NU Human Grasping Database

Zhanibek Rysbek, Batyrzhan Alikhanov and Aidyn Torgayev

***Abstract*—The study of grasp insights is essential in developing grasp planning algorithms and control systems of upper-limb prosthetic devices. As no any significant work has been done in related field this research takes the initiative on creating the grasping database. NU Human Grasping Database is intended to contain upper body kinematics, depth and RGB image streams for over 3800 instances of grasps collected from 13 subjects engaging in activities of daily routines. The 226 GB database is annotated based on the GRASP taxonomy.**

***Index Terms*—Robotics, Mechatronics, grasps, database, kinematics.**

I. Introduction

DESPITE the technological advancement, nowadays there is still no adequate implementation of hand prosthesis for people with upper limb disabilities, which would reproduce the functional range of the native hand. The development of high-precision grippers and anthropomorphic robot hands [5], which can be comparable to human hands, makes it mechanically possible. However the main problem is the development of control system, especially the models for grasping and trajectory manipulation.

Grasp planning algorithms are essentially depend on object geometric descriptions, artificial hand configuration, actuation mechanism, and sensory input, which is often a multimodal component (e.g. tactile feedback, visual data, inertial measurements). In this work we want to emphasize on advantages of multimodal sensory data synchronized with annotated video material. We focus our effort on human hand motions as applied for the activities of daily living (ADLs), as these are recurring with people on a regular basis and with relatively high frequency. The contribution of our work is in activity dataset generated using three sensing modalities, namely head-mounted RGB camera, depth sensor attached to the hand and upper-body XSENS MVN motion capture suit. The data sequences are synchronized and annotated for grasp and task type applied by human subjects during performance of an experimental routine, which involved ADLs commonly practiced during cooking, housekeeping, clothes folding and ironing. Based on the comprehensive related work survey, we suggest that the presented data is structurally correct and useful for versatile analysis and application in the areas of grasp planning, prediction and evaluation, object recognition and classification models on a multimodal scientific database.

II. Methods

This work delivers an activity dataset consisting of annotated video sequence synchronized with depth images and inertial motion data. Additionally, we make available the Matlab software package developed to process, synchronize and annotate the data files acquired from different streams. This section describes the acquisition setup and protocol, experimental procedures and annotation followed by human subjects and software architecture.

A. Study Participants

We invited 13 human subjects to participate in the data acquisition procedures. The age of 5 female and 8 male participants was in the range of 19 to 42. At the moment of the experiments, all participants had no known hand or arm injury, or other issues which could affect their performance. All 13 participants were right-handed. Before engaging into experiments, each subject was comprehensively briefed about the procedure, the hardware involved and potential risks. Additionally, we provided written description of the experiments and required participants to sign an informed consent form. The study and experiments were carried out in accord with principles of Declaration of Helsinki [8], and approved by the Institutional Research Ethics Committee of Nazarbayev University, Kazakhstan.

B. Data Acquisition Protocol

Activities of Daily Living is a commonly used term in rehabilitation and occupational therapy, referring to set of everyday tasks critical for unassisted living. Recently, the term gained wider usage among Robotics community with similar connotation to enable evaluation of an artificial systems performance for daily tasks. ADLs were then sub-categorized to suitably accommodate robotic application domains. Domestic activities of daily living (DADL) were encompassed tasks regularly performed in human living environments, e.g. housekeeping and cooking. On the contrary, extradomestic activities of daily living (EADL) covered tasks systematically performed outside of home.

