6	1.0	1.0	0.99(0.81;1.00)	0.98(0.75;1.00)
Sitting far										
Max Ext_X	2.1	2.9	0.5	0.4	4.5	6.1	1.0	1.0	0.98(0.93;1.00)	0.97(0.83;0.99)
Max Ext_Y	1.3	2.4	0.3	0.3	2.4	4.0	1.0	1.0	1.00(0.99;1.00)	0.98(0.95;0.99)
Max Ext_Z	1.0	2.8	0.4	0.4	1.9	5.0	1.0	0.9	0.97(0.94;0.99)	0.93(0.85;0.96)
Max Ext_3D	1.8	2.7	0.3	0.4	2.6	3.9	1.0	0.8	0.97(0.89;0.99)	0.81(0.42;0.92)
Max Speed	3.3	4.7	0.5	1.0	2.4	2.8	1.0	1.0	1.00(1.00;1.00)	1.00(0.99;1.00)
Mean Speed	2.4	2.0	0.3	0.3	12.9	11.0	1.0	1.0	1.00(0.99;1.00)	0.98(0.94;0.99)
Standing close										
Max Ext_X	1.8	2.7	0.3	0.3	3.5	5.3	1.0	1.0	0.99(0.98;1.00)	0.98(0.96;0.99)
Max Ext_Y	1.2	2.3	0.4	0.3	1.9	3.7	1.0	0.9	0.99(0.98;1.00)	0.96(0.91;0.98)
Max Ext_Z	1.0	2.7	0.4	0.4	1.6	4.2	0.8	0.8	0.88(0.74;0.95)	0.83(0.37;0.94)
Max Ext_3D	1.6	2.5	0.3	0.5	2.3	3.5	0.9	0.6	0.90(0.68;0.96)	0.74(0.45;0.88)
Max Speed	7.6	7.4	1.2	1.4	4.1	4.3	1.0	1.0	0.99(0.99;1.00)	0.99(0.98;1.00)
Mean Speed	1.4	1.7	0.3	0.3	6.5	7.5	1.0	1.0	0.99(0.93;1.00)	0.99(0.94;0.99)
Standing far										
Max Ext_X	2.7	2.9	0.5	0.3	4.8	4.9	1.0	0.9	0.99(0.96;1.00)	0.93(0.74;0.97)
Max Ext_Y	1.0	1.4	0.2	0.2	1.7	2.4	1.0	1.0	0.99(0.99;1.00)	0.99(0.98;1.00)
Max Ext_Z	1.1	2.9	0.4	0.4	1.7	4.6	0.9	0.9	0.94(0.85;0.97)	0.92(0.69;0.97)
Max Ext_3D	2.5	2.6	0.5	0.4	3.5	3.6	0.7	0.6	0.74(0.32;0.89)	0.68(0.18;0.86)
Max Speed	5.9	8.5	1.0	1.2	3.3	3.9	1.0	1.0	1.00(0.99;1.00)	0.99(0.97;0.99)
Mean Speed	2.1	2.0	0.3	0.3	9.5	9.0	1.0	0.9	0.97(0.93;0.99)	0.97(0.94;0.99)
Standing step										
Max Ext_X	2.6	3.0	0.4	0.3	4.9	5.4	1.0	1.0	0.99(0.97;1.00)	0.97(0.89;0.99)
Max Ext_Y	0.9	2.3	0.2	0.4	1.6	3.5	1.0	0.9	0.99(0.99;1.00)	0.97(0.93;0.98)
Max Ext_Z	1.8	2.8	0.4	0.4	2.8	4.5	0.9	0.9	0.88(0.62;0.95)	0.90(0.41;0.97)
Max Ext_3D	2.3	2.5	0.3	0.4	3.4	3.5	0.8	0.7	0.87(0.69;0.94)	0.75(0.46;0.88)
Max Speed	6.6	4.9	1.1	1.0	3.8	2.3	1.0	1.0	1.00(1.00;1.00)	1.00(0.99;1.00)
Mean Speed	1.6	1.7	0.3	0.3	4.9	5.5	1.0	1.0	1.00(0.99;1.00)	0.99(0.96;1.00)

Game										
Max Ext_X	1.5	3.0	0.3	0.3	2.8	5.0	1.0	0.9	0.99(0.98;1.00)	0.91(0.59;0.97)
Max Ext_Y	1.8	2.8	0.4	0.4	3.2	4.7	1.0	0.9	0.98(0.96;1.00)	0.96(0.91;0.98)
Max Ext_Z	2.8	2.7	0.4	0.5	4.2	4.4	0.7	0.8	0.70(-0.10;0.89)	0.87(0.73;0.94)
Max Ext_3D	2.4	2.9	0.5	0.5	3.4	4.1	0.8	0.6	0.81(0.51;0.92)	0.63(0.18;0.83)
Max Speed	8.1	9.3	1.6	1.7	4.7	4.5	1.0	1.0	0.99(0.98;1.00)	0.99(0.97;0.99)
Mean Speed	3.5	3.7	0.6	0.7	9.2	9.0	0.9	0.9	0.97(0.93;0.98)	0.97(0.94;0.97)

*Note: 'L' and 'R' means left hand and right hand. Max_Ext means the maximum hand of extent.

3.4.3 Discussion

Mystic Isle, similar to other rehabilitation-focused games and software, has been shown feasible as an intervention for people with stroke with the Microsoft Kinect V2 camera being used as an input device. Before using the Kinect V2 and the Mystic Isle software as an assessment tool in a clinical setting, it is necessary to validate the accuracy of the Kinect V2's tracking capabilities. Therefore, the purpose of this study was to determine the spatial accuracy and measurement validity of the Microsoft Kinect V2 sensor in comparison to a gold-standard marker-based motion capture system (ViconTM). We have demonstrated that Mystic Isle provides an accurate measurement of movement relative to the Vicon system; however there are some movements and planes of measurement in which the accuracy is considerably lower. The findings from this study are similar to findings from other comparison studies between the Kinect V1 and the Vicon. This study is different from prior work in that we tracked movements of the upper limbs during unrestrained full-body movements (versus just the lower limbs during walking) and the participants were not limited to specific planes of movements and could choose to use either hand during a reach [109, 111, 112]. The movements in this study more closely mimic real-world performance; this has significant implications for clinical rehabilitation practice. Each of these points are discussed below. We conclude with limitations and next steps for research and clinical practice.

With regards to exploring measurement validity, we found that the errors of hand extension and speed metrics from the right hand were larger than the errors of the left hand, but the higher error rate is still close enough for relevant clinical assessments. Also, we also observed that the percentage errors of mean velocity of sitting trials were larger than the error from standing trials. In sitting, participants tended to move slower than in standing; thus overall velocity was lower in sitting. However, the absolute errors were similar across all trials.

The sixth trial, the sorting game, had the largest percentage errors of all trials. There are two reasons for this. First, the sorting game trial was the longest trial. The longer a person is engaged with the task, the greater the potential for noise to be introduced. Second, the required movements for game success were different than the other trials. Participants "dragged" a virtual target from one side of the screen to another in order to "sort" the virtual objects. Further, some participants had to bend at the knee in order to "place" the virtual object in the correct spot. The bending position likely introduced some noise and limited the tracking capability of the Kinect V2, particularly at the hip joints.

When considering spatial accuracy of the tracked joints, the joints of the arm were highly correlated between both systems and had high SNR values. The joints of the hip had much lower SNR values and fewer correlations over 0.90. Other researchers have reported similar findings with regards to the lower body [111]. Mentiplay et al. found poor agreement between Kinect V2 and a Vicon system in peak hip flexion [113]. These lower correlation values have often been interpreted as a consequence of the optimization of

Kinect SDK for gesture-based games [111]. Thus, The Kinect SDK appears to provide higher tracking abilities on upper body joints.

Despite the decreased spatiotemporal accuracy of the Kinect for tracking lower body joints, researchers have shown that the Kinect is able to track walking paths and provide data for calculating gait-related variables with relatively high accuracy (e.g., stride length, walking speed) [114, 115]. Guess et al. showed the Kinect can accurately measure hip and knee flexion angles for a vertical drop jump [112]. One of the first evaluations of the Kinect for upper body tracking demonstrated similar percentage errors [109]. In this study, we explored full-body movements that involved reaching, sitting, stepping, and cross-body movements. These results add a richness to the primarily gait-related literature validating the use of the Kinect for tracking upper body kinematics during full body movement. Allowing participants more freedom in a reaching movement (e.g., choice of hand, allowing cross-body reaches) mimics daily activity much more closely than other studies [109, 111, 112]. This may limit the internal validity of the study; however it greatly increases the external validity of the findings. We are the first to validate the Kinect V2 in this scenario.

Additionally, these findings support the use of the Kinect V2 in a clinical rehabilitation setting. We have shown that the Kinect V2 is an accurate tool for tracking movement; the clinical measurements we can obtain (e.g., extent of reach) are repeatable and valid. Reliability of standard clinical assessment tools for range of motion (goniometers) vary across clinical populations and joints measured [116, 117]. The ICCs between raters and between tools in prior studies range from 0.50 to 0.98 [118, 119]. Therefore, the Kinect V2 has the potential to be utilized in clinical practice and home-based rehabilitation to complement existing outcome assessments. Furthermore, these data can be collected by the Kinect V2 in remote settings, such as a patient's home, and provide clinicians with a look at performance over time. Health insurance companies are demanding more data and metrics to support clinical decision making. With a validated sensor, this system has the potential to provide rehabilitation clinicians and insurers with high quality, performance-based data and outcomes.

This study has a few limitations. First, the sample is relatively homogenous, young, and a majority of females, which limits a generalization of the findings to an older population, which is common in stroke rehabilitation. Our previous studies with the Mystic Isle game have involved stroke patients [27, 39]; however, we have not yet validated the assessments in this population. Our ongoing research is investigating this further in an older population. Second, there were some error differences in tracking the left and right hands, although these were not statistically significant, and the errors are within acceptable rates for clinical use for both left and right sides [115, 120]. Lastly, people with stroke have different movement patterns and postures as compared to healthy individuals. Flexor synergy patterns and spasticity might make it more difficult for the Kinect V2 to reliably track the more affected extremity; however, we have had much success in our prior work [27, 39].

Our preliminary research has shown that motor function and daily activity performance of stroke patients can improve through the use of Mystic Isle as an in-home intervention [121]. The next step for our research is to use the Kinect V2 to complete assessments of movements in people with stroke. Additionally, we are building an in-home monitoring system that utilizes the Kinect V2 for ambient tracking of movement and as an assessment of upper extremity movement performance. These studies will further test the use of the Kinect V2 as a valid tool for tracking movements in rehabilitation populations.

3.5 Kinematic assessment on stroke participants

To explore the use of VR-based rehabilitation game to assess upper extremity movement for individuals post-stroke, movement data captured with the Microsoft Kinect® from four separate studies were aggregated for analysis (n = 8 individuals post-stroke, n = 30 individuals without disabilities). Kinematic measurements, normalized jerk, movement path ratio, and average path sway, were used to evaluate the smoothness and efficiency of the hand movements. Data from the 30 healthy individuals created a normative baseline for the three kinematic variables. The assessment outcome of individuals post-stroke was compared with the normative baseline.

3.5.1 Participants and study design

Demographic data for all studies are combined and reported in Table 7. Respective Institutional Review Boards approved all studies and all participants provided written Informed Consent prior to enrollment.

*FMA-UE: Fugl-Meyer Assessment- Upper Extremity baseline score

<u>Study 1: Healthy Individuals</u>

Table 7. Demographics for participants across the four studies included in the analysis.

		AGE	SEX	HANDEDNESS	SIDE IMPACTED BY STROKE	FMA-UE*
STUDY 1		24.2 ± 6.6	M: 6, F: 24	R: 28, L: 2	--	Healthy
STUDY 2	P1	55	M	R	L	24
	P2	54	F	R	R	45
	P3	56	M	R	L	25
STUDY 3	P1	61	F	R	L	64
	P2	47	M	R	R	66
	P3	36	M	R	L	64
STUDY 4	P1	56	F	L	L	ND
	P2	57	M	R	R	ND

For this study [40], thirty individuals without any hemiparesis or disability in the arm or hand interacted with the Mystic Isle game while being simultaneously tracked by the Vicon motion capture system. Each individual played 12 games.

Study 2: In-Home Post-stroke (moderate impairments)

For this study [39], three individuals with moderate upper extremity impairments post-stroke played the Mystic Isle game in their home for 6 weeks. Each of the games was designed based on participant-identified goals using the Canadian Occupational Performance Measure (COPM) [122]. All three participants reported improvements in performance of their daily occupations and there were minimal barriers to use of the system in the home setting [39].

Study 3: In-Home Post-stroke (minimal impairments)

The study followed the same methods as Study 2; however all three participants had mild upper extremity impairments post-stroke. No outcome data have been formally reported for this study.

Study 4: LSVT®BIG home exercises (moderate impairments)

For this study, two participants completed the Lee Silverman Voice Treatment®-BIG (LSVT®BIG) intervention 4 days per week for 4 weeks. The LSVT®BIG intervention has an associated home program that is completed once on intervention days and twice on non-intervention days. The two participants completed the home program via the Mystic Isle game. The exercises and functional tasks involved sustained reaches and movements, interacting with virtual objects in a similar way to the prior studies. Both participants had

Table 8. Average values for efficiency metrics across the four studies.

		PATH RATIO		**AVERAGE SWAY DISTANCE (MM)**	
Study 1		0.76 ± 0.11		80.15 ± 30.08	
		More affected	**Less affected**	**More affected**	**Less affected**
Study 2	P1	0.47 ± 0.25	0.63 ± 0.18	106.3 ± 65.9	146.0 ± 105.5
	P2	0.55 ± 0.22	0.51 ± 0.23	182.2 ± 87.7	115.4 ± 61.9
	P3	0.20 ± 0.12	0.15 ± 0.14	367.6 ± 102.8	271.4 ± 193.1
Study 3	P1	0.56 ± 0.17	0.61 ± 0.11	151.9 ± 52.5	131.6 ± 40.8
	P2	0.47 ± 0.23	0.51 ± 0.21	170.9 ± 88.8	156.3 ± 86.0
	P3	0.53 ± 0.23	0.54 ± 0.21	182.4 ± 115.2	175.8 ± 98.8
Study 4	P1	0.35 ± 0.24	0.38 ± 0.26	253.9 ± 126.0	233.0 ± 133.0
	P2	0.50 ± 0.20	0.51 ± 0.17	112.0 ± 76.6	122.0 ± 67.1

improvements in upper extremity motor function and self-rated occupational performance [123].

3.5.2 Statistical analysis

For the healthy individuals in Study 1, a within subjects t-test was used to determine if there was a statistically significant difference in right vs. left upper extremity reaches. If there was no difference, the data were collapsed into one group and averaged for all kinematic variables. Given the small sample size and diversity within the study participants post-stroke, the data were averaged across all reaches for each kinematic variable.

3.5.3 Assessment results

There were no differences in path ratio (p = 0.56), average sway distance (p = 0.45), and normalized jerk (p = 0.62, 0.49 and 0.86 for reaches on target paths, reaches without outliers and reaches on valid paths, respectively) between the right and left upper extremities for the healthy individuals. Therefore, the data were averaged across the right and left upper extremities for all three kinematic variables. Table 8 and Table 9 display the results for all kinematic variables for healthy individuals as well as each post-stroke participant by sides more and less impacted by their stroke.

Table 9. Average values for normalized jerk across the four studies.

		Reaches on target paths		Reaches without outliers		Reaches on valid paths	
Study 1		$6.0E^{4} \pm 8.9E^{5}$		$4.3E^{3} \pm 6.0E^{3}$		$4.3E^{4} \pm 5.7E^{5}$	
		More affected	Less affected	More affected	Less affected	More affected	Less affected
Study 2	P1	$1.0E^{7} \pm 1.4E^{8}$	$7.4E^{6} \pm 7.4E^{0}$	$1.5E^{5} \pm 2.5E^{5}$	$2.6E^{5} \pm 3.8E^{5}$	$1.2E^{7} \pm 1.8E^{8}$	$1.8E^{7} \pm 1.4E^{8}$
	P2	$2.6E^{5} \pm 5.7E^{6}$	$4.4E^{5} \pm 1.1E^{7}$	$5.5E^{3} \pm 8.8E^{3}$	$6.3E^{3} \pm 9.7E^{3}$	$1.2E^{5} \pm 1.6E^{6}$	$8.8E^{4} \pm 1.0E^{6}$
	P3	$8.9E^{8} \pm 5.6E^{9}$	$1.0E^{9} \pm 5.8E^{9}$	$3.5E^{6} \pm 6.3E^{6}$	$6.9E^{6} \pm 1.3E^{7}$	$3.3E^{8} \pm 1.2E^{9}$	$8.8E^{8} \pm 4.6E^{9}$
Study 3	P1	$3.1E^{6} \pm 2.1E^{7}$	$2.6E^{6} \pm 1.9E^{7}$	$2.5E^{5} \pm 3.2E^{5}$	$2.4E^{5} \pm 3.3E^{5}$	$2.0E^{6} \pm 1.7E^{7}$	$1.6E^{6} \pm 1.9E^{7}$
	P2	$1.4E^{6} \pm 4.4E^{7}$	$3.7E^{6} \pm 1.3E^{8}$	$1.7E^{4} \pm 2.3E^{4}$	$1.7E^{4} \pm 2.3E^{4}$	$1.5E^{6} \pm 4.8E^{7}$	$4.9E^{6} \pm 1.7E^{8}$
	P3	$1.7E^{6} \pm 7.2E^{6}$	$2.0E^{6} \pm 9.9E^{6}$	$2.2E^{5} \pm 3.2E^{5}$	$2.2E^{5} \pm 3.2E^{5}$	$1.1E^{6} \pm 6.1E^{6}$	$1.4E^{6} \pm 7.4E^{6}$
Study 4	P1	$1.4E^{5} \pm 1.4E^{6}$	$1.3E^{5} \pm 1.0E^{6}$	$1.7E^{4} \pm 2.4E^{4}$	$4.2E^{3} \pm 6.0E^{3}$	$1.2E^{5} \pm 1.0E^{6}$	$9.8E^{4} \pm 6.8E^{5}$
	P2	$1.6E^{6} \pm 4.4E^{7}$	$1.8E^{6} \pm 5.4E^{7}$	$3.2E^{4} \pm 4.4E^{4}$	$3.2E^{4} \pm 4.3E^{4}$	$1.8E^{6} \pm 4.9E^{7}$	$1.9E^{6} \pm 5.9E^{7}$

Kinematic variable: Path ratio

The average movement path ratio for healthy individuals was 0.76 ± 0.11. For all participants post-stroke, even those with mild impairments, average path ratio was smaller (indicating less efficient movements). Most participants post-stroke also had a smaller path ratio for their more affected side post-stroke shown in Figure 9, Table 8.

Kinematic variable: Average sway distance

The average sway distance for healthy individuals was 80.15 ± 30.08. Most participants post-stroke had larger sway distances (indicating a less efficient path taken by the upper extremity). Again, those with mild impairments had large average sway

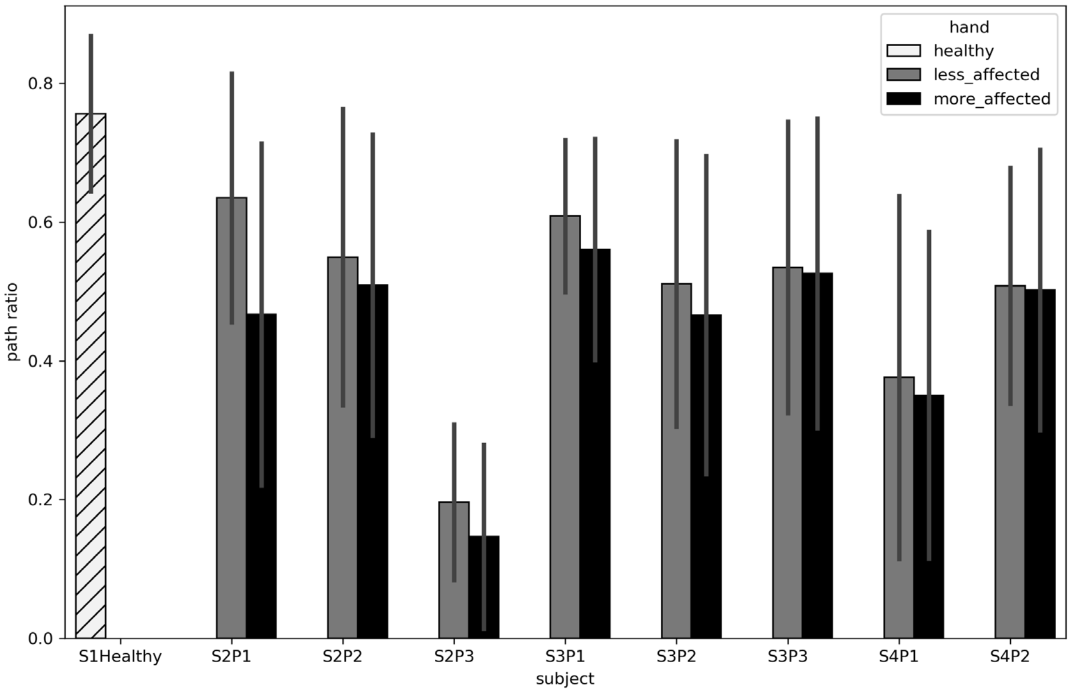


Figure 9. The mean and standard deviation of path ratio of healthy individuals and participants post-stroke.

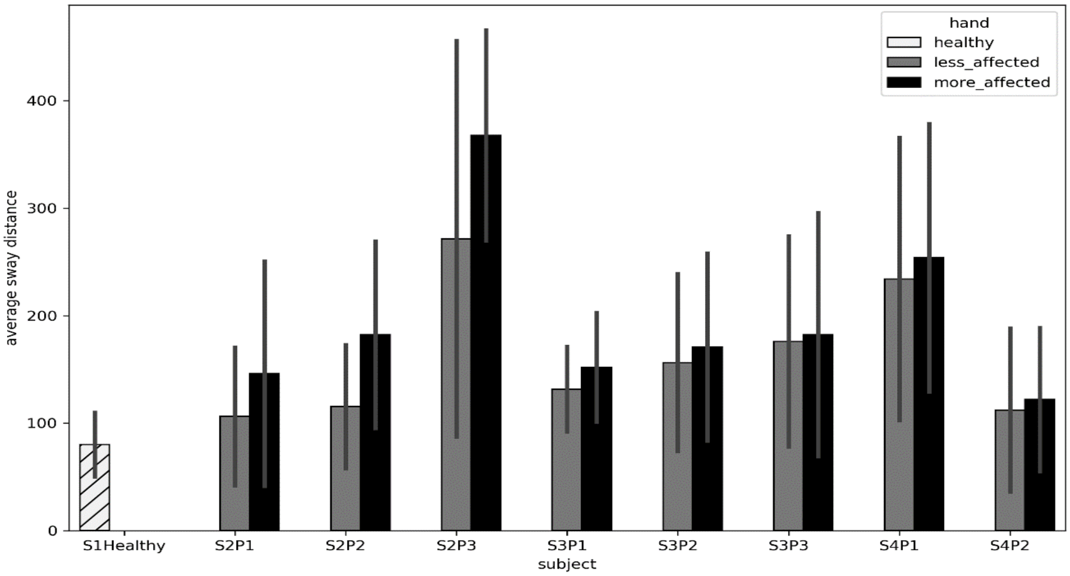


Figure 10. The mean and standard deviation of average sway distance of healthy individuals and participants post-stroke.

distances and large standard deviations (variance) in their movements shown in Fgiure 10, Table 8.

<u>Kinematic variable: Normalized jerk</u>

We report normalized jerk values for all data analysis approaches. The average normalized jerk on target paths for healthy individuals was 6.0E4 ± 8.9E5 for reaches on target paths, 4.3E3 ± 6.0E3 for reaches with outliers removed, and 4.3E4 ± 5.7E5 for reaches on valid paths. For the participants post-stroke, most values were 2-5 orders of magnitude larger than healthy individuals, even on the less affected side post-stroke shown in Figure 11, Table 9.

3.5.4 Discussion

The purpose of this study was to explore the utility of the Microsoft Kinect® in assessing clinically relevant measures of movement quality for individuals post-stroke.

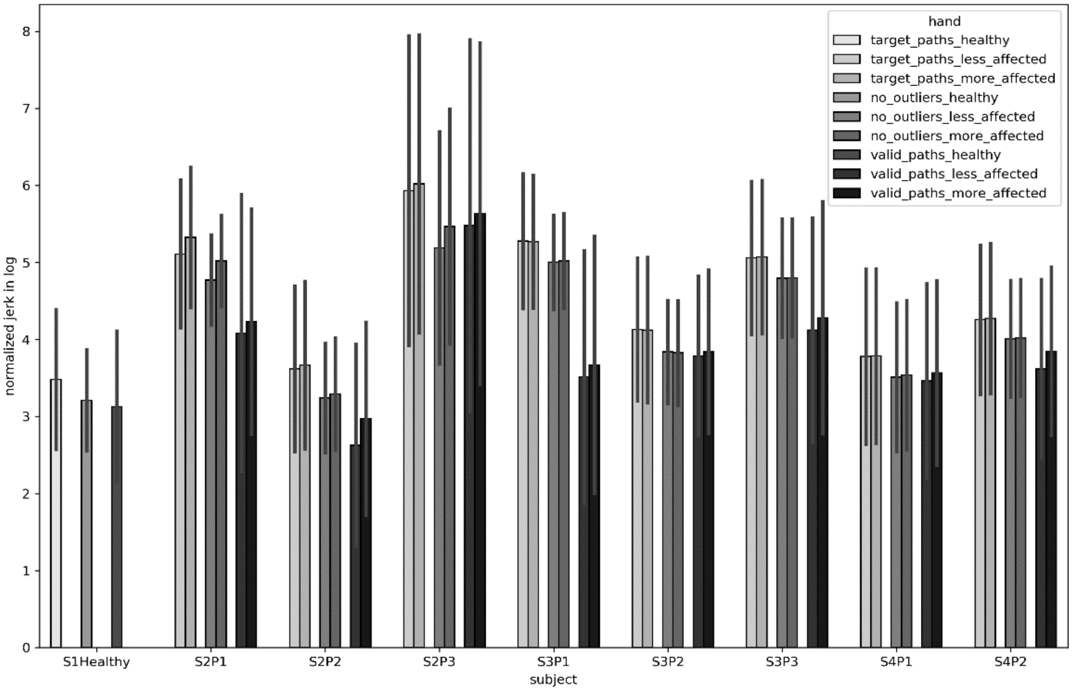


Figure 11. The mean and standard deviation of normalized jerk of healthy individuals and participants post-stroke.

We can calculate new assessments of movement quality, such as smoothness of movement and movement efficiency. Further, these kinematic variables are potentially sensitive in detecting less efficient movement in individuals with mild motor impairments post-stroke. We discuss each of these points below along with limitations and future research.

The clinical kinematic variables calculated in this study provide insight into the quality of movement for post-stroke individuals. For rehabilitation games played using a 3-dimensional depth sensor, the therapist programs the location of the virtual objects in 3-dimensional space [39, 124]. Individuals are only successful in the *Mystic Isle* game if they touch the virtual object at that full extent of reach. Therefore, traditional measures of range of motion or extent of reach provide little clinical insight. Additionally, for post-stroke individuals, reaching velocity is often not indicative of movement quality. For those with spasticity or increased tone, asking them to move faster leads to compensatory movement strategies during reaching [125]. Movement efficiency and smoothness are measures of the quality of a reach and have implications for designing interventions in the clinic and home settings. A therapist can include treatment components that challenge movement efficiency and smoothness within a functional task. As an assessment tool, these clinical kinematic variables can be used to document performance over time and supplement existing clinical outcome measures.

For the mild stroke population, these kinematic variables are potentially sensitive to performance impairments that are unobservable, even with a trained therapist eye. Therapists often treat individuals with mild stroke and other mild brain injuries (e. g., post-concussive syndrome) who report difficulty in motor-based tasks that have an added cognitive or balance component. For example, individuals post-stroke who completed dual tasking during walking demonstrated poorer performance than walking alone [126]. Therapists could use these kinematic variables to detect deficits in performance and intervene as appropriate. This is especially important for those with mild stroke who return sooner to work and other community activities [127].

This study has some limitations. First, the sample size is very small; however, this is to be expected given the exploratory nature of the study. The next step is to recruit a larger sample of individuals with both mild and moderate motor impairments post-stroke to determine sensitivity and specificity of the kinematic variables. Second, the Microsoft Kinect® is a fairly robust tracking system; however, it is subject to variations in tracking quality based on the individual’s environment, lighting, and distance from the sensor. This impacts the overall reliability of the data. Additionally, the Microsoft Kinect® is no longer being produced; however currently available depth sensors use similar methods for skeletal tracking and the results shown here have applications regardless of the technology. Lastly, this study explored three new kinematic variables over short time periods. Future research will explore these variables over longer intervention periods.

More rehabilitation games are including a depth sensor as the input/tracking device [128]. It is imperative that these systems include clinically applicable and useful assessments. These kinematic variables within game-based rehabilitation systems using depth sensors become more necessary as telehealth becomes more widespread and insurance companies demand measures of patient progress for reimbursement. The portability of this system and pairing with engaging rehabilitation-specific games adds new avenues for in-home stroke rehabilitation as both a stand-alone and adjunct to existing rehabilitation.

