Towards Object Prediction based on Hand Postures for Reach to Grasp Interaction

Florian Daiber
German Research Institute for Artificial Intelligence (DFKI)
Saarbrücken, Germany
florian.daiber@dfki.de

Antonio Krüger
German Research Institute for Artificial Intelligence (DFKI)
Saarbrücken, Germany
krueger@dfki.de

Dimitar Valkov
University of Münster
Münster, Germany
dimitar.valkov@uni-muenster.de

Frank Steinicke
University of Würzburg
Würzburg, Germany
frank.steinicke@uni-wuerzburg.de

Klaus Hinrichs
University of Münster
Münster, Germany
khh@uni-muenster.de

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
Recently, traditional multi-touch surfaces are extended by stereoscopic displays and 3D tracking technology. While reaching and pointing tasks have a long tradition in human-computer interaction (HCI), the hand pre-shaping which usually accompanies them has rarely been considered. The *Reach to Grasp* task has been widely investigated by many neuropsychological and robotic research groups over the last few decades. We believe that subtle grasping hand postures in combination with stereoscopic multi-touch displays have the potential to improve multi-touch 3D user interfaces. We present a study that aims to identify if the intended object can be predicted in advance, relying only on detection of the hand posture.

Author Keywords
Input and Interaction Technologies, Tabletop and Large Wall Displays, Stereoscopic Display

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces. - Graphical user interfaces

General Terms
Multi-touch Interaction, 3D Interaction, Hand Postures, Gestures, User Studies

Motivation and Background

In recent years, the interaction with 3D data became more and more popular, but current 3D user interfaces (e.g. virtual reality systems) are often expert systems with complex user interfaces and high instrumentation. Stereoscopic displays allow users to perceive 3D data in an intuitive and natural way. On stereoscopic displays objects might be displayed with different parallax paradigms, i. e. negative, zero, and positive parallax, resulting in different stereoscopic effects. Objects may appear behind (positive parallax), at the screen/surface level (zero parallax), or in front (negative parallax) of the screen. Recently, several approaches extended traditional multi-touch surfaces by 3D stereoscopic output [2, 3, 4, 16] as well as different technologies that allow to detect gestures before the user actually touches the surface [5, 17, 19, 20].

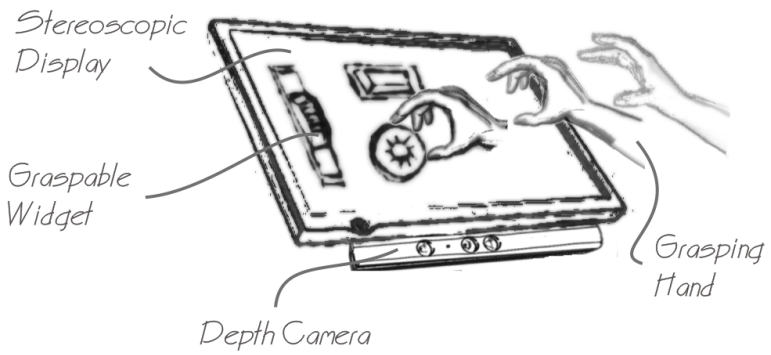

Figure 1: Design concept of a multi-touch enabled stereoscopic surface equipped with additional depth sensors that can predict the user's intention during grasping movements.

Multi-touch technology can be used in order to allow a rich set of interactions without any instrumentation of the user. Schöning et al. have considered some of the challenges of multi-touch interaction with stereoscopically rendered projections [14]. One limitation of these approaches is that the interaction and visualization is often constrained to almost zero parallax. Although the combination of multi-touch technology, depth cameras and stereoscopic display promise interesting and novel user interfaces (see Figure 1), the benefits, possibilities and limitations of using this combination have not been examined in-depth and are so far not well understood [15].

