VicoVR-based Wireless Daily Activity Recognition and Assessment System for Stroke Rehabilitation

Mengxuan Ma*, Benjamin J. Meyer†, Le Lin‡, Rachel Proffitt§ and Marjorie Skubic¶
*†¶Electrical Engineering & Computer Science, §Occupational Therapy, University of Missouri, Columbia, MO
‡Electrical Engineering & Computer Science, Syracuse University, Syracuse, NY
*mmrnc@mail.missouri.edu, †bjmgy4@mail.missouri.edu, ‡llin19@syr.edu, §proffittrm@health.missouri.edu, ¶SkubicM@missouri.edu

***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.**

Keywords—VicoVR, Wireless, Android, Daily Activity Recognition, Assessment, Stroke Rehabilitation

I. INTRODUCTION

Nearly 800,000 people each year experience strokes in the U.S. [1]. Moreover, about 50% report hemiparesis, or weakness of one side of the body afterwards [1]. Stroke patients can recover through rehabilitation. To make the rehabilitation treatment effective, it is essential for a therapist to personalize and refine the rehabilitation plan for each patient. This requires the therapist to monitor the patient's health status and recovery progress continuously. Traditional rehabilitation involves patients performing exercises in a clinic or at home, monitored by a therapist [2, 3]. A patient usually receives treatment only a few hours per week, and evaluations of progress are typically only done at the beginning and end of an episode of care. As a result, the health information and feedback from the treatment are limited. High demands are placed on the therapist's professional knowledge to identify the most effective and appropriate methods of treatment for the individual patient.

We propose a new daily activity recognition and assessment system (DARAS) using a wireless depth sensor VicoVR to perform activity recognition and assessment in real-home settings. The system is comprised of three modules: a data logging module, an action localization module and an action assessment module. The data logging module was built using a VicoVR and an android application. It records the depth videos and skeletal data of a patient's daily activities. It utilizes a customized convolution-deconvolution neural network [4] which learns the spatial features of videos, and preserves the temporal information and recognizes the actions from untrimmed videos. The action assessment module quantifies the motion performance using evaluation metrics based on skeletal data, such as hand extent of reach and movement speed.

This paper makes three contributions: (1) To the best of our knowledge, we are the first to provide a video-based system to observe the daily activities of a stroke patient in a kitchen and quantitatively evaluate the upper body motions. To reduce the layout size and connection overhead of the system, we utilize the wireless VicoVR sensor paired with a mobile device. (2) The proposed system is tested on realistic video records, while most of the existing activity recognition approaches are trained and tested on datasets collected in well-controlled laboratory environments. (3) This work studies the effective use of temporal action localization of untrimmed depth videos in everyday stroke rehabilitation. Specifically, we customized the convolution-deconvolution (CDC) neural network so that it can be used with untrimmed depth videos as input.

This paper is organized as follows. We survey the related work in Section II. We describe the implementation of the data logging system, the temporal action localization algorithm and the activity assessment in Sections III, IV and V, respectively. We present the experiments and results in Section VI and conclude our work in Section VII.

II. RELATED WORK

A. Daily Activity Recognition

Recognizing daily activities is an important technology in pervasive computing. It benefits many real-life, human-centric problems such as stroke rehabilitation [5]. Various sensors have been investigated to capture and log human activities.

Videos have been widely studied for activity recognition. With the advance of computing ability and the improvement of sensor techniques, various data modalities including RGB data, depth data and skeletal data have been introduced. 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 [6]. Without using deep learning, hand-crafted features need to be extracted. A large set of gradient-based descriptors have appeared for action

This material is based upon work supported by the National Science Foundation under Award Number: CNS-1659134.

recognition. Examples are histogram of oriented gradients (HOG) [7], cuboid descriptor [8] and scale-invariant feature transform (SIFT) [9]. In recent years, with the development of Convolutional Neural Networks, features can be learned and extracted by a network [6]; the deep-learning convolutional features generally outperform the hand-crafted features. Previously, actions were manually segmented for training and testing. In recent years, researchers have started investigating temporal action localization, which detects the start and end of the actions in input video streams.

B. Daily Activity Assessment

Walking and gait measurement are vital metrics for health and rehabilitation assessments. However, the assessment on walking-related motions focuses on lower body movement only, whereas the quality of upper-body movement is also important for patients with stroke. In [20] researchers 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 an incomplete picture of rehabilitation status. Assessing each action category will better depict the rehabilitation status, and thus, is preferred by therapists.