We developed a scenario comprised of three experiments. Each was designed to encompass a wide range of common activities performed at a household, e.g. cleaning routines, dealing with cutlery, food items, clothes and special equipment such as an iron. Each experiment hence included DADL tasks, which were generalized to cooking, housekeeping and ironing/clothes folding activities. The exact sequence of domestic activities of daily living experiments is as follows:

- cooking breakfast and cleaning the involved kitchen areas
- housework activities, e.g. wiping the dust and vacuum cleaning
- clothes folding and ironing

Experiments took place in a two-room apartment. All activities took place in a large room which combines living room and kitchen. Before the experiment, human subjects were introduced to the general outline of the three experiments,

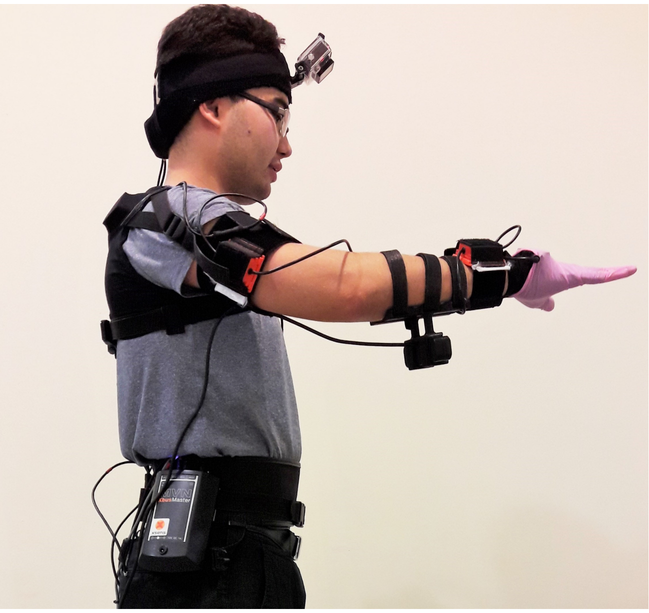

Fig. 1. Researcher wearing the setup: head-mounted camera GoPro Hero 4 Silver, upper-body XSENS MVN suit and hand strapped DS325 depth camera.

provided with high-level task characterization and thorough instructions about the hardware involved. The duration of each of the experiments and approach to its accomplishment were decided by participants (e.g. subjects were asked to cook a breakfast, but decided themselves on what the meal will be and which cutlery to use). Each participant then provided age, gender and a dominant hand information, signed a consent form and was given some time to feel himself/herself comfortable. It took every subject any time in the range of 30 - 50 minutes to finish data acquisition. Duration of all experiments reaches almost 9 hours of data from each of the three sensing modalities.

C. Data Acquisition Setup

The multisensory setup consists of three modules enabling acquisition of video, depth and confidence images and upper body inertial motion data. Confidence images are acquired per each depth image from the RGB-Depth camera. They express the magnitude of IR light inflicting on the detector, hence expressing the level of reliability of depth measurement in the corresponding pixel of a depth map.

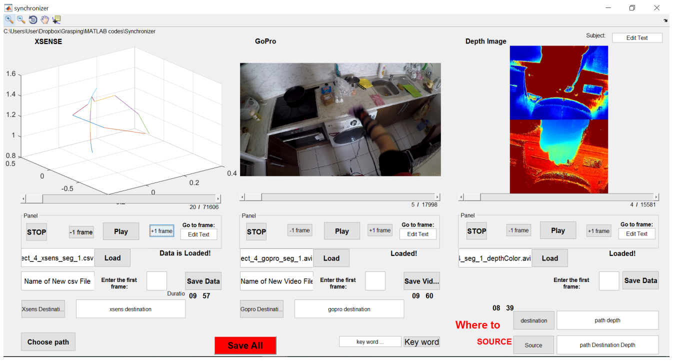

Fig. 2. Interface of Synchronization Software containing relevant frames from three channels: XSENS stick man (left), GoPro (middle) and pseudo colored depth/confidence map (right).

Video is recorded using a GoPro Hero 4 Silver action camera, weight 83 grams, 59 mm long and 41 mm in height, at 1280 x 720 resolution and 30 frames per second rate. Depth images were acquired using DepthSense RGB-Depth camera, model DS325, of 105 mm x 30 mm x 23 mm dimensions, at 320 x 240 resolution and 30 frames per second rate. The nominal operating range of the RGB-Depth camera is 0.15 m 1 m. The action camera was mounted to a participants head and DepthSense camera was strapped to the dominant hand, hence providing us with high-resolution video capturing wider scene range and high-fidelity depth data more focused on the hand operating range.