Psychological research on the *Reach to Grasp* task has shown that the pre-shaping phase of the human hand allows a prediction of the object a human is going to grab. Multiple studies were conducted in this direction including physical objects as well as memorized and virtual objects that had to be reached and grasped [1, 9, 12]. Research in this direction has shown evidence, that only a few variables have an impact on that prediction [9, 11, 13, 18]. These insights from neuropsychological and robotic research are promising and we believe that the information of the reach to grasp phase can substantially improve interaction with stereoscopic multi-touch displays. While reaching and pointing tasks have a long tradition in the HCI field, the hand pre-shaping has rarely been investigated.

However, due to the availability of low-cost algorithms, simple off-the-shelf hardware and low instrumentation are now sufficient to track the human hand above the interactive surface. Depth cameras provide the possibility to recognize hand gestures and postures. Furthermore, when tracking the user's grasp postures above a multi-touch display her intented interactions might be

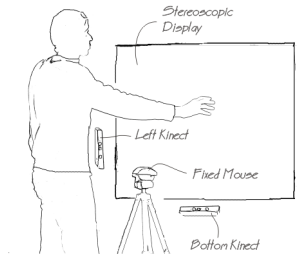

Figure 2: Experiment setup: Illustration of the setup including two kinect cameras and a fixed mouse used as hand rest and starting point for each trial.

Figure 3: A subject performing grasp gestures with a 3D user interface widget during the experiment.

Figure 4: Sample interface widgets that have to be grasped by subjects during the study.

predicted before she or he actually touches the surface. Such knowledge has the potential to improve the user interface of stereoscopic multi-touch surfaces, for example, by snapping desired objects to the touch surface. The main contribution of this work is to show the feasibility of the intended object prediction based on simple features and light-weight pattern recognition algorithms. With the knowledge of the user's intention, the touch-based user interface can then be adapted before the user finally reaches the interactive surface with respect to the user's intention, i. e., what the user is planning to do next and with which objects the user interface (see Figure 1). In order to verify this approach, we performed an experiment in which we analyzed hand postures above and on the interactive surface. The aim of this experiment was to examine whether or not the hand posture allows an early prediction of the objects the user is intending to interact with.

Data Aquisition Study

To collect a corpus of grasping postures we have set up a data aquisition study. In this study we investigated if a stereoscopic rendered object could be detected in advance while the user reaches to grasp it, based only on hand posture and determine the parameters that affect this detection. Therefore participants had to perform typical *Reach to Grasp* tasks using different virtual stereoscopic displayed objects as visual stimuli and recorded their hand motions with multiple depth cameras.

The setup for the study is shown in Figure 2. The study was performed using a prototype stereoscopic multi-touch projection wall. For the back projection a projector with native resolution of 1400×1050, using a frame-sequential stereoscopic projection at 120 Hz was used. The projection uses only a portion of the touch enabled screen with dimensions $136cm \times 102cm$, that resulted in an effective pixel size of approximately $1mm$ (645 pixel per in^2). Although we could track the subjects' head positions, it was not needed, since the subjects remained in the same position during the entire study. The position of the virtual camera and its viewing frustum were adjusted to match the subject's height. Hand motions were recorded with two Microsoft Kinect depth sensors as RAW video streams with resolution of 640×480 at 30 frames per second (fps). Both sensors were arranged (one at the left side of the projection and one below it) in such a way, that the user's hand was in her field of view during the whole time of each trial. To indicate the start of each trial subjects used a common computer mouse, which was mounted on a camera tripod and also adjusted to each participant's height. The study was run on PC with Intel Core i7 Processor with 8GB of RAM and nVidia GeForce GTX470 graphics card.

22 participants (3 female), naive to the experimental conditions, took part in this study. The subjects were between 22 and 56 years old ($M = 28$, $SD = 6.9$) and none has reported any visual or stereopsis disruptions. All subjects were members of our institute or university students and reported, in a 5-Point Likert scale, between good and excellent experience with touch devices. All participants were right handed. The entire study took about 30 minutes. The subjects were allowed to take breaks at any time during the study. In addition the subjects had to take mandatory two minute breaks at regular intervals to minimize errors due to exhaustion or poor concentration.