CHAPTER 4 DAILY ACTIVITY RECOGNITION AND ASSESSMENT SYSTEM

There are three main parts of the daily activity recognition and assessment system (DARAS) which are an action data logging system part, action recognition and temporal action localization part and an action assessment part. In the data logging system part, different depth sensors have been investigated to provide an efficient and convenient way to collect daily motion data. Two versions of action recognition algorithms have been explored to provide accurate action recognition results from manually segmented videos to real-life unsegmented video streams. Kinematic assessments are performed for each recognized action using joint data.

4.1 Action logging system

The action logging system of in the DARAS records depth and skeletal data. Depth data are made of pixels that contain the distance from the camera plane to the objects. Skeletal data are a series of entries of 3D Cartesian points, specifying the location of joints in 3D space during recorded time.

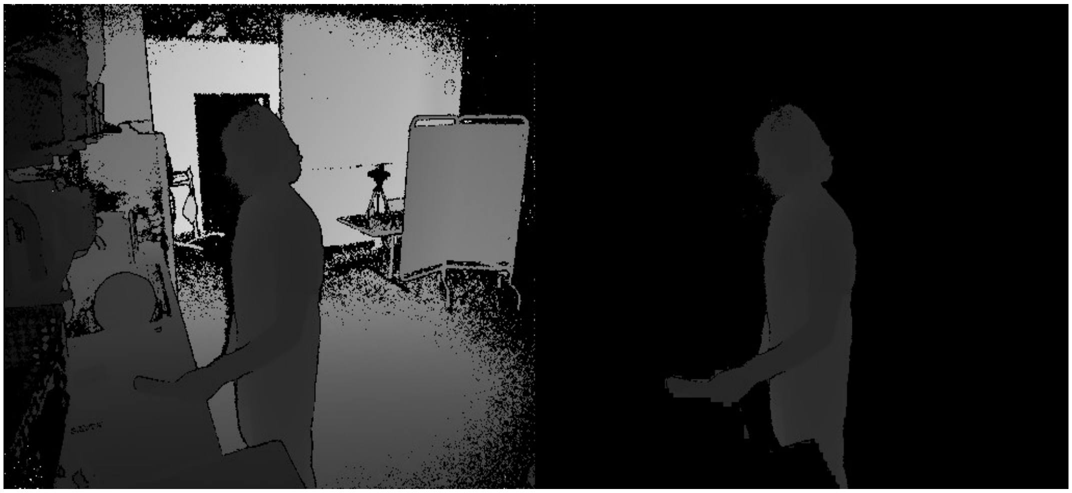


Figure 12. Unprocessed (left) and processed (right) depth data of a subject cutting with a knife.

4.1.1 Kinect-based Window application

Using the Kinect sensor, the logger records unprocessed depth data as well as depth data with a patient segmented out shown in Figure 12 [129]. In addition to depth data, skeletal data are also collected. Each recording has an associated .csv file with x, y, and z coordinates for all the joints tracked by the Kinect. The units are in meters, with the z coordinate encoding depth.

The skeletal data are utilized for assessment of range of motion. With x, y, and z coordinates of all joints, we calculate kinematic measures such as mean velocities, max extension, symmetry of hand movements, and chest sway.

All data collected, processed and unprocessed, are stored locally by a computer hosting the program.

4.1.2 VicoVR-based mobile application

VicoVR sensor [130] has been investigated for collecting depth and skeletal data in the data logger system. Figure 13 shows the module diagram. The main components are a

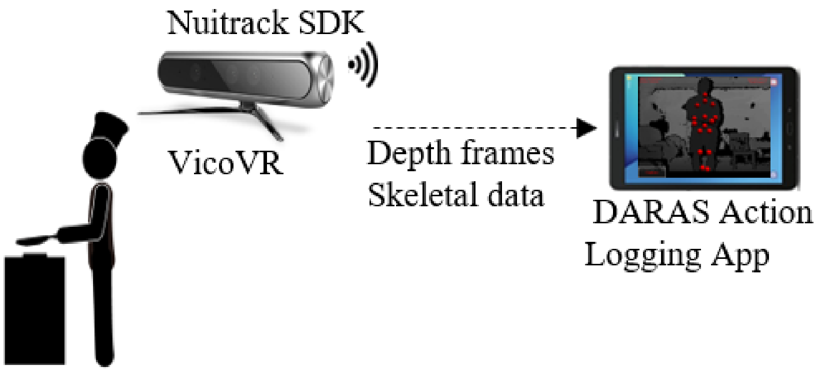


Figure 13. The diagram of action logging module. The VicoVR sensor with Nuitrack SDK recorded the cooking actions in both depth and skeletal data. The developed action logger app on a tablet enabled the data collection, received and stored the data, and visualized the received data in real time.

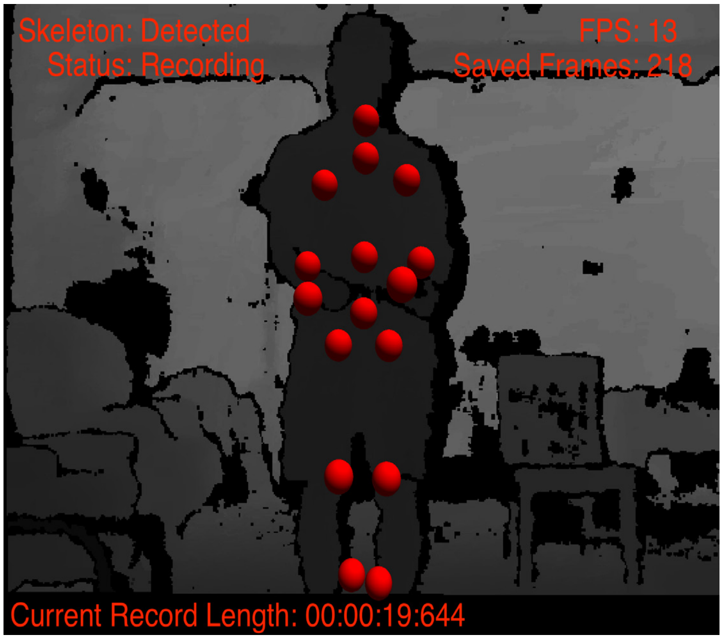
VicoVR sensor and an android-based application. The VicoVR sensor is able to collect depth data continuously. To set up a reliable connection, the VicoVR broadcasts the depth data over a private wifi network (WiFi 802.11n). The data stream includes three-dimensional coordinates of skeletal joints, and a raw depth map with a maximum resolution of 640x480 at 30 frames per second [37]. The Android device connects to the WiFi hotspot and runs a lightweight application built with Unity and the Nuitrack SDK. The application records the depth frames at maximum possible transfer rate. The skeletal joints' three-dimensional positions are recorded in synchronization with each corresponding frame. The recording interface is shown in Figure 14. By saving this data to an external SD card on the android device, skeletal data or depth data can be transferred off site to be used with temporal action localization and assessment.

4.1.2.1 VicoVR sensor

VicoVR is a Wi-Fi and Bluetooth accessory that provides full body and positional tracking to Android and iOS smart devices. With the wireless connection, wires for data transferring can be removed. Nuitrack SDK has been integrated for Unity3D in the design to develop a mobile application to retrieve the data from the VicoVR sensor. The SDK provides up to 19 skeletal joints in 3D coordinates with maximum 30 frames per second [37] to perform full body tracking and gesture control [37]. The built-in depth sensor has an ideal range of .5 to 4.5 meters but can measure depth up to approximately 6.6 meters. With a horizontal field of view of 60 degrees, VicoVR is comparable to a Kinect V2.

4.1.2.2 Android-based daily action observation application

An android-based application has been developed using Unity 3D to automatically enable the data collection when a person is in the view of the VicoVR sensor. We map the



Skeleton: Detected
Status: Recording
FPS: 13
Saved Frames: 218
Current Record Length: 00:00:19:644

Figure 14. The interface of daily action logging App with the real-time depth and skeletal joint data displayed.

view of the sensor to the interface of the application for display. The application also monitors the wireless connection between the VicoVR sensor and an android device. Since tablet PC usually provides a bigger interface compared with smartphones, we choose the Samsung Galaxy Tab S3 to run the application.

The application was designed to be as simple and lightweight as possible. Compared to a traditional setup, the tablet has less storage and slower processing speeds. To compensate this, all data is saved as a one-dimensional byte array compressed into a binary file. Each byte represents a decimal value between 0 and 66536 which corresponds to the distance value of each pixel in the depth map. The data is later converted to a series of .png format images for use with temporal action localization. With each depth frame, a new skeletal data entry is saved as a csv for action assessment. The skeletal joints chosen are head, neck, torso, waist, left/right shoulder, left/right elbow, left/right hand, left/right hip, left/right knee, and left/right ankle. Each joint has a 3D coordinate value as well as a

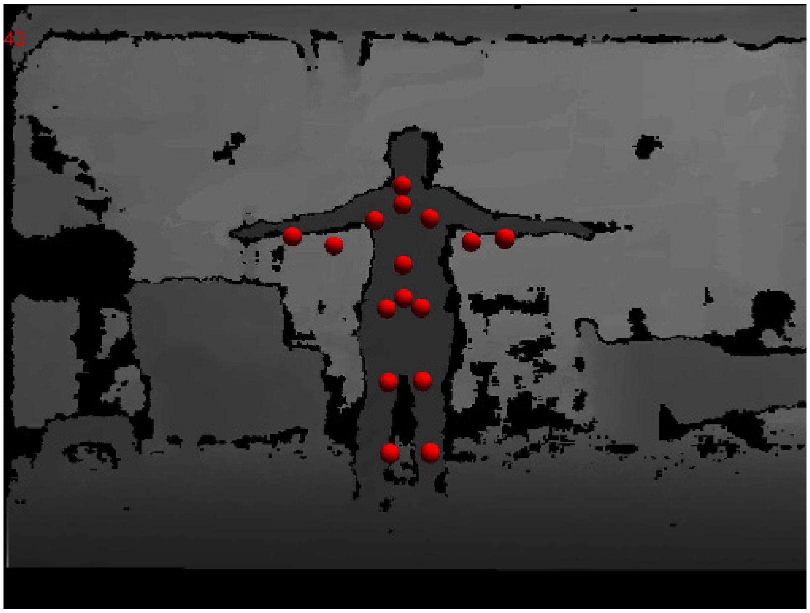


Figure 15. A depth frame and the corresponding skeletal joint data collected from the TVico sensor.

confidence score. The Nuitrack middleware currently only supports values of 0 and .75 which represent not detected and detected joints, respectively. For action assessment, we are able to ignore 0 values.

4.1.3 TVico-based mobile application

TVico [24] is an interactive Android computer with 3D sensor and RGB camera, a product developed jointly by Orbbec and 3DiVi Inc. It tracks up to 19 skeletal joints using Nuitrack Body Tracking SDK. I integrated Nuitrack SDK for Unity3D in the design to develop a mobile application to retrieve the data from the TVico sensor. The recording rate for both skeletal joint data and depth frames are 30 frame per second. The real-time data stream is displayed on the screen shown in Figure 15.

4.1.4 Foresite based system

The Foresite system is a standalone system [25]. The hardware modules including an Astra depth camera, processor, Wi-Fi module, and memory have been integrated into a small box. An optimized action logger application with Astra SDK efficiently collects depth and joint data of a patient and family members and stores the data in either local disk or remote cloud storage. The system is able to perform data collection for 24 hours. When the logger crashes, the operating system automatically restarts the logger application. The status of the data collection and data storage have been monitored by the operating system. Email notifications were sent to a selected researcher. The main functions of the logger system are shown in Figure 16.

4.1.4.1 Hardware overview

The hardware modules of the Foresite system includes a processor, caches, memory an Astra depth sensor, a Wi-Fi module, and a USB drive shown in Figure 17. All the modules are integrated in one small box.

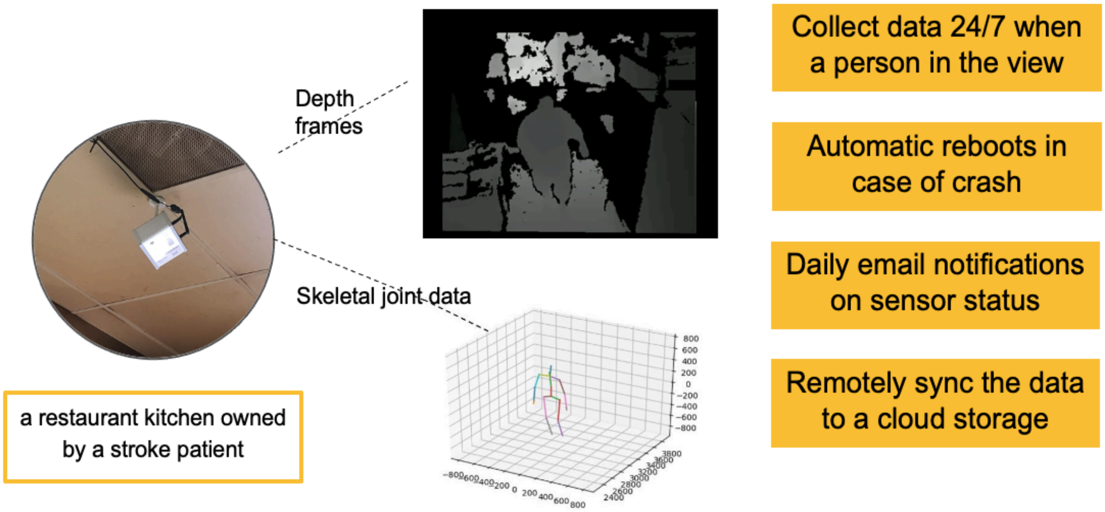

Depth frames
Skeletal joint data
a restaurant kitchen owned by a stroke patient
Collect data 24/7 when a person in the view
Automatic reboots in case of crash
Daily email notifications on sensor status
Remotely sync the data to a cloud storage
800
600
400
200
0
−200
−400
−600
−800
−800 −600 −400 −200 0 200 400 600 800
2400 2600 2800 3000 3200 3400 3600 3800

Figure 16. The main functionalities of the Foresite based action logger system.

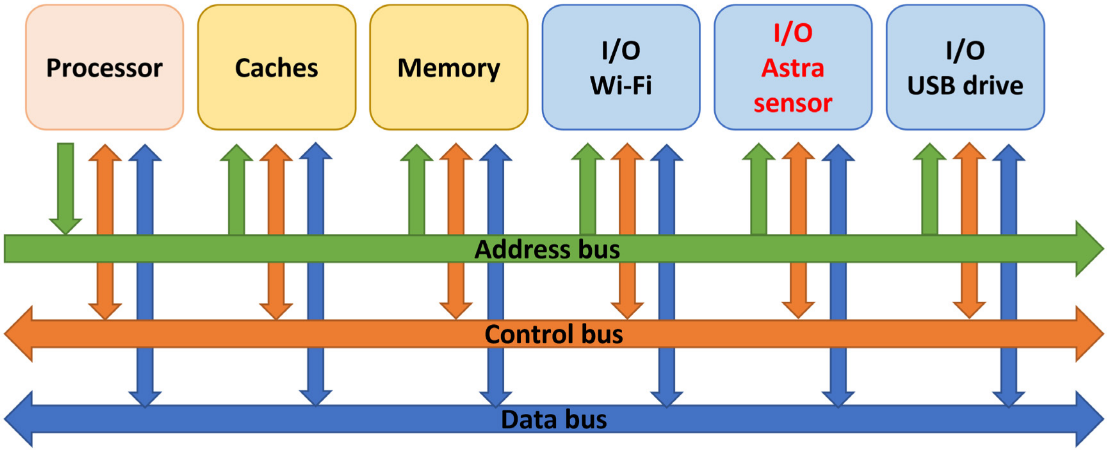


Figure 17. The hardware modules embedded in the Foresite system.

4.1.4.2 Action Logger software

An action logger application has been developed to initialize and utilize the Astra SDK to retrieve the depth data and joint data. The depth data are matrices with the depth values, which represent the distance of the objects in the view to the depth sensor. The joint data are a list of joint locations of the detected persons in the view in three dimensions. After retrieving the data, the logger application converts the raw depth data to a format which can be easily interpreted. Finally, the logger saves the depth data and joint data to a memory space.

The Astra SDK provides low level depth stream and higher-level joint body stream. The depth stream contains depth data from the sensor. The data array included in each DepthFrame contains values in millimeters for each pixel within the sensor's field of view. Body data is computed from the depth data. It contains 2D position and 3D position of 19 joints shown in Figure 18, the mask of users on depth data and floor info. It supports max to 5 people. The Astra SDK provides free access to the depth data and 30-minute access to the joint data. A license is required to get the unlimited access to the joint data.

It shows that it requires at least 6 frame per second (fps) for an action recognition model to provide accurate recognition results. The higher the depth frame sample rate, the

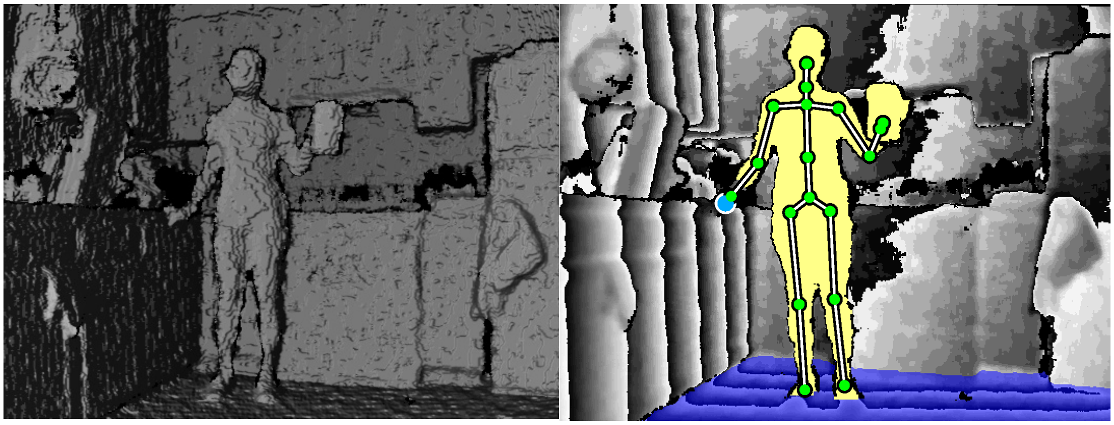


Figure 18. The depth image (left) collected from the Foresite sensor. The corresponding depth image with skeletal joints (right) generated using the Astra SDK.

more details will be provided of an action. Thus, the goal of the action logger application is to efficiently record and store the depth date and joint data in at least 6 frames per second. The sample rate of depth data is about 30 fps. It consumes time to store the depth frame to memory. We need to ensure the saving rate of the depth frame is above 6 fps.

In the initial version of the action logger application. The raw depth data are converted to grey-level pixel values and each frame are converted to image frame in .png using ImageMagic library. Finally, the png frames were saved in the memory space. In this version, the depth data saving rate is about 3 fps.

After tracking the processing time of each function in the initial version, the two most time-consuming processes are converting a pixel matrix to a .png image and saving the images to the memory, because the size of an image frame is large.

To achieve the goal of depth frame collection rate at more than 6 fps, optimizations have been performed in three directions. First, the png image conversion rate is low. It turns out that the png frame generation can be conducted in post-processing. As a result, the time of converting pixel matrices to png frames can be reduced, and the overall time of processing the raw image to the desired format can be reduced. Secondly, the time of writing the depth data to a memory space is proportion to the size of the data. To reduce

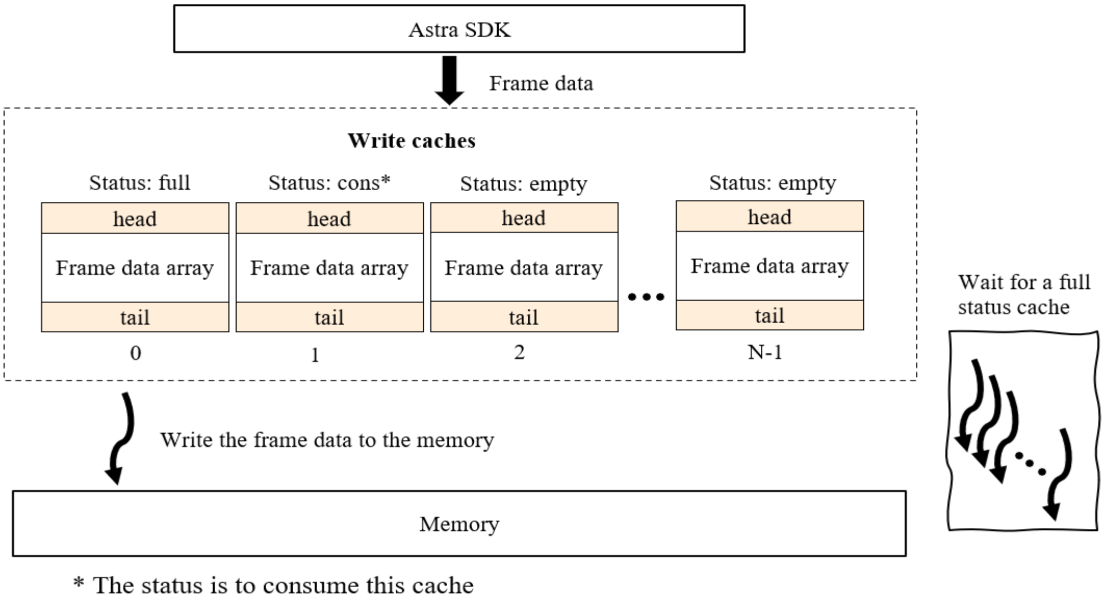
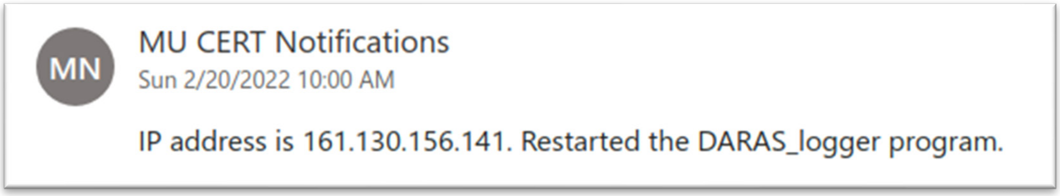
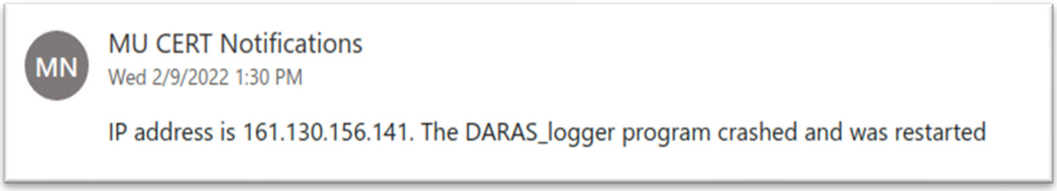


Figure 19. Data collection optimization mechanism. Multiple caches were allocated to store the income data frames. Several threads were created to store the data in batch whenever a cache is full. The frame rate with the optimization mechanism has been increased to 8 fps.

the size of data to be stored, the depth matrices have been compressed. Thirdly, the bottleneck is the process of writing the depth matrices to memory. To optimize the memory writing process, a multi-thread saving mechanism has been implemented shown in Figure 19. Specifically, a customized data structure named frame_data is created to store a depth matrix, and a customized data structure named write_cache is created to save a batch of frame data wrapped with header and tail for validation. The size of the write_cache is set to 30 frames. The status variable is also included in a write_cache to indicate the writing status of this cache. The main process looks for an empty status write_cache and writes the depth matrix into the cache. While writing the cache, the status of that cache has been changed to ‘consume’. Whenever the cache is full, the status of the cache will also be updated to ‘full’. Multiple threads have been created to handle the process of writing cache data to memory. Whenever a cache’s status changes to full, a thread will compress the data in the cache and write the compressed data to memory. The semaphore mechanism has


MU CERT Notifications
Sun 2/20/2022 10:00 AM

IP address is 161.130.156.141. Restarted the DARAS_logger program.


MU CERT Notifications
Wed 2/9/2022 1:30 PM

IP address is 161.130.156.141. The DARAS_logger program crashed and was restarted

been applied to organize the threads. With the optimization, the frame rates of the depth frame and joint data have been increased to at least 8 frames per second (fps).

4.1.4.3 Self-recovery functions

4.1.4.3.1 Restart

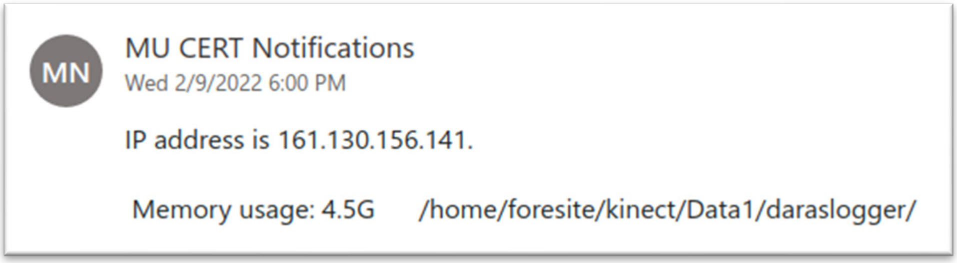
The operating system kills the existing processes of the action logger application. Then, the operating system executes the logger application. At the same time, an email notification will be sent to the assigned email address as shown in Figure 20. The output of the logger application will be echoed to an assigned text file.

4.1.4.3.2 Monitor the logger and fix the crash

The operating system verifies if the logger process is running every ten minutes. If the logger process doesn’t exist, the operating system will execute the logger application. The crash information will be logged in a text file and also be sent to the assigned email address as shown in Figure 21.

4.1.4.3.3 Monitor memory usage

The amount of data that has been collected is monitored. The usage of the memory is checked daily, and the memory status is sent to the assigned email address as shown in Figure 22.



MN

MU CERT Notifications
Wed 2/9/2022 6:00 PM

IP address is 161.130.156.141.

Memory usage: 4.5G /home/foresite/kinect/Data1/daraslogger/

Figure 23. An email notification on memory usage of the system.

```
astra_initialize(); //initialize astra sdk
//to use body tracking, you must set license after initializing
const char* licenseString = "<INSERT LICENSE KEY HERE>";
orbbec_body_tracking_set_license(licenseString);
```

Figure 22. Include the Astra SDK license to the DARAS logger source code.

4.1.4.4 Astra SDK setup and license

The SDK license needs to be added in the logger source shown in Figure 23. The free version only supports 30-minute joint data service.

4.1.5 Comparison of the developed systems

The primary objective of the action logger system is to collect the depth and joint data of a patient. Besides the primary goal, whether the system is portable, easy to install, stable, and efficient in data collection and transmission will also be considered.

The Table 10 below lists the advantages and disadvantages of each developed system using different depth senser.

As we can see from the table, the Kinect system collects depth and joint data with the highest quality, and it is the most stable system. However, using the Kinect system brings the following drawbacks. First of all, the installation requires connecting the Kinect sensor to a computer and several converters and chargers. Second, it can be difficult to find a place to set up the system, since there are several components and multiple wires. As a result, it turns out to be not feasible to use the system for long-term data collection. Though

Table 10. The comparison of the developed action logger system using different depth sensors. The Kinect V2 device, the VicoVR, TVico sensor, and the Foresite system were investigated. The Foresite system is the best choice.

	Kinect	***VicoVR/ TVico***	***Foresite***
Software	Kinect SDK	Nuitrack SDK	Astra SDK
Portability	Need a high-performance computer. A Kinect sensor needs to be charged by multiple adapters.	VicoVR needs to pair with a smart device.	Standalone system.
Installation	A developed application running on a computer.	An app was developed for an Android device.	A web application was developed.
Sample rate	Fast (15/30 fps)	Slow (~6 fps)	~10 fps with multi-thread optimization
Stability	Stable	The Bluetooth connection might get lost	Stable
Storage	Local	local	Local/Remote
Monitoring	None	None	Email notification and auto-recovering features

the VicoVR system with the Nuitrack SDK generates depth data and joint data, the Bluetooth connection is not stable.

Compared with the other systems, the Foresite system with Astra SDK meets the primary goal, of generating depth and joint data. The Foresite system is standalone, which is easy to install. A web-based application has been developed for wirelessly adjusting the view of the camera, which makes the installation easier. Developed functions for monitoring the status of data collection, data transmission, and memory usage have been created for the system. As a result, the Foresite system is the best choice.

4.2 Data collection

4.2.1 View experiment and installation strategy

Preliminary experiments were conducted to test the capability of continuously collecting data for three months. In addition, we tested various sensor placements to

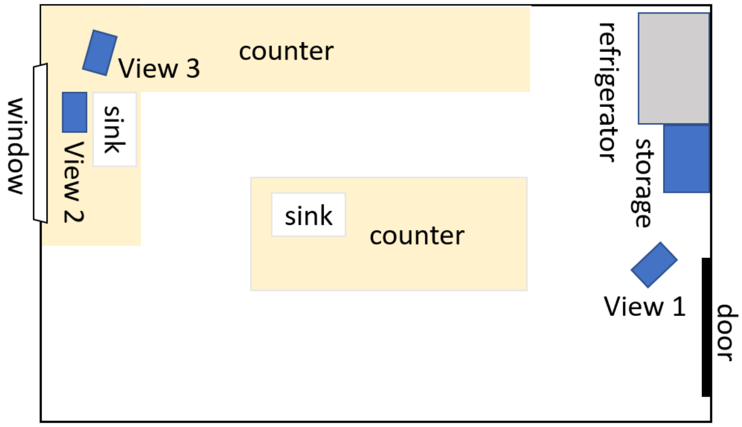


Figure 24. The kitchen layout of the view experiment.