Inertial motion data was recorded using XSENS MVN suit. The suit is comprised of 17 sensors dividing the body into 23 segment labels and 22 joints (see fig. 1). In this work we only used the upper body configuration of the XSENS MVN, which includes 11 sensors and omits segments and joints related to legs, feet and toes. Along with gyroscope, accelerometer and magnetometer values, these sensors enable obtaining the position, angular velocity and quaternion data for each body segment. Data from the inertial measurement unit sensors was acquired at 120 frames per second and transmitted wirelessly to the HP EliteBook 8560W laptop (Windows 7 operating system, Intel Core i7 processor).

Data was recorded in chunks, segments of duration of less than 10 minutes for further annotation convenience. Both laptops were located within 2 - 3 meters range from human participants and did not cause any disturbance or interference with the process. The process was continuously monitored by one of the authors this paper.

D. Data Mapping and Annotation

Three data channels had been recorded asynchronously. Hence, segments of each stream started at different time

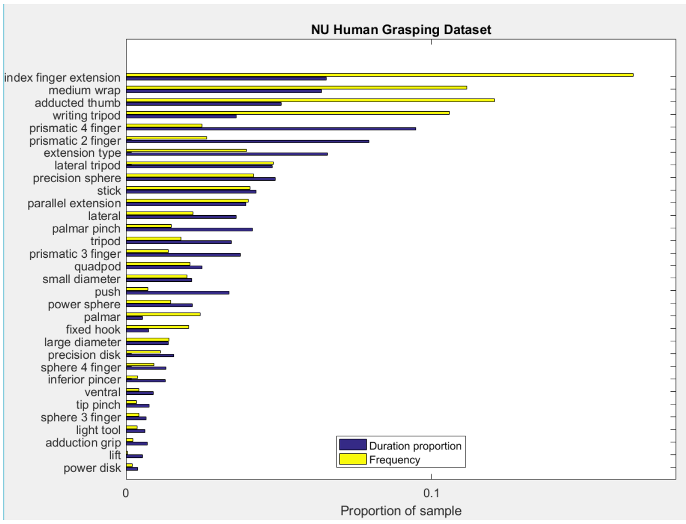

Fig. 3. Frequency and Duration proportion of all grasp actions in the NU Grasp Database.

instant. Additionally, some depth frames were dropped which resulted in non-systematic frame shift. Hence, manual mapping of events by a human expert was required. We created a Matlab graphical user interface program (see fig. 2) which allowed visualization of all three streams and the means to access specific frames in each of the channels. The depth stream was visualized using the Matlabs Jet color map implementation. Annotation was performed by three researchers with engineering background.

We identified discrete grasp actions, similar to the concept of elementary grasp actions used in [7]. Based on our observation of human grasping, we collected multiple grasp actions performed consequently on the same object to the sequence of grasps. Grasp sequence is chronologically numbered stack of grasps. Starting point of each grasp action was established from the moment a human expert detects a contact to acquire an object. The end point of the action was established when subject released an object or switched to another static grasp. Grasp events identified in the action camera videos were mapped to corresponding frames in the depth and inertial motion data. Each actions start and end frames were then recorded, respectively. Only the right hand (dominant in each participant) was considered. Time-line outside of the grasps was left untagged.

There is the total of 3826 grasps identified from the recorded material among which 2922 (4 h 50 m), 564 (1 h 5 m) and 340 (28 m) are performed during cooking, housekeeping and laundry activities, respectively. These absolute values reflect both the duration and nature of each activity.

In figures 3 and 4 are depicted statistical plot of occurrence and duration proportion of all grasps and grasps with 0 sequence number (first grasp in a sequence). The distribution of grasps in both cases has similar pattern. Note that, first grasp in sequence usually deployed to pick up the object and to switch to task related grasp. Therefore, in latter case frequency and durations of precision grasps increased (e.g. prismatic 4/3/2 finger and palmar pinch).