In this study subjects were asked to grasp virtual objects that are graspable counterparts to standard user interface widgets (see Figure 3). These widgets were designed to

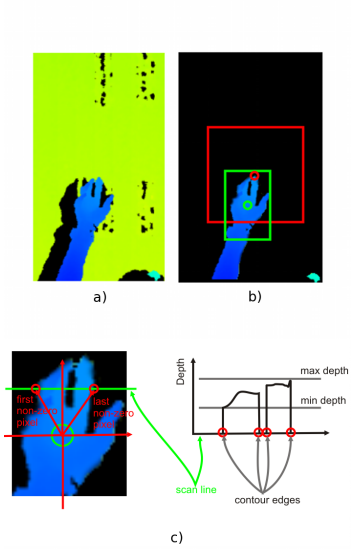

Figure 5: Feature extraction pipeline: (a) the raw depth image from Kinect; (b) after background subtraction, with regions for hand segmentation; (c) illustration of the feature extraction for a row scane line.

have approximately the same size and are meant to be interacted via grasp-gestures. Subjects were positioned in front of the projection screen at a distance of approximately $3/4$ parts of their arm length, such that they could conveniently perform all grasp gestures during the study with their dominant arm. All trials had to be performed with the dominant hand. To guarantee a consistent initial start position at the beginning of each trial the subject had to press the left button on the mouse mounted at convenient distance (ca. $25cm$) on her right side. As visual stimuli ten different stereoscopic rendered virtual objects shown on Figure 4 were projected at five different object positions: $(-a, 0)$, $(a, 0)$, $(0, 0)$, $(0, -a)$, $(0, a)$ with a being the half arm length of the subject and $(0, 0)$ being adjusted to match the orthogonal projection of the participant's right shoulder on the surface.

Four trials had to be performed for each object at the five different positions resulting in a total of 200 trials per subject. Five trials (not included in the evaluation) have to be performed at the beginning to ensure that the subjects understood the task and received some initial training. After all trials were completed the subjects were asked to fill out a short questionnaire about their subjective experience with the interface, visualization issues, and exhaustion during the performance of the study.

Analysis

The collected RAW video streams are not applicable for direct evaluation. Thus we first pre-processed the data set to extract an image-based set of representative features for each frame. Afterwards we split the video streams in time segments, filtered the frame feature sets within time segment from redundant information and evaluated the results with correlation based algorithms.

Feature Extraction

For each frame in each video sequence we first removed the background by clamping all values to zero above a threshold (i. e. too far away from the camera) and subtracting a static captured background image from the resulting frame as shown on Figure 5 (b). This pixel was used as reference to determine a rectangular subregion of 200×200 pixels that contained the user's hand (the red rectangle on 5 (b)). We then found the weight center of the region as the mean of the pixel-coordinates of all non-zero pixels within the region and built up a new subregion of 100×150 pixels centered in the weight center (marked with a green circle on 5 (b)). The hand contour and the distribution of the depth extrema within a depth image of the hand have been successfully used in multiple works as features for hand-gesture recognition [7, 6]. In our approach similar parameters have been used as well as the outer contour of the hand. In order to evaluate the frame data we have extracted from each region some representative parameters, which seemed to contain meaningful data about the current hand posture. Nevertheless, we did not use the hand contour and topology directly, but extracted from the segmented subregions some unified representative parameters, that were more appropriate for direct comparison.

The following parameters have been considered to be most useful as feature vector (cf. Figure 5 (c)): the number of depth minimums, as well as their mean, minimal and maximal values; the mean depth of the region; the number of non-zero pixels within the region; for each row - the unprojected positions of the first and the last non-zero pixel, relative to the unprojected coordinates of the regions center; the number of contour edges; the number of non-zero pixel and the mean, maximal and minimal depths of the row; for each column

– the same parameters as for the rows. This leads to a 2206-dimensional feature vector (6 global image features, 11 features per row and 11 features per column) which contains, for our considerations, the essential information of a frame. Such ad-hoc features extraction may indeed contain a lot of redundant information that can not be easily determined based on local features. We therefore performed additional filtering on the entire data set as described in the next subsection.