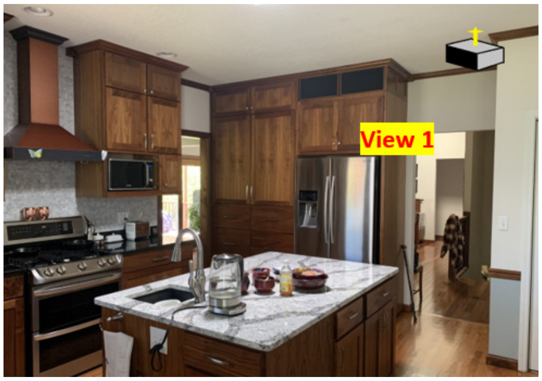
maximize the number of frames with low noisy and good views of a participant. Within each setup we explored the height of the device, the view angle of the device, full-body view or half-body view, and the distance between the device to a participant.

Before investigating the data quality in different views, the system was installed in a healthy participant's home for three months to ensure that the system can collect data for a long period of time. The system ran as expected, which collected depth and joint data of the participant and their family members. The system was able to recover automatically when it crashed. Developers received daily notifications of memory usage and action logger status.

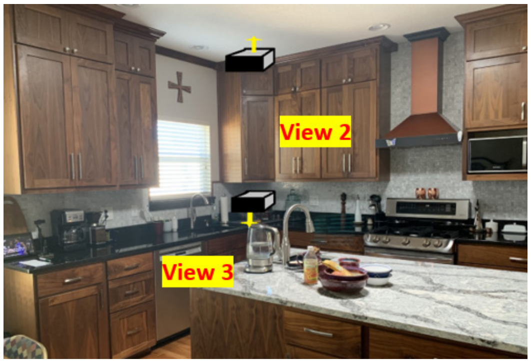
After verifying the stability of the system, a view experiment was conducted in a healthy participant’s kitchen. The layout of the kitchen is displayed in Figure 24. Three different views were investigated marked in Figure 24 and Figure 26. The pictures of the kitchen were shown in Figure 25.

The description of the views is listed below.

a. View 1. The system was installed on the ceiling near the entry door of the kitchen. The participants’ depth and joint data were captured in a diagonal view. The system had the furthest distance to the participants, about 5 meters.



(a)



(b)

Figure 25. The pictures of the kitchen where the view experiments were conducted. (a) A system was installed on the ceiling near the entry to record data from View 1. (b) A system was installed on the ceiling in the corner of the kitchen to record data from View 2. A system was placed on the countertop to record data from View 3.

b. View 2. The system was installed on the ceiling in a corner of the kitchen. It captured the participant's data from a side view. The distance between the system and the participants was about 3 meters.

c. View 3. The system was placed on the countertop, and the location was close to the second view.

The depth data and joint data were compared among three views shown in Table 11.

After analyzing the quality of the data collected from different views, we found that a side view or diagonal view with about a three-meter distance provide less noisy joint data and more full-body view frames.

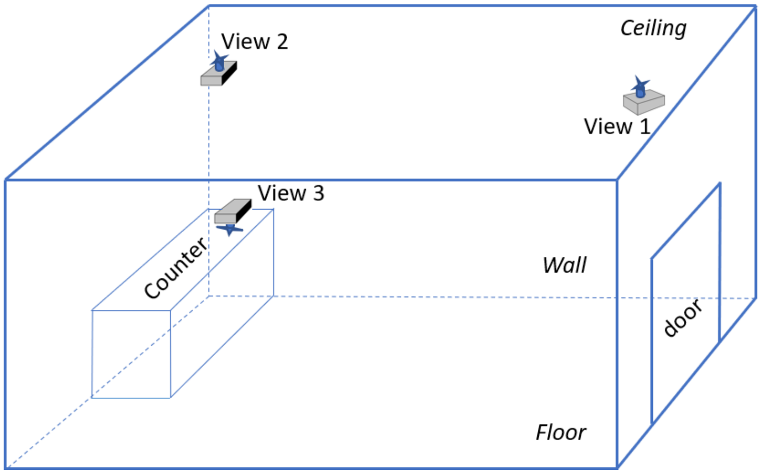
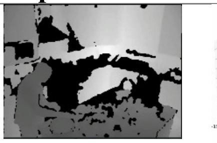
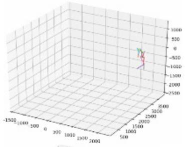
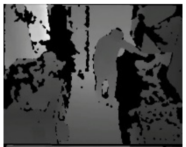
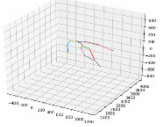
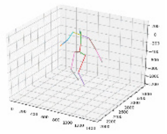


Figure 26. The 3D visualization of the kitchen with installed systems in the kitchen.

4.2.2 Stroke participant data collection

To develop and refine an action recognition algorithm for detecting specific functional activities in the home setting of stroke patients, in-home activity data in the kitchen area were collected. The IRB protocol is included in Appendix 9.1. Potential subjects from the MU Stroke Registry will be contacted. The inclusion and exclusion criteria are listed below:

Table 11. The details of three views for investigating the optimal data collection view.

	Height	Direction	Distance	Depth data	Joint data
VIEW 1	On the ceiling	Diagonal view	About 5m		
VIEW 2	On the ceiling	Side view	About 3m		
VIEW 3	Under the cabinet	Side view	About 3m		

- Inclusion criteria:
    a. Over the age of 18
    b. Conversational in English
    c. Able to ambulate with or without an assistive device
    d. At least mild hemiparesis in the arm (NIH Stroke Scale score > 6)
- Exclusion Criteria:
    a. Under the age of 18
    b. Unable to converse in English

A Foresite depth sensor system was installed in one participant's kitchen and recorded data over the course of 3 months. The participant will be asked to not deviate from their normal activities within the home. Depth and skeletal data will be transmitted from the Foresite sensor to the Foresite managed secure server via the participant's Wi-Fi/high speed internet. For homes where the internet connection is not available for data transmission, the data were stored in a USB drive locally.

Totally, 16 stroke subject were participated in the study. One week of subject data provides approximately 10GB of data for algorithm training. Over a 3-month period, each subject will provide approximately 120GB of data for training and testing the algorithm.

The demographic information and device information of the participants were included in the appendix 9.2.

4.3 Datasets

4.3.1 Single action dataset

A single action dataset was collected to mimic the daily cooking actions in a simulated kitchen environment in the Occupational Therapy department, where the counter and appliances were on one wall shown in Figure 27. To address a lack of forward-facing data,

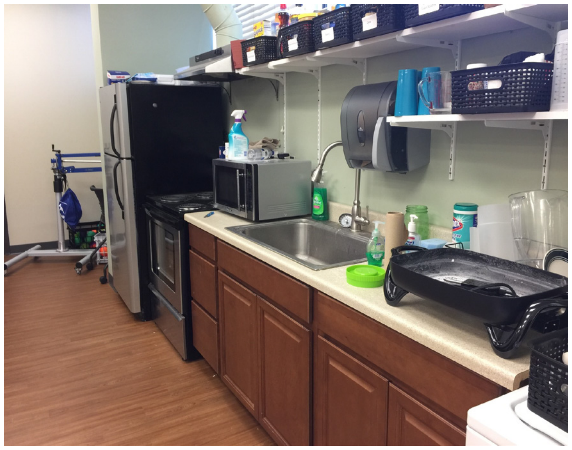


Figure 27. A simulated kitchen environment in the Occupational Therapy department.

our Kinect was placed at a side angle. As a result, our kitchen environment was highly susceptible to self-occlusion.

The dataset includes 9 different healthy students as subjects, 2 recordings of each action, and 28 different actions that correspond to general kitchen tasks. This amounts to 504 different image sequences. The 28 actions can be considered as 5 categories which are washing, meal preparation, kitchen gadget manipulation, general picking tasks, and walking. In the washing tasks, the participants were asked to put an item into the sink (WashSink), wash and rinse a dish (WashRinse) and place an item into the dishwasher (WashDishwasher). In the meal preparation tasks, the participants cut (PrepCut) and stirred (PrepStir) food. They were also asked to open (PrepOpen) and close (PrepClose) a container. In the manipulation tasks, participants performed actions including using the stove (ManipulateStove), microwave (ManipulateMicrowave) and refrigerator (ManipulateFridge), and turning on and off the sink faucet (ManipulateSinkOn/Off). In general picking tasks, the participants were asked to pick up and place an object onto a counter or a cabinet (PickUpCounter, PutDownCounter, PickUpTop, PutDownTop, PickUpBottom, PutDownBottom). They also opened and closed a cabinet (OpenBottomCabinet, CloseBottomCabinet, OpenTopCabinet, CloseTopCabinet). In the walking tasks, the participants walked in the kitchen either holding an object or without holding an object.

4.3.2 Simulated kitchen dataset

The video clips in the single action dataset only contains one action, which means that the actions have been potentially segmented. In reality, a sequence of actions will be recorded in a video clip. To make it possible to investigate temporal action localization to segment the clinically relevant actions out, the cooking scenario datasets was collected.

Simulated kitchen dataset. We collected some kitchen action videos in a simulated kitchen environment at the Occupational Therapy department. The VicoVR sensor was selected to record kitchen actions. The sensor was placed close to the corner of the wall to capture the full view of the kitchen. Eleven healthy students were recruited to perform three pre-designed action scenarios at least three times. In total, 100 continuous, untrimmed action video sequences were logged. Three scenarios of continuous actions were designed and described following:

- Scenario 1: Walk into the kitchen carrying the gallon of milk and put it in the fridge. Get out the peanut butter and jelly from the overhead cabinet. Get out the knife from the drawer. Get out the cutting board from the cabinet below. Walk out of the kitchen.

- Scenario 2: Walk into the kitchen. Get out the pasta from the overhead cabinet. Get out the strainer from the cabinet below. Rinse off the strainer in the sink and put it on the counter. Use the towel to dry it. Walk out of the kitchen.
- Scenario 3: Walk into the kitchen. You notice that someone has spilled some cereal on the floor! Get the broom and dustpan and sweep it up. Carry the swept-up cereal to the trashcan. Come back into the kitchen and sit in the chair.

Based on the input from occupational therapies, each frame has been labeled to one of these 8 action categories which are walking, sitting, reaching above the head, reaching forward, reaching below the waist, hand manipulation, sweeping and none of the above.

4.3.3 Stroke dataset

The cooking scenario dataset includes sequences of cooking activities from young participants. The aim of developing DARAS is to recognize the clinically relevant actions of stroke patients. There is no daily action dataset containing stroke subjects. As a result, sixteen stroke participants were recruited in our IRB-approved data collection. Three-month daily activity data were collected from each participant in realistic cooking environments.

The collected data were firstly post-processed from the raw binary format to the png format for depth data and the CSV format for joint data. To train the action recognition and temporal localization algorithm, the per-frame labels were provided manually. To assess the motions of each participant, subject identification labels were provided for each action segment.

The stroke dataset was formed by the data from six participants, which are participants 1, 2, 3, 4, 5, and 10.

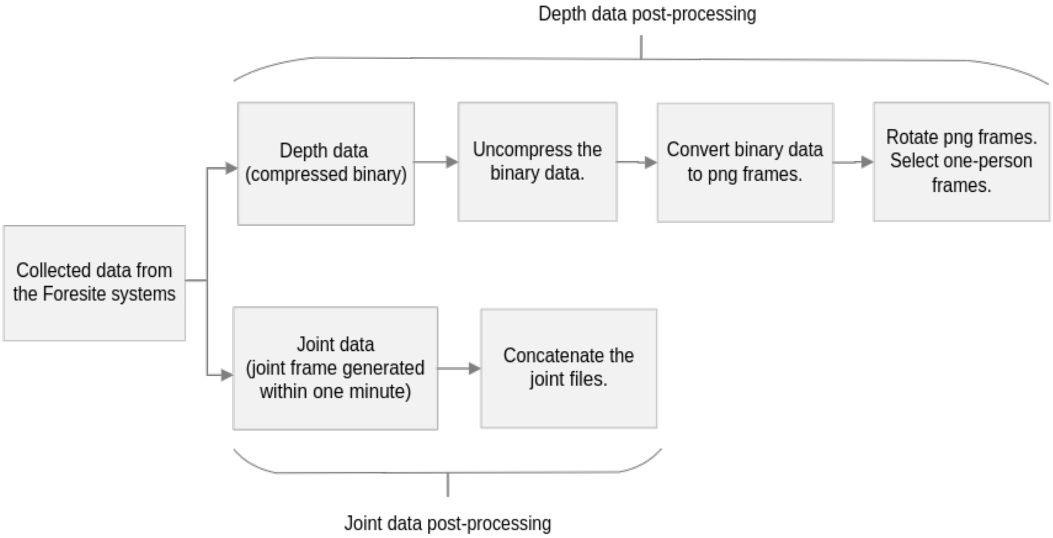

Depth data post-processing
Depth data (compressed binary)
Uncompress the binary data.
Convert binary data to png frames.
Rotate png frames. Select one-person frames.
Collected data from the Foresite systems
Joint data (joint frame generated within one minute)
Concatenate the joint files.
Joint data post-processing

Figure 28. The procedure of post-processing on the collected data.

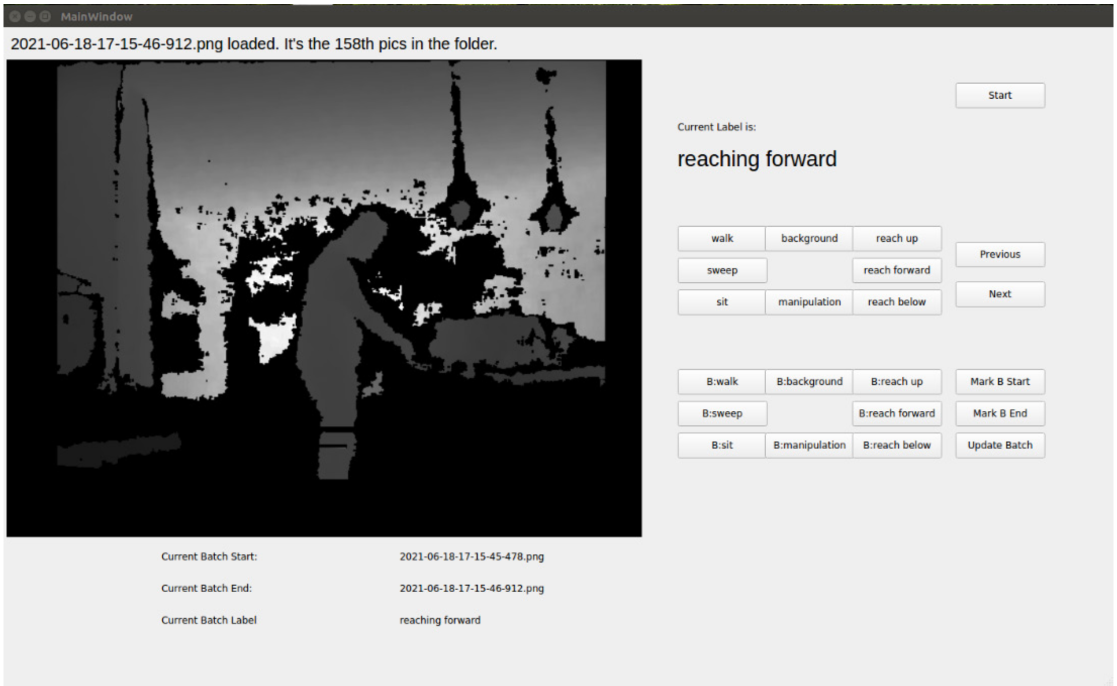
4.3.3.1 Post-processing

To optimize the data saving process, the Foresite system stores the compressed binary depth data and the joint data in CSV format. Each CSV file contains the joint frames generated within a minute. The compressed binary data and joint data require a further process to be used for action recognition and assessment.

The procedure of the post-processing is shown in Figure 28. For depth data, the binary data were uncompressed and converted to the .png frames. The .png frames were rotated. The depth frames were filtered to ensure no more than one person in the view. For joint data, all the joint data have been concatenated into one file for each participant.

4.3.3.2 Data labeling and clinically relevant actions

In order to train and test the action recognition model, a dataset with per-frame level ground-truth is necessary. Each frame was labeled to one of the clinically relevant action categories: are walking, sitting, reaching above the head, reaching forward, reaching below the waist, hand manipulation, sweeping and none of the above. A data labeling tool is developed shown in Figure 29 to label the depth frames in batch.



2021-06-18-17-15-46-912.png loaded. It's the 158th pics in the folder.

Current Label is:

reaching forward

Current Batch Start: 2021-06-18-17-15-45-478.png

Current Batch End: 2021-06-18-17-15-46-912.png

Current Batch Label reaching forward

Figure 29. The interface of the data labeling toolbox.

The label tool shows the current data frame and the corresponding label. Two options are provided in the tool. One option is to change the label of the current frame. Another option is to label a batch of frames by selecting the start frame and the end frame.

In order to perform assessments for each individual, the subject identification labels are necessary for each action segment. After the per-frame action labels were reviewed, the ground truth action segments were identified by grouping the same consecutive per-frame labels together. The subject identification label was provided manually for each action segment.

4.3.3.3 Data sets and EDA

The labeled data were split into different sets. Half of the whole set was used as a training and testing set. Another half was used as a validation and assessment set. To make the two sets have a similar distribution. The data were first grouped by date as small buckets.

The data in each bucket were split in half. One for training and testing, and the other for validation and assessment.

The exploratory data analysis (EDA) was conducted to understand the training and test dataset and to perform the necessary data cleaning.

Table 12. Statistical analysis of action segments.

Action	*Count*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
None of the above	522	119	612	**1**	5	10	37	**7503**
Walk	261	12	10	**1**	4	9	17	59
Reach Up	19	15	11	**2**	5.5	16	24.5	35
Reach forward	258	12	12	**1**	7	10	15	137
Reach below	47	34	28	5	14	26	45.5	121
Manipulation	161	52	72	**1**	9	24	65	421

Each frame in the dataset has been labeled to one of the clinically relevant actions or none of the above. Action segments can be located by grouping the frames with the same labels. The length of segments for each action category was analyzed. The results are shown in Table 12. The number of segments, average length, and the maximum length of each action category are listed in the table. For example, there are 522 segments that are none of the above segments. The average length among those actions is 119 and the maximum length is 7503.

Short action segments whose length is less than 5 frames need to be removed since temporal information is limited from the short segments. The long none-of-the-above segments need to be removed because the long none-of-the-above segments make the action segments unbalanced. When analyzing the total number of frames of each action category, the proportion of some action categories is much larger than the rest, which

Table 13. The summary of the pre-processed training and testing set. The total number of frames of each action category for each participant were counted.

	P1	P2	P3	P4	P5	P10
None of the above	4000	4000	4000	1000	6000	6182
Walk	5000	3122	5000	378	6171	8885
Reach up	5840	297	8479	430	612	366
Reach forward	5000	3313	5000	1091	6121	4080
Reach below	1171	1625	4926	773	3060	3516
Manipulation	4000	4000	4000	701	6000	1395
	25011	16357	31405	4373	27964	24424

makes the dataset unbalanced. As a result, the following three data-cleaning procedures were performed:

a. Any segment whose frame length is less than 5 frames was removed.

b. Find all the long none-of-the-above segments whose lengths are longer than 150 frames. Keep the first 50 frames and remove the rest.

c. To balance the number of frames among action categories, the total number of frames of action categories that have large proportions has been reduced.

The summary of the pre-processed training and testing set was shown in Table 13.

4.4 Action recognition on manually segmented depth videos

4.4.1 HON4D descriptor for action recognition

It is important to have a rigorous global descriptor of a sequence of depth images so that actions that resemble each other are distinguishable. Many kitchen actions have an opposite, and only differ in the temporal sense by reversing the movement, such as opening and closing various items. By including temporal information, a histogram of normal 4d algorithm (HON4D) can be constructed [82]. We implemented a novel C# version that uses the first half of creating a HON4D descriptor, without adding projectors to the already-calculated histogram [82].

The first step in the methodology is to calculate normals for every pixel in a given set of depth images $I = \{i_1, i_2, \cdots, i_k\}$, as in Figure 30. The components of the normals are the changes in depth, which is summarized by

$$\boldsymbol{n} = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial z}{\partial t}, -1\right). \tag{11}$$

Once all normals are computed, they are normalized since only the direction matters for bin contribution. A polychoron is then initialized uniformly in 4D space, where the vertices are treated as vectors and called projectors to be used as the bins of the histogram [82]. The contributions are calculated as the dot product of every normal and projector, as in Figure 31. After every dot product is calculated and added to the proper bin of the histogram, it is normalized.

To enhance the uniqueness of a HON4D descriptor, it is essential to subdivide the sequence of images into cells [82]. The Kinect sensor records depth information at a resolution of 512 × 424. Our cells are 4 × 4 × 3 (w × h × d). As normals are computed, they are placed into their proper cells. Once all normals are placed, a separate HON4D is calculated for each cell. Afterwards, they are concatenated and normalized. This produces a 120-bin histogram.

```
procedure CalculateNormals(images)
    for k = 0; k < images.Count − 1; k++ do
        img1 ← images[k]
        img2 ← images[k+1]
        for x = 0; x < img1.Width; x++ do
            for y = 0; y < img1.Height; y++ do
                currentPixel = img1.GetPixel(x, y)
                nextPixel = img2.GetPixel(x, y)
                rightPixel = img1.GetPixel(x + 1, y)
                leftPixel = img1.GetPixel(x − 1, y)
                upPixel = img1.GetPixel(x, y − 1)
                downPixel = img1.GetPixel(x, y + 1)
                x = rightPixel - leftPixel
                y = downPixel - upPixel
                z = currentPixel - nextPixel
                normalList.Add(x, y, z, −1)
            end for
        end for
    end for
    return normalList
end procedure
```

Figure 30. Pseudocode to generate a list of oriented 4D normals for a sequence of images.

4.5 Temporal action localization on untrimmed depth videos

The depth videos and skeletal joint data of daily activities were collected by using the action logging system. Specifically, the collected data contains many continuous untrimmed action sequences. For example, the “taking a drinking” video includes the actions of fetching a cup, pouring the water, drinking and putting back the cup continuously in a sequence. According to the need from clinicians, performing qualitative assessments on clinically relevant actions, such as the arm reaching above the head, is more desired. To

```
procedure CreateHON4D(proj, normList, hon4d)
    for k = 0; k < proj.Count; k++ do
        for n = 0; n < normList.Count; n++ do
            hon4d[k]  +=  max(0,  dotP(proj[k],norm- List[n]))
        end for
    end for
    return hon4d
end procedure
```

Figure 31. Pseudocode to generate a histogram of oriented 4D normals, where proj is the list of projectors, normalList is the list of normals calculated from Figure 28 and hon4d is the histogram.

perform such assessments, it is necessary to recognize the specified actions and locate these actions from the untrimmed videos. Such process is often referred to as temporal action localization [131]. An ensemble network was proposed and implemented to predict action at the frame level. Thus, the recognized actions can be segmented based on the per-frame action labels.

4.5.1 Ensemble network architecture

It is a challenging task to recognize and localize the clinically relevant actions from a realistic environment. Three networks that outperformed in the RGB dataset were selected and customized to depth-based collected datasets. To have a more accurate prediction outcome, an ensemble network [132] has been proposed to fully utilize the collected data in different data types. The network includes three networks, which are the 3D convolutional-de-convolutional network, Region Convolutional 3D Network, and Region Hierarchical Co-occurrence Network. The architecture of the proposed ensemble network is shown Figure 32.

4.5.2 Convolutional-De-Convolutional (CDC) network

4.5.2.1 3D Convolution and CDC filter

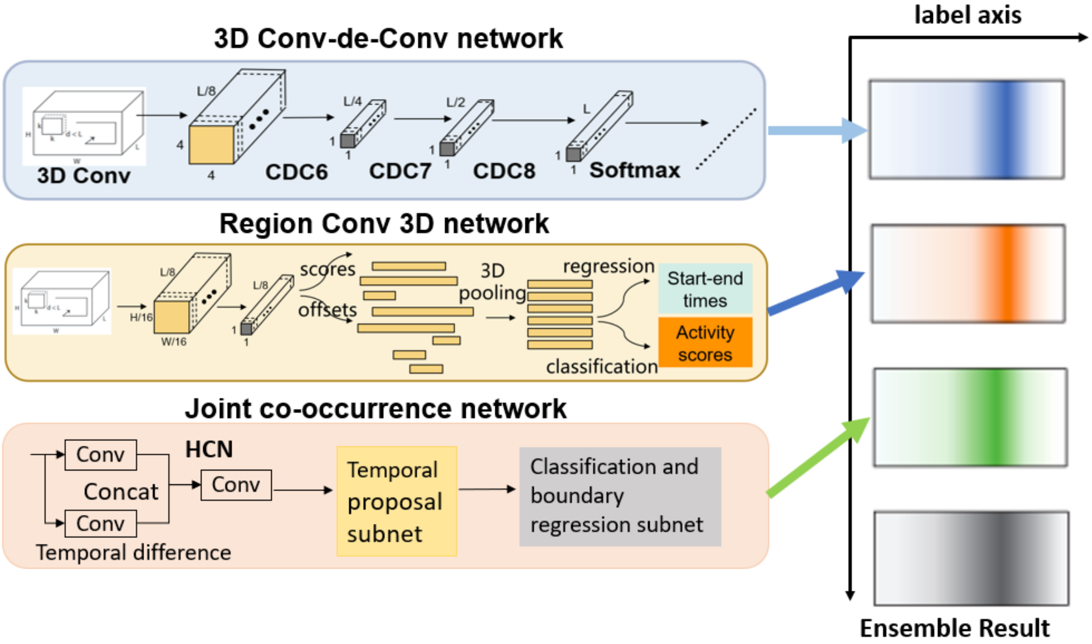


Figure 32. The architecture of the ensemble network.

Shou et al. [133] proposed a Convolutional-De-Convolutional (CDC) network which places CDC filters on top of 3D ConvNets. The CDC network performs spatial down-sampling to extract the action semantics and temporal up-sampling to preserve the time information for each frame. Thus, it provides the prediction score at each frame, which can be used to locate the actions.

Convolution neural networks (CNN), where the dimension of the convolution kernel is two-dimensional, have been widely used in image classification, detection, segmentation and other tasks. For video analysis, the temporal features need to be preserved. However, 2D convolution cannot capture the timing information very well. So, 3D convolution neural networks (3D CNN), which consists of 3D ConvNets followed by three fully connected layers, were proposed in [134]. Although the 3D CNN can learn the advanced semantic

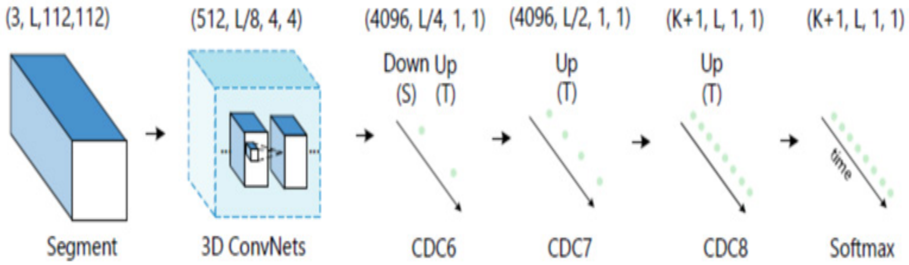


Figure 33. Architecture of a typical CDC network.

abstraction of time and space, the output of video time sequence length is decreased 8 times. Thus, the fine-grained time has been lost.

For timing location problems, the timing output should be consistent with the input video, but the output size should be reduced to 1*1. Motivated by pixel level semantic segmentation, Shou et al. [133] proposed a CDC filter that generates two 1*1 points for each input feature map. As a result, the filter performs convolution in space (for semantic abstraction) and de-convolution in time (for frame level resolution) simultaneously. It is unique in jointly modeling the spatial-temporal interactions between summarizing high-level semantics in space and inferring fine-grained action dynamics in time.

4.5.2.2 Network architecture

Figure 33 shows the CDC network. The input video segment is 112*112*L, a continuous L frame 112*112 image. After C3D network, L is sampled down to L/8 in the time domain, and the image size in space is sampled from 112*112 to 4*4. Then the time domain is sampled up to L/4 in CDC6 and the image size is continuously down sampled to 1*1 in the spatial domain. The time domain is sampled up to L/2 in the CDC7. Next in CDC8, the time domain is sampled up to L, and 4096*K+1 is used in the full connection layer, where K is the number of classes. The last layer is the softmax layer. The final output is (K+1, L, 1, 1), where K+1 stands for K action categories plus the background class.

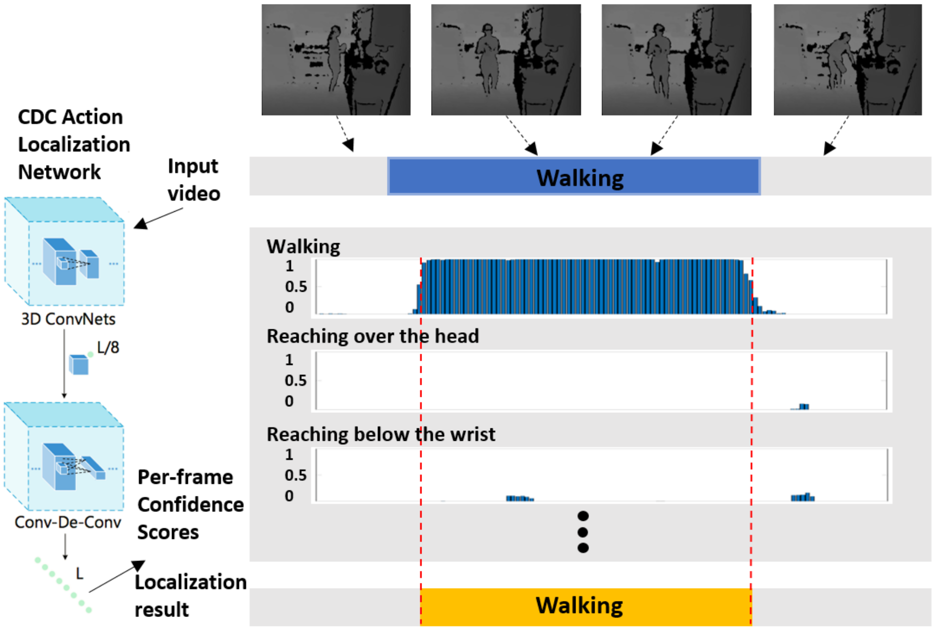


Figure 34. A framework for positioning of temporal action recognition and localization.