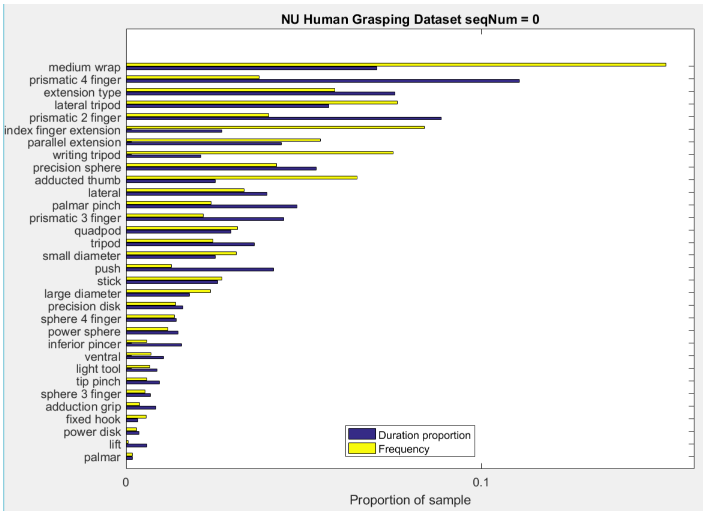

Fig. 4. Frequency and Duration proportion of first grasp action in all sequences

E. Grasp taxonomy and assumptions

For our list of grasps, we followed the taxonomy provided by Feix et al in [4], where authors carried out a comprehensive survey and analysis of human grasps. Authors identified 33 distinct types and showed that this list can be reduced to 17 general grasps, when considering only hand configuration. In addition, we added two common non-prehensile grasps - Push and Lift, which do not appear in [4].

Along with the taxonomy, we followed two assumptions from [4], during tagging grasp actions. Specifically, we omitted bi-manual tasks, when an action is possible only with application of two hands (e.g. two-hand bedsheet folding). And in-hand motion actions, i.e. movements causing object motion within the hand.

F. Data description

As a result of aforementioned procedures, we generated and structured experimental data and corresponding annotation file. The dataset is organized as per human subject such that video, depth, inertial motion data and annotation obtained for each participant are grouped together in a separate folder. Thus, our dataset contains 13 folders with names of participant the data was acquired from. Furthermore, each subject directory contains separate folders for GoPro video files, depth images and inertial motion data. GoPro videos are stored as Audio Video Interleaved (AVI) files, depth and confidence images in the Portable Network Graphics (PNG) format, and inertial motion data as comma separated files. As mentioned in the Methods section, the length of experiments varied across participants and acquisition was performed in 10-minute long segments. This way, for every sensing modality, there might be different number of files available. E.g. if subject 1 experiment lasted for 30 minutes, there will be 3 AVI files in the GoPro videos folder of that subject. At the same time, if subject 2 experiment lasted 40 minutes, there will be 4 AVI video files in the corresponding folder.

Data acquired from the XSENS motion capture suit, stored in a comma-separated file, contains 227 columns each representing specific inertial measurement taken from an upper body of a subject. The full body XSENS suit contains 17 inertial measurement units (IMU) each comprised of a gyroscope, accelerometer and magnetometer. There are 22 labeled joints and 23 enumerated body segment labels indicated the origin of the inertial motion data. In our experiments, we used the upper body configuration of the suit, hence omitting number sensors, joints and segment labels, such as upper and lower legs, feet and toes. Upper body configuration provides with 227 sensor and body segment labels, arranged as 227 data columns in the comma-separated file. Specifically, first 150 columns contain the data from 15 segment labels, including Pelvis, Right Hand, Right Forearm, Right Upper Arm, Right Shoulder, Neck, Head, etc. For each segment, the system provides 10 fields - 3 position, 4 quaternion and 3 angular velocity values. Subsequently, there are 77 columns, which contain 3 gyroscope and 4 quaternion values for each of 11 sensors of the upper body suit, such as Pelvis, Head, Right Hand, Right Upper Arm, etc.