Feature Sets
Since the hand runs through the same phases while performing reach-to-grasp task (or reaching task in general), the whole motion can be normalized by the time [12, 18]. The progress data should be temporally scaled for each trial, such that the trial begins at "time" 0 and ends at "time" 1. Such a normalization is usually made to enable direct comparison of the progress relevant features among all subjects and conditions.

We normalized the trial performance times for each video sequence in such a way that the mouse click (which indicated the beginning of the trial) is at "time" 0 and at "time" 1 the subject's hand was $1cm$ away from the virtual object to be grasped. Each frame, and also each feature vector was labelled with its normalized time. We split the set of feature vectors into six groups based on their normalized time. In the first half of the motion, the grasp pre-shaping and the wrist transport are in too early stage, which makes a prediction in this case a very challenging task. Indeed, in common settings, the wrist path is unpredictable until the transport phase reaches its peak velocity, usually at time 0.5 [8]. Thus, the frames from the set $[0, 0.5]$ were excluded because we were more interested in robust object prediction in a short interval before grasping that object.

To reduce information redundancy in the extracted feature vectors, features with constant values or very low variance within the datasets of each time segment were removed. Afterwards the data sets were transformed with algorithms for principle component analysis (PCA) and the transformed feature vectors were constrained to the first n principle components, with n determined such, that at least 99% of the information was contained in the components.

Results

None of the participants has appraised the study as being too long or the task as too difficult, thus we took all the data acquired into account. The participants were asked to grasp in a natural way, with moderate but realistic speed from the resting position to the surface. The mean task performance time was $1584ms$ ($SD = 363.38ms$).

Since we were interested in the influence of different parameters on the correlation between captured frames and the visual object we used a very simple correlation based classification algorithm, i. e. the *Naive Bayes* classifier. This classifier is based on maximization of the cross-correlation within the group of measurements (represented as multidimensional feature vectors) and minimization of the between-groups cross-correlation. We have tested four clustering variables: only the object type (OT); object type and position (OTP); object type and user (OTU); object type, position and user (OTPU). For each clustering variable and each training set a classifier was trained with 80% of the feature vectors, and its prediction rate was tested with the other 20% of the set. This process was repeated ten times and the calculated prediction rates were further evaluated with statistical methods. The achieved mean prediction rates in percent for the left (LEFT) and the bottom (BOTTOM) sensors

are shown in Tables 1 and 2 respectively.

normalized time	OT	OTP	OTU	OTPU
0.5-0.6	30.61	45.25	76.75	97.89
0.6-0.7	30.56	45.23	76.98	97.91
0.7-0.8	30.37	45.05	76.26	97.95
0.8-0.9	30.36	45.32	76.44	97.94
0.9-1.0	29.99	45.53	76.44	97.90

Table 1: Mean prediction rates in percent for the LEFT sensor.

normalized time	OT	OTP	OTU	OTPU
0.5-0.6	24.51	44.78	55.23	93.59
0.6-0.7	26.74	48.09	61.07	96.19
0.7-0.8	24.30	43.97	58.09	96.11
0.8-0.9	22.19	32.45	52.56	92.79
0.9-1.0	21.90	29.55	49.03	90.65

Table 2: Mean prediction rates in percent for the BOTTOM sensor.

The data was analyzed with a factorial analysis of variance (ANOVA), in order to test the within-group effects of the time set, sensor position and clustering variable. The analysis revealed a significant main effect for the sensor position ($F = 16556$, $p < 0.001$) as well as for the time set ($F = 684.99$, $p < 0.001$) and clustering variable ($F = 169820$, $p < 0.001$). The subsequent post-hoc analysis with the Tukey test revealed significant difference for all the tested conditions and values (with $p < 0.01$).