4.5.2.3 CDC networks on collected depth-based datasets

The CDC network has been evaluated using THUMOS' 14, an untrimmed RGB sport action video dataset. The evaluation results show that the model outperforms state-of-the-art methods in video per-frame action labeling. Due to the privacy requirement, a network that can perform temporal action localization on depth kitchen action videos is desired in DARAS. However, the proposed CDC network was designed for RGB videos. So, we first adopted the CDC network for depth videos and then fine-tuned the network using a new collected depth video dataset.

Given a piece of untrimmed depth video as shown in Figure 34, it is input into the CDC network, in which the 3D convolution neural network is used to extract semantics, and the CDC network is used to predict the dense frame number level scores. Since a depth image only has one grey channel compared to an RGB image, the input of the network is adjusted for depth videos. The time boundary of action instances is identified by grouping the same labels of frames.

4.5.3 Region Convolutional 3D network

Region Convolutional 3D Network (R-C3D) [91] recognizes and detects actions from untrimmed continuous videos. The key innovations include effectively extracting spatio-temporal features using the 3D ConvNet [134] and extending the 2D RoI pooling in Faster R-CNN to 3d RoI pooling to extract features from proposals with various lengths.

4.5.3.1 Network architecture

The R-C3D Network consists of three components: a shared 3d ConvNet feature extractor, a temporal proposal subnet, and an activity classification and refinement subnet.

4.5.3.2 3D Convolutional feature extractor

Both spatial and temporal features are essential for representing action sequences. To make the action recognition and localization accurate, it is important to extract meaningful spatio-temporal features. A 3D ConvNet encodes rich spatio-temporal features by extending the 2d convolutional layer to 3d, the temporal information has also been preserved while learning the spatial information. Specifically, the features are learned from the convolutional layers (conv1a to conv5b) of C3D. The conv5d activations have been used as the input to the temporal proposal subnet.

4.5.3.3 Temporal proposal subnet

The potential action segments are initially by anchor segments. The anchor segments are the segments that are uniformly distributed throught the input with different pre-defined scales. A 3D convolutional filter is used on the top of the conv5d to extend the temporal receptive field for the anchor segments. Then a 3D max-pooling filter is used to down-sample the spatial feature to produce a temporal-only feature map. The output of the 3D max-pooling layer has been used as the feature vector to predict a relative offset to the central location and the length of each anchor segment by adding two convolutional layers.

4.5.3.4 Activity classification and refinement subnet

There are three main functions: 1) selecting proposal segments using a greedy Mon-maximum Suppress strategy to eliminate highly overlapping and low confidence proposals 2) extracting fixed-size features from selected proposals using 3D region of interest pooling and 3) activity classification and boundary regression for the selected proposal using the pooled features from a series of two fully connected layers.

4.5.3.5 Optimization

Both the classification and regression subnets are optimized by the objective function.

$$Loss = \frac{1}{N_{cls}} \sum_i L_{cls}(a_i, a_i^*) + \lambda \frac{1}{N_{reg}} \sum_i a_i^* \, L_{reg}(t_i, t_i^*) \qquad (12)$$

The softmax loss function is used for classification and the smooth L1 loss function was used for regression. The notation is explained in the Table 14.

The window regression and coordinate transformation are calculated using equation below:

$$t_x = \frac{x - x_a}{w_a}, t_w = \log\left(\frac{w}{w_a}\right) \qquad (13)$$

$$t_x^* = \frac{x^* - x_a}{w_a}, t_w^* = \log\left(w^*/w_a\right) \qquad (14)$$

where x is the predicted window, x_a is the anchor window, and x* is the ground truth window.

Table 14. The notation of the optimization function where the i is the anchor/proposal segments index in a batch, and lambda is the trade-off parameter.

Classification	Regression
N_{cls}: number of batches	N_{reg}: number of anchor/proposal segments
a_i: the predicted probability of the proposal or activity	$t_i=\{t_x,t_w\}$: predicted relative offset to anchor segments or proposals
a_i^*: the ground truth. (1 if the anchor is positive, and 0 if the anchor is negative.)	$t_i^*=\{t_x^*,t_w^*\}$: the coordinate transformation of ground truth segments to anchor segments or proposals

4.5.4 Region Hierarchical Co-occurrence network

4.5.4.1 Skeleton motion

Temporal features are important for recognizing the underlying actions. The temporal representation of skeleton motion was computed and explicitly fed into the network.

For the skeleton of a person in frame t, it is formulated as $S^t = \{J_1^t, J_2^t, \ldots, J_N^t\}$ where N is the number of joint and $J = (x, y, z)$ is a 3D joint coordinate. The skeleton motion is defined as the temporal difference of each joint between two consecutive frames:

$$M^t = S^{t+1} - S^t = \{J_1^{t+1} - J_1^t,\ J_2^{t+1} - J_2^t, \ldots, J_N^{t+1} - J_N^t\} \qquad (15)$$

4.5.4.2 Joint feature extractor

The hierarchical co-occurrence network [135, 136] as shown in Figure 35 was employed to learn the joint co-occurrences and the temporal features jointly. The inputs are a skeleton sequence X with dimension T*N*D and its corresponding skeleton motion with the same shape as X. They are fed into two point-level learning layers, since the kernel sizes along the joint dimension are forced to 1 to learn the point-level representation of each joint. The transform layer switches the joint dimension with the coordinate dimension. The global co-occurrence features from all joints were extracted, and the features from the joint and motion are concatenated. Finally, the feature maps are flattened, and the further features are extracted by two fully connected layers.

4.5.4.3 Network architecture

The region hierarchical co-occurrence network as shown in Figure 36 extracts the spatial-temporal features from the input joint sequence. The temporal proposal subnet and

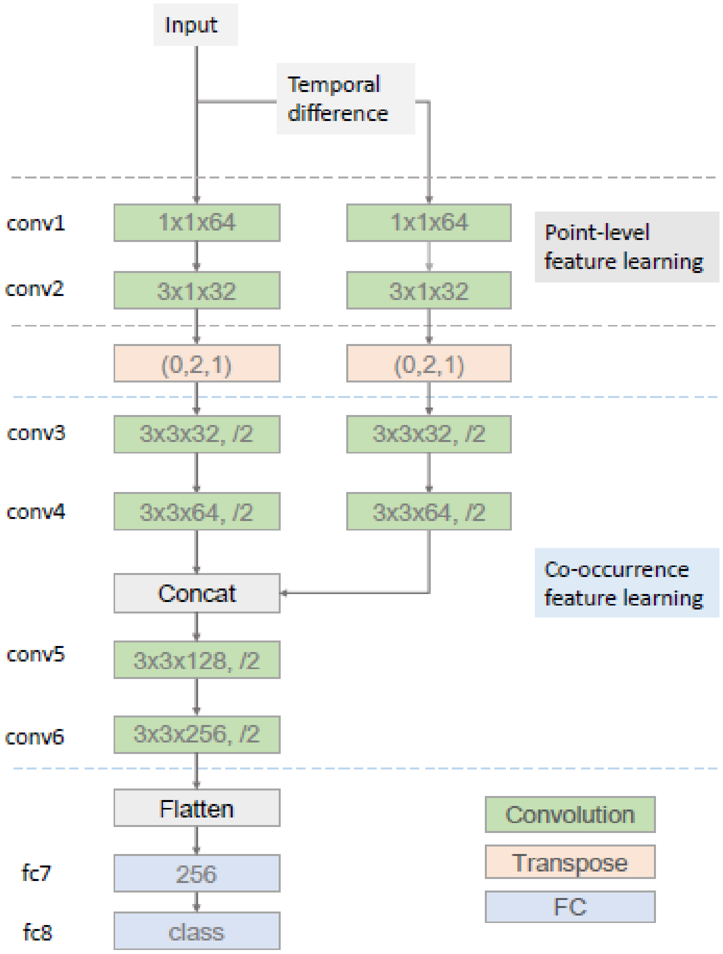


Figure 35. The architecture of the hierarchical co-occurrence network.

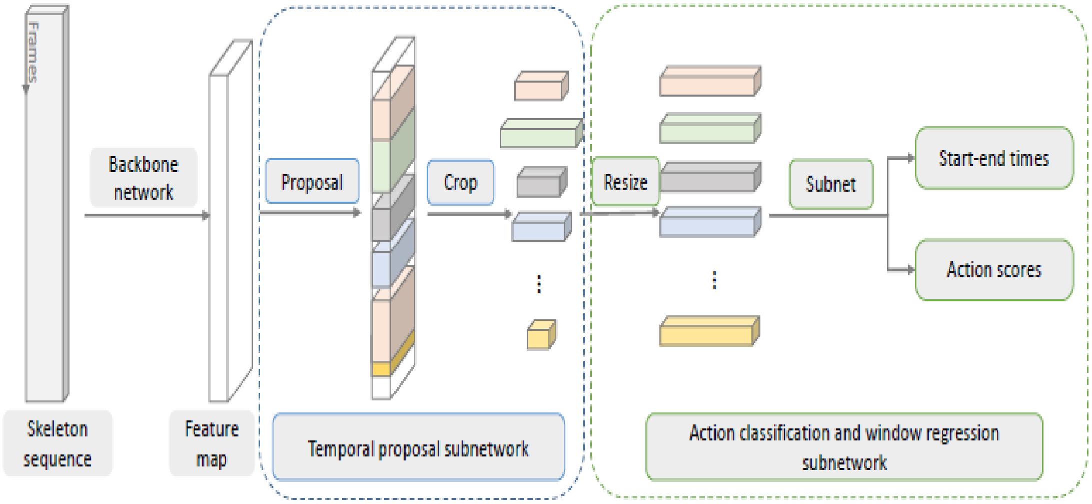


Figure 36. The architecture of the proposed region hierarchical co-occurrence network.

the action classification subnet used in the R-C3D network were employed to perform action recognition and detection.

4.5.5 Ensemble network action recognition and detection

The goal of action recognition and detection system is to recognize the clinically relevant actions and segment the detected action out. Specifically, the per-frame action label needs to be generated. The ensemble network consists of three networks. All of them output the start frame, end frame and predicted label of a detected segment. The per-frame label can be easily generated using the outputs of the networks. As a result, for each frame, three labels were given by the three networks in the ensemble network.

The final per-frame label was fused using the labels predicted by the three separate networks. If two of the networks vote the same action for a frame, the corresponding label of that frame was assigned to an action label with most vote counts. Otherwise, the frame was considered as none of the above.

4.6 Action recognition results

4.6.1 Single action recognition

Descriptor HON4D extracted a histogram of 3D and time information from a sequence of depth frames for each action segment. An SVM with a quadratic kernel was then chosen to classify actions using histograms. Since DARAS was made to work in a challenging home environment, some data can be confused with one another. Prep Cut, Prep Open, and Prep Close look similar to the algorithm since they involve standing in the same spot with similar hand movements. To this end, the action recognition algorithm was tested by subdividing the dataset into smaller datasets.

Table 15. Datasets Generated and the corresponding action recognition accuracies.

Dataset name	Actions	SVM accuracy
Prep2	Prep Cut, Prep Stir Prep Open + Close	66.7%
Manipulate1	All Manipulate Actions, Separated	73.3%
Manipulate2	Manipulate Stove Manipulate Microwave Manipulate Sink On + Off Manipulate Fridge	77.1%
Wash	Wash Rinse Wash Sink Wash Dishwasher	72.2%
Walk2	WIK Hold + Not Hold WAK Hold + Not Hold WOK Hold + Not Hold	41.7%
Walk Hold	WIK Hold WAK Hold WOK Hold	69.3%
Walk Not Hold	WIK Not Hold WAK Not Hold WOK Not Hold	52.8%
Pick Put1	All Pick Up and Put Down Actions, Separated	37.5%
Pick Put2	Pick Up Counter + Put Down Pick Up Top + Put Down Pick Up Bottom + Put Down	69.4%

Open Close1	Open Top Cabinet Close Top Cabinet Open Bottom Cabinet Close Bottom Cabinet	54.2%
Open Close2	Open Top Cabinet + Close Open Bottom Cabinet + Close	75.0%
Mixed1	Manipulate Fridge Close Bottom Cabinet WIK Hold	97.2%
Mixed2	Wash Sink WAK Not Hold Prep Stir	97.2%
Mixed3	Open Top Cabinet + Close Prep Open + Close + Cut WIK + WOK, Hold + Not Hold Manipulate Microwave Manipulate Stove Manipulate Fridge	86.9%

Initial datasets were put together under their theme. After running a raw category, actions that are similar are combined. In each instance accuracy improved. The last three datasets have members that are from mixed categories to better imitate the diverse actions possible in a kitchen. Accuracy is important, so therapists can have a better understanding of what actions a stroke patient is performing.

The cross-validation method chosen was the holdout method with 4 items in the test set. The rest of the data were assigned to a training set. This was performed 3 times. All the datasets and their accuracies are summarized in Table 16.

The confusion matrix in Figure 37 was produced from a dataset of 50 random actions from each group, which amounts to 250 items. 20 actions from each category of this new subset were put into a test set, totaling 100. The rest, 150, were used as a training set.

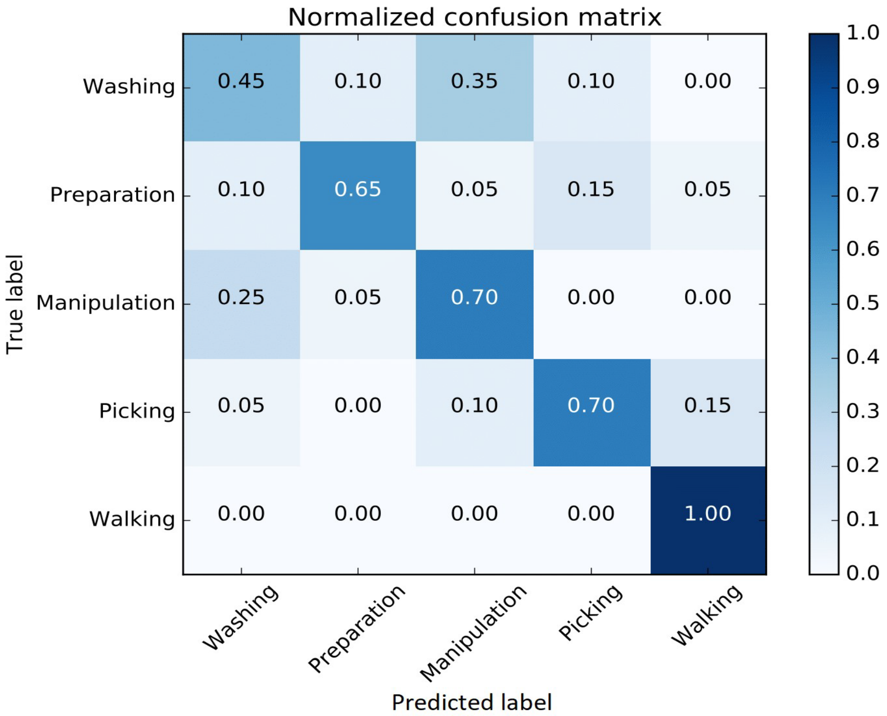


Figure 37. A confusion matrix generated from a 100-item test set and 150 training set.

Overall, the precision is 70%. Individually, Walking, Preparation, Picking, Manipulation, and Washing have the precisions 83.3%, 81.3%, 73.7%, 58.3%, and 53.0%, respectively. Manipulation and washing have the lowest precisions due to confusion from the algorithm because of similar hand movements and standing positions. Manipulation is incorrectly labeled as Washing 25% of the time, and Washing is falsely labeled as Manipulation 35% of the time.

4.6.2 Action recognition and temporal localization on simulated data

4.6.2.1 CDC

We first trained and evaluated the CDC network using the simulated kitchen dataset described in section 4.3.2. Although the CDC filter can be applied to the input of arbitrary size, due to memory limitation, we applied a 16-frame sliding window to segment the videos without overlapping. Each window with a label of each frame was then fed into the

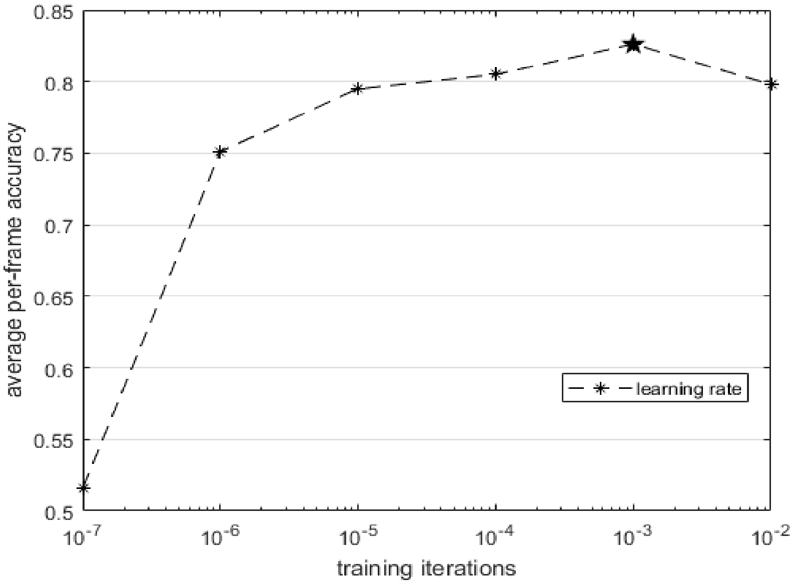


Figure 38. An experiment of selecting suitable hyperparameter, learning rate. The network was trained on randomly selected 90 videos from the simulated dataset and tested on the rest of 10 videos for three times under different learning rates. The average per-frame precisions were calculated. The accuracy was highest when the learning rate was 0.001.

CDC network. Note that the frames in one window can have different action labels. The CDC network was initialized by the model trained on the THUMOS dataset and trained on the collected dataset. The stochastic gradient descent was applied for optimization. Following conventional settings, the momentum was set to 0.9 and the weight decay was set to 0.005. I randomly selected 90 videos as the training set and the remaining 10 videos as the test set. To find the suitable initial learning rate, we trained the network using different learning rates ranging from 0.0000001 to 0.01. The network was trained three times at the same learning rate and the training iteration was set to be 10000. The average per-frame recognition accuracies of different learning rates were shown in Figure 38. The learning rate of 0.001 generated the best per-frame precision.

After the best initial learning rate was found, we initialized the learning rate as the optimal value of 0.001, and then decreased it by 0.1 for each 5000 iterations, resulting in a total of 30,000 iterations. To evaluate the ability of detecting the actions, the per-action precision was calculated. To test the ability of localizing the actions, the per-frame precision was calculated. The recognition and localization results of the simulated dataset are shown in Table 16. The normalized confusion matrix of per-action recognition is shown in Figure 39. None of the above category was excluded for per-action precision performance. Reaching over head, reaching below the waist and sweeping categories show exceptional recognition results.

Table 16. Per-frame precisions and per-action precisions of test videos from CDC on simulated kitchen dataset.

	Average precision	
	Per frame	**Per action**
Simulated kitchen	85.1%	92.1%

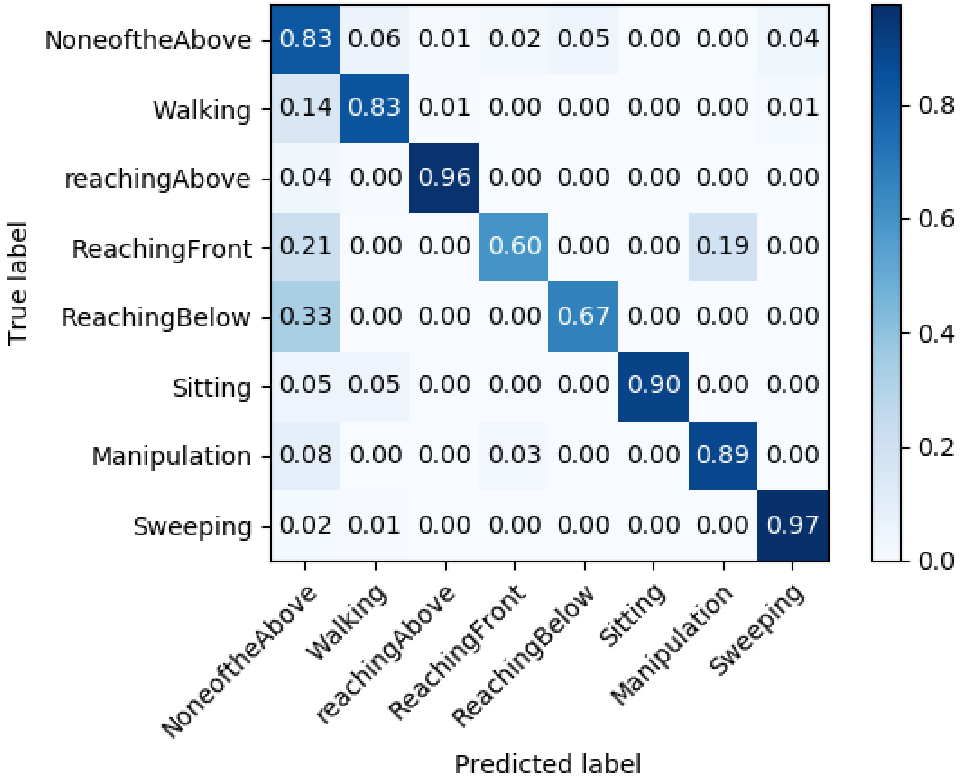


Figure 39. Normalized confusion matrix of recognizing actions from CDC network on simulated kitchen dataset.

4.6.2.2 R-C3D

Ten anchor segments were chosen with the scale value of [2, 4, 5, 6, 8, 9, 10, 12, 14, 16]. For training, the positive/negative labels need to be assigned to the anchor segments. Following the standard practice in object detection [91], a positive label was selected if the anchor segment 1) overlaps with some ground-truth activity with Intersection-over-Union (IoU) higher than 0.7, or 2) has the highest IoU overlap with some ground-truth activity. If the anchor has IoU overlap lower than 0.3 with all ground-truth activities, then it is given a negative label. The search scope of the hyperparameter is shown in Table 18. The learning rate was 1e-14, and the learning weight decay was 0.0005. The network output the start frame id, end frame id, prediction label, and the confidence score of the detected action segments as shown in Table 17. A threshold was set to select segments with high confidence.

Table 17. The output of the R-C3D network. The start frames, end frames, action predict labels, and the confidence scores of detected segments were generated.

Video name	Start frame	End frame	Predict label	Confidence score
16-17-35	58	69	walking	0.3679
16-17-35	69	107	sweeping	0.9470
		...		
14-58-53	229	239	Reaching below	0.0052

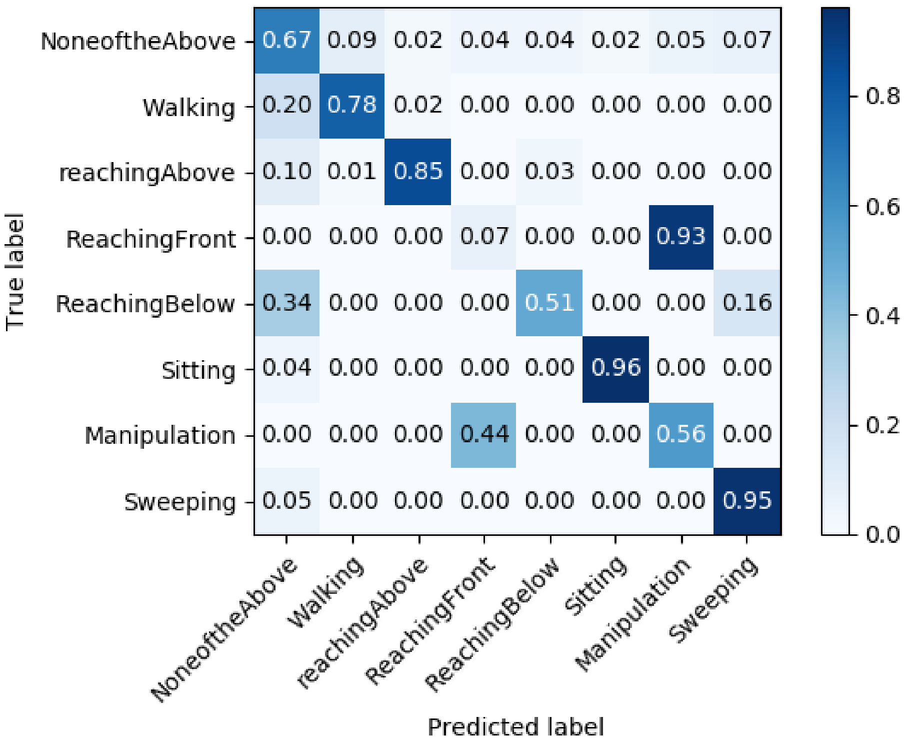
Table 18. Search scope of the hyperparameters of the R-C3D network.

Anchor scales	Optimizer	learning rate	lr weight decay	lr decay step	training max epochs
[2,4,5,6,8,9,10,12,14,16]	sgd	1.00E03	0.1	3	5
	adam	1.00E-04	0.05	4	6
		1.00E-05		5	7
		1.00E-06		6	8
				7	9
					10

Table 19. Per-frame precisions and per-action precisions of test videos from R-C3D on simulated kitchen dataset.

Threshold	Per-frame precision	Per-action precision
0.4	0.726	0.882
0.5	0.772	0.942
0.6	0.793	0.941

Per-action precision and per-frame action precisions were used to evaluate the accuracy of action recognition and detection. The results with different thresholds were shown in Table 19. The confusion matrix was shown in Figure 40. The per-frame precision and the per-action precision were 0.793 and 0.941, respectively. Some reaching front actions were mis-predicted as hand manipulation actions based on the confusion matrix.



True label \ Predicted label	NoneoftheAbove	Walking	reachingAbove	ReachingFront	ReachingBelow	Sitting	Manipulation	Sweeping
NoneoftheAbove	0.67	0.09	0.02	0.04	0.04	0.02	0.05	0.07
Walking	0.20	0.78	0.02	0.00	0.00	0.00	0.00	0.00
reachingAbove	0.10	0.01	0.85	0.00	0.03	0.00	0.00	0.00
ReachingFront	0.00	0.00	0.00	0.07	0.00	0.00	0.93	0.00
ReachingBelow	0.34	0.00	0.00	0.00	0.51	0.00	0.00	0.16
Sitting	0.04	0.00	0.00	0.00	0.00	0.96	0.00	0.00
Manipulation	0.00	0.00	0.00	0.44	0.00	0.00	0.56	0.00
Sweeping	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.95

Figure 40. Normalized confusion matrix of recognizing actions from R-C3D network on simulated kitchen dataset.

4.6.2.3 R-HCN

The anchor segments were initially with scales of [50, 100, 200, 400] in the temporal proposal. The search scope of the hyperparameter in the R-HCN network is the same as the R-C3D network. The stroke dataset was split for training and testing. 80 percent of the data were selected as a training set. The best test result was generated with the learning rate 1e-14, and the learning weight decay 0.0005. Since the skeletal joint data are more sensitive to noise, only detected walking actions were considered.

4.6.2.4 Ensemble net

The per-frame label was generated by fusing the labels predicted by the three networks. The per-frame precision and per-action precision were used to evaluate the accuracy of the ensemble network on the simulated kitchen dataset shown in Table 20. The confusion table was shown in Figure 41. The per-frame precision and per-action precision of the ensemble net were 0.916 and 0.942. Though the accuracy of predicting reaching forward actions increased, about 30 percent of reaching forward actions were predicted as hand manipulation.

Table 20. Per-frame precisions and per-action precisions of test videos from the ensemble network on simulated kitchen dataset.

Models	Per-frame precision	Per-action precision
R-C3D (t=0.4)	0.726	0.882
CDC	0.849	0.890
CDC+R-C3D	0.882	0.931
CDC + R-C3D + R-HCN (walking)	0.916	0.942

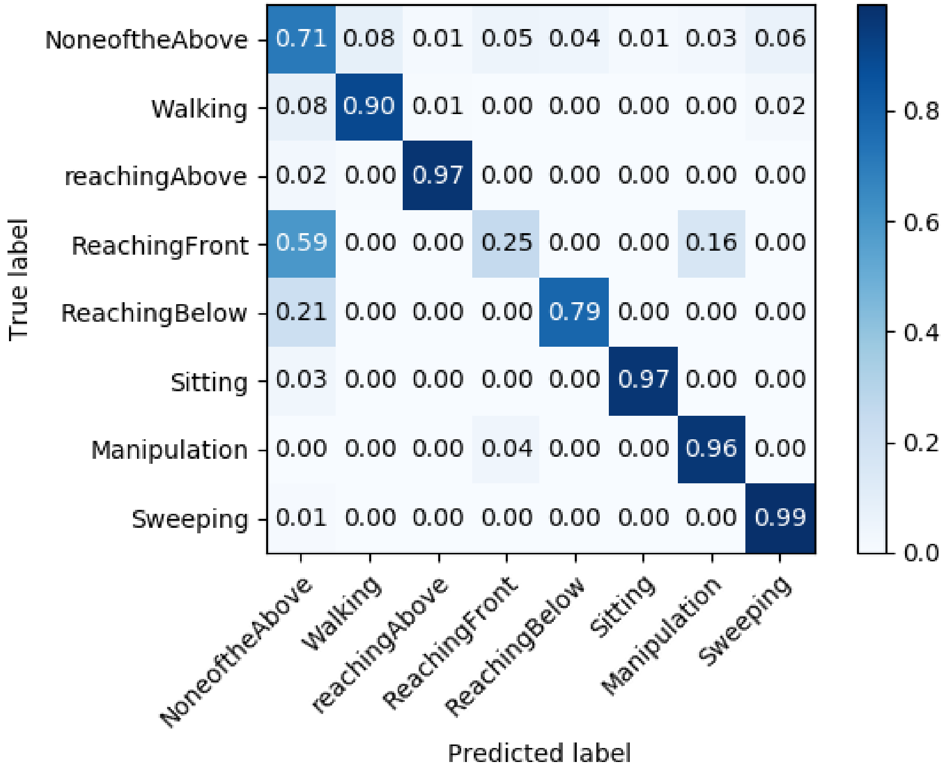


Figure 41. Normalized confusion matrix of recognizing actions from the ensemble network on simulated kitchen dataset.

