Towards an Understanding of Grasping using a Multi-Sensing Approach

Jonathan Maycock, Kai Essig, Robert Haschke, Thomas Schack and Helge Ritter

Abstract—To gain a deeper understanding of human manual dexterity, we investigate an everyday manual interaction, recording hand kinematics, the involved forces and eye-movements. Using a multi-sensing approach we hope to develop a more complete understanding of what we have termed manual intelligence, with a view to enabling robots carry out complex tasks that we take for granted. A preliminary experiment was carried out which illuminated important aspects of manual intelligence such as where to look to find suitable grasping points, as well investigating differences in hand kinematics and the applied forces while carrying out a simple manual interaction.

I. INTRODUCTION

A major goal of current robotics research is to try to mimic the excellent abilities of the human hand, and advances in mechatronics, sensing and control are facilitating this endeavor [1], [2], [3]. Achieving this goal will require a deep understanding of manual interactions performed by humans at the level of kinematic control, contact and force physics [4], [5], [6] and attention [7], [8]. In this paper we use motion capture, tactile sensing and eye-tracking to investigate an everyday manual interaction in which a tray is moved from one shelf to another.

At the kinematic control level it has been proposed that the control of human hand movement is organized in a modular way, comprising higher levels that combine and couple several DOF into functional groups, or motor synergies, thus simplifying the execution of hand postures [9]. Principal component analysis (PCA), and more recently Gaussian Process Latent Variable Models [10], allows the extraction of these synergies and a large body of work supports their existence [9], [11].

The role of contact and forces in manual interactions needs to be considered if we are to build up a full picture of the processes involved [6]. Tactile sensors can be integrated into gloves to acquire information about how humans manipulate objects, or as we have done in this experiment the object to be manipulated can be fitted with tactile sensors. This has the added benefit that the experiment can be performed in a more natural way than could be achieved wearing a glove. Analyzing the resulting patterns and forces is important if we are to understand and model human manipulation strategies and reproduce them in robotic systems. We have measured the forces applied to a tray object using an in-house developed tactile sensor system providing a high spatial (5mm) and temporal (600Hz) resolution [12].

The importance of attention in guiding manual actions has seen renewed interest of late [7], [8]. In order to achieve good eye-hand coordination, hands and eyes must work together in smooth and efficient patterns. Johansson et al. [13] analyzed the coordination between gaze behavior, fingertip movements, and movements of a manipulated object. Obligatory gaze targets were those regions where contact occurred. They concluded that gaze supports hand movement planning by marking key positions to which the fingertips are subsequently directed. We have also analyzed this phenomenon in the present study. Furthermore, we investigate whether the Eye-Mind Hypothesis [14] is supported by the captured data. This states that the number and the distribution of fixations reflect the degree of cognitive processing, required for the understanding of particular scene regions. Long fixation durations and short saccade lengths signify a fine and deep processing of a particular image region, indicating that the understanding of its visual information is quite difficult. In contrast, long saccade lengths and short fixation durations indicate a fast and coarse scanning of a particular scene region, signaling that the information content of that particular image region is easy to process or less important for the current task.

The work in this paper is part of a larger project in which we complement the current, strongly control- and physics-based approach for the synthesis of robot manual actions with an observation-driven approach [15]. To facilitate this a multi-modal database is being constructed in which direct recordings of humans performing various interactive manual tasks is its main focus. Besides requiring the allocation of a larger number of degrees of freedom to the representation of the human hand, a manual interaction database can benefit from additional modalities recording eye gaze data and tactile forces by revealing important aspects details such as where to look to find suitable grasping points, as well as the adjustments necessary to applied forces in manual interactions. This multi-sensing approach makes the design of a manual interaction database a highly non-trivial task that is coupled with the design of suitable lab environment. We briefly discuss the lab we have built to meet these goals in the next section.

II. THE MANUAL INTELLIGENCE LAB (MILAB)

There are some freely available motion capture databases such as the CMU Graphics Lab Motion Capture Database [16], the Karlsruhe Human Motion Library [17] or the recent TUM Kitchen Data Set [18] focusing on different
aspects and settings of human motion data. However, as their common focus is on full body movements, they only allocate a very small number of degrees of freedom to the hands, enabling only a very coarse representation of manual actions. Thus, to date there does not exist a database of fine detailed human manual interaction situations and therefore the Manual Intelligence Lab (MILAB) was purposefully built to capture such details at a high level of spatio and temporal coherence.

