S. Boll, S. Maaß & R. Malaka (Hrsg.): Mensch & Computer 2013
München: Oldenbourg Verlag, 2013, S. 191– 200

User Tracking for Collaboration on Interactive Wall-Sized Displays

Moritz Wiechers[1], Alexander Nolte[1], Michael Ksoll[1], Thomas Herrmann[1], Andrea Kienle[2]

Lehrstuhl für Informations- und Technikmanagement, Ruhr Universität Bochum[1]
Fachbereich Informatik, FH Dortmund[2]

Abstract

To support collaboration on wall-sized interactive displays we developed a system that is capable of distinguishing multiple users collaboratively interacting with a large surface at the same time. In order to allow for seamless switches between different modes of collaboration, the system uses camera based tracking thus requiring no additional hardware. The system also allows exploiting the position of a user in front of the screen display in order to show information about the users' context directly in front of them. This information could e.g. indicate which item is currently used by whom during a collaborative session, so that all participants can coordinate their actions. We present a study in which we assess the quality of the distinction mechanism, show possibilities for improvement and describe how awareness of the actions of others could enhance collaboration.

1 Introduction

Researchers as well as practitioners have tried to exploit the unique advantages of large interactive displays for collaboration in the past few years. These dispalys are expected to create more involvement as they allow direct interaction with the displayed materials by all participants, e.g. during a workshop. One major challenge in order to achieve this is the need to distinguish between multiple users who interact with the material at varying positions in front of the screen. For example, a simple copy and paste action can not be carried out via drag and drop since users may have to walk around each other, or a copy operation by a user can be followed by another user copying an element before the paste operation is completed. Consequently in order to assure that the correct elements are pasted, the system has to assign the actions to the correct user. There are several approaches supporting this e.g. *MERL DiamondTouch* by Dietz and Leigh (Dietz, & Leigh, 2001) which distinguishes users by their respective capacitive resistance. Other systems distinguish users by hand contour analysis (Schmidt et al. 2010), by analyzing the dorsal hand region (Ramakers et al. 2012), by tracking mobile phones (Schöning et al. 2008) or using digital pens (Rekimoto 1997). Apart from most of these systems being built for horizontal displays (e.g. tabletops), it also takes time to

set them up or they require additional hardware to work. Focusing on enabling collaboration at any moment during a workshop, we aimed at developing a system that is capable of distinguishing users without any preparation and without any additional hardware such as markers or mobile phones.Camera based systems such as Microsoft´s Kinect™ proved to be a good starting point as they are capable of distinguishing and tracking multiple people without equipping them with any additional hardware. As they are also cheap to buy, these cameras steadily gained popularity among practitioners and researchers alike. The main focus of research so far has been allowing natural physical interactions with a digital device as e.g. described by Stellmach et al. (Stellmach et al. 2012). However, there is also a multitude of other scenarios in which these cameras have been used as navigation systems for blind people (Mann et al. 2011) or 3D modeling of physical objects (Xu et al. 2012). They were also used to support collaboration on horizontal touch interfaces by supporting user distinction in combination with RFID-chips (Jung et al. 2011). The use of depth-based camera systems has proven to be very beneficial for user distinction and tracking especially within the area of cooperative work on large displays, however, research on this area especially with respect to collaboration on vertical displays is still missing to a large extent.

First attempts have shown that it is challenging to reliably distinguish users interacting with a large display using camera based tracking (Turnwald et al. 2012). This is due to the necessity to place the camera at a certain distance to the display in order to cover its whole width. However as these cameras only provide a certain depth resolution, their capability to reliably detect users deteriorates when the camera is placed too far away. Furthermore due to users frequently crossing each other, it is difficult for a single camera to continuously track them. So we came up with the idea to combine multiple Kinect™ cameras, thus developing a scalable system that is capable of covering walls of any size. There are some approaches combining multiple depth-based cameras in order to cover larger spaces or to solve the problem of overlapping viewpoints e.g. for body scanning (Tong et al. 2012) or tracking a person moving into another room (Schönauer & Kaufmann 2011). We however aim at combining multiple Kinect™ cameras with a large interactive display in order to support collaboration, which, to the best of our knowledge, has not been attempted so far.

Awareness is a crucial factor in multi-user collaboration especially in co-located settings (Herrmann et al. 2013). As users have to be aware of the actions of others, continuous operations, such as moving virtual items by dragging them over the whole width of the screen, proved to be not feasible due to the necessity of people moving around each other (Figure 2). Instead our mechanism allows picking up items by touching them and placing them elsewhere by pointing on the desired position. Although this prevents people running into each other, it also reduces the possibility to track a user's interaction with a certain item.