4.6.3 Action recognition and temporal localization on stroke data

The stroke dataset contains the labeled data from participants 1,2,3,4,5 and 10. In order to develop a model which can predict clinically relevant actions accurately, as well as preserve sufficient data to perform the kinematic assessment, the stroke dataset has been firstly split into two halves. One for algorithm training and test set, and another one for validation and assessment set. Specifically, the data were grouped into buckets by dates. The data in each bucket were split into two halves. The training and test set was randomly selected from those two halves. The second half was used as the validation and assessment set. As a result, the training and test set has a similar histogram to the validation and assessment set. In addition, the assessment set covers all the collection days of a participant.

4.6.3.1 CDC

The CDC network was also evaluated using the stroke dataset. To feed the video sequences to the network while maximizing the memory usage, a 16-frame sliding window was applied to segment the videos without overlapping. The CDC network was initialized by the model trained on the simulated kitchen dataset. The stochastic gradient decent was applied for optimization. Following conventional settings, we set the momentum to 0.9 and the weight decay to 0.005.

The training dataset was formed by 80% of the training and test set and the remaining data was used as the test set. The best test results were generated with the learning rate as 0.001 and decreased by 0.1 for each 5000 iterations. The recognition and localization results of the simulated dataset are shown in Table 21. The highest predict accuracies were from the full dataset with all six participants' data. The best per-frame and Per-action precisions were 0.859 and 0.871, respectively.

Table 21. Per-frame precisions and per-action precisions of test videos from the CDC network on stroke dataset.

Dataset	Precision	Trial1	Trial2	Trial3	Mean	Std.
P1-3	Per frame	0.765	0.542	0.576	0.628	0.120
	Per action	0.807	0.574	0.562	0.647	0.113
P1-4	Per frame	0.682	0.688	0.526	0.632	0.092
	Per action	0.702	0.685	0.581	0.656	0.053
P1-5	Per frame	0.728	0.614	0.737	0.693	0.068
	Per action	0.752	0.635	0.739	0.709	0.052
P1-5, 10	Per frame	0.758	0.769	0.859	0.795	0.055
	Per action	0.767	0.814	0.871	0.817	0.043

4.6.3.2 R-C3D

The same training strategy was used to train R-C3D models for both stroke dataset and the simulated kitchen dataset. For training on the stroke dataset, ten anchor segments were chosen. The scale value was [2, 4, 5, 6, 8, 9, 10, 12, 14, 16]. To take advantage of what have been learning from the healthy subjects, the model trainable parameters were initialized by the model trained on the simulated kitchen dataset.

For the training and test stroke dataset, 80% of the data were randomly selected for training and the rest of the data were used for testing. The different combination of hyperparameter were investigated. The optimal results were generated where learning rate was 1e-14, learning rate weight decay was 0.0005, learning rate decay step was 7, and the maximum epochs was set to be 10. Per-action precision and per-frame action accuracy were used to evaluate the accuracy of action recognition and detection. The per-action precision and per-frame precision were shown in Table 22 with the threshold as 0.9.

Table 22. Per-frame precisions and per-action precisions of test videos from the R-C3D network on stroke dataset.

Dataset	Precision	Trial1	Trial2	Trial3	Mean	Std.
P1-3	Per frame	0.684	0.742	0.666	0.697	0.040
	Per action	0.785	0.804	0.754	0.781	0.025
P1-4	Per frame	0.713	0.705	0.673	0.697	0.021
	Per action	0.794	0.733	0.722	0.750	0.039
P1-5	Per frame	0.805	0.708	0.665	0.726	0.072
	Per action	0.828	0.792	0.732	0.784	0.048
P1-5, 10	Per frame	0.811	0.891	0.880	0.861	0.044
	Per action	0.837	0.895	0.860	0.864	0.029

4.6.3.3 R-HCN

We followed the same training strategy for both the simulated kitchen dataset and the stroke dataset. Specifically, the anchor segments were initially with scales of [50, 100, 200, 400] in the temporal proposal. The stroke dataset was split for training and testing. 80 percent of the data were selected as training set. The best test result was generated with the learning rate 1e-14, and the learning weight decay 0.0005. Since the skeletal joint data are more sensitive to noise, only detected walking actions were considered. The recognition and localization results of the simulated dataset are shown in Table 23.

4.6.3.4 Ensemble net

The per-frame label was generated by fusing the labels predicted by the three networks. The per-frame precision and per-action precision were used to evaluate the accuracy of ensemble network on the stroke dataset shown in Table 24.

Table 23. Per-frame precisions and per-action precisions of test videos from the R-HCN network on stroke dataset.

Dataset	Precision	Trial1	Trial2	Trial3	Mean	Std.
P1-3	Per frame	0.569	0.544	0.573	0.562	0.016
	Per action	0.543	0.585	0.554	0.561	0.022
P1-4	Per frame	0.623	0.517	0.787	0.642	0.136
	Per action	0.657	0.594	0.794	0.682	0.102
P1-5	Per frame	0.721	0.809	0.588	0.706	0.112
	Per action	0.782	0.808	0.598	0.730	0.115
P1-5, 10	Per frame	0.822	0.652	0.685	0.719	0.090
	Per action	0.815	0.691	0.702	0.736	0.068

The performance of the model with the highest accuracies in Table 24 was evaluated using the validation set. The confusion matrices of the ensemble network on the validation set were shown in Figure 42 and Figure 43. The per-frame precision and the per-action precision were 0.819 and 0.838 on the validation set, respectively.

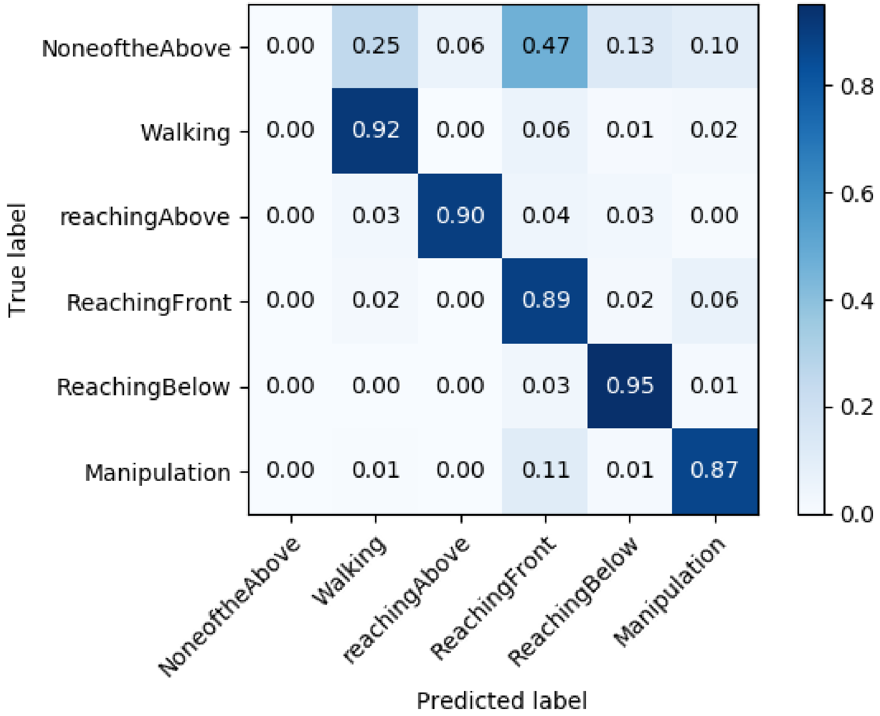


Figure 42. Normalized confusion matrix of recognizing actions from the ensemble network on validation stroke dataset.

Table 24. Per-frame precisions and per-action precisions of test videos from the ensemble network on stroke dataset.

Dataset	Precision	Trial1	Trial2	Trial3	Mean	Std.
P1-3	Per Frame	0.785	0.791	0.679	0.752	0.063
	Per Action	0.831	0.801	0.761	0.798	0.035
P1-4	Per Frame	0.790	0.772	0.789	0.784	0.010
	Per Action	0.801	0.792	0.724	0.773	0.042
P1-5	Per Frame	0.818	0.817	0.742	0.792	0.044
	Per Action	0.839	0.801	0.791	0.810	0.025
P1-5, 10	Per Frame	0.824	0.904	0.881	0.869	0.041
	Per Action	0.827	0.911	0.867	0.868	0.042

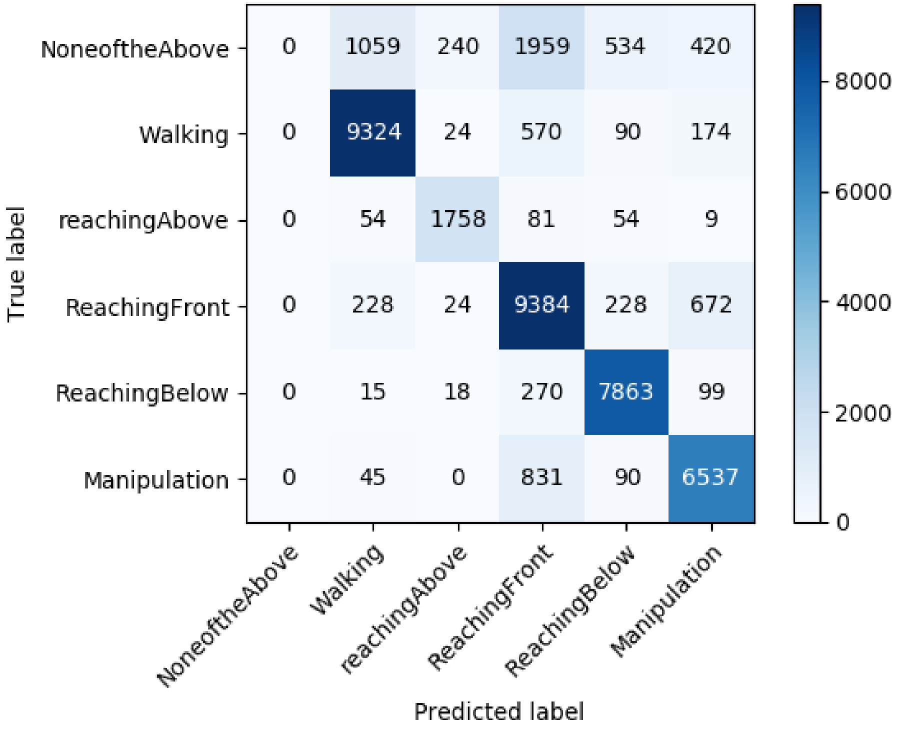


Figure 43. Confusion matrix of recognizing actions from the ensemble network on the validation stroke dataset.

4.7 Kinematic assessments

Stroke patients can suffer from weakness or paralysis on a side or the whole body [2]. If a clinician can track improvements and declines, critical intervention points can be identified, and care can be adjusted accordingly. Quantitative measures of movement quality are important for expressing the outcomes and clinically important changes in functional status of stroke patients. Kinematic metrics in relation to joint displacements, analysis of hand trajectories and velocity profiles have been commonly used to perform quantitative measures. For this reason, maximum extent of reach and speed related metrics of hands are calculated in the DARAS system. To evaluate the smoothness of hand movement, normalized jerk metric are also calculated. Each piece of information can be used to track improvement over time, or indicate a decline where intervention is needed.

The Foresite system senses a user in 3D space and provides the 3D coordinates of up to 19 skeletal joints for a user. The maximum sample rate is 30 frames per second. The DARAS system recorded the skeletal joint samples and the timestamps in a csv file. The kinematic assessments were performed using the recorded joint data. The confidence values of each joint are also recorded to indicate the quality.

4.7.1 Preprocessing

The start frame names, end frame names, and the corresponding labels of the predicted actions were generated by the proposed ensemble network on the validation and assessment set. Before performing the kinematic assessment, it is necessary to crop the recognized joint action segment, remove the outliners and reduce the noise.

Crop the action segments. The recognized actions were cropped from the validation joint set using the ensemble output, which includes segment start names and segment end names.

Remove the outliners. If the total number of frames of a segment is less than 4 frames, the segment is too short for assessing the normalized jerk measure. Thus, short segments with frames less than 4 were removed. A bone length filter was applied to each segment. The bone length filter first computed the lengths of all the bones using the joint data. Then the filter computed the average standard deviation of all the bones among all the frames in a segment. If the average standard deviation was larger than 10 centimeters, the segment was considered an outlier and was removed.

Reduce the noise. The Butterworth filter designed for the data in the stroke rehabilitation game was applied to the joint data for action segments to reduce the noise.

4.7.2 Kinematic measures

4.7.2.1 Extent of reach

Extent of reach was calculated for each recognized action. Extent of reach was defined as the distance from the hand joint to the shoulder center, where the shoulder center is the middle of the left and right shoulder joints. Suppose the hand joint and the shoulder center are represented by $j_{hand} = \{h_x, h_y, h_z\}$and $j_{shoulderC} = \{s_x, s_y, s_z\}$, then the extent of reach for each frame is calculated by

$$Extent\ of\ Reach = \sqrt{(h_x - s_x)^2 + (h_y - s_y)^2 + (h_z - s_z)^2} \qquad (16)$$

4.7.2.2 Speed

We also calculated maximum and mean velocities for each action. Suppose the hand joint of the ith frame is represented by $j_{hand} = \{x_i, y_i, z_i\}$, the velocity of this frame is calculated by

$$Hand\ Velocity = \frac{\sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2}}{t_{i+1} - t_i} \tag{17}$$

where t, timestamp, is measured and stored automatically with each frame by the system.

4.7.2.3 Normalized jerk

Jerk is the third derivative of position. Reference [76] presented that the time-integrated squared jerk decreases as the smoothness of the movement increases. The integrated squared jerk has dimensions of (squared length)/(5th power of duration) [77]. To make the measurement dimensionless, normalized Jerk is computed using (18):

$$\text{Normalized Jerk} = \sqrt{\frac{1}{2} \times \frac{d^5}{l^2} \int_{t_i}^{t_f} J^2(t)dt} \tag{18}$$

where d denotes the overall movement duration, and l denotes the overall movement length, and J denotes the jerk function, the third derivative of position.

4.7.3 Trajectory visualization

For each recognized action segment, it is necessary to visualize the motion in multiple ways, because that we can straightforwardly review how the subject performs this action. In the DARAS, there are two main approaches to visualize an action segment. One is the visualization of full-body skeletal joints frame by frame as a video. Another is the plot of left- and right-hand trajectory.

For left-right trajectory visualization, the middle spine joint is used as the origin. This is accomplished by subtracting the middle spine joint coordinates from any chosen joint's coordinates.

4.7.4 Statistical analysis

After the assessment was performed, a set of kinematic measures were calculated for each recognized action segment. With the kinematic outcomes of the all the action for a subject, statistical analysis can be conducted to investigate the following two aspects.

1. Can we quantitively describe the movement qualities between stroke participants and healthy participants?
2. Can we identify the change of the movement for a stroke participant using the assessment data?

To investigate the first question, the average values of each action category of each hand were computed for both stroke and non-stroke participants. Also, the frequency and type of movement were evaluated to show how active a person is.

To investigate the second question, the histogram of a metric of both hands were generated for stroke participants and the over-time trend of a metric of both hands were plotted.

4.8 Assessment results

The DARAS also quantitatively evaluated the kitchen actions using assessment metrics including hand movement speeds, maximum hand extent of reach, and smoothness. In this section, the assessment results of two single actions are shown first. Then, statistical analysis results of stroke participants and healthy participants are presented.

4.8.1 Assessment outcomes on a single action

The hand extent of reach in x, y, z, and 3d space was computed to evaluate the range of the hand movement, hand mean speed and hand maximum speed were computed to evaluate the hand movement velocity, and the normalized jerk metric was computed to evaluate the smoothness of a hand movement. The assessment outcomes of a reaching forward action and a hand manipulation action were presented.

For the reaching forward action, the replay of the skeletal model and hand trajectories were visualized in Figure 44 and Figure 45. The outcomes of the kinematic measures were listed in Table 25.

The participant reached an object using his right hand. The extent of reach values of the right hand was larger than the values of the right hand. The assessment outcome also showed that the right hand moved faster than the left hand.

Table 25. The assessment outcome of the reaching forward action using the proposed metrics.

	Left	**Right**
Extent of reach		
X	67.7	316.04
Y	283.41	436.73
Z	208.88	236.27
3D	352.15	522.89
Speed and smoothness		
Max speed	348.41	1171.08
Mean speed	166.27	579.32
Norm. jerk	82.27	36.29

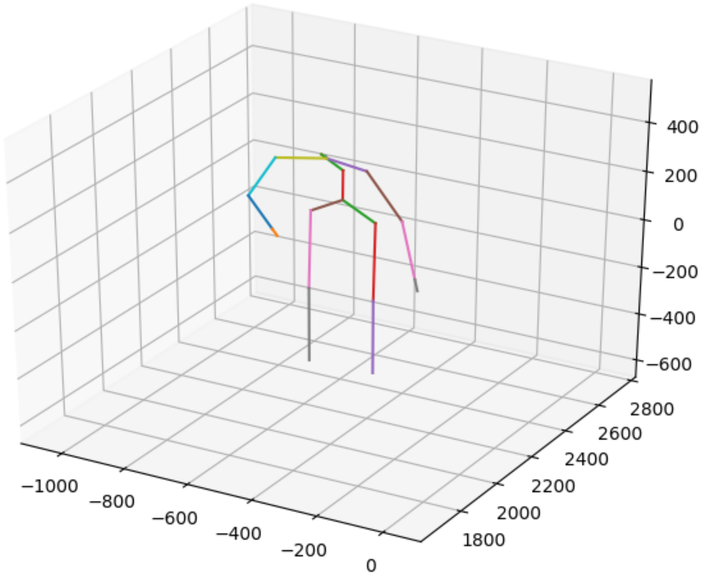


Figure 44. Visualization of the skeletal joints of a reaching forward action from a healthy subject. The right hand reached forward.

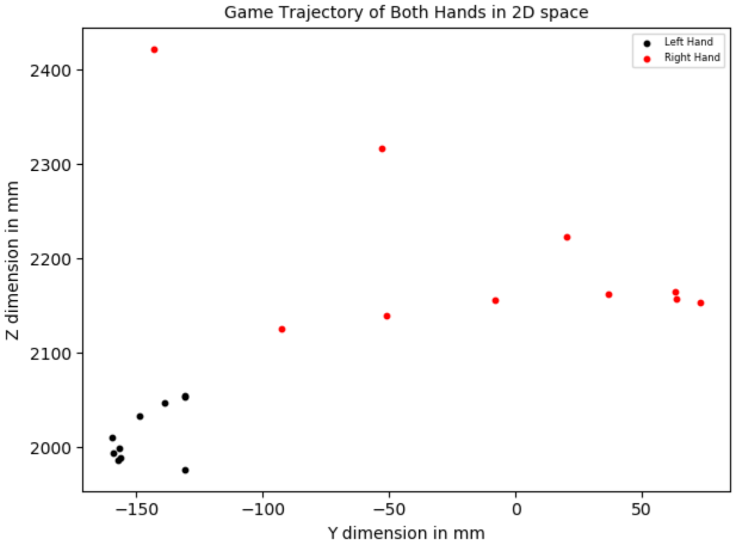


Figure 45. The trajectories of both hands of a reaching forward action in the lateral-vertical space. The right hand had a larger movement range compared to the left hand.

For the hand manipulation action, the replay of the skeletal joins and hand trajectories were visualized in Figure 46 and Figure 47. The outcomes of the kinematic measures were listed in Table 26.

The extent of reach values and the speed values were similar for both hands. The hand manipulation movement usually involves manipulating objects in front of the chest, such as cutting vegetables. Thus, the movements of both hands can be similar, which matched with the assessment outcome.

Table 26. The assessment outcome of the manipulation action using the proposed metrics.

	Left	**Right**
Extent of reach		
X	101.1	179.58
Y	377.06	442.28
Z	141.49	377.97
3D	395.6	527.02
Speed and smoothness		
Max speed	296.61	222.15
Mean speed	93.87	106.27
Norm. jerk	906.05	627.83

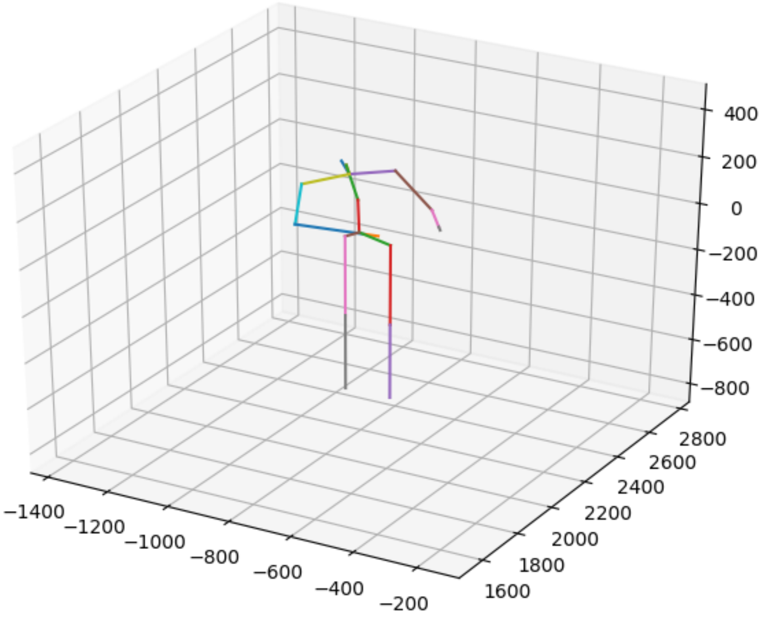


Figure 46. Visualization of the skeletal joints of a hand manipulation action from a healthy subject. Both hands were occupied in this hand manipulation action.

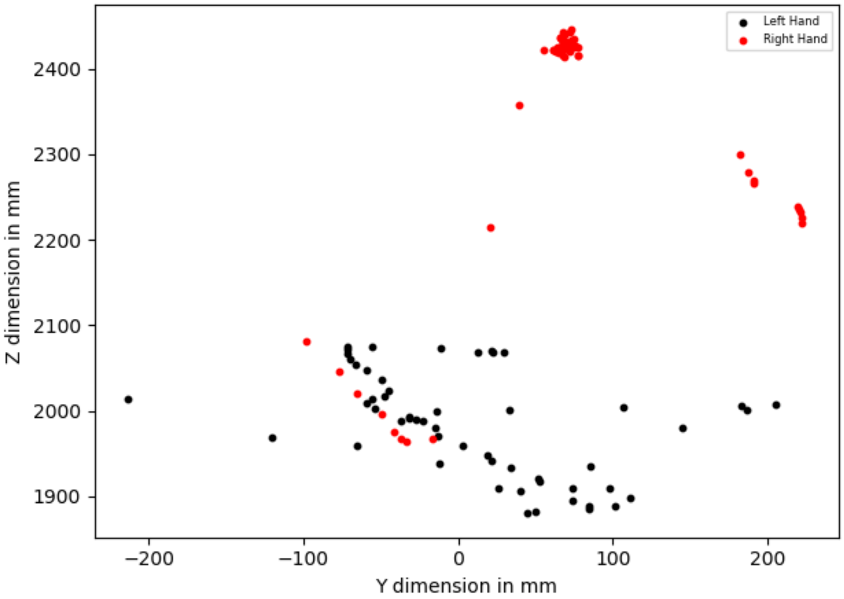


Figure 47. The trajectories of both hands of a hand manipulation action in the lateral-vertical space. The movement ranges of both hands look similar.

4.8.2 Comparison between both hands

The assessment outcome from a single action cannot represent the character of a participant's movement. Clinicians are more interested in investigating the overall movement quality, the behavior pattern, and the change or trend of movement quality of a stroke individual. Statistical analysis approaches have applied to assessment results of recognized action segments.

The assessment results were first grouped by different participants. There were usually a stroke participant and several non-stroke participants in a stroke participant's home. For non-stroke participants, the most appeared person was selected in the analysis in each location. The assessment data were also grouped by different clinically relevant actions. Finally, the average values were calculated for the left and right hands of a non-stroke participant and for impaired side and less affected side of a stroke participant. The average values were summarized in Table 27.

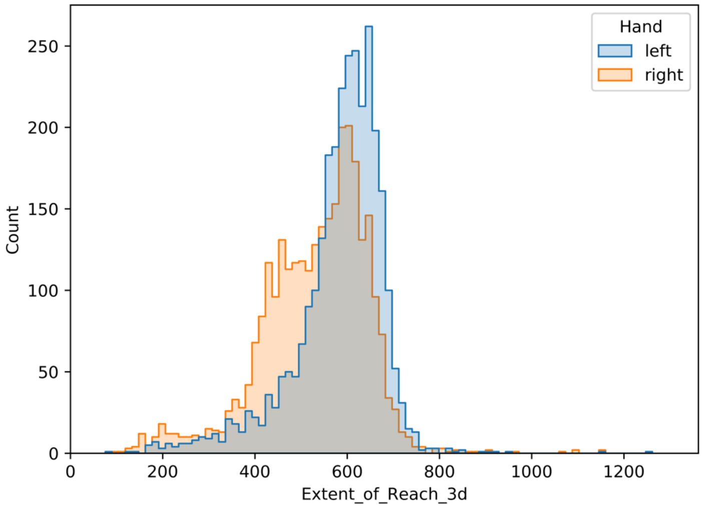


Figure 48. Histogram of the extent of reach in 3d of left and right hands.

To compare the difference between the movement quality of left and right hands of participants, histogram graphs were applied in Figure 48.

4.8.3 Active level analysis

To show how active a stroke participant is in the cooking area, the number of recognized clinically relevant actions were counted. The average number of recognized actions detected per day of all six participants in the three-month data collection period

Table 27. Quantitatively assessment of clinically relevant actions performed by subjects in their kitchens. Assessment metrics includes extent of reach in millimeter, and speed metrics in millimeter/s. All the recognized trials over the three-month collection period were selected to perform the assessment. The average values were calculated for each metric of each action category.

metrics	**actions**	**non-stroke**		**stroke**	
		left	right	impaired	less affected
Extent in 3d	manipulation	508.20	506.87	517.54	509.18
	Reach below	559.34	538.84	488.04	534.72
	reach forward	549.16	511.74	557.31	566.20
	reach up	516.32	502.80	533.10	548.71
	walk	561.53	550.47	588.04	570.48
Max speed	manipulation	1169.60	611.83	393.19	572.33
	Reach below	1346.66	1329.16	579.95	512.25
	reach forward	1651.40	1408.08	1038.93	1034.63
	reach up	2200.71	1594.95	2068.23	1646.90
	walk	1951.12	2176.05	1788.86	1560.78
Mean speed	manipulation	415.77	258.86	135.84	179.89
	Reach below	422.64	361.96	213.35	188.17
	reach forward	550.79	497.93	330.89	357.11
	reach up	807.38	625.04	647.75	476.41
	walk	955.04	1040.82	725.56	668.92
Norm. Jerk	manipulation	3454.02	2412.95	25749.73	22255.37
	Reach below	2408.91	1623.91	25007.95	25063.26
	reach forward	1399.34	6550.83	16529.75	16036.36
	reach up	1777.19	698.39	2934.12	2631.72
	walk	6333.16	4455.02	3521.29	4257.34

Table 28. The average number of actions recognized per day. For most stroke participants, the number of recognized actions per day is smaller than non-stroke participants.

	Kitchen1	Kitchen2	Kitchen3	Kitchen4	Kitchen10
Stroke participant	11	5	**495**	7	48
Non-stroke participant	**210**	**12**	34	**39**	**89**

were shown in Table 28. For most of stroke participants, the average number of actions per day of a stroke participant was less than the value of the non-stroke participant in the same location, which means that a post-stroke individual is usually less active than a healthy individual.