III. Implementtion of Action Logging Module

The action logging module of our daily activity recognition and assessment system (DARAS) records depth and skeletal data from a VicoVR sensor [10]. Fig. 1 shows the module diagram. The main components are a VicoVR sensor and an android-based application. The VicoVR sensor is a Wi-Fi accessory that provides wireless full body and positional tracking to Android devices. To set up a reliable connection, the VicoVR broadcasts the depth data over a private wifi network. 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 [10]. 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. In this test implementation, data were saved to an external SD card on the android device, for transfer offsite, to be used with temporal action localization and assessment.

IV. Action Localization on Depth Videos

The depth videos and skeletal joint data of daily activities were collected by using the action logging app. The collected data consisted of continuous untrimmed video. In order to perform real world assessments, a process of recognizing the specific actions and locating these actions from the untrimmed videos is desired. Shou et al. [4] proposed a convolutional-de-convolutional network to recognize the actions at the frame level. Thus, the recognized actions can be located based on the per-frame action labels.

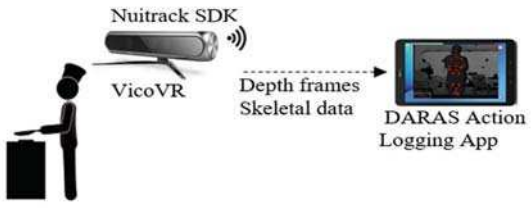

Fig. 1. The diagram of action logging module.

A. Convolutional-De-Convolutional (CDC) network

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) were proposed in [11]. Although the 3D CNN can learn the advanced semantic abstraction of time and space, the output of video time sequence length is decreased. Thus, the fine-grained time has been lost. Shou et al. [4] 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. Thus, it provides the prediction score at each frame, which can be used to locate the actions.

B. CDC networks on depth kitchen videos

The CDC network has been evaluated using THUMOS' 14, an untrimmed RGB sport 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 action videos is desired in DARAS. However, the proposed CDC network was designed for RGB videos. So, we 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 Fig. 2), 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 a 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.

V. Action Assessment

If therapists can track patient progress regularly, care can be adjusted accordingly. Quantitative measures of movement quality are key metrics when reporting on the 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. Each piece of information can be used to track improvement over time, or indicate a decline where intervention is needed.

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 shoulder center is the middle of the left and right shoulder joints. We also calculated maximum and mean velocities for each action. For a healthy user, the ratio between mean and maximum velocity should be close to 1.0, but

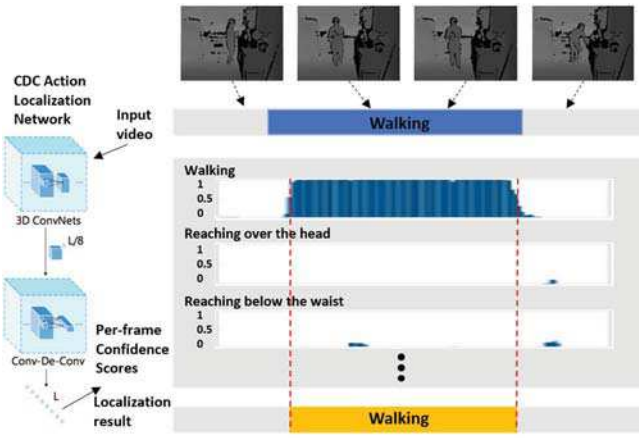

Fig. 2. A framework for positioning of temporal action recognization and localization.

in the presence of movement disorders this metric could detect changes during the movement pattern related to various acceleration and deceleration periods.

VI. Experiments, Results and Analysis

There are three main modules of the DARAS system. The action data logging app, action recognition and action assessment. In this section, we first present the training and prediction results of the CDC network using a simulated kitchen dataset and the recognition results on real-life cooking data. Since there are more upper-body movements among cooking activities in a kitchen, kitchen environments were chosen to test the DARAS system. Finally, we present the assessment results for each recognized action category.

A. Action recognition and localization

a) Datasets: We collected a cooking action dataset in a simulated kitchen to train the action recognition and localization network, CDC, and collected a cooking dataset in real kitchens to test the prediction accuracy.

Simulated kitchen dataset. We collected kitchen action videos in a simulated kitchen to train the CDC network. Though there are public depth-video datasets containing kitchen related actions, they were recorded using the Kinect sensor. Due to the difference of the depth images generated from these two types of sensors, we decided to record a new dataset using the proposed DARAS app to keep consistency of depth format in both training and test datasets. Eleven subjects were recruited to perform three pre-designed action scenarios at least three times. In total, 100 continuous, untrimmed action video sequences were logged. The scenarios of continuous actions are described below:

- Scenario 1: Walk into the kitchen carrying a 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.
- 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 and sit in the chair.