To meet the challenge of combining user distinction and awareness of user operations, we developed a system that detects the position of users in front of the screen and continuously visualizes the items they are currently interacting with in front of them. The system also may provide users with additional content tailored especially for them right in front of them such as interface components or context menus, which would otherwise have to be triggered by extra touches. Further explorations of these proxemic interactions for collaboration (Greenberg et al. 2011) show additional possibilities for improvement.

In what follows, we present an approach that combines user distinction and awareness. It is capable of distinguishing multiple users operating a large interactive display and can be used without any further preparation. Moreover, it determines the position of the user in front of

the screen and continuously visualizes the items the user is currently interacting with (Figure 2). Section 2 describes the usage scenario and requirements for the system while section 3 deals with its realization. Section 0 describes a preliminary evaluation which shows promising results as well as means for improvement. The paper concludes with a summary and provides an outlook on future work (section 5).

2 Scenario and requirements

The underlying scenario deals with tasks where a large number of items have to be sorted or clustered. A typical example is the collaborative modeling of processes where activities have to be assigned to larger units. This is often feasible if a modeling session has been started by gathering contibutions from users without sorting or clustering them immediately (Andersen & Richardson 1997), so that the creative phase is not disturbed (Herrmann 2009; Herrmann 2012). This results in the necessity to deal with a lot of items that have to be compared and aligned to each other, which can be very time-consuming if carried out by a single person. This led us to the idea of allowing multiple people to simultaneously cluster items on a large interactive display using simple pick & drop interactions (Turnwald et al. 2012).

During the evaluation of a preliminary prototype we observed a number of problems. First it was not possible to track participants continuously with a single Kinect™ camera due to the aforementiones reasons. Furthermore, the participants´ awareness of the operations of others was restricted as users picked items up, moved to another position and dropped them there. This led to participants being confused by items suddenly moving or disappearing which would not be the case when a participant moves an item by dragging it on the screen as s/he is easily observable by the others during that process.

These observations led to the following requirements, providing the basis for our system:

1 Continuous User-Tracking: The system has to be able to identify situations in which multiple users are simultaneously interacting with the screen and also to differentiate them. Additionally, it has to handle situations in which people stand close to each other or overlap. Manual error correction also has to be supported.

2 Smooth Error correction and efficiency: Users must be able to correct tracking errors efficiently. All in all, being automatically tracked must significantly reduce the number of necessary dialogue steps compared to the test case where each interaction task has to be accompanied by a step with which the users identify themselves.

3 Providing Awareness of other Participants' Work: To foster users' awareness while they are collaboratively working on the screen, the system has to provide features that allow users to be constantly aware about who is performing an action on which item. For this, it has to be able to determine the user's position in front of the display.

4 Marker free operation: As it may be feasible during workshops to spontaneously switch between front facilitated settings and multi-user collaboration in front of the screen, the system has to be instantly functional, thus at best requiring no time to prepare at all. Therefore, the system has to be able to be used on the fly, without attaching any markers to users or assigning additional hardware to them such as digital pens.

5 Flexibly adaptable System: The system has to be able to work with any vertical display regardless its size or form. This includes its flexibility to add additional cameras or even other detection systems when necessary.

3 Realization

3.1 User Interface

Based upon the aforementioned requirements, we created a prototype that is part of a software for the SeeMe[12] modeling notation. The system allows users to assign elements of the modeling notation to categories by selecting them with a single click and placing it within a designated cluster (Figure 1). To indicate that an element has been selected by a user, it follows the user based upon her position in front of the screen until it has been dropped. To achieve this, the system automatically detects each user operating the wall, assigning touch events to them and detecting their position in front of the screen. As errors are likely to occur, the system provides a mechanism that allows users to correct them manually. For that, each user is assigned a color and a letter and a menu is displayed next to the position where the element is dropped. This menu allows the user to simply select her color and letter on a clock-face (Figure 1) thus placing the right element there.

3.2 Setting and Hardware

The prototype is designed to be operated within a special facilitation collaboratory. Its centerpiece is a large, high-resolution interactive display (4.8m x 1.2m; 4200x1050px), which allows for seamless interaction over the whole width of the wall. In order to observe the whole area in front of the wall, we used two Kinect™ cameras. The system however is capable for operating a number of cameras thus allowing for future enhancement. We placed the cameras 3.7 meters away from the screen and about 2.8 high right below the ceiling of the room. The cameras are set at an angle of 60 degrees and have an intersection of roughly 7 square meters. Before coming up with this setup, we tested several camera positions aiming at maximizing the observed area and minimizing the previously described situation where users are overlapping (Figure 3). We also tested whether two structured light based cameras would interfere with each other but found no significant negative effect on the tracking.