Discussion

As initially expected, the results show that the hand posture reflects the object to be grasped. Thus the object type could be anticipated in advance based on features extracted from the captured hand posture. Given the best prediction rates residing in the time segment $[0.6, 0.7)$ and the mean task performance time of $1584ms$ this gives us about $500ms$ in advance for use of this information by the interface. Although, the participants in our study performed the task slightly slower, then they may do this in a real user interface, the $500ms$ is a reasonable amount of time for an user-interface to adapt to the user's intention or to execute complicated background tasks, reducing the overall latency of the interface. One of the interesting results is that the prediction rates do not constantly increase with the hand approaching the visual object as initially expected, but have its peak values in the time cluster $[0.6, 0.7)$ and are then falling. We have currently no explanation for this fact, and will address it in detail in future research.

Not surprisingly the object type by its own is not sufficient as clustering variable. Indeed, the hand posture depends on the personal preference of the user. This may have led to the significantly better prediction rates in the condition OTU. Nevertheless, it is currently unclear, if there is a (perhaps broader) set of typical hand postures which could be mapped on a single object to compensate for the personal differences. Surprisingly, the object position has also a significant effect on the prediction rates, although its effect is not as strong as the personal preferences. This might be due to our initial feature extraction, which does not fully compensate for different hand orientations. We expect that using more advanced feature extraction techniques will reduce or eliminate this effect. Indeed, more evolved feature set, which compensate for different hand orientations and sizes, could be extracted from each frame as well as from the frame sequence. Such feature extraction would then make the recognition user-independent.

In general, in our implementation the recognition of an

object to be grasped depends on different parameters including the users' personal properties and habits, which may make a robust mapping of objects to grasp posture a challenging task. Nevertheless, our approach shows the feasibility of the task at hand and provides an easily reproducible procedure for establishing an initial corpus of training data. Based on the reported prediction rates, which have been achieved even with this very simple algorithm, we believe that a complex alternative method (cf. [10]) feed with our training corpus may achieve remarkable, in many cases user-independent, prediction rates.

The findings of the experiment have also shown that the affordance of an object plays an important role. Because there are often multiple ways to grasp an object a careful design of the UI items has to be taken into account. Hence, it sounds reasonable to design objects with unambiguous affordances that reduces the variability of possible grasps to a single gesture. It is now possible to design user interfaces that can be dynamically adapted based on predictions of the user's intention. Adaptation means that stereoscopically displayed 3D objects serve as virtually graspable objects of their real counterparts, which respond to the user's grasping behavior. Thus, an immersive interaction experience can be realized by "touching" virtual objects together with haptic feedback through the physical border of the interactive surface.

Conclusion and Future Work

In this paper we presented a study in which we collect a corpus of grasp postures for stereoscopic objects. The analysis of the gathered data shows that a recognition of the grasp posture during the *Reach to Grasp* phase is feasible a certain amount of time ($500ms$) before the user reaches the surface. This can be used to improve interaction and gives rise to novel user interfaces. These findings show that the objects the user wants to interact with can be predicted unambiguously before the user actually touches these objects. Following this, information about the grasp intention now allows the adaptation of the user interface to improve interaction. With such knowledge the potential of novel interaction techniques and improvements in UIs might be tremendous. In the workshop we like to discuss the novel design space of multi-touch interaction which derive if the grasp phase above the touch-sensitive surface is taken into account.

As next step we plan to develop and thoroughly evaluate an adaptive user interface system that makes use of the techniques and concepts proposed in this paper. There are different potential domains that can benefit from this UI and interaction concepts. A good example for cluttered interfaces is a (virtual) 3D UI for DJs that emulates a real mixer console. The browsing and interaction in large image databases visualized in 3D might be another interesting direction to investigate.

References

[1] S. Chieffi and M. Gentilucci. Coordination between the transport and the grasp components during prehension movements. *Experimental Brain Research*, 94:471–477, 1993.