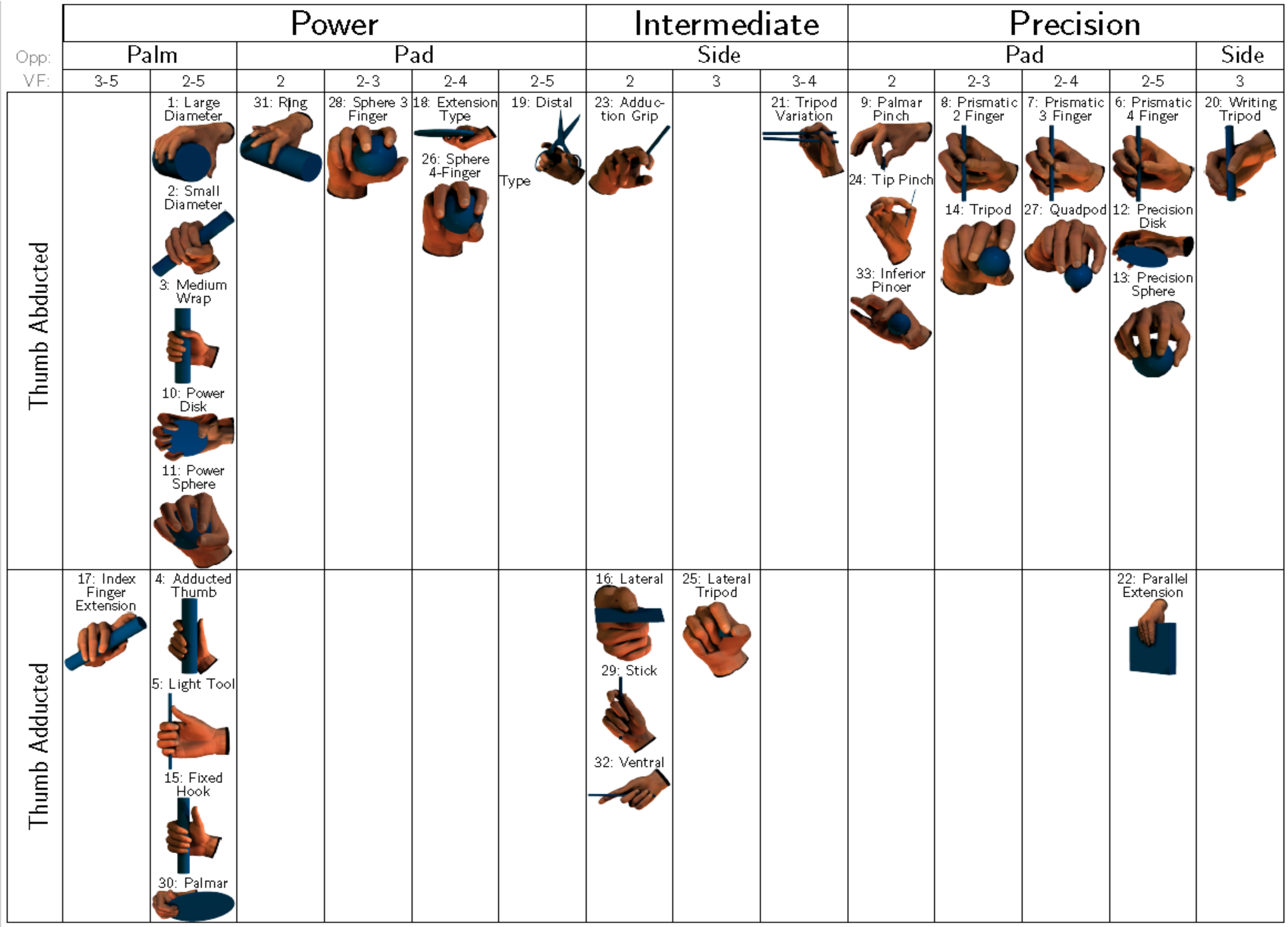

Fig. 5. GRASP taxonomy developed by Thomas Feix, Yale University. The grasps are classified in the columns according to their assignment into power, intermediate and precision grasp, the opposition type, and the Virtual-Fingers assignment. The assignment of the rows is done be the position of the thumb that can be in an abducted or adducted position.