[12] Visit http://seeme-imtm.de for further information.

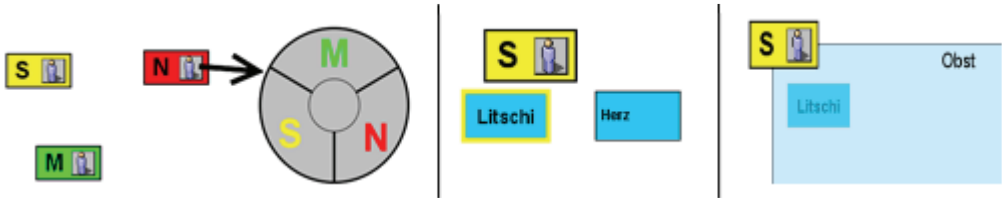

Figure 1: User menus and the clock-face (left) where users can select their literal; Selecting a cluster element with one touch (middle) and moving it to a certain cluster (right) with an additional touch. The yellow user menu offers the possibility of correcting a wrongly distinguished user.

3.3 Software prototype

At first we had to find a way to determine whether a person that is tracked by one camera is the same person that is tracked by the other camera. Once detected, we also had to relate the respective person to the touch events on the screen. To achieve this, we used a combination of the frameworks OpenNI and NITE. While OpenNI allows accessing the data of a Kinect™ camera, NITE is capable of analyzing that data to distinguish people by creating a point map of the scene and complementing each point with a user ID. Having established this for one camera, we then matched them together by mapping both coordinate systems into one. This requires some calibration which has to be done only once as the cameras are set at static positions. The calibration involves a chessboard pattern which both cameras observe simultaneuosly, using OpenCV to calculate a transformation which minimizes the reprojection error between the corners of the chessboard squares recorded from one camera and the corresponding corners recorded from the other camera. Afterwards we had to analyze each point map generated by each camera in order to find out whether two point maps represent the same person or not. For this we divided every bounding box of a point map into a set of smaller ones calculating whether they contain points of both point maps and then calculating the distance between their respective center positions.

After some performance tweaks we arrived at a solution that was ready to be tested. In order to relate the touch interactions on the screen to the people in front of it we used the same architecture as described in Turnwald et al. 2012: 2D coordinates of the touch points on the screen are sent to a user distinction server that also operates the Kinect™ cameras (Figure 2, right, step 2). Then, these coordinates are transferred into 3D coordinates and used to detect the nearest user in front of the screen (Figure 2, right, step 3). Afterwards, the original 2D coordinates are enriched with the corresponding user ID and sent back to the screen (Figure 2, right, step 4), where the intended operation is executed.

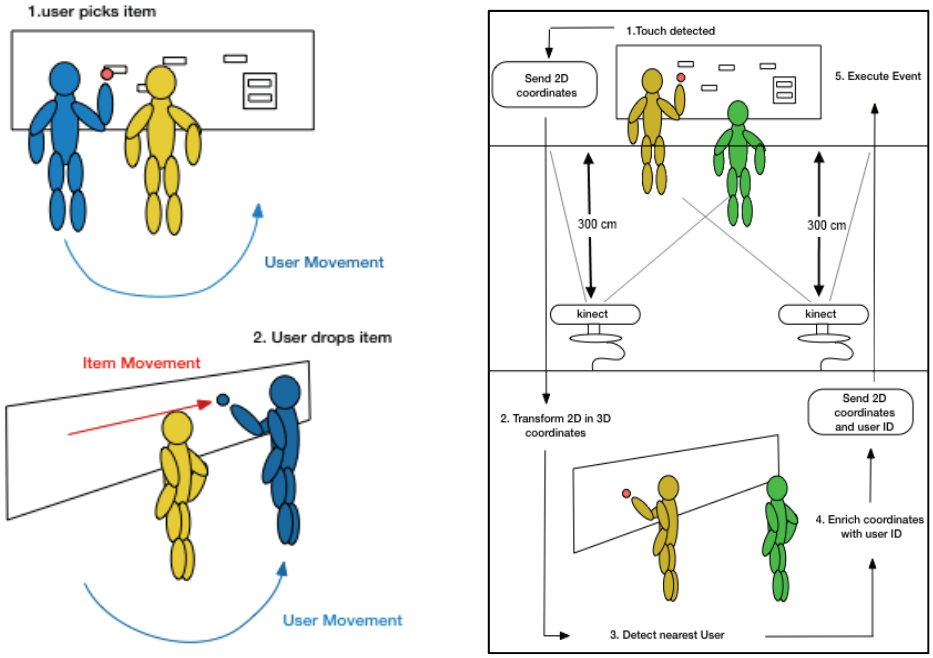

Figure 2. Item and User Movement, while changing the position of an item (left); Client-Server Architecture for multi-user distinction on a large interactive display using multiple cameras (right).