In-home kitchen dataset. We collected real-life cooking videos from three subjects as a test dataset. In order to test the ability of the CDC network in different levels of difficulty, we also designed two types of test sequences performed in the home kitchen, as well as actual cooking.

- Level 1: Test Scenarios. The subject was asked to perform the test scenarios which are exactly the same as the actions recorded in the training set. In this level, we tested whether the system can recognize the actions in videos where only the backgrounds are different.
- Level 2: Action Combinations. The subject was asked to perform actions from the test scenarios in a random sequence. In this level, we tested whether the system can recognize the actions in randomly selected sequences with different backgrounds.
- Level 3: Cooking. We recorded videos when the subject was cooking. One subject made salad and the other subjects made sandwiches for themselves.

Based on the input from occupational therapists, each frame has been labeled to one of these 8 action categories which are background (no user/no classification), walking, sitting, reaching above the head, reaching forward, reaching below the waist, hand manipulation and sweeping.

b) Training and prediction via simulated dataset: We first trained and evaluated the CDC network using the simulated kitchen dataset. Although the CDC filter can be applied to input of arbitrary size, due to the memory limitation, we applied a 32-frame sliding window to segment the videos without overlapping. We then fed each window with per-frame labels into the 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 0.9 and the weight decay was 0.005. We 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 for three times at the same learning rate and the training iteration was 10000. The average per-frame recognition accuracies of different learning rates were shown in Fig. 3. Learning rate 0.001 generated the best per-frame accuracy.

After the best initial learning rate was found, we initialized the learning rate as the optimal value 0.001, and then decreased 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 accuracy was calculated. To test the ability of localizing the actions, the per-frame accuracy was calculated. The recognition and localization results of the simulated dataset are shown in Table I. The normalized confusion matrix of per-action recognition is shown in Fig. 4. The background category

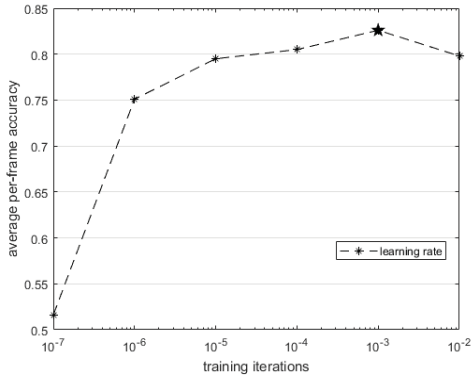

Fig. 3. An experiment of selecting suitable hyperparameter, learning rate. The network was trained on ramdomly selected 90 videos from the simulated dataset and tested on the rest of 10 videos for three times under different learning rates. The avarage per-frame accuracies were calculated. The accuracy was highest when the learning rate was 0.001.

was excluded for per-action accuracy performance. Reaching over head, sitting and hand manipulation categories show exceptional recognition results.

c) Prediction via in-home dataset. The CDC model has been trained and tested using the simulated kitchen dataset. Actions can be detected and localized by grouping the per-frame labels. But our aim is to provide a system to recognize actions from real-life cooking videos. As a result, the in-home kitchen dataset was collected to evaluate the trained model, using actual cooking sequences from three participants in three different kitchens (*Cooking*). One participant prepared a salad, and two made sandwiches. There were many differences between the training videos and the real-life cooking sequences, e.g., the background. In order to investigate which factors affected the recognition accuracy, two additional types of videos were collected: (1) the test scenarios performed in each participant's kitchen (*Test Scenarios*), and (2) a different combination of actions from the test scenarios (*Action Combinations*).

Table II shows the per-frame accuracies and per-action accuracies of *Test Scenarios*, *Action Combinations* and *Cooking* in different kitchens. For each participant, we randomly selected ten videos for *Test Scenarios* and three videos for *Action Combinations*. In addition, the cooking videos we collected last at least two minutes. For action localization, the highest per-frame accuracies of pre-designed actions and cooking actions were 90.7% and 84.3%, respectively. For action detection, the highest per-action accuracies of pre-designed actions and cooking actions were 90.9% and 87.5%, respectively. Comparing the recognition results between the simulated kitchen sets and the test scenario sets, the change of background did not affect the recognition accuracy. However, the performance dropped on the action combinations tests, which indicates that the sequence of action could be a feature of recognition. The normalized confusion matrix of per-action recognition is shown in Fig. 5. The background category was excluded for per-action accuracy evaluation. The recognition accuracies of walking, reaching below the waist and sweeping categories were above 90%.