Figure 49 visualized the numbers of actions of each action category for a stroke participant and a non-stroke participant in a kitchen of a week period. The figures also

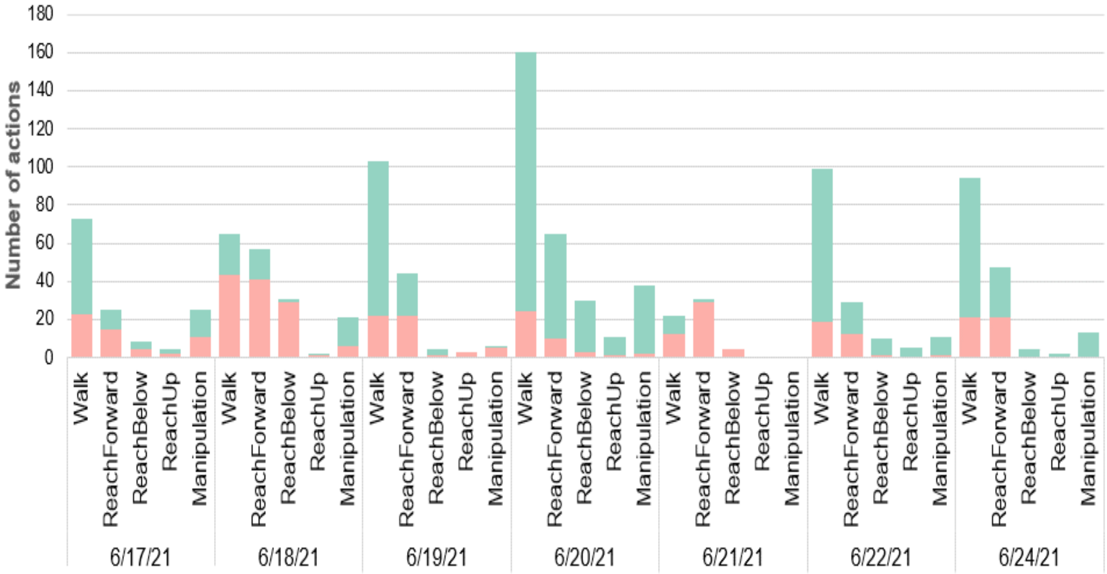


Figure 49. Count number of actions in different categories by each date for a stroke subject and a healthy subject.

showed that the stroke participant was less active. In addition, the stroke participant tends to do more walking and reaching forward actions among all the clinically relevant actions.

4.8.4 longitudinal analysis

Tracking the change of the movement quality over time of a stroke individual is important because it directly reflects the effectiveness of the treatment outcomes. Clinicians usually provide scale-based assessments periodically to evaluate the status of a post-stroke individual and compare the outcomes of multiple assessments. Due to the limitation of healthcare resources, scale-based assessments can't be conducted frequently.

The metrics were computed for each recognized action. Thus, the DARAS can evaluate the quality of every recognized clinically relevant action of a stroke individual.

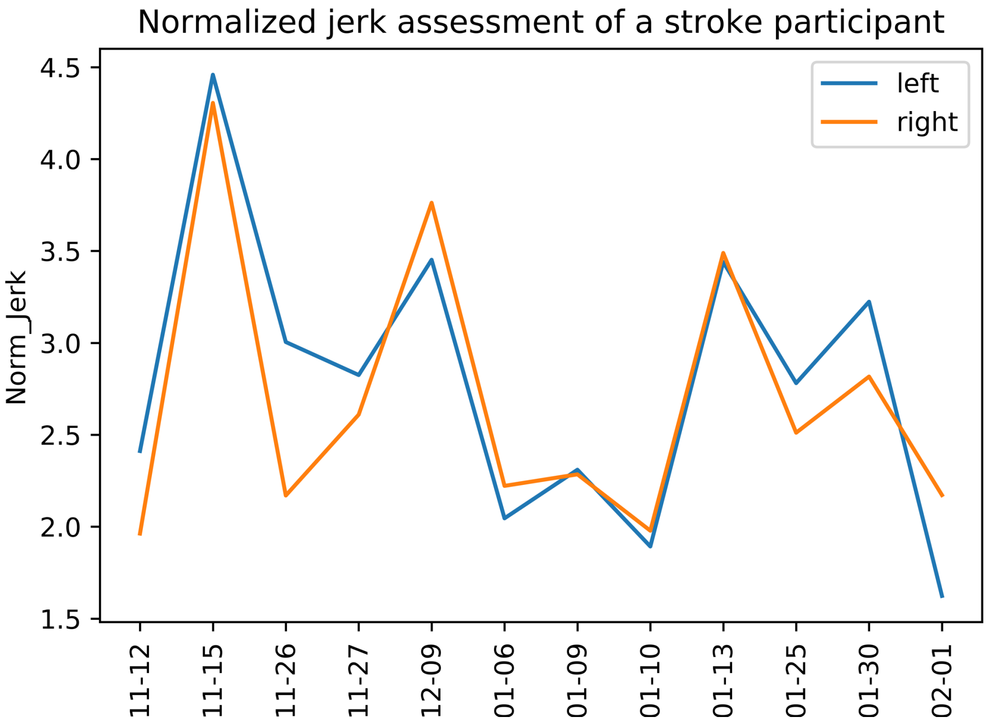


Figure 50. Average normalized jerk by date over time of both hands.

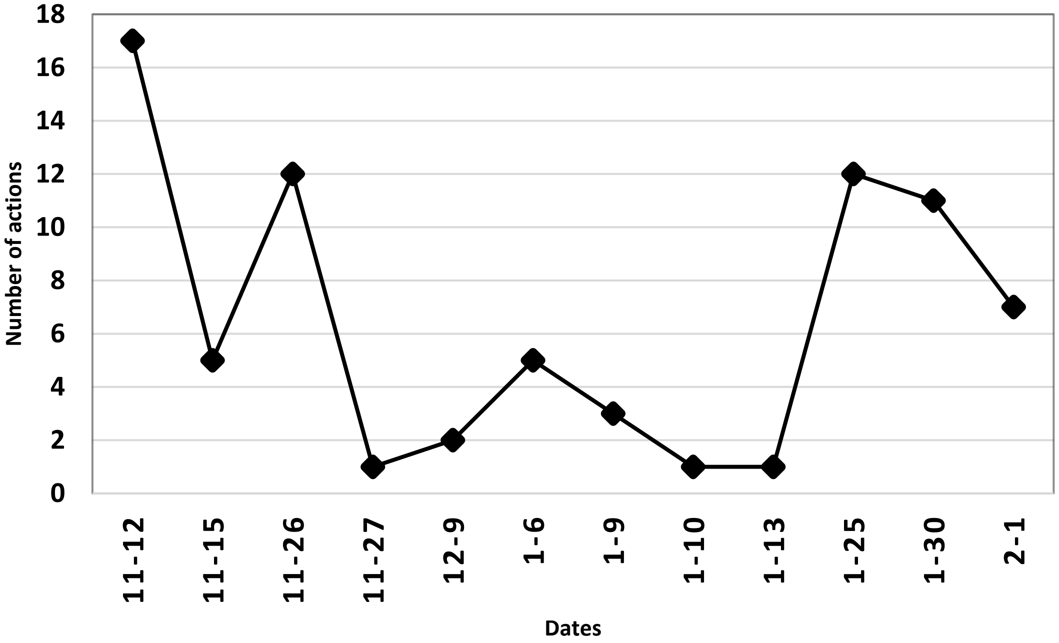


Figure 51. The changes of the number of actions over the three months of a stroke participant.

By analyzing the change of the values for metrics, the longitudinal analysis can be more sensitive. Also, clinicians can detect changes of the movement of a stroke individual quickly. Figure 50 and Figure 51 show the over-time change of normalized jerk values of both hands and change of the number of recognized actions by dates of a stroke participant.

4.9 Discussion

Post-stroke individuals can regain their motion abilities and achieve functional independence from efficient rehabilitation treatment. Due to the limitations of healthcare resources, post-stroke individuals usually can only visit a clinic at most a few hours per week. Clinicians may not be able to provide a personalized effective rehabilitation plan for each individual based on the feedback from in-clinic visits. Patients spend most of their time doing daily activities at home. If a system can evaluate the quality of a post-stroke individuals' daily actions and provide the evaluation reports to clinicians, clinicians can have a better understanding of the health status of the post-stroke individual. As a result, a more personalized rehabilitation plan can be provided. Currently, no system has been developed to track the daily activities of stroke individuals and perform assessments on the clinically relevant actions.

The daily activity recognition and assessment system (DARAS) has been proposed and implemented to fill this gap. The DARAS contains three main components which are action logging system, action recognition and temporal localization part, and action assessment part. The development of DARAS has been summarized in three phases.

Phase 1. A Kinect-based DARAS was implemented to investigate if cooking-related actions can be collected and recognized. Then the assessment was performed on the recognized actions. Twenty-eight different general kitchen actions were designed, such as rinsing a dish. 504 action segments were collected from a simulated kitchen environment. HON4D was applied to extract the spatial-temporal feature for depth data action segments. An SVM classifier was trained to recognize actions. The prediction results showed that the model had difficulties recognizing all 28 actions. But when the action segments were grouped as action categories, the prediction accuracy increased to about 80%. The action categories were walking, preparation, picking, manipulation, and washing. The kinematic assessment can be performed using the skeletal joint data generated by the Kinect SDK.

In this phase, the prediction results showed that the cooking action categories can be distinguished and recognized by the HON4D based model. Skeletal joint data made the assessments possible. The limitation of this version of DARAS is that it was inconvenient to set up a Kinect-based system. In addition, in a realistic situation, a sequence of actions is recorded in a video clip, so temporal action localization is necessary.

Phase 2. A VicoVR-based DARAS was implemented to reduce the size of the whole system and simplify the set-up process. The Nuitrack SDK was selected to retrieve the depth and skeletal joint from the VicoVR sensor. In this phase, I investigated if actions can be recognized and segmented from action sequences. The CDC network was customized to perform action recognition and temporal action localization from depth sequences. In order to make the action as close as possible to daily cooking activities, three cooking scenarios were designed. Totally, 100 action sequences were collected using the VicoVR-based logger. Each frame was labeled to be one of seven clinically relevant action categories or none of the above. The clinically relevant action categories included walking, sitting, reaching above the head, reaching forward, reaching below the waist, hand manipulation, and sweeping. The trained CDC model provided the per-frame action prediction. The per-frame prediction labels were grouped to locate the action segments. The per-frame and Per-action precisions were 85% and 92%, which showed that the trained model can recognize and segment actions for assessment. The kinematic assessments were performed using the skeletal joint data.

In this version, I demonstrated that the VicoVR-based DARAS can recognize and segment clinically relevant actions from untrimmed continuous action sequences. But the data were collected from young non-stroke participants in a simulated kitchen environment.

Phase 3. A Foresite-based DARAS system was implemented. The Foresite system is a standalone system that integrates an ubuntu computer, a depth sensor, and necessary accessories in a small box. The Foresite-based DARAS system is easy to install and set up.

Multiple features were added to the logger software, including the self-recovering functions. Sixteen post-stroke participants were recruited for the data collection. Three-month daily cooking activity data were recorded from each participant. To improve the accuracy, an ensemble network was proposed and implemented for recognizing actions and temporal localizing actions. A dataset was formed using the data from six participants. Each frame in the dataset was manually labeled to the clinically relevant action defined in Phase 2. The ground-truth action segments were located by grouping the same adjacent per-frame action labels. The subject identification label was provided manually for assessing the actions of each person in a participant's home. The prediction accuracies of the trained ensemble network on the validation set were 0.819 and 0.838, respectively. The results showed that the ensemble network trained on a dataset with more data tended to provide higher prediction accuracies. The ensemble network sometimes mis-predicted hand manipulation action and the reaching forward action. The view of the system affected the prediction accuracy. The prediction accuracy drops for the model to predict action using the occluded body-view or back-view frames.

The kinematic assessments were performed on recognized actions of the validation set. The statistical analysis of the assessment outcomes showed that stroke participants were less active compared with non-stroke participants. Most of the recognized action types of stroke participants were walking and reaching forward. The speed of the motions of stroke participants was slower than that of non-stroke participants. The over-time trends were visualized for each kinematic measure of each stroke participant.

The limitations of the current DARAS are that training the ensemble network requires the per-frame label and the assessments required a subject identification label. However, it is expensive to manually label the data. In the future, networks with semi-supervised learning approaches can be investigated to reduce the burden of data labeling.

CHAPTER 5 CONCLUSION

For stoke rehabilitation, the more clinicians collect the health information and understand the recovery status of patients, the better customized and personalized treatment plans can be made, which helps patients recover efficiently and return quickly to independent living.

Traditionally, patients are asked to go to clinic to undergo scale-based assessments up to a few hours a week, then are prescribed with in-home exercise. Clinicians acquire patients' information from the in-clinic assessments and from the self-reports of in-home exercise. Throughout this process, not only the collected data in clinics are limited, but the measurement approaches, scale-based assessments and even self-reports could be subjective.

In this work, I proposed and developed a toolchain that comprises a kinematic assessment tool designed in a rehabilitation game and a system for daily activity recognition and assessment, aiming to help clinicians to gather sufficient and reliable health information.

5.1 Kinematic assessment in a rehabilitation game

The proposed kinematic assessment tool was designed for a rehabilitation game called Mystic Isle which is a Kinect-based virtual reality video game that targets balance training and provides upper limb reaching exercises for people with orthopedic and neurological injury or impairments, including stroke. The game tracked the 3D positions of joints and exported the skeletal joint data in a rate of either 15 or 30 frames per second (fps).

To prove the spatial accuracy and the measurement validity of the data collection on Kinect V2, I propose to compare the data collected by Kinect V2 with those collected from a gold standard motion capture system named Vicon™. The Vicon system uses markers on body and is proved to be accurate. Thirty healthy participants were recruited to play six pre-designed game trials that cover sitting and standing full-body movements.

For the spatial accuracy of tracking joints, the data of the arm joints were highly correlated between both systems with high SNR values. For the joints of the hip, although they reported less number of correlations over 0.90 with lightly lower SNR values, we still found data from two systems are closely correlated. Thus, we conclude that the Kinect V2 is sufficiently accurate for tracking movements, and the clinical measurements we obtained (e.g., extent of reach) are repeatable and valid, which supports the use of the Kinect V2 in a clinical rehabilitation setting.

To explore the use of VR-based rehabilitation game to assess upper extremity movement for post-stroke individuals, I aggregated and analyzed movement data captured with Kinect V2 from four separate studies including 8 post-stroke individuals and 30 individuals without disabilities. Kinematic measurements, normalized jerk, movement path ratio, and average path sway, were used to evaluate the smoothness and efficiency of the hand movements. Data from the 30 healthy individuals created a normative baseline for the three kinematic variables. The assessment outcomes of individuals post-stroke were compared with the normative baseline.

For the mild stroke population, these kinematic variables are potentially sensitive to performance impairments that are even unobservable with a trained therapist eye.

Individuals post-stroke were less efficient and had more jerky movements in both upper extremities as compared with healthy individuals.

The founding showed that it is feasible to use a movement sensor paired with a VR-based intervention to quantify and qualify upper extremity movement for post-stroke individuals.

5.2 Daily activity recognition and assessment system (DARAS)

DARAS comprises three main parts. The data logging system part, action recognition and temporal localization part and action assessment part. In the data logging system part, different depth sensors have been investigated to provide an efficient and convenient way to collect daily motion data. Two versions of action recognition algorithms have been explored to acquire accurate action recognition results from both manually segmented videos and real-life unsegmented video streams. The assessments were performed for each recognized action using joint data.

Multiple depth sensing systems, including the Kinect sensor, the VicoVR sensor, the TVico system, and the Foresite system, were investigated. Among them, the Foresite system was selected for the action logging system, because the Foresite system not only provides the accurate depth and skeletal joint data, but also make the logger system stable, reliable, and portable. In addition, the logger has supporting functions for monitoring the status of the data collection, memory usage, and data transmission.

The ability to log the daily actions has been proved by an IRB-approved data collection from 16 stroke participants. Over a 3-month period, each subject will provide approximately 120GB of data for training the proposed action recognition and temporal localization network and for assessments.

The depth videos and skeletal joint data of daily activities were collected by using the action logging app. Specifically, the data included many continuous untrimmed actions. According to the need of clinicians, it is desired to qualitatively evaluate clinically relevant actions, such as the arm reaching above the head. To perform such kinematic assessments, the specified actions need to be recognized and segmented first.

It is a challenging task to recognize and segment clinically relevant action from a realistic environment. To make accurate prediction outcomes, an ensemble network has been proposed and implemented for action recognition and temporal action localization using collected different data sources. The ensemble network includes three networks which are CDC, R-C3D and R-HCN. The CDC and R-C3D learn the spatial-temporal motion features from depth data, and the R-HCN network extracts features from joint data. The prediction results from all three subnetworks were fused to make the final prediction. The prediction accuracies of the trained ensemble network on the validation set were 0.819 and 0.838, respectively. The results also showed that the ensemble network trained on a dataset with more data tended to provide higher prediction accuracies.

The kinematic assessments were performed on recognized actions of the validation set. The assessment measures include extent of reach, speed, smoothness of hand motions. The statistical analysis of the assessment outcomes showed that stroke participants were less active compared with non-stroke participants. Most of the recognized action types of stroke participants were walking and reaching forward. The speed of the motions of stroke participants was slower than that of non-stroke participants. The longitudinal analysis was visualized for each kinematic measure of each stroke participant.

This work demonstrates that the proposed kinematic assessment toolchain provides an objective and sensitive approach to quantitatively evaluate the motions from stroke. The movements in the Mystic Isle rehabilitation game can be tailored to each post-stroke individual. With the help of the assessments in the game, clinicians can track the quality of the predesigned motions longitudinally. With the help of the DARAS, clinicians can gather sufficient information from daily behaviors. As a result, a more personalized treatment plan can be provided for each post-stroke individual with the help from the assessment outcome from the toolchain.

5.3 Future work

Though the DARAS has been proved to monitor daily motions of a stroke participant, recognize and segment the clinically relevant actions, and finally perform kinematic assessment on the recognized actions, the system can be improved from the following aspects.

1. **Process the data from the rest of the stroke participants in the IRB study**. The action recognition and temporal localization results showed that the ensemble network trained on a larger dataset is likely to have a higher prediction accuracy. A more generalized model can be created by training the network on a dataset with more participants. 16 stroke participants were recruited in the IRB study. Currently, data from six participants have been utilized for developing a network and for assessments. With more data processed, we can not only build a more accurate and

generalized action recognition and temporal localization model but also, we can perform kinematic assessments on action motions for more stroke participants. As a result, potential behavior patterns for stroke participants can be found via assessments.

2. **Investigate algorithms to reduce the workload of data labeling**. Processing the raw collected data involves converting the binary data to a readable format and providing the per-frame action labels. It usually takes at least one second to label a frame. The data collection rate is at least 6 fps. So, it takes at least 6 minutes to label the data collected within a minute. The action logger system is running 24/7 for 3 months for a stroke participant. Thus, the data labeling process is the most time-consuming task. Algorithms that don't require per-frame labels, such as active learning or semi-supervised learning, can dramatically reduce the workload of data labeling.
3. **Stroke participant identification algorithms**. Before performing kinematic assessments on the motions of a stroke participants, it is necessary to identify the stroke participant and select the action segments performed by the stroke participant. Currently, I manually labeled the person in each recognized action and grouped the action segments by different persons in a kitchen. A stroke participant identification algorithm is required to automatically provide subject-level label to the participant in each recognized action segment. Then, the kinematic assessment can be conducted on each stroke individual.

CHAPTER 6 PROJECT PROGRESS

Table 29. Project progress summary

Mystic Isle Stroke Game Assessment			
Task	*Started*	*Progress*	*Completed*
Skeletal joint data pre-processing (transformation, filling missing samples, filtering, cropping) [40]	✓	100%	✓
Outlier detection	✓	100%	✓
Definition of upper-extremity assessment metrics [26, 40]	✓	100%	✓
Statistical analysis on assessment results [26]	✓	100%	✓
Daily Activity Recognition and Assessment System (DARAS)			
Task	*Started*	*Progress*	*Completed*
System Development			
Kinect-based action logging system using a Windows computer [129]	✓	100%	✓
Wireless VicoVR-based action logging system using an android device [130]	✓	100%	✓
Wireless Tvico-based action logging system	✓	100%	✓
Foresite-based action logging system	✓	100%	✓
Action Recognition and Temporal Localization Algorithm			
Action recognition on manually segmented videos using HON4D algorithm [129]	✓	100%	✓
Action localization and recognition on continuous videos using a customized CDC algorithm [130]	✓	100%	✓
New algorithm to improve the action recognition accuracy	✓	100%	✓
Refine the model for stroke patients' actions	✓	100%	✓
Action Assessment			
Skeletal joint data pre-processing [40]	✓	100%	✓
Outlier detection	✓	100%	✓
Assessment using defined metrics [26, 40]	✓	100%	✓
Perform over-time assessment for each subject	✓	100%	✓
Evaluation			
Evaluation of confidence level	✓	100%	✓
Test the system in a simulated kitchen on healthy subjects [129, 130]	✓	100%	✓
Test the system in real kitchens on healthy subjects [130]	✓	100%	✓
Test the system in real kitchens on stroke subjects	✓	100%	✓

CHAPTER 7 PUBLICATIONS

7.1 Published papers

M. Ma, R. Proffitt, and M. Skubic, "Quantitative Assessment and Validation of a Stroke Rehabilitation Game," in 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017, pp. 255-257.

Abstract: We explore a quantitative assessment for a Microsoft Kinect-based stroke rehabilitation virtual reality (VR) video game, Mystic Isle, by evaluating three assessment metrics of player hand movement- maximum range (extension), peak velocity and mean velocity. We also analyze the left-right hand symmetry by visualizing trajectories of both hands throughout the game. Assessment metrics obtained by the Kinect-based game have been validated using a Vicon motion capture system. The percentage errors of maximum range and mean velocity are less than 10%. The peak velocity metric is more sensitive to noise and sampling rate with a percentage error up to 18%.

J. Collins, J. Warren, M. Ma, R. Proffitt, and M. Skubic, "Stroke patient daily activity observation system," in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 844-848.

Abstract: Stroke is a leading cause of long-term adult disability. Stroke patients can recover through rehabilitation programs prescribed by occupational therapists (OT); however, an individualized rehabilitation program can reduce recovery times compared to traditional ones. In this paper, we propose a daily activity observation system (DAOS) that uses a Kinect v2 sensor to collect and retrieve motion data. The DAOS has a robust interface to extract depth and skeleton data and supports data collection in an unstructured kitchen environment. Depth data are used to perform action recognition and track problematic movements, while skeleton data are used to calculate mean velocities of hand joints, max extensions, symmetry of hand movements, and other assessment metrics for therapists. Histogram of oriented 4D normals is used for action recognition. The action recognition accuracy is 97% on a multi-class kitchen action dataset. Through action recognition and accurate assessment, we present a novel system that can assist therapists and their ability to provide quality care to stroke patients.

M. Ma, R. Proffitt, and M. Skubic, "Validation of a Kinect V2 based rehabilitation game," PLOS ONE, vol. 13, no. 8, p. e0202338, 2018.

Abstract: Interactive technologies are beneficial to stroke recovery as rehabilitation interventions; however, they lack evidence for use as assessment tools. Mystic Isle is a multi-planar full-body rehabilitation game developed using the Microsoft Kinect® V2. It aims to help stroke patients improve their motor function and daily activity performance and to assess the motions of the players. It is important that the assessment results generated from Mystic Isle are accurate. The Kinect V2 has been validated for tracking lower limbs and calculating gait-specific parameters. However, few studies have validated the accuracy of the Kinect® V2 skeleton model in upper-body movements. In this paper, we evaluated the spatial accuracy and measurement validity of a Kinect-based game Mystic Isle in comparison to a gold-standard optical motion capture system, the Vicon system. Thirty participants completed six trials in sitting and standing. Game data from the Kinect sensor and the Vicon system were recorded simultaneously, then filtered and sample rate synchronized. The spatial accuracy was evaluated using Pearson's r correlation coefficient, signal to noise ratio (SNR) and 3D distance difference. Each arm-joint signal had an average correlation coefficient above 0.9 and a SNR above 5. The hip joints data had less stability and a large variation in SNR. Also, the mean 3D distance difference of joints were less than 10 centimeters. For measurement validity, the accuracy was evaluated using mean and standard error of the difference, percentage error, Pearson's r correlation coefficient and intra-class correlation (ICC). Average errors of maximum hand extent of reach were less than 5% and the average errors of mean and maximum velocities were about 10% and less than 5%, respectively. We have demonstrated that Mystic Isle provides accurate measurement and assessment of movement relative to the Vicon system.

M. Ma, B. J. Meyer, L. Lin, R. Proffitt, and M. Skubic, "VicoVR-Based Wireless Daily Activity Recognition and Assessment System for Stroke Rehabilitation," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 1117-1121.

Abstract: Stroke is the leading cause of long-term disability. Stroke patients can recover faster with personalized therapy treatments. This requires both clinical assessments and in-home assessments of daily activities. In this paper, we propose a daily activity recognition and assessment system for stroke patients. Our system is able to classify daily activities in real home environments and quantitatively evaluate upper body motions while preserving privacy by utilizing depth videos. Specifically, our system collects the depth videos and skeletal joint data of daily activities using a VicoVR sensor. It then recognizes and localizes clinically relevant actions from continuous untrimmed depth videos using a customized convolutional de-convolutional network. In addition, it assesses the extent of reach and speed metrics of both hands using skeletal joint data. The system has been tested on simulated cooking videos and real-life cooking videos in various kitchens with different room layouts and light conditions. The action recognition accuracies for simulated and real-life videos can reach 90.9% and 87.5%, respectively. With the valuable assessment feedback of our system, therapists can make better personalized treatments for stroke patients.

Z. Moore, C. Sifferman, S. Tullis, M. Ma, R. Proffitt, and M. Skubic, "Depth Sensor-Based In-Home Daily Activity Recognition and Assessment System for Stroke Rehabilitation," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 18-21 Nov. 2019 2019, pp. 1051-1056.

Abstract: Stroke is a leading cause of long-term adult disability. Many stroke patients participate in rehabilitation programs prescribed by an occupational therapist to aid in recovery; however, occupational therapists rely on in-clinic assessments and often-unreliable self-assessments at home to track a patient's progress, limiting their ability to monitor how patients perform outside of a clinical setting. Our Daily Activity Recognition and Assessment System collects depth and skeletal data passively from within the patient's home to assess long-term recovery and provide metrics to an occupational therapist to allow for more individualized rehabilitation plans. Using data from a wall-mounted depth sensor, we adapt a hierarchical co-occurrence network to identify actions from pre-segmented skeletal data. We then perform assessments on the classified actions to track key recovery metrics: normalized jerk, speed of motions, and extent of reach. We also introduce novel filters to identify high quality data for analysis. Our sensor was installed in a stroke patient's kitchen for seven days, generating the first action recognition data set from a stroke patient in a naturalistic environment. We use this data in conjunction with the NTU-RGB-D data set to validate our recognition and assessment algorithms. We achieved 90.1% accuracy by replicating the results of the NTU-RGB-D data set and a maximum of 59.6% accuracy on our kitchen data set.

R. Proffitt, M. Ma, and M. Skubic, "Novel clinically-relevant assessment of upper extremity movement using depth sensors," Topics in Stroke Rehabilitation, pp. 1-10, 2022, doi: 10.1080/10749357.2021.2006981.

Abstract:

Background: For individuals post-stroke, home-based programs are necessary to deliver additional hours of therapy outside of the limited time in the clinic. Virtual reality (VR)-based approaches show modest outcomes in improving client function when delivered in the home. The movement sensors used in these VR-based approaches, such as the Microsoft Kinect® have been validated against gold standards tools but have not been used as an assessment of upper extremity movement quality in the stroke population.

Objectives: The purpose of this study was to explore the use of a movement sensor paired with a VR-based intervention to assess upper extremity movement for individuals post-stroke.

Methods: Movement data captured with the Microsoft Kinect® from four separate studies were aggregated for analysis (n = 8 individuals post-stroke, n = 30 individuals without disabilities). For all participants, the skeletal data (x, y, z coordinates for 15 tracked joints) for each game play session were processed in

MatLab and movement variables (normalized jerk, movement path ratio, average path sway) were calculated using an OPTICS density-based cluster algorithm.

Results: Data from the 30 healthy individuals created a normative baseline for the three kinematic variables. Individuals post-stroke were less efficient and had more jerky movements in both upper extremities as compared to healthy individuals.

Conclusion: It is feasible to use a movement sensor paired with a VR-based intervention to quantify and qualify upper extremity movement for individuals post-stroke.

7.2 Planned papers

An ambient in-home activity recognition and assessment system for stroke

In the paper, our newly developed Foresite-based daily activity recognition and assessment system will be presented. The novelty of this paper is that the Foresite system will be introduced to log the daily actions of a stroke participant, and an ensemble network for action recognition and temporal action localization will be proposed. The ensemble network has been trained in a dataset that contains daily action data collected from six stroke participants. The per-frame precision and the per-action precision were 0.819 and 0.838 on the validation set, respectively.

Clinically analysis of the kinematic assessment outcomes on daily actions of post-stroke individuals

In this paper, the kinematic assessment outcomes from the DARAS on the recruited stroke participants will be presented. The statistical analysis of the assessment results showed that stroke participants were less active compared with non-stroke participants. Most of the recognized action types of stroke participants were walking and reaching forward. The speed of the motions of stroke participants was slower than that of non-stroke participants.

APPENDIX

A1. Social/Behavioral/Educational Research Protocol

Project Title: Development and Acceptability of an Ambient In-Home Activity
Assessment Tool for Stroke
IRB Number: 2017864
Version Number: 2
Version Date: 11/13/2020
Principal Investigator: Rachel Proffitt, Marjorie Skubic
Funding Source: National Institutes of Health

I. Research Objectives/Background

1. The purpose of this study is to develop and refine an action recognition algorithm for detecting specific functional activities in the home setting.
2. Nearly 80,000 people each year suffer from a stroke in the U.S. [1]. Moreover, about 50% report hemiparesis which affects their ability to live independently [1]. Traditional rehabilitation for stroke involves patients performing exercises in a hospital or clinic monitored by a therapist [2]. To make rehabilitation treatment effective, it is essential for therapists to personalize and refine rehabilitation plans. This requires monitoring patient health status and recovery progress. To collect such information, the therapist can either observe the patient in the clinic or use patient/caregiver self-report. However, constant visits to the clinic are not convenient for most and self-report is unreliable. Neither provides assessment on how the rehabilitation translates to everyday activities. To accomplish this, a system that can provide daily activity assessment in the home is needed. To our knowledge, there is currently no such system for tracking rehabilitation progress as it relates to everyday movements at home, for example, tracking range of motion of actions in the kitchen.