4 Study

4.1 Setting

In order to assess whether the system meets the intended requirements, we conducted five workshops with three participants each. The group composition was mainly based upon the figure of the respective users as we intentionally mixed larger and smaller people within a group to create a maximum amount of friction for the system. During the study the participants were given a task to sort a number of elements into a set of predefined clusters. We placed the elements on the left and the clusters on the right side of the screen to force the participants to switch positions frequently over a long distance (Figure 3). This caused additional friction and increased the probability for distinction errors. After a short briefing and a warm-up phase in which the users could become familiar with the system, we asked them to start clustering. For later analysis, we tracked the number of interactions with the display and especially the number of manual corrections by the users. Additionally, we also conducted interviews and asked the participants to fill out an AttrakDiff[13] survey after each workshop in order to gain additional insight into the participants´ perception of the system thus evaluating its hedonic and pragmatic quality.

[13] http://www.attrakdiff.de

Figure 3: The study parcipants working with the prototype (left); Arrangement of elements used in the clustering study (right)

4.2 Results

All in all 15 participants clustered 522 items. The participants were divided into 5 groups as we considered 3 participants to be a reasonable realistic group size. They needed 1392 recorded interactions to do so while each clustering phase lasted for about 9 minutes. Taking into account the number of corrections (296), this leads to a correction rate of 21%, which of course is too high for a real world setting but is still surprisingly low considering the setup of the study which was aimed at causing a maximum amount of friction. The correction rate tends to fluctuate minimally throughout the 5 groups mainly due to different constellations of people. Comparing this to a system where users have to identify themselves after each interaction, our system saves them a lot of touches (711). In total they saved about 34% of the touches with fluctuations depending on how often they actually had to correct errors.

The analysis of the interviews revealed that the participants experienced the prototype as a great support for cooperative tasks. Additionally, most of the participants emphasized that the prototype was “easy to understand and autodidactically to settle in”. Its usability had been considered to be “very functional, innovative and cutting-edge”. Furthermore, the correction mechanism was perceived as “comfortable, intuitive and easy to use” and the color allocation, despite lagging from time to time, was considered “clearly interpretive”. False assignments were “not seen as being disruptive”, but rather “arousing the participants’ curiosity and even being fun” for them at the same time.

Though positive aspects predominated within the participants’ interviews, there also was some criticism for example about the reaction time of the system. This caused some irritation as the participants sometimes did not know whether the system had processed an input or not. This behavior could often be attributed to the lack of the interactive display´s missing multi-touch functionality: Despite the users being distinguished correctly, the touch functionality of the screen sometimes provided wrong data e.g when multiple people touched the display simultaneously. This, in turn, led to partitipants having to pay strong attention to their inputs in order to make sure, that they had been processed correctly. Consequently, the users experienced an unpleasant distraction from time to time.

During these distractive situations, we observed that users sometimes lost awareness about who is doing what and needed additional coordination by talking to each other. By contrast, we realized through this observation that there was no need for additional coordinative communication if the system worked properly. Furthermore, we observed that the participants

easily started discussing – because of standing closely together – about items which could not be unambiguously assigned to a certain cluster.

The analysis of the AttrakDiff survey revealed that the users liked the way the user interface was implemented (HQ: “rather desired”), despite the users expecting it to be implemented differently with respect to its range of application (PQ: “consequently there is room for improvements in terms of usability”). Furthermore, the survey revealead that the users consider the prototype being stylish, presentable and premium, thus enchancing the group experience of an individual user (HQ-I). It also showed, that by being especially characterized as inventive, innovative and novel, the prototype can improve the user’s experience regarding her knowledge and skills (HQ-S), although it requires some improvements within these scopes (“Should you wish to bind the user more strongly to the product, you must aim at improvement.” and “Should you wish to motivate, enthrall and stimulate users even more intensely, you must aim at further improvement”). In conlusion, the survey conveys that the prototype is considered to be a very attractive solution (“the overall impression of the product is very attractive (ATT)”), which matches and supports the aforementioned results extracted from the single interviews.