[2] D. Coffey, N. Malbraaten, T. Le, I. Borazjani, F. Sotiropoulos, and D. F. Keefe. Slice WIM: a multi-surface, multi-touch interface for overview+detail exploration of volume datasets in virtual reality. In *Proc. of I3D '11*, pages 191–198. ACM, 2011.

[3] D. M. Coffey and D. F. Keefe. Shadow WIM: a multi-touch, dynamic world-in-miniature interface for exploring biomedical data. In *SIGGRAPH Posters*,

2010.

[4] A. Cohé, F. Dècle, and M. Hachet. tBox: A 3D Transformation Widget designed for Touch-screens. In *Proc. of CHI '11*, pages 3005–3008. ACM, 2011.

[5] O. Hilliges, S. Izadi, A. D. Wilson, S. Hodges, A. Garcia-Mendoza, and A. Butz. Interactions in the air: adding further depth to interactive tabletops. In *Proc. of UIST '09*, pages 139–148. ACM, 2009.

[6] H. Lahamy and D. Litchi. Real-time hand gesture recognition using range cameras. In *Proc. of the 2010 Canadian Geomatics Conference and Symposium of Commission I*, 2010.

[7] X. Liu and K. Fujimura. Hand gesture recognition using depth data. In *FGR' 04: Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*, pages 529–534. IEEE Computer Society, 2004.

[8] C. L. MacKenziea, R. G. Marteniuka, C. Dugasa, D. Liskea, and B. Eickmeiera. Three-dimensional movement trajectories in fitts' task: Implications for control. *The Quarterly Journal of Experimental Psychology Section A*, 39(4):629–647, 1987.

[9] J. Maycock, B. Blaesing, T. Bockemühl, H. J. Ritter, and T. Schack. Motor synergies and object representations in virtual and real grasping. In *Proc. of ICABB '10*. IEEE Computer Society, 2010.

[10] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:677–695, 1997.

[11] M. Santello, M. Flanders, and J. F. Soechting. Postural hand synergies for tool use. *J. Neurosci.*, 18(23):10105–10115, 1998.

[12] M. Santello, M. Flanders, and J. F. Soechting. Patterns of hand motion during grasping and the influence of sensory guidance. *J. Neurosci.*, 22(4):1426–1435, 2002.

[13] M. Santello and J. F. Soechting. Matching object size by controlling finger span and hand shape. *Somatosens Mot Res*, 14(3):203–212, 1997.

[14] J. Schöning, F. Steinicke, D. Valkov, A. Krüger, and K. H. Hinrichs. Bimanual interaction with interscopic multi-touch surfaces. In *Proc. of INTERACT '09*, Lecture Notes in Computer Science (LNCS). Springer, 2009.

[15] F. Steinicke, K. H. Hinrichs, J. Schöning, and A. Krüger. Multi-touching 3d data: Towards direct interaction in stereoscopic display environments coupled with mobile devices. In *AVI '08: PPD Workshop*, pages 46–49, 2008.

[16] S. Strothoff, D. Valkov, and K. Hinrichs. Triangle Cursor: Interactions With Objects Above the Tabeltop. In *ITS '11: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, pages 111–119. ACM, 2011.

[17] Y. Takeoka, T. Miyaki, and J. Rekimoto. Z-touch: An infrastructure for 3d gesture interaction in the proximity of tabletop surfaces. In *Proc. of ITS '10*. ACM, 2010.

[18] P. H. Thakur, A. J. Bastian, and S. S. Hsiao. Multidigit movement synergies of the human hand in an unconstrained haptic exploration task. *J. Neurosci.*, 28(6):1271–1281, 2008.

[19] A. D. Wilson. Simulating grasping behavior on an imaging interactive surface. In *Proc. of ITS '09*, pages 125–132. ACM, 2009.

[20] A. D. Wilson and H. Benko. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proc. of UIST '10*, pages 273–282. ACM, 2010.