II. Recruitment Process

1. Potential subjects from the MU Stroke Registry will be contacted by phone. If recruitment numbers are low at month 6 of recruitment, we will mail a postcard to those we have not yet contacted or those we were unable to contact during our initial calls. The postcard will contain brief information about the study and contact information. We will also post an informational flier across MU campus, in the School of Health Professions newsletter, MU Information e-blast, and distribute fliers at MU Outpatient facilities with patient populations of at least 50% stroke.

III. Consent Process

1. If the subject is interested in the study, they will be provided with a consent form to review. A time to consent will be scheduled. During the consent process, the subject will review the consent form with the researcher. They will be asked various questions to validate their comprehension such as "Is your participation voluntary?" and "How long will you be in this study if you agree to participate?". If they fully comprehend the study, they will be asked to sign the consent form. If the participant has expressive aphasia, they will be asked only yes/no questions to assess understanding.

IV. Inclusion/Exclusion Criteria

1. Inclusion criteria:
    a. Over the age of 18
    b. Conversational in English
    c. Able to ambulate with or without an assistive device
    d. At least mild hemiparesis in the arm (NIH Stroke Scale score > 6)
2. Exclusion Criteria:
    a. Under the age of 18
    b. Unable to converse in English
3. The MU Stroke Registry will be the initial point of screening for eligibility (age). The recruitment process will be in English, screening out those who do not converse in English. The NIH Stroke Scale will be used to screen out those without any upper extremity impairments. Participants will be asked over the phone if they are able to ambulate with or without an assistive device.

V. Number of Subjects

1. Anticipated enrollment: 20
2. One week of subject data provides approximately 10GB of data for algorithm training. Over a 3-month period, each subject will provide approximately 120GB of data for training and testing the algorithm. Therefore, a sample of 20 subjects will allow for adequate training and testing, especially if any subjects are outliers.

VI. Study Procedures/Study Design

1. Home Visit 1: The research team will set up the Foresite depth sensor in the subject’s kitchen. The exact placement will be determined by the available space. The research assistant will complete the Demographic Questionnaire and Fugl-Meyer Assessment- Upper Extremity with the participant. The Demographic Questionnaire asks questions about stroke history, daily activities, and experience with technology. The Fugl-Meyer Assessment- Upper Extremity is a standardized clinical assessment of upper extremity motor function. After the sensor is in place, the participant will don an ActiGraph accelerometer on both wrists. The participant will then complete 3 tasks from the Performance Assessment of Self-

Care Skills (PASS) while the sensor and ActiGraph record data. The research team will collect contact information for a designated person in case of a fall. This visit will take no longer than 2 hours.

2. Home Monitoring (3 months): The Foresite depth sensor system will record data over the course of 3 months. The participant will be asked to not deviate from their normal activities within the home. Depth and skeletal data will be transmitted from the Foresite sensor to the Foresite managed secure server via the participant's WiFi/high speed internet.
3. Home Visit 2: The research team will collect all the equipment from the study participant's home. Before removal of the sensor, the participant will don an ActiGraph accelerometer on both wrists. The participant will then complete 4-5 tasks from the Performance Assessment of Self-Care Skills (PASS) while the sensor and ActiGraph record data. This visit will take no longer than 1 hour.
4. Interview: The participant will participate in an interview. The interview will occur via videoconferencing. All interview will be audio-recorded. Total time will be no more than 2 hours.
5. All procedures are research-only.

VII. Potential Risks

1. The subjects may feel that having the sensor in their home is an invasion of privacy. The data will be reviewed with the participant both before and after the study so that they can feel comfortable with the nature of the data collected.
2. Subjects will be provided with researcher contact information and asked to call if any adverse events occur. All unanticipated problems will be reported to the IRB within 5 days of knowledge of the problem or event.
3. Subjects will provide the research team with contact information for a family member/caregiver/friend who will act as a designated point of contact for fall detection. The Foresite sensor system has a built-in fall detection algorithm that is monitored in real-time. Detection of a fall by the system will trigger an alert to the designated individual as well as the research team. The research team will provide assistance to the designated individual in securing emergency services, if necessary. The research team cannot provide direct medical assistance.

VIII. Anticipated Benefits

1. There are no direct benefits to study participants. The potential benefits is that we may be able to develop and ambient sensor for activity recognition in the stroke population. This could add to our outcome assessments in this population and gain a glimpse into the home environment.

IX. Compensation

1. Participants will be provided with $100 as compensation for their time and effort. Participants will receive $50 at the completion of the 3-month in-home data collection. Participants will receive a second $50 at the completion of the interview.

XI. Costs

1. This study is funded by an NIH R21 grant through the *Eunice Kennedy Shriver* National Institute on Child Health and Human Development. All study costs are covered by the grant. There are no costs to the participants.

X. Data Safety Monitoring Plan

1. N/A- not an intervention study

XI. Multiple Sites

1. N/A- single site

XII. References

1. R. Proffitt and B. Lange, “Considerations in the efficacy and effectiveness of virtual reality interventions for stroke rehabilitation: Moving the field forward,” Physical Therapy, vol. 95, no. 3, p. 441448, 2014.
2. K. Nair and A. Taly, “Stroke rehabilitation: traditional and modern approaches,” Neurology India 50, pp. S85 – 93, 2002.

A2. Data collection and participants

It is essential to have a dataset with stroke patients' data activities to train an action recognition and temporal localization network. However, there is no public dataset which includes any stroke individuals. As a result, we conducted an IRB approved data collection to collect stroke individual's kitchen-related daily actions. The IRB is included in Appendix 9.1.

sixteen stroke participants were recruited in the study. The demographics of participants is listed in Table 29 and the device information is shown in Table 30.

Table 30. The demographics of the recruited participants

Participant	Gender	Age*	Time*	Hand dominance	Impaired side	FMA-UE
1	F		1 year 6 months	R	R	31
2	F	57	6 year 7 months	L	R	40
3	M	59	2 years 2 months	R	L	36
4	M	50	2 years 10 months	R	R	26
5	F	73	1 year 6 monts	R	L	63
6	M	64	1 year 7 months	R	R	57
7	F	76	1 year 7 months	R	R	47
8	F	49	2 years 5 months	R	L	15
9	F	78	1 year 10 months	R	R	54
10	M	64	12 years 9 months	R	R	11
11	F	82	7 years 7 months	R	R	59
12	F	71	1 year 11 months	R	L	47
13	F	39	5 years 4 months	R	L	64
14	F	61	3 years 5 months	R	L	62
15	M	77	3 years 2 months	R	L	56
16	M	45	4 years 3 months	R	R	54

*Age means the age of a participant at time of Install. The time means the time since the first stroke

A Foresite DS5 systems was installed in each participant's kitchen to collect data from three month. If the devices had the online access, the data were eventually synchronized to the HTC server. The specific directory is listed below:

/storage/htc/eldercare/data/sensordata/cert/<device id>/kinect/Data1/daraslogger

Otherwise, the data were stored locally and manually uploaded to the shared folder in OneDrive.

Table 31. Data collection log of the R21 Participants

Participant	Device serial number	Install date	Removal date	Data destination
1	DS5-RN-18000192	6/24/20	10/2/20	online
2	DS5-RN-18000289	7/20/20	10/12/20	online
3	DS5-RN-19000744	8/11/20	11/22/20	online
4	DS5-RN-19000725	11/6/20	3/2/21	local
5	DS5-RN-18000192	12/4/20	3/26/21	online
6	DS5-RN-19000743	12/11/20	3/29/21	online
7	DS5-RN-18000289	12/18/20	4/16/21	local
8	DS5-RN-18000191	3/29/21	6/29/21	online
9	DS5-RN-19000743	4/20/21	8/23/21	online
10	DS5-RN-18000190	6/15/21	10/1/21	online
11	DS5-RN-17000192	7/17/21	10/20/21	online
12	DS5-RN-18000191	7/6/21	10/20/21	local
13	DS5-RN-19000743	9/18/21	2/20/22	online
14	DS5-RN-18000289	9/30/21	4/5/22	local
15	DS5-RN-18000298	10/13/21	4/5/22	online
16	DS5-RN-19000743	4/6/22	10/12/22	local

The data were not successfully transferred to the desired cloud location of participant six due to upgrade the data transmission approach. As a result, the data from the participant six were missing. Totally, data from 15 participants were collected and stored properly.

A3. Data preprocessing and labeling

The DARAS logger outputs depth data as compressed binary files to increase the data sample rate and save the memory storage. The binary files need to be converted to image frames as both png or jpeg formats for training the networks. Thus, after retrieving the data from the cloud storage, the first step is to convert the binary files to image files.

The per-frame action labels are required for training the proposed ensemble network, and the subject identification labels for each actions segment are required for performing kinematic assessment for each individual at each home. The per-frame label files contain the frame name and the corresponding action category for each frame. In this study, we are focusing the single-person action recognition. Thus, the multiple-person frames were filtered out while labeling frames. To speed up the labeling process, undergraduate students and students from the Occupational Therapy department have helped to label per-frame action. I also reviewed the action label files, because of the definition of the boundary of action. We labeled a sequence of motions as a reaching action starting from raising an arm to putting down the arm. It can be subjective process of determining the start and end frame. I reviewed the labeled file to ensure the definition of the start and end frame is consistent for the whole dataset. An action-label review tool was developed to easily view a frame and its corresponding label. The label can be quickly modified in the tool if it is necessary.

Table 32. The summary of the labeled data.

Participant	Dates
1	All
2	All
3	All
4	All
5	2020-12-4 to 2021-2-28
10	2021-6-16 to 2021-6-24

To perform the kinematic assessment for each individual, especially the stroke participant, the individual identification label is required. Two label tools have been developed to speed up the labeling process. After the per-frame action labels were ready, the action segments with their corresponding action category labels were generated. I provided the identification label for each action segment manually. The labeled data were summarized in Table 31.

A4. DARAS logger and the supporting software

Scripts for data processing

- **Synchronize the data.**

  Synchronize the collected data from the HTC server to the local storage.

- **Group data by type.**

  Automatically separate the data by types to different folder.

- **Batch binary to png converter.**

  Convert all the binary files in a given folder to png frames

- **Select one-person frames.**

  Loop through all the frames of a participant and select single-person frames only.

- **Rotate depth frames.**

  Rotate the depth frames if they are up-side down.

- **Joint visualization.**

  Visualize the joint data as an avatar

- **Bone length filter.**

  Evaluate the standard deviate of the bone-length different among adjacent frames for a segment.

Software for data labeling

A tool to review the label files. After the label file and the corresponding depth frame folder are imported to the tool, the selected frame and its label will be display in the interface. Users can loop backward or forward of the frames in the selected folder. The label of a specific frame or the labels of a batch of frames can be modified in the label tool.

Source code of the ensemble network

CDC: https://github.com/pgtgrly/Convolution-Deconvolution-Network-Pytorch

R-C3D: https://github.com/sunnyxiaohu/R-C3D.pytorch

HCN: https://github.com/huguyuehuhu/HCN-pytorch

Source code of assessment

It includes the source code of bone-length filter, per-action assessment tool, and scripts of statistical analysis.

All the scripts and source code can be found in the GitLab repo of the DARAS project.

Supporting documents

The supporting documents of my dissertation have been uploaded to the following shared folder on

OneDrive: R21_daras_data/Dissertation_supporting_documents_Mengxuan_Ma.

Data labels

The per-frame labels and subject id labels can be found in the following shared folder: *R21_daras_data/labels_of_participant1-5and10*.

A5. Action recognition and temporal localization of each participant

The performance of the proposed ensemble network was also evaluated by each participant using the per-frame precision and per-action precision. The ensemble network model with the best test results in the experiment was selected to predict and segment actions from the validation set of each participant. The per-frame precision and per-action precision are listed in Table 32, and the confusion matrices are presented in Figure 52 to Figure 55.

Table 33. Per-frame and per-action precisions on the validation sets of each participant

Precision	P1	P2	P3	P4	P5	P10
Per frame	0.845	0.821	0.792	0.652	0.786	0.865
Per action	0.870	0.850	0.810	0.556	0.837	0.851

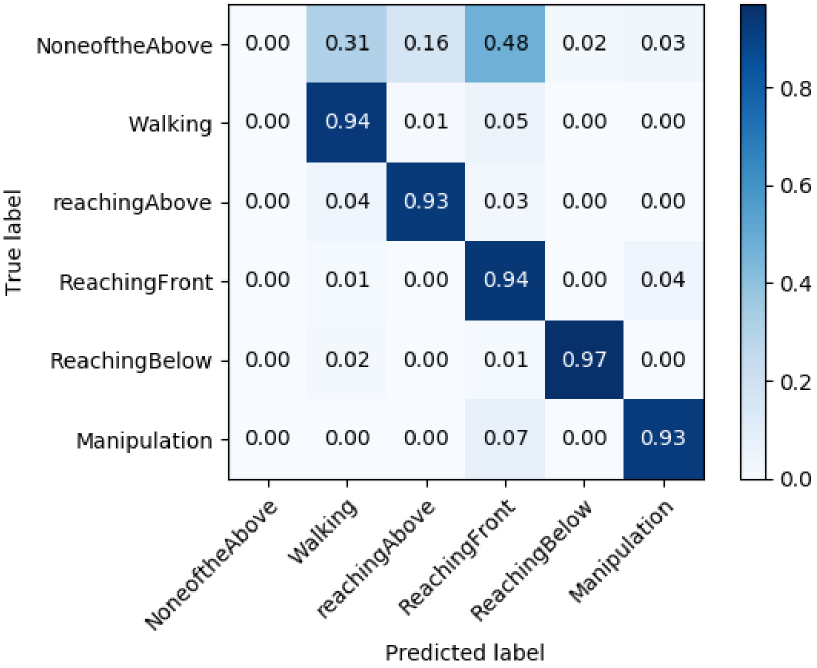


Figure 52. Confusion matrix of the validation set of the participant 1.

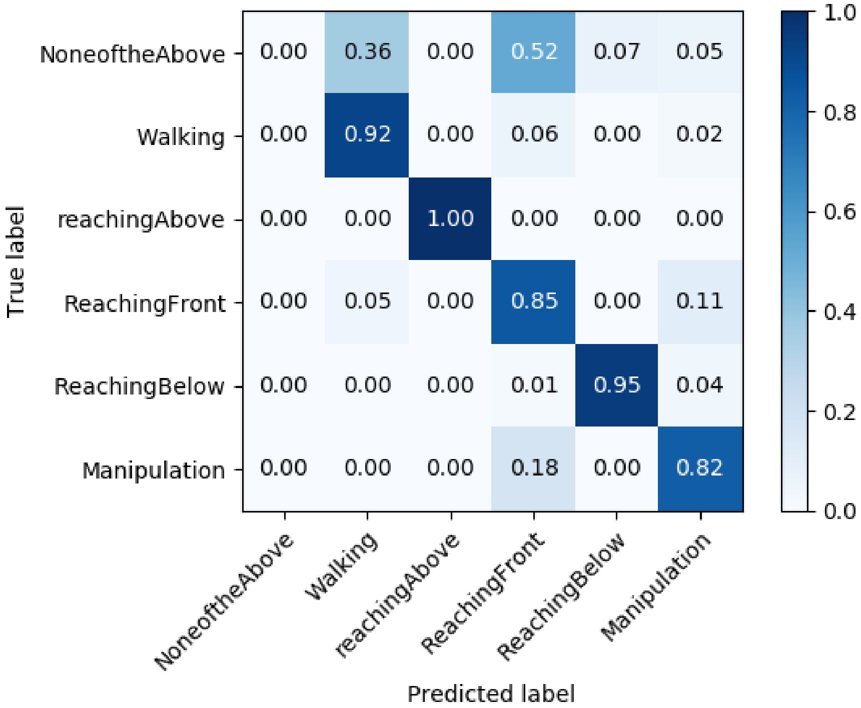


Figure 53. Confusion matrix of the validation set of the participant 2.

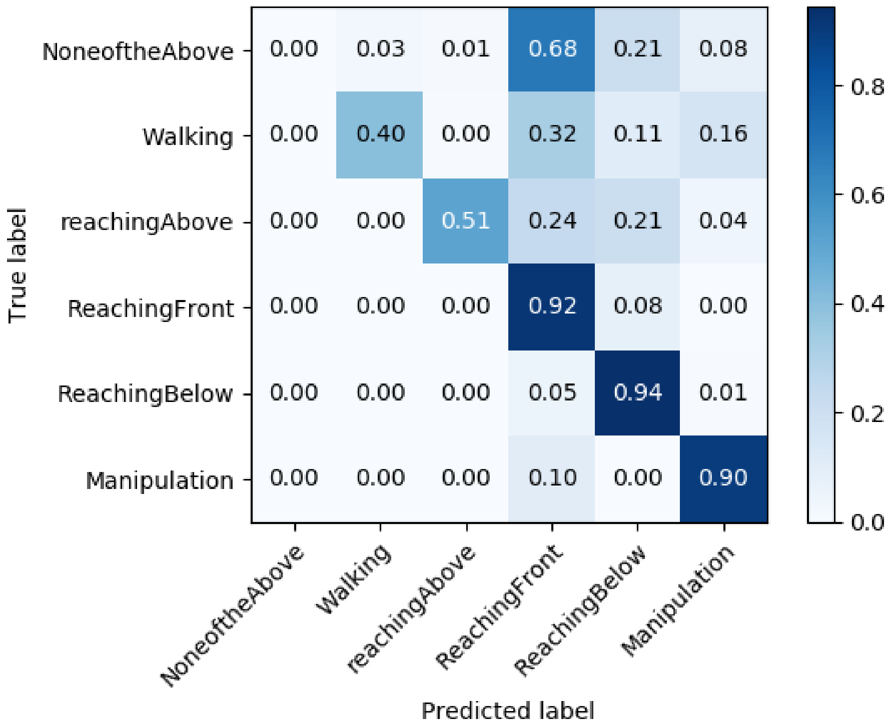


Figure 54. Confusion matrix of the validation set of the participant 3.

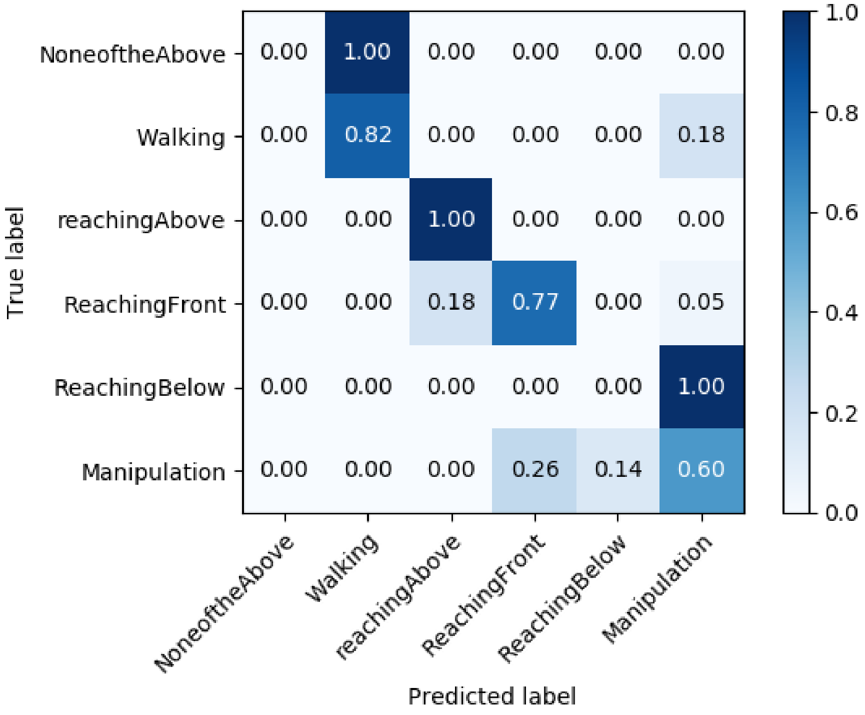


Figure 55. Confusion matrix of the validation set of the participant 4.

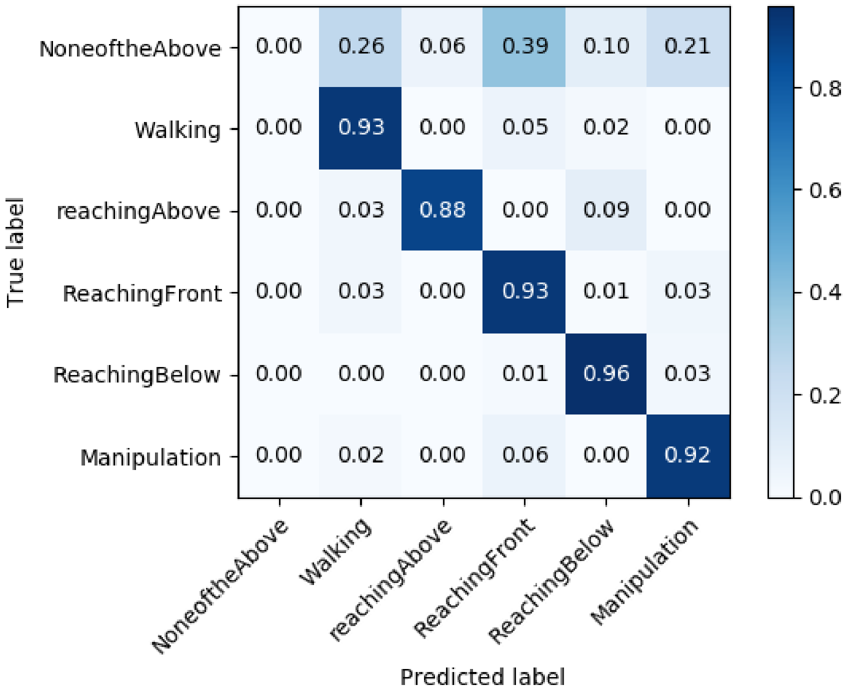


Figure 56. Confusion matrix of the validation set of the participant 5.

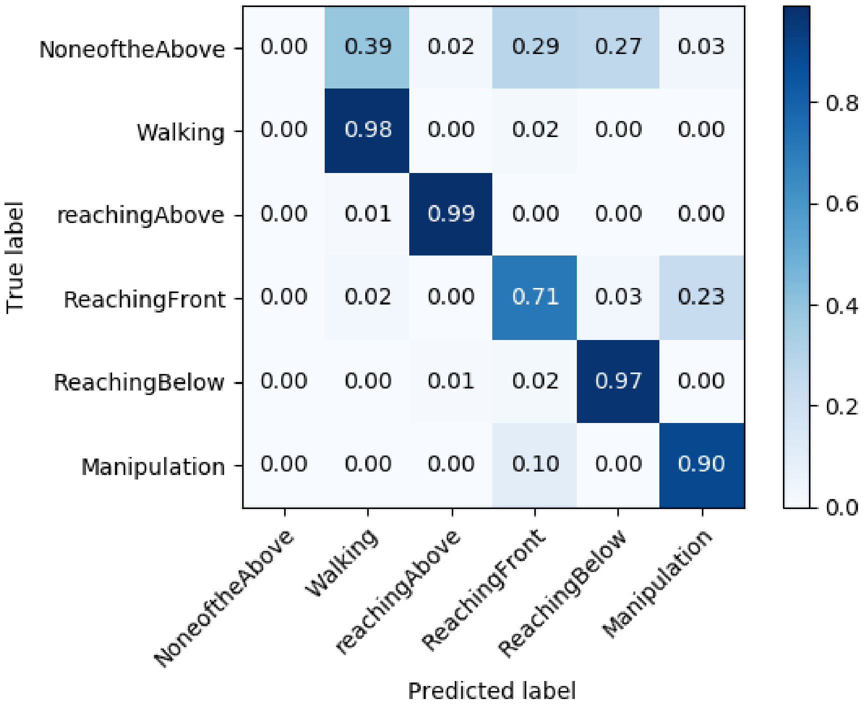


Figure 57. Confusion matrix of the validation set of the participant 10.

REFERENCES

[1] C. W. Tsao *et al.*, "Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association," *Circulation,* vol. 145, no. 8, pp. e153-e639, 2022/02/22 2022, doi: 10.1161/CIR.0000000000001052.

[2] S. S. Virani *et al.*, "Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association," (in eng), *Circulation,* vol. 143, no. 8, pp. e254-e743, Feb 23 2021, doi: 10.1161/cir.0000000000000950.

[3] A. de los Reyes-Guzman, I. Dimbwadyo-Terrer, F. Trincado-Alonso, F. Monasterio-Huelin, D. Torricelli, and A. Gil-Agudo, "Quantitative assessment based on kinematic measures of functional impairments during upper extremity movements: A review," *Clin Biomech (Bristol, Avon),* vol. 29, no. 7, pp. 719-27, Aug 2014, doi: 10.1016/j.clinbiomech.2014.06.013.

[4] C. f. D. C. a. Prevention, "Outpatient rehabilitation among stroke survivors--21 States and the District of Columbia, 2005," (in eng), *MMWR Morb Mortal Wkly Rep,* vol. 56, no. 20, pp. 504-7, May 25 2007.

[5] C. J. Winstein *et al.*, "Guidelines for Adult Stroke Rehabilitation and Recovery: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association," *Stroke,* vol. 47, no. 6, pp. e98-e169, Jun 2016, doi: 10.1161/STR.0000000000000098.

[6] M. T. Jurkiewicz, S. Marzolini, and P. Oh, "Adherence to a home-based exercise program for individuals after stroke," *Top Stroke Rehabil,* vol. 18, no. 3, pp. 277-84, May-Jun 2011, doi: 10.1310/tsr1803-277.

[7] J. H. Morris and B. Williams, "Optimising long-term participation in physical activities after stroke: exploring new ways of working for physiotherapists," *Physiotherapy,* vol. 95, no. 3, pp. 228-34, Sep 2009, doi: 10.1016/j.physio.2008.11.006.

[8] S. Nicholson *et al.*, "A systematic review of perceived barriers and motivators to physical activity after stroke," *Int J Stroke,* vol. 8, no. 5, pp. 357-64, Jul 2013, doi: 10.1111/j.1747-4949.2012.00880.x.

[9] S. A. Billinger *et al.*, "Physical activity and exercise recommendations for stroke survivors: a statement for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke,* vol. 45, no. 8, pp. 2532-53, Aug 2014, doi: 10.1161/STR.0000000000000022.

[10] A. Pourmand, S. Davis, D. Lee, S. Barber, and N. Sikka, "Emerging Utility of Virtual Reality as a Multidisciplinary Tool in Clinical Medicine," (in eng), *Games Health J,* vol. 6, no. 5, pp. 263-270, Oct 2017, doi: 10.1089/g4h.2017.0046.

[11] K. R. Anderson, M. L. Woodbury, K. Phillips, and L. V. Gauthier, "Virtual reality video games to promote movement recovery in stroke rehabilitation: a guide for clinicians," *Arch Phys Med Rehabil,* vol. 96, no. 5, pp. 973-6, May 2015, doi: 10.1016/j.apmr.2014.09.008.

[12] Y.-W. Chow, W. Susilo, J. G. Phillips, J. Baek, and E. Vlahu-Gjorgievska, "Video games and virtual reality as persuasive technologies for health care: An overview," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications,* vol. 8, pp. 18-35, 09/01 2017, doi: 10.22667/JOWUA.2017.09.30.018.

[13] K. Baheux, M. Yoshizawa, A. Tanaka, K. Seki, and Y. Handa, "Diagnosis and rehabilitation of hemispatial neglect patients with virtual reality technology," *Technol Health Care,* vol. 13, no. 4, pp. 245-60, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16055973.

[14] J. Broeren, A. Björkdahl, R. Pascher, and M. Rydmark, "Virtual reality and haptics as an assessment device in the postacute phase after stroke," (in eng), *Cyberpsychol Behav,* vol. 5, no. 3, pp. 207-11, Jun 2002, doi: 10.1089/109493102760147196.

[15] F. Abdulsatar, R. G. Walker, B. W. Timmons, and K. Choong, ""Wii-Hab" in critically ill children: a pilot trial," *J Pediatr Rehabil Med,* vol. 6, no. 4, pp. 193-204, Jan 1 2013, doi: 10.3233/PRM-130260.

[16] G. Acar, G. P. Altun, S. Yurdalan, and M. G. Polat, "Efficacy of neurodevelopmental treatment combined with the Nintendo((R)) Wii in patients with cerebral palsy," *J Phys Ther Sci,* vol. 28, no. 3, pp. 774-80, Mar 2016, doi: 10.1589/jpts.28.774.

[17] S. Chanpimol, B. Seamon, H. Hernandez, M. Harris-Love, and M. R. Blackman, "Using Xbox kinect motion capture technology to improve clinical rehabilitation outcomes for balance and cardiovascular health in an individual with chronic TBI," *Archives of Physiotherapy,* vol. 7, no. 1, p. 6, 2017/05/31 2017, doi: 10.1186/s40945-017-0033-9.

[18] C. Chien-Yen *et al.*, "Towards pervasive physical rehabilitation using Microsoft Kinect," in *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, 21-24 May 2012 2012, pp. 159-162.