MILAB consists of fourteen Vicon [19] MX3+ cameras that can track reflective markers attached to both subjects and objects. Having so many cameras in such a small space allows us to track the human hand, which due to self occlusions and occlusions caused by objects is a difficult task. Vicon has a marker tracking accuracy of below 1 mm and this provides us with good ground truth data for our other vision modalities. In order to get insights into where humans look as they perform everyday tasks with their hands, we use a SMI IViewX (monocular) mobile eye tracking system [20] for our experiments. It has a sampling rate of 200 Hertz, with a gaze position accuracy of \(0.5° - 1°\). Each scene video has a resolution of 376 x 240 pixels at 25 fps. Using an in-house developed object fitted with tactile sensors, we were able to gain information about the positional contact and the forces required to lift and move the object. Other devices that are supported in MILAB, but not used in the current experiment, are stereo-vision cameras, Swiss Ranger time-of-flight cameras, Cyber Glove II datagloves, smaller fingertip tactile sensors and microphones. For a more detailed discussion of MILAB please see [15]. The interested reader is also directed towards recent work carried out in the lab involving recognition of manual actions [21] and the kinematics of grasping real versus virtual spherical objects [22].

### III. SYNCHRONIZATION OF DIFFERENT SENSORS

The challenges of manual action capture are many and varied, but in most cases reduce to the problem of ensuring that there is a high degree of both spatial and temporal coherence amongst the different input channels. At the core of MILAB is a 14 camera Vicon system providing us with high precision 3D positional data to which the other modalities need to be aligned. For modalities not directly supported by Vicon, i.e., the eye-tracker and the tactile sensors used in this experiment, it was necessary to develop custom made solutions in order to ensure spatial coherence. Equally important is the need to ensure temporal coherence across all modalities. The Multiple Start Synchronizer (MSS) was developed to ensure that recording of all data streams start at the same time, and run for a specific duration or until a stop command is sent. MSS first checks if all computer clocks are synchronized correctly using the Network Time Protocol. Pressing start sends all relevant data (i.e., start and stop time, experiment details) via Open Sound Control protocol [23] to the listening clients. The client applications then wait for the starting time and begin capturing until the end time is reached or a stop signal from MSS is received.

### IV. EXPERIMENT AND SOME INITIAL RESULTS

To test our multi-sensing setup a simple experiment was designed in which the task was to pick up a rectangular block object, conceptualized as a tray, from a shelf and place it on another shelf. There were two constraints placed on the task. The first was that the subject could only grasp the top and bottom of the tray and the second was that in trials in which a cup was placed on the tray, the movement had to be done ensuring that the cup did not fall down. Each trial can be segmented into an approach and grasp phase and a transportation of the tray phase. 10 trials were performed, 5 with no cup on the tray (condition 1) and 5 with a cup on the tray (condition 2). The subject’s hand movement, forces applied, and eye movements were all tracked. As this was a preliminary study data for only one subject was captured, but based on these promising results more experiments are currently being planned.

We analyzed hand kinematics using PCA. The 26 markers on the subject’s hand were mapped to a human hand model with 22 DOF. These joint-angles were used as input to the PCA algorithm. For the approach and grasp phase of the experiment results indicate that in the first condition 78.6% of the variance is described by just a single principal component, while in the second condition 72.1% of the variance is described by the first principal component. This would indicate that the second condition contains more complex movements whose variance extends into higher components. Figure 1 shows that if the 22 joint-angles are projected onto just 3 principal components it is possible to determine which condition (tray with or without cup) is being executed from an extremely early stage in the movement (about a quarter of the time into the movement). After PCA was computed on the raw data, linear interpolation was used to ensure that all trials had the same number of data points and then the mean of the trajectories in each condition were computed. The result is that we can recognize grasp types not only by their end position configuration, but at much earlier points along the trajectory of the movement. This is in line with a previous study we carried out that involved the grasping of real and virtual objects [22]. Looking at the transportation of the tray phase reveals that the first PC describes 79.4% and 71.91% of the variance for conditions 1 and 2, respectively. This could indicate that the grasp is not as fixed in the second condition as it is in the first. Small adjustments could be needed to ensure that the cup does not fall from the tray. Analyzes of the motion tracking data revealed that the average time taken in condition 1 was 3761ms and in condition 2 was considerably longer at 5806ms. While moving the tray the subject was slightly quicker in condition 1 with a maximum velocity of 0.45m·s\(^{-1}\) compared with only 0.39m·s\(^{-1}\) in condition 2.

We analyzed the gaze data with regard to the number of fixations, fixation duration, number of saccades and saccadic length in x- and y-direction. With regard to the fixation duration we found a higher number of fixations as well as longer fixation durations for condition 2 (8.3 and 9.4, as
2: Tray with a cup.