The annotation is provided as a csv file in the root directory of the dataset. There are 13 columns in the annotation file. The file indicates the id of the grasp and the participant it belongs to. In addition, the file identifies the grasp type as identified from the taxonomy, task type as annotated by human experts and start and end frames of the episode to locate them in the GoPro video files. Finally, the annotation describes the grasp properties including opposition type, power level and thumb position. From first to last, the columns represent the values for Grasp ID, Participant ID, Grasp Type, Task Type, Video File Name, Start Frame, End Frame, Opposition Type, Power level, Virtual Finger, Thumb configuration, Sequence Number, Start Frame (Depth) and End Frame of Depth images. The overall size of the annotated data is 226 GB.

III. Mining Depth data

Depth images were applied to a machine learning task. The goal was to learn the state of the hand, whether it is grasping or not. Figure 6 shows pseudo-colored two classes of depth image that trained model supposed to distinguish. First, raw depth images were supplied to Cubic SVM and Bagged Trees classifiers with five fold cross validation which resulted in 83% and 89% level of accuracy. Meanwhile, confidence map based filtered depth images improved the accuracy level to 89% and 92% respectively.

A. *Feature Extraction*

Annotated file of RGB video streams provides the moment of contact of the hand with an object. Thus, we assume that frames before the moment of contact are images of non-grasping hand and after are images of grasping hand. Although, number of grasping instances are limited to 3800 (2500 usable, because of grasp sequencing) for this specific task we can supply significantly large amount of data. As a result over 25000 of frames were taken as a training data.

Features from depth images were acquired by dividing it into uniformly sized rectangular regions in quadtree manner. Furthermore, four statistical values minimum, maximum, mean and standard deviation from each region were taken into account. Depth image is divided into regions in an iterative manner: first level is the full 320 x 240 pixels image. This region is then divided into four uniform regions of 160 x 120 pixels. Each of the four obtained regions is divided into another four uniform rectangles, which provides 16 regions

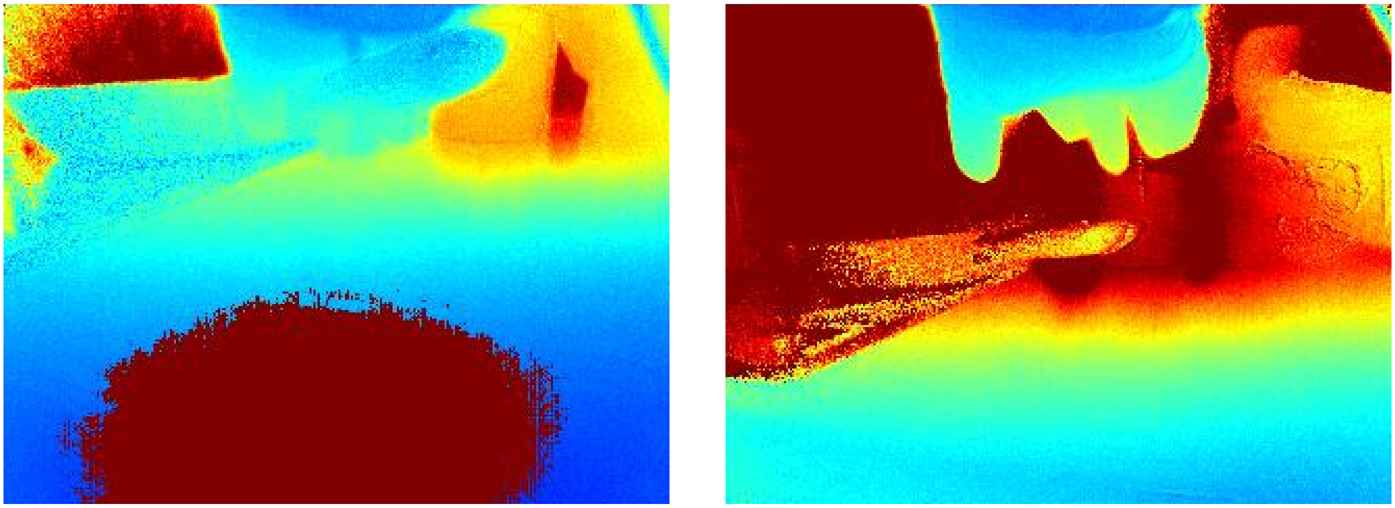

Fig. 6. Pseudo-colored depth images showing grasping(left) and non-grasping hand(right).