TABLE I. PER-FRAME ACCURACIES AND PER-ACTION ACCURACIES OF TEST VIDEOS FROM SIMULATED KITCHEN DATASET

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

TABLE II. PER-FRAME ACCURACIES AND PER-ACTION ACCURACIES OF TEST SCENARIO, COMBINATION AND COOKING TEST VIDEOS FROM THREE DIFFERENT KITCHENS.

	Test Scenarios		**Action Combinations**		**Cooking**	
	Per frame	***Per action***	***Per frame***	***Per action***	***Per frame***	***Per action***
Kitchen 1	88.5%	87.3%	79.2%	85.7%	81.0%	80.0%
Kitchen 2	87.9%	89.2%	85.9%	86.4%	84.1%	82.4%
Kitchen 3	90.4%	90.1%	90.7%	90.9%	84.3%	87.5%

B. Assessments

We obtained the action segments by grouping continuous frames of the same per-frame labels together. The assessments were performed for each recognized action using joint data. Specifically, we use the timestamps of the first and last frames of an action to locate the corresponding joint data samples. With five trials selected for each action, we averaged the results of each assessment metric. The assessment outcomes of maximum extent of reach and speed metrics are presented in Table III.

VII. DISCUSSION

The aim of this study is to provide an automatic daily activity recognition and assessment system that can provide sufficient quantitative assessments of daily activities of stroke patients to occupational therapists so that they can design more personalized treatment plans to help patients recover faster. We are the first to design and test a depth video based model in real-life cooking videos. We have shown the assessment results on a set of action categories, which has significant implications for clinical rehabilitation practice. We discuss each of these points below. We conclude with limitations and next steps for research.

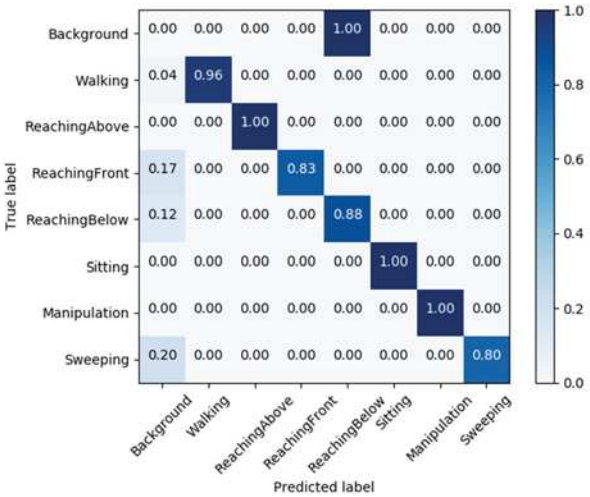

Fig. 4. Normalized confusion matrix of recognizing actions from simulated kitchen dataset.

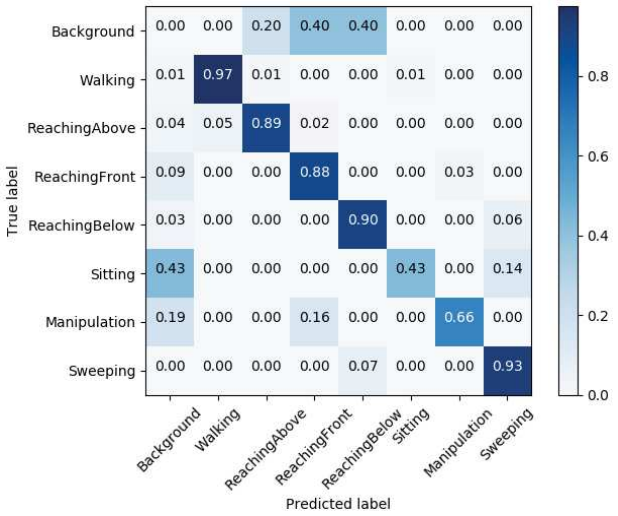

Fig. 5. Normalized confusion matrix of recognizing actions from in-home kitchen dataset.