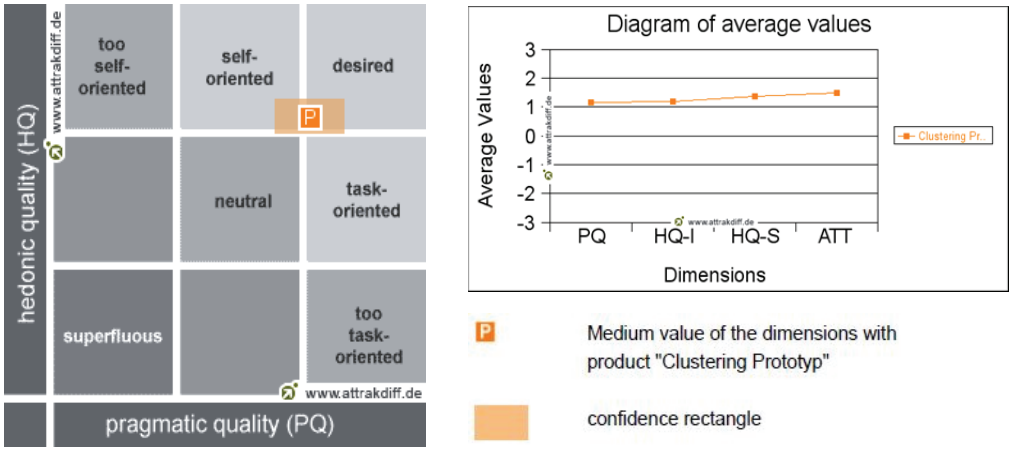

Figure 4: Portfolio with average values of the dimensions PQ and HQ and confidence rectangle of the product "Clustering Prototype"

5 Conclusion and Outlook

The results of the study indicate that coupling two Kinect™ cameras with a large interactive display allows automatic user distinction without the need of any further preparation or hardware. Furthermore, despite intentionally producing critical situations resulting in users having to correct errors at times, the system was capable of dealing with the issue of people overlapping each other to a large extent. The number of necessary interactions was significantly reduced compared to the theoretical test case of combing each interaction with an explicit step of identification. After having this established, the goal will now be to try to

further extend the setting, thus supporting bigger screens and using more cameras in order to improve the system.

The study also revealed the possibility and the usefulness of exploiting the position of a user in front of a large interactive display. Blending the position with the detection of touches and the aforementioned user distinction system allowed us to enhance the tracking of other participants' elements, even without the requirement to continuously hold contact to the wall by dragging. We observed that less coordination is necessary compared to those situations where the automatic tracking was disturbed. We assume that the immediate awareness of the interaction of others allows the participants to focus on communication which is related to the content of their tasks. This kind of beneficial effects of awareness, compared to cases where the sorting and clustering is conducted from a distance via laptops etc. is planned as a subject of further research.

Furthermore, we are also aiming at complementing the system with means to detect the size and shape of a person and also analyze the color and texture of their clothes in order to reduce the error rate. Adding these features would make the system more robust and would also allow it to reidentify a person after it has lost track of her/him due to other users covering each other while interacting with the screen or due to them leaving the observed area in front of the screen and reentering it afterwardsIn addition, we also plan on extending the functionality of the user interface with respect to its capabilities as a full scale modeling software that also allows creating, manipulating and relating elements to each other.

References

Andersen, D.F. & Richardson, G.P. (1997). Scripts for group model building. *System Dynamics Review*. *13*(2), 107–129.

Dietz,, P. & Leigh,, D. (2001). DiamondTouch: a multi-user touch technology. *UIST '01: Proceedings of the 14th annual ACM symposium on User interface software and technology*., 219–226. New York, NY, USA: ACM.

Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R. & Wang, M. (2011). Proxemic interactions: the new ubicomp? *interactions*. *18*(1), 42–50.

Herrmann, T. (2009). Design Heuristics for Computer Supported Collaborative Creativity. *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*., 1–10.

Herrmann, T. (2012). *Kreatives Prozessdesign*. Berlin Heidelberg: Spriger Verlag.

Herrmann, T., Nolte, A. & Prilla, M. (2013). Awareness support for combining individual and collaborative process design in co-located meetings. *Computer Supported Cooperative Work (CSCW)*. *22*(2), 241–270.

Jung, H., Nebe, K., Klompmaker, F. & Fischer, H. (2011). Authentifizierte Eingaben auf