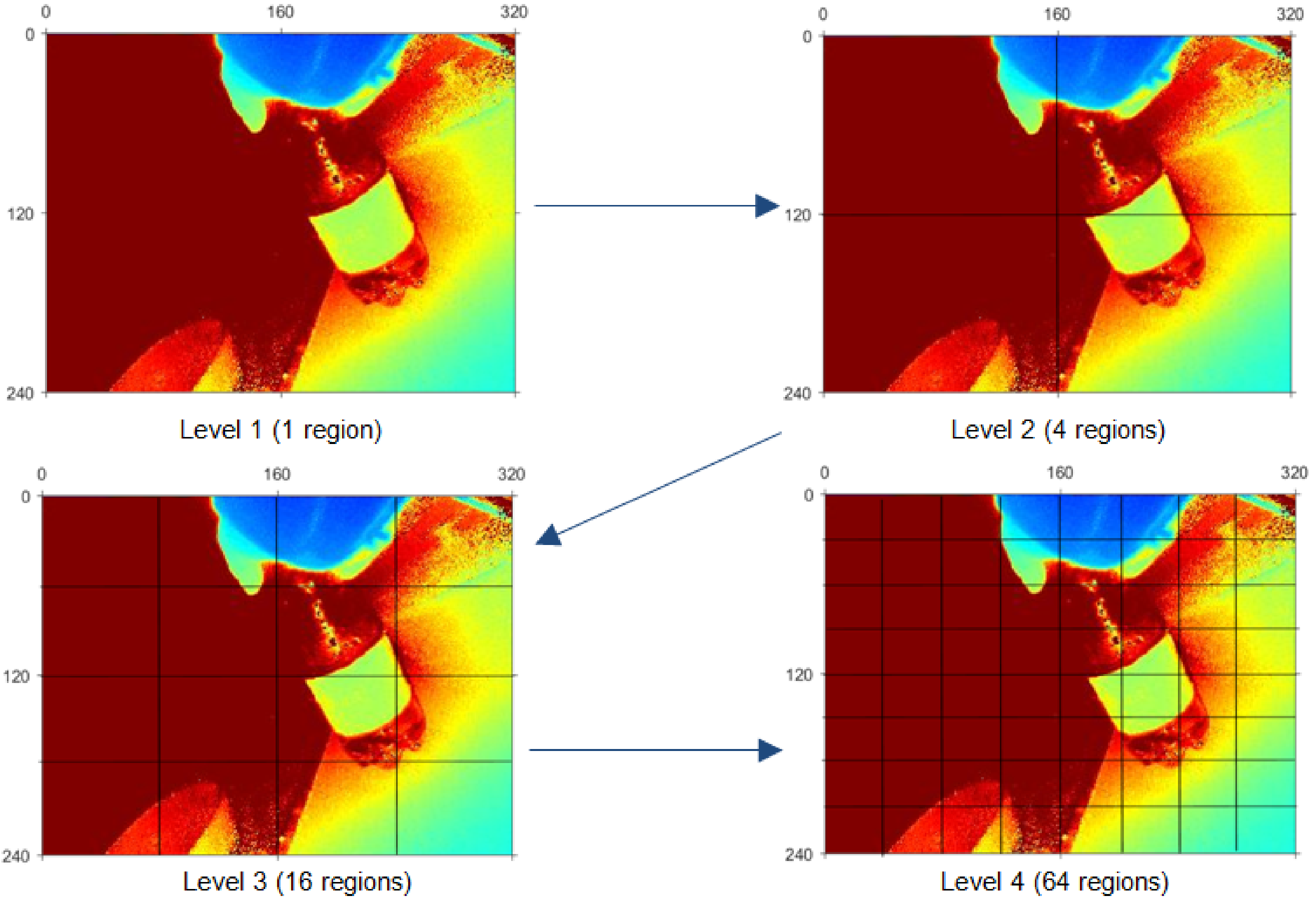

Fig. 7. Recursive quadtree images used for feature extraction

and 16 + 4 + 1 = 21 total regions at level 3. This process continues and for levels 4 there are a total of 85, respectively (see Figure 7). At every stage statistical data of each region is extracted and stored in a vector. With four features obtained from a region, iterations 2, 3, and 4 generate feature vectors of 20, 84, and 340 lengths, respectively.

B. Filtering Depth Images

Unfortunately, depth cameras have technical limitations in the area they are able to cover. There are cases when data reliability drops as the object goes beyond the the sensibility of the camera. Mainly distortions are introduced due to the reflectivity of the materials being observed. Luckily, DepthSense camera provides us with the confidence map together with the depth images. By assigning values of infrared specter from confidence maps to the depth data, there is a way to test reliability of the recordings. The issue was solved by using normalized confidence map as a filtering mask for corresponding vales in the depth image to obtain higher levels of denoising. This solution was preferred as it is rather fast and requires minimum amount of computations.[6] The used formula is as follows:

$$D_{filt}(i,j) = \frac{\sum_{k_i=-l}^{l}\sum_{k_j=-l}^{l} D(i+k_i, j+k_j)\cdot C(i+k_i, j+k_j)}{\sum_{k_i=-l}^{l}\sum_{k_j=-l}^{l} C(i+k_i, j+k_j)} \tag{1}$$

Where D is the depth image and C is a confidence map matrix.

This approach requires that for each depth frame there is a corresponding confidence map acquired simultaneously. Usage of filtering improved results of mining the depth data.

IV. Conclusion