We demonstrated the application's capability to provide continuous data collection by creating the training and testing datasets in different kitchen environments. The CDC neural network was customized for recognizing and localizing actions from continuous depth videos. The results showed that the considered actions can be recognized from real-cooking videos efficiently. We found that the accuracy of recognizing the manipulation category decreased in real-kitchen test. In the training set, actions of rinsing off and drying the strainer were considered as the manipulation category, while in cooking test set, other actions, such as cutting vegetables, were also considered in the manipulation category. We also found that the accuracy of recognizing sitting decreased in real-kitchen test. Since the training set was collected in one room with chair position fixed and the sitting action doesn't contain much upper-body movement, the recognition result may decrease when testing in different rooms with different setting. As a result, the manipulation and sitting categories in the training set is not complex enough for recognizing these actions from real-life videos. We demonstrated that the quantitative assessment can be performed on clinically relevant action categories. The tools used in test scenarios were placed on the left side of the cabinets. From our observation, most subjects turned to open the cabinets using their left arms, which matches the assessment result of reaching above the head and reaching below the waist actions.

This study has a few limitations. First, the Wi-Fi module of the Samsung tablet S3 does not have enough capability to receive all the depth frames sampled from the sensor. Second, the current system can only handle the situation with one person in the view. The next step for our research is to set up the system in kitchens with both healthy subjects and stroke patients. The collected data will be used to create a more comprehensive training set for a more robust model. In addition, we will investigate the assessment results for both healthy and pathologic subjects. At last, we will improve our algorithm to handle the situations with multiple persons in the view.

REFERENCES

[1] R. Proffitt and B. Lange, "Considerations in the efficacy and effectiveness of virtual reality interventions for stroke rehabilitation: moving the field forward," Phys Ther, vol. 95, pp. 441-8, Mar 2015.

[2] K. P. S. Nair and A. B. Taly, "Stroke rehabilitation: traditional and modern approaches," Neurology India, vol. 50, pp. S85 – 93, 2002.

[3] 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.

[4] 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," CoRR, vol. abs/1703.01515, 2017.

[5] E. Kim, S. Helal, and D. Cook, "Human Activity Recognition and Pattern Discovery," IEEE Pervasive Computing, vol. 9, pp. 48-53, 2010.

[6] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A Review of Human Activity Recognition Methods," Frontiers in Robotics and AI, vol. 2, 2015-November-16 2015.

[7] 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), 2005, pp. 886-893 vol. 1.

[8] 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, 2005, pp. 65-72.

[9] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110, 2004/11/01 2004.

[10] VicoVR. (August 29th). Product description and technical data. Available: https://vicovr.com/user-guide/product-description-and-technical-data

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. (2014, December 01, 2014). Learning Spatiotemporal Features with 3D Convolutional Networks. ArXiv e-prints. Available: https://ui.adsabs.harvard.edu/ - abs/2014arXiv1412.0767T

TABLE III. QUANTITATIVELY ASSESSMENT OF CLINICALLY RELEVENT ACTIONS PERFORMED BY SUBJECTS IN THEIR KITCHENS. ASSESSMENT METRICS INCLUDES EXTENT OF REACH IN MILLIMETER, AND SPEED METRICS IN MILLIMETER/S. FIVE RECOGNIZED TRIALS OF EACH ACTION CATEGORY WERE SELECTED TO PERFORM THE ASSESSMENT. THE AVERAGE VALUES WERE CALCULATED FOR EACH METRIC.

Actions	$Extent_x$ [b]		$Extent_y$		$Extent_z$		$Extent_{3d}$		Max speed		Mean speed		Max/Mean	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*
Reaching above [a]	280.1	184.0	326.1	237.2	243.6	246.1	351.8	315.7	3146.5	3497.5	498.4	470.4	6.3	7.5
Reaching forward	134.8	152.4	205.1	198.4	245.6	219.3	341.2	297.6	1841.5	1543.7	716.2	673.5	2.6	2.2
Reaching below	222.4	226.5	248.0	194.4	291.6	216.5	344.6	295.7	2282.4	1993.4	694.8	604.0	3.1	2.8
Manipulation	236.7	175.7	148.1	74.1	202.3	244.2	303.2	289.9	615.7	476.6	129.9	109.8	4.6	3.5
Walking	218.5	204.4	168.4	159.4	234.8	245.7	314.2	329.1	3014.7	2970.4	985.6	997.1	3.1	3.0
Sitting	80.1	59.3	189.5	188.4	199.0	233.7	283.2	299.2	187.6	254.6	45.2	51.7	4.6	5.2
Sweeping	246.3	182.4	246.7	239.2	246.6	219.8	327.6	305.3	2118.7	2581.7	745.3	853.2	3.5	3.0

[a] Reaching above represents reaching above the head actions and reaching below represents reaching below the waist actions. [b] Extent x, y, z means the hand maximum extent of reach projection in depth, lateral and vertical dimensions. Extent 3d means the hand maximum extent of reach in 3d space. Max/Mean represents the ratio between maximum and mean velocities.