Fig. 1. Visualization of hand kinematics trajectories through PC space for the approach and grasp phase. Condition 1: Tray without a cup, Condition 2: Tray with a cup.

as well as 411.55ms and 661.37ms for condition 1 and 2 for the number of fixations and fixation duration, respectively. The number of fixations is slightly higher and the fixation duration is approximately 1.5 as long for condition 2. There are also differences with regard to the saccade lengths in the x- and y-directions between the two conditions. In condition 2, the saccade length in the horizontal direction is double the amount compared to condition 1 (122.62 and 59.89 pixels for conditions 1 and 2, respectively). The difference for the saccade length in vertical direction is smaller because the shelves had nearly the same height (83.54 and 69.47 pixel for condition 1 and 2, respectively). But also here, the saccade length is smaller (at approximately 15 pixels) when the cup is on the tray. The overall results are in line with the Eye-Mind Hypothesis: the number and distributions of fixations reflect the degree of cognitive processing required to understand the scene. Long fixation durations and short saccade lengths signify a fine and deep processing of a particular scene region. When the cup is on the tray, it is harder to grasp and stabilize it adequately (i.e., to find suitable landmark points) to ensure that the cup does not fall off while grasping the tray and moving it from one shelf to the other. Thus, and in accordance with the findings of Johansson et al. [13], we found that the gaze supports hand movement planning by marking key positions to which the fingers are subsequently directed. When the cup is on the tray, the subject has to scan the tray in much more detail in order to provide suitable grasping points for the subsequent sensomotoric motor control. The uncertainty of the subject to find suitable grasping points is also reflected in the longer average time taken in condition 2.

An attention map [24] can be calculated to highlight which areas where looked at and how intensely they were studied [see Figure 2]. In an attention maps each fixated pixel is assigned a value between 0 and 1 depending on how long the subjects focused on the particular pixel (values closer to 1 indicate higher attention by the subject). Color-spots indicate which areas were attended and how intensive the subject looked at them. The attention maps for condition 1 show that the subject focuses mainly on the back of the hand, whereas in condition 2 the focus is mainly on the fingertips.

We calculated the number of pixels with an attention value of 0.2 or higher, i.e. those areas receiving attention by the subject. The size of the attention area in condition 1 is 13753 pixels or 3.98% of the image size, whereas in condition 2 the area is 18212 pixels or 5.27%. These results clearly indicate that the subject searched a larger area for suitable landing points for his fingers when a cup was present on the tray, while in condition 1 attention is directed to the hand only to check whether it is in the right position - not for finding landing points. This data matches well with the PCA analysis of the hand kinematic data.

Preliminary results for the tactile data clearly show the different grasp strategies that are necessary for the two conditions [see Figure 3]. In condition 2, when a cup was placed on the tray, the subject first slid the tray out from the shelf [see Figure 3 (b)] allowing for a new grasp to be formed [see Figure 3 (c)], ensuring that the cup did not fall off the tray violating the second constraint. The tactile tray used in this experiment has a 20mm thick border casing, which meant that all the forces applied to the object were not measured. This is evident in the difference between the third grid of Figures 3 (a) and (c), in which the surface area is noticeably smaller in the first. A newer model with reduced border casing of 4mm will allow a fuller description of the forces applied. We also observed average force profiles produced by the fingers during the transportation phase and note that condition 2 contained more fluctuations than condition 1. This could be an indicator for the slight adjustments in force required to keep the tray level and ensure the cup does not fall off [see Figure 4] and would correlate with what we found in the analysis of hand kinematics and eye movement data.

V. CONCLUSION AND OUTLOOK

Initial results from our multi-sensing approach have given important insights into manual interaction and perception processes when humans perform grasping tasks of different complexities. Correlations were found between the different modalities that indicated a higher cognitive load was required for the more complex task. From an early stage in the movement hand kinematics can be easily distinguished. Furthermore, eye movement parameters reveal the searching pattern for suitable grasping points when a cup was placed on the tray. Finally, a jagged pattern in the average forces
applied by the fingers to the tray when the cup was present could indicate that slight adjustments in the grasp were necessary to ensure the tray was kept level. Ultimately our goal is to enable robots interact with the world in a natural way. By observing humans performing everyday actions with a multi-sensing approach, we hope to be able to gain a deeper knowledge of the complexities and move a step closer to this goal.

Looking to the future we wish to carry our further trials with multiple subjects performing a large volume of manual interaction tasks. Important questions will arise from these experiments. What is the variability in terms of the trajectory of hand movements among different subjects and indeed across trials carried out by the same subject? Given the current scenario, what effect does the height of the shelf have, relative to the height of the subject, on the type of grasp chosen?

VI. ACKNOWLEDGMENTS

This research was supported by the DFG CoE 277: Cognitive Interaction Technology (CITEC).

REFERENCES