[19] A. T. Garcia, A. L. d. S. Kelbouscas, L. L. Guimarães, S. A. V. e. Silva, and V. M. Oliveira, "Use of RGB-D Camera for Analysis of Compensatory Trunk Movements in Upper Limbs Rehabilitation," in *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, 20-23 July 2020 2020, vol. 1, pp. 243-248, doi: 10.1109/INDIN45582.2020.9442112.

[20] H. Kolivand, I. Mardenli, and S. Asadianfam, "Review on Augmented Reality Technology," in *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, 7-10 Dec. 2021 2021, pp. 7-12, doi: 10.1109/DeSE54285.2021.9719356.

[21] B. Silva, R. Matos, A. Rehem, and J. Araujo, "A systematic review on the development of interactive environments for rehabilitation from injuries caused by stroke," in *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, 13-16 June 2018 2018, pp. 1-6, doi: 10.23919/CISTI.2018.8399321.

[22] D. Huamanchahua, J. Ortiz-Zacarias, Y. Rojas-Tapara, Y. Taza-Aquino, and J. Quispe-Quispe, "Human Cinematic Capture and Movement System Through Kinect: A Detailed and Innovative Review," in *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 1-4 June 2022 2022, pp. 1-7, doi: 10.1109/IEMTRONICS55184.2022.9795801.

[23] V. M. S. Ltd. "Vicon Life Science." https://www.vicon.com/applications/life-sciences/ (accessed 10/05/2022, 2022).

[24] D. Inc. "TVico Interactive Android Box." https://tvico.io/ (accessed 10/05/2022, 2022).

[25] F. Healthcare. "Intelligent Design for Preventive Healthcare, Foresite." https://www.foresitehealthcare.com/ (accessed 10/10/2022, 2022).

[26] R. Proffitt, M. Ma, and M. Skubic, "Novel clinically-relevant assessment of upper extremity movement using depth sensors," *Topics in Stroke Rehabilitation,* pp. 1-10, 2022, doi: 10.1080/10749357.2021.2006981.

[27] B. Lange *et al.*, "Designing informed game-based rehabilitation tasks leveraging advances in virtual reality," *Disabil Rehabil,* vol. 34, no. 22, pp. 1863-70, 2012, doi: 10.3109/09638288.2012.670029.

[28] C. National Collaborating Centre for Chronic, "National Institute for Health and Clinical Excellence: Guidance," in *Stroke: National Clinical Guideline for Diagnosis and Initial Management of Acute Stroke and Transient Ischaemic Attack (TIA)*. London: Royal College of Physicians (UK), 2008.

[29] A. G. Rudd, A. Bowen, G. R. Young, and M. A. James, "The latest national clinical guideline for stroke," (in eng), *Clin Med (Lond),* vol. 17, no. 2, pp. 154-155, Apr 2017, doi: 10.7861/clinmedicine.17-2-154.

[30] B. H. Dobkin and A. Dorsch, "New evidence for therapies in stroke rehabilitation," *Curr Atheroscler Rep,* vol. 15, no. 6, p. 331, Jun 2013, doi: 10.1007/s11883-013-0331-y.

[31] B. H. Dobkin, "Strategies for stroke rehabilitation," *Lancet Neurol,* vol. 3, no. 9, pp. 528-36, Sep 2004, doi: 10.1016/S1474-4422(04)00851-8.

[32] J. Young and A. Forster, "Review of stroke rehabilitation," *BMJ,* vol. 334, no. 7584, pp. 86-90, Jan 13 2007, doi: 10.1136/bmj.39059.456794.68.

[33] R. W. Teasell, N. C. Foley, S. K. Bhogal, and M. R. Speechley, "An evidence-based review of stroke rehabilitation," *Top Stroke Rehabil,* vol. 10, no. 1, pp. 29-58, Spring 2003, doi: 10.1310/8YNA-1YHK-YMHB-XTE1.

[34] O. World Health, "International classification of functioning, disability and health : ICF," ed. Geneva: World Health Organization, 2001.

[35] A. Luchetti *et al.*, "Multidimensional assessment of daily living activities in a shared Augmented Reality environment," in *2022 IEEE International Workshop on Metrology for Living Environment (MetroLivEn)*, 25-27 May 2022 2022, pp. 60-65, doi: 10.1109/MetroLivEnv54405.2022.9826952.

[36] K. Ashwini, R. Amutha, K. K. Nagarajan, and S. A. Raj, "Kinect based Upper Limb Performance Assessment in Daily Life Activities," in *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 21-23 March 2019 2019, pp. 201-205, doi: 10.1109/WiSPNET45539.2019.9032717.

[37] VicoVR. "Product description and technical data." https://vicovr.com/user-guide/product-description-and-technical-data (accessed 10/05/2022, 2022).

[38] E. Burdet and T. E. Milner, "Quantization of human motions and learning of accurate movements," *Biol Cybern,* vol. 78, no. 4, pp. 307-18, Apr 1998, doi: 10.1007/s004220050435.

[39] R. Proffitt and B. Lange, "Feasibility of a Customized, In-Home, Game-Based Stroke Exercise Program Using the Microsoft Kinect® Sensor," (in eng), *Int J Telerehabil,* vol. 7, no. 2, pp. 23-34, Fall 2015, doi: 10.5195/ijt.2015.6177.

[40] M. Ma, R. Proffitt, and M. Skubic, "Validation of a Kinect V2 based rehabilitation game," (in eng), *PLoS One,* vol. 13, no. 8, p. e0202338, 2018, doi: 10.1371/journal.pone.0202338.

[41] J. Hee-Tae, K. Hwan, J. Jugyeong, J. Bomin, R. Taekeong, and K. Yangsoo, "Feasibility of using the RAPAEL Smart Glove in upper limb physical therapy for patients after stroke: A randomized controlled trial," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2017, pp. 3856-3859, Jul 2017, doi: 10.1109/embc.2017.8037698.

[42] D. Zhang, J. Zhou, M. Guo, J. Cao, and T. Li, "TASA: Tag-Free Activity Sensing Using RFID Tag Arrays," *IEEE Transactions on Parallel and Distributed Systems,* vol. 22, no. 4, pp. 558-570, 2011, doi: 10.1109/TPDS.2010.118.

[43] J. Barman *et al.*, "Sensor-enabled RFID system for monitoring arm activity: reliability and validity," (in eng), *IEEE Trans Neural Syst Rehabil Eng,* vol. 20, no. 6, pp. 771-7, Nov 2012, doi: 10.1109/tnsre.2012.2210561.

[44] J. Barman, G. Uswatte, N. Sarkar, T. Ghaffari, and B. Sokal, "Sensor-enabled RFID system for monitoring arm activity in daily life," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 30 Aug.-3 Sept. 2011 2011, pp. 5219-5223, doi: 10.1109/IEMBS.2011.6091291.

[45] L. Wang, T. Gu, X. Tao, and J. Lu, "Toward a Wearable RFID System for Real-Time Activity Recognition Using Radio Patterns," *IEEE Transactions on Mobile Computing,* vol. 16, no. 1, pp. 228-242, 2017, doi: 10.1109/TMC.2016.2538230.

[46] L. Yao *et al.*, "Compressive Representation for Device-Free Activity Recognition with Passive RFID Signal Strength," *IEEE Transactions on Mobile Computing,* vol. 17, no. 2, pp. 293-306, 2018, doi: 10.1109/TMC.2017.2706282.

[47] R. Yared, M. E. Negassi, and L. Yang, "Physical activity classification and assessment using smartphone," in *2018 IEEE 9th Annual Information Technology, Electronics and*

Mobile Communication Conference (IEMCON), 1-3 Nov. 2018 2018, pp. 140-144, doi: 10.1109/IEMCON.2018.8615065.

[48] W. Y. Cheng *et al.*, "Smartphone-based continuous mobility monitoring of Parkinsons disease patients reveals impacts of ambulatory bout length on gait features," in *2017 IEEE Life Sciences Conference (LSC)*, 13-15 Dec. 2017 2017, pp. 166-169, doi: 10.1109/LSC.2017.8268169.

[49] A. Coni, S. Mellone, J. M. Leach, M. Colpo, S. Bandinelli, and L. Chiari, "Association between smartphone-based activity monitoring and traditional clinical assessment tools in community-dwelling older people," in *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, 7-9 Sept. 2016 2016, pp. 1-4, doi: 10.1109/RTSI.2016.7740575.

[50] S. Mekruksavanich, P. Jantawong, N. Hnoohom, and A. Jitpattanakul, "ResNet-based Network for Recognizing Daily and Transitional Activities based on Smartphone Sensors," in *2022 3rd International Conference on Big Data Analytics and Practices (IBDAP)*, 1-2 Sept. 2022 2022, pp. 27-30, doi: 10.1109/IBDAP55587.2022.9907111.

[51] J. Morales and D. Akopian, "Physical activity recognition by smartphones, a survey," *Biocybernetics and Biomedical Engineering,* vol. 37, no. 3, pp. 388-400, 2017/01/01/ 2017, doi: https://doi.org/10.1016/j.bbe.2017.04.004.

[52] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "A Survey of Online Activity Recognition Using Mobile Phones," *Sensors*, vol. 15, no. 1*,* pp. 2059-2085doi: 10.3390/s150102059.

[53] Xsens. "inertial-sensor-modules." https://www.xsens.com/inertial-sensor-modules (accessed 10/04/2022, 2022).

[54] H. Nguyen, K. Lebel, S. Bogard, E. Goubault, P. Boissy, and C. Duval, "Using Inertial Sensors to Automatically Detect and Segment Activities of Daily Living in People With Parkinson's Disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 26, no. 1, pp. 197-204, 2018, doi: 10.1109/TNSRE.2017.2745418.

[55] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar, "A Survey on Activity Detection and Classification Using Wearable Sensors," *IEEE Sensors Journal,* vol. 17, no. 2, pp. 386-403, 2017, doi: 10.1109/JSEN.2016.2628346.

[56] C. C. Wang *et al.*, "Development of a Fall Detecting System for the Elderly Residents," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, 16-18 May 2008 2008, pp. 1359-1362, doi: 10.1109/ICBBE.2008.669.

[57] M. R. Narayanan, S. R. Lord, M. M. Budge, B. G. Celler, and N. H. Lovell, "Falls management: detection and prevention, using a waist-mounted triaxial accelerometer," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2007, pp. 4037-40, 2007, doi: 10.1109/iembs.2007.4353219.

[58] N. Noury, A. Galay, J. Pasquier, and M. Ballussaud, "Preliminary investigation into the use of Autonomous Fall Detectors," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2008, pp. 2828-31, 2008, doi: 10.1109/iembs.2008.4649791.

[59] M. Song and J. Kim, "An Ambulatory Gait Monitoring System with Activity Classification and Gait Parameter Calculation Based on a Single Foot Inertial Sensor," *IEEE Transactions on Biomedical Engineering,* vol. 65, no. 4, pp. 885-893, 2018, doi: 10.1109/TBME.2017.2724543.

[60] I. C. Gyllensten and A. G. Bonomi, "Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life," (in eng), *IEEE Trans Biomed Eng,* vol. 58, no. 9, pp. 2656-63, Sep 2011, doi: 10.1109/tbme.2011.2160723.

[61] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer

for ambulatory monitoring," *IEEE Transactions on Information Technology in Biomedicine,* vol. 10, no. 1, pp. 156-167, 2006, doi: 10.1109/TITB.2005.856864.

[62] C. Sangil, R. LeMay, and Y. Jong-Hoon, "On-board processing of acceleration data for real-time activity classification," in *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, 11-14 Jan. 2013 2013, pp. 68-73, doi: 10.1109/CCNC.2013.6488427.

[63] H. A. Imran, "Khail-Net: A Shallow Convolutional Neural Network for Recognizing Sports Activities Using Wearable Inertial Sensors," *IEEE Sensors Letters,* vol. 6, no. 9, pp. 1-4, 2022, doi: 10.1109/LSENS.2022.3197396.

[64] J. P. Gomez-Arrunategui, J. J. Eng, and A. J. Hodgson, "Monitoring Arm Movements Post-Stroke for Applications in Rehabilitation and Home Settings," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 30, pp. 2312-2321, 2022, doi: 10.1109/TNSRE.2022.3197993.

[65] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications,* vol. 76, 02/01 2017, doi: 10.1007/s11042-015-3177-1.

[66] Y. Nam, S. Rho, and C. Lee, "Physical Activity Recognition using Multiple Sensors Embedded in a Wearable Device," *ACM Transactions on Embedded Computing Systems (TECS),* vol. 12, 02/01 2013, doi: 10.1145/2423636.2423644.

[67] S. Hafeez, A. Jalal, and S. Kamal, "Multi-Fusion Sensors for Action Recognition based on Discriminative Motion Cues and Random Forest," in *2021 International Conference on Communication Technologies (ComTech)*, 21-22 Sept. 2021 2021, pp. 91-96, doi: 10.1109/ComTech52583.2021.9616668.

[68] A. Doherty *et al.*, "Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity," *The international journal of behavioral nutrition and physical activity,* vol. 10, p. 22, 02/13 2013, doi: 10.1186/1479-5868-10-22.

[69] L. Meng *et al.*, "Automatic Upper-Limb Brunnstrom Recovery Stage Evaluation via Daily Activity Monitoring," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 30, pp. 2589-2599, 2022, doi: 10.1109/TNSRE.2022.3204781.

[70] K. Taylor, U. A. Abdulla, R. J. N. Helmer, J. Lee, and I. Blanchonette, "Activity classification with smart phones for sports activities," *Procedia Engineering,* vol. 13, pp. 428-433, 2011/01/01/ 2011, doi: https://doi.org/10.1016/j.proeng.2011.05.109.

[71] Y. Prathivadi, J. Wu, T. R. Bennett, and R. Jafari, "Robust activity recognition using wearable IMU sensors," in *SENSORS, 2014 IEEE*, 2-5 Nov. 2014 2014, pp. 486-489, doi: 10.1109/ICSENS.2014.6985041.

[72] L. Atallah, B. Lo, R. King, and G. Z. Yang, "Sensor Placement for Activity Detection Using Wearable Accelerometers," in *2010 International Conference on Body Sensor Networks*, 7-9 June 2010 2010, pp. 24-29, doi: 10.1109/BSN.2010.23.

[73] K. Kunze and P. Lukowicz, "Sensor Placement Variations in Wearable Activity Recognition," *IEEE Pervasive Computing,* vol. 13, no. 4, pp. 32-41, 2014, doi: 10.1109/MPRV.2014.73.

[74] R. Saeedi, J. Purath, K. K. Venkatasubramanian, and H. Ghasemzadeh, "Toward seamless wearable sensing: Automatic on-body sensor localization for physical activity monitoring," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* pp. 5385-5388, 2014.

[75] A. Hadjidj, M. Souil, A. Bouabdallah, Y. Challal, and H. Owen, "Wireless sensor networks for rehabilitation applications: Challenges and opportunities," *Journal of Network and Computer Applications,* vol. 36, no. 1, pp. 1-15, 2013/01/01/ 2013, doi: https://doi.org/10.1016/j.jnca.2012.10.002.

[76] K. Connelly *et al.*, "Evaluation framework for selecting wearable activity monitors for research," *mHealth,* vol. 7, pp. 6-6, 2021, doi: 10.21037/mhealth-19-253.

[77] B. Liang and L. Zheng, "A Survey on Human Action Recognition Using Depth Sensors," in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 23-25 Nov. 2015 2015, pp. 1-8, doi: 10.1109/DICTA.2015.7371223.

[78] M. Vrigkas, C. Nikou, and I. Kakadiaris, "A Review of Human Activity Recognition Methods," *Frontiers in Robotics and Artificial Intelligence,* vol. 2, 11/16 2015, doi: 10.3389/frobt.2015.00028.

[79] D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," *The Visual Computer,* vol. 32, no. 3, pp. 289-306, 2016/03/01 2016, doi: 10.1007/s00371-015-1066-2.

[80] I. Laptev and T. Lindeberg, *Space-time interest points*. 2003, pp. 432-439 vol.1.

[81] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 20-25 June 2005 2005, vol. 1, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[82] O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 23-28 June 2013 2013, pp. 716-723, doi: 10.1109/CVPR.2013.98.

[83] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 15-16 Oct. 2005 2005, pp. 65-72, doi: 10.1109/VSPETS.2005.1570899.

[84] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004/11/01 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[85] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, 20-25 June 2011 2011, pp. 3169-3176, doi: 10.1109/CVPR.2011.5995407.

[86] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic Image Networks for Action Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 3034-3042, doi: 10.1109/CVPR.2016.331.

[87] T. Banerjee, J. M. Keller, M. Popescu, and M. Skubic, "Recognizing complex instrumental activities of daily living using scene information and fuzzy logic," *Computer Vision and Image Understanding,* vol. 140, pp. 68-82, 2015/11/01/ 2015, doi: https://doi.org/10.1016/j.cviu.2015.04.005.

[88] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *2013 IEEE International Conference on Computer Vision*, 1-8 Dec. 2013 2013, pp. 3551-3558, doi: 10.1109/ICCV.2013.441.

[89] H. Wang and C. Schmid, "Lear-inria submission for the thumos workshop," presented at the ICCV workshop on action recognition with a large number of classes, 2013, 8.

[90] E. Vahdani and Y. Tian, "Deep Learning-based Action Detection in Untrimmed Videos: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1-20, 2022, doi: 10.1109/TPAMI.2022.3193611.

[91] H. Xu, A. Das, and K. Saenko, "R-C3D: Region Convolutional 3D Network for Temporal Activity Detection," 03/22 2017.

[92] G. Gong, L. Zheng, and Y. Mu, "Scale Matters: Temporal Scale Aggregation Network For Precise Action Localization In Untrimmed Videos," in *2020 IEEE International*

Conference on Multimedia and Expo (ICME), 6-10 July 2020 2020, pp. 1-6, doi: 10.1109/ICME46284.2020.9102850.

[93] L. Wang, Y. Xiong, D. Lin, and L. V. Gool, "UntrimmedNets for Weakly Supervised Action Recognition and Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 6402-6411, doi: 10.1109/CVPR.2017.678.

[94] M. Rashid, H. Kjellström, and Y. Lee, *Action Graphs: Weakly-supervised Action Localization with Graph Convolution Networks*. 2020.

[95] P. Lee, Y. Uh, and H. Byun, *Background Suppression Network for Weakly-supervised Temporal Action Localization*. 2019.

[96] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly Supervised Action Localization by Sparse Temporal Pooling Network," *arXiv e-prints,* p. arXiv:1712.05080, 2017. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2017arXiv171205080N.

[97] P. Nguyen, D. Ramanan, and C. Fowlkes, "Weakly-Supervised Action Localization With Background Modeling," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 Oct.-2 Nov. 2019 2019, pp. 5501-5510, doi: 10.1109/ICCV.2019.00560.

[98] K. Wang *et al.*, "Inertial measurements of free-living activities: assessing mobility to predict falls," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2014, pp. 6892-5, 2014, doi: 10.1109/embc.2014.6945212.

[99] M. Zhang, B. Lange, C. Y. Chang, A. A. Sawchuk, and A. A. Rizzo, "Beyond the standard clinical rating scales: fine-grained assessment of post-stroke motor functionality using wearable inertial sensors," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2012, pp. 6111-5, 2012, doi: 10.1109/embc.2012.6347388.

[100] A. Aguirre, "Evaluating upper-extremity (dys)function using measurement unit technology and its applications to resource-constrained settings," in *2016 IEEE Global Humanitarian Technology Conference (GHTC)*, 13-16 Oct. 2016 2016, pp. 640-645, doi: 10.1109/GHTC.2016.7857346.

[101] A. Merlo *et al.*, "Upper limb evaluation with robotic exoskeleton. Normative values for indices of accuracy, speed and smoothness," (in eng), *NeuroRehabilitation,* vol. 33, no. 4, pp. 523-30, 2013, doi: 10.3233/nre-130998.

[102] H. L. Teulings, J. L. Contreras-Vidal, G. E. Stelmach, and C. H. Adler, "Parkinsonism reduces coordination of fingers, wrist, and arm in fine motor control," (in eng), *Exp Neurol,* vol. 146, no. 1, pp. 159-70, Jul 1997, doi: 10.1006/exnr.1997.6507.

[103] J. Yan, R. Hinrichs, V. Payne, and J. Thomas, "Normalized Jerk: A Measure to Capture Developmental Characteristics of Young Girls' Overarm Throwing," *Journal of Applied Biomechanics,* vol. 16, pp. 196-203, 05/01 2000, doi: 10.1123/jab.16.2.196.

[104] C. E. Lang *et al.*, "Deficits in grasp versus reach during acute hemiparesis," (in eng), *Exp Brain Res,* vol. 166, no. 1, pp. 126-36, Sep 2005, doi: 10.1007/s00221-005-2350-6.

[105] M. K. Rand, Y. Shimansky, G. E. Stelmach, V. Bracha, and J. R. Bloedel, "Effects of accuracy constraints on reach-to-grasp movements in cerebellar patients," (in eng), *Exp Brain Res,* vol. 135, no. 2, pp. 179-88, Nov 2000, doi: 10.1007/s002210000528.

[106] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander, *OPTICS: Ordering Points to Identify the Clustering Structure*. 1999, pp. 49-60.

[107] Microsoft. "JointType Enumeration." https://learn.microsoft.com/en-us/previous-versions/windows/kinect-1.8/hh855342(v=ieb.10) (accessed 10/05/2022, 2022).

[108] Vicon. "Full body modeling with Plug-in Gait." https://docs.vicon.com/display/Nexus213/Full+body+modeling+with+Plug-in+Gait (accessed 10/05/2022, 2022).

[109] M. E. Nixon, A. M. Howard, and Y.-P. Chen, "Quantitative evaluation of the Microsoft KinectTM for use in an upper extremity virtual rehabilitation environment," *2013 International Conference on Virtual Rehabilitation (ICVR),* pp. 222-228, 2013.

[110] M. Ma, R. Proffitt, and M. Skubic, "Quantitative Assessment and Validation of a Stroke Rehabilitation Game," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 17-19 July 2017 2017, pp. 255-257, doi: 10.1109/CHASE.2017.90.

[111] K. Otte *et al.*, "Accuracy and Reliability of the Kinect Version 2 for Clinical Measurement of Motor Function," *PLOS ONE,* vol. 11, no. 11, p. e0166532, 2016, doi: 10.1371/journal.pone.0166532.

[112] T. M. Guess, S. Razu, A. Jahandar, M. Skubic, and Z. Huo, "Comparison of 3D Joint Angles Measured With the Kinect 2.0 Skeletal Tracker Versus a Marker-Based Motion Capture System," (in eng), *J Appl Biomech,* vol. 33, no. 2, pp. 176-181, Apr 2017, doi: 10.1123/jab.2016-0107.

[113] B. F. Mentiplay *et al.*, "Gait assessment using the Microsoft Xbox One Kinect: Concurrent validity and inter-day reliability of spatiotemporal and kinematic variables," (in eng), *J Biomech,* vol. 48, no. 10, pp. 2166-70, Jul 16 2015, doi: 10.1016/j.jbiomech.2015.05.021.

[114] S. Springer and G. Yogev Seligmann, "Validity of the Kinect for Gait Assessment: A Focused Review," (in eng), *Sensors (Basel),* vol. 16, no. 2, p. 194, Feb 4 2016, doi: 10.3390/s16020194.

[115] E. Stone and M. Skubic, "Evaluation of an inexpensive depth camera for in-home gait assessment," *JAISE,* vol. 3, pp. 349-361, 01/01 2011, doi: 10.3233/AIS-2011-0124.

[116] M. M. Horger, "The reliability of goniometric measurements of active and passive wrist motions," (in eng), *Am J Occup Ther,* vol. 44, no. 4, pp. 342-8, Apr 1990, doi: 10.5014/ajot.44.4.342.

[117] D. C. Boone, S. P. Azen, C. M. Lin, C. Spence, C. Baron, and L. Lee, "Reliability of goniometric measurements," (in eng), *Phys Ther,* vol. 58, no. 11, pp. 1355-60, Nov 1978, doi: 10.1093/ptj/58.11.1355.

[118] K. Mitchell, S. B. Gutierrez, S. Sutton, S. Morton, and A. Morgenthaler, "Reliability and validity of goniometric iPhone applications for the assessment of active shoulder external rotation," (in eng), *Physiother Theory Pract,* vol. 30, no. 7, pp. 521-5, Oct 2014, doi: 10.3109/09593985.2014.900593.

[119] R. J. van de Pol, E. van Trijffel, and C. Lucas, "Inter-rater reliability for measurement of passive physiological range of motion of upper extremity joints is better if instruments are used: a systematic review," (in eng), *J Physiother,* vol. 56, no. 1, pp. 7-17, 2010, doi: 10.1016/s1836-9553(10)70049-7.

[120] B. Hotrabhavananda, A. K. Mishra, M. Skubic, N. Hotrabhavananda, and C. Abbott, "Evaluation of the microsoft kinect skeletal versus depth data analysis for timed-up and go and figure of 8 walk tests," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2016, pp. 2274-2277, Aug 2016, doi: 10.1109/embc.2016.7591183.

[121] R. Proffitt and B. Lange, "Considerations in the efficacy and effectiveness of virtual reality interventions for stroke rehabilitation: moving the field forward," (in eng), *Phys Ther,* vol. 95, no. 3, pp. 441-8, Mar 2015, doi: 10.2522/ptj.20130571.

[122] E. H. Cup, W. J. Scholte op Reimer, M. C. Thijssen, and M. A. van Kuyk-Minis, "Reliability and validity of the Canadian Occupational Performance Measure in stroke patients," (in eng), *Clin Rehabil,* vol. 17, no. 4, pp. 402-9, Jul 2003, doi: 10.1191/0269215503cr635oa.

[123] R. M. Proffitt, W. Henderson, S. Scholl, and M. Nettleton, "Lee Silverman Voice Treatment BIG(®) for a Person With Stroke," (in eng), *Am J Occup Ther,* vol. 72, no. 5, pp. 7205210010p1-7205210010p6, Sep/Oct 2018, doi: 10.5014/ajot.2018.028217.

[124] B. Lange, C. Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, "Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2011, pp. 1831-4, 2011, doi: 10.1109/iembs.2011.6090521.

[125] L. Mandon, J. Boudarham, J. Robertson, D. Bensmail, N. Roche, and A. Roby-Brami, "Faster Reaching in Chronic Spastic Stroke Patients Comes at the Expense of Arm-Trunk Coordination," (in eng), *Neurorehabil Neural Repair,* vol. 30, no. 3, pp. 209-20, Mar 2016, doi: 10.1177/1545968315591704.

[126] P. Patel and T. Bhatt, "Task Matters: Influence of Different Cognitive Tasks on Cognitive-Motor Interference During Dual-Task Walking in Chronic Stroke Survivors," *Topics in Stroke Rehabilitation,* vol. 21, no. 4, pp. 347-357, 2014/07/01 2014, doi: 10.1310/tsr2104-347.

[127] T. J. Wolf, C. Baum, and L. T. Conner, "Changing face of stroke: implications for occupational therapy practice," (in eng), *Am J Occup Ther,* vol. 63, no. 5, pp. 621-5, Sep-Oct 2009, doi: 10.5014/ajot.63.5.621.

[128] K. E. Laver, B. Lange, S. George, J. E. Deutsch, G. Saposnik, and M. Crotty, "Virtual reality for stroke rehabilitation," (in eng), *Cochrane Database Syst Rev,* vol. 11, no. 11, p. Cd008349, Nov 20 2017, doi: 10.1002/14651858.CD008349.pub4.

[129] J. Collins, J. Warren, M. Ma, R. Proffitt, and M. Skubic, "Stroke patient daily activity observation system," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 13-16 Nov. 2017 2017, pp. 844-848, doi: 10.1109/BIBM.2017.8217765.

[130] M. Ma, B. J. Meyer, L. Lin, R. Proffitt, and M. Skubic, "VicoVR-Based Wireless Daily Activity Recognition and Assessment System for Stroke Rehabilitation," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3-6 Dec. 2018 2018, pp. 1117-1121, doi: 10.1109/BIBM.2018.8621151.

[131] H. Xia and Y. Zhan, "A Survey on Temporal Action Localization," *IEEE Access,* vol. 8, pp. 70477-70487, 2020, doi: 10.1109/ACCESS.2020.2986861.

[132] R. Proffitt, M. Ma, and M. Skubic, "Daily Activity Recognition and Assessment System for Stroke Rehabilitation," *The American Journal of Occupational Therapy,* vol. 74, no. 4_Supplement_1, pp. 7411500046p1-7411500046p1, 2020, doi: 10.5014/ajot.2020.74S1-PO6729.

[133] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos," 03/04 2017.

[134] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Learning Spatiotemporal Features with 3D Convolutional Networks*. 2015, pp. 4489-4497.

[135] C. Li, Q. Zhong, D. Xie, and S. Pu, *Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation*. 2018, pp. 786-792.

[136] Z. Moore, C. Sifferman, S. Tullis, M. Ma, R. Proffitt, and M. Skubic, "Depth Sensor-Based In-Home Daily Activity Recognition and Assessment System for Stroke Rehabilitation," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 18-21 Nov. 2019 2019, pp. 1051-1056, doi: 10.1109/BIBM47256.2019.8983376.

VITA

Mengxuan Ma was born in Kunming, Yunnan, China. She received her Bachelor of Science degree in Electrical Engineering at Beijing Jiaotong University in 2013. She received her Master of Science degree in 2015 and Doctor of Philosophy degree in 2022 from the Department of Electrical Engineering and Computer Science at University of Missouri-Columbia. She was a graduate research assistant in the Center to Stream HealthCare in Place, in the Department of Electrical Engineering and Computer Science at University of Missouri-Columbia. Her research interests include computational intelligence, machine learning, and computer vision.