Primary contribution of this work is providing sensor rich Human Grasping database consisting arm kinematics and depth image streams. The dataset is available on-line and at the ARMS lab server. 9 hours of RGB video streams were annotated for grasping matter. Moreover, depth streams were synchronized with head mounted camera. Provided the over 3800 grasping instances this database has potential to answer various questions regarding human grasping behavior related to the arm movements and object shapes.

References

[1] Bullock, Ian M., Thomas Feix, and Aaron M. Dollar. *"The Yale human grasping dataset: Grasp, object, and task data in household and machine shop environments."* The International Journal of Robotics Research 34.3 (2015): 251-255.

[2] Bullock, Ian M., et al. *"Grasp frequency and usage in daily household and machine shop tasks."* IEEE transactions on haptics 6.3 (2013): 296-308.

[3] Cutkosky, Mark R. *"On grasp choice, grasp models, and the design of hands for manufacturing tasks."* IEEE Transactions on robotics and automation 5.3 (1989): 269-279.

[4] Feix, Thomas, et al. *"The GRASP taxonomy of human grasp types."* IEEE Transactions on Human-Machine Systems 46.1 (2016): 66-77.

[5] Grebenstein, Markus, et al. "The hand of the DLR hand arm system: Designed for interaction." The International Journal of Robotics Research 31.13 (2012): 1531-1555.

[6] Saudabayev,Artur, Farabi Kungozhin, Damir Nurseitov, and Huseyin Atakan Varol, Locomotion Strategy Selection for a Hybrid Mobile Robot Using Time of Flight Depth Sensor, Journal of Sensors, vol. 2015, Article ID 425732, 14 pages, 2015. doi:10.1155/2015/425732

[7] Vergara, Margarita, et al. *"An introductory study of common grasps used by adults during performance of activities of daily living."* Journal of Hand Therapy 27.3 (2014): 225-234.

[8] World Medical Association. "World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects." Journal of postgraduate medicine 48.3 (2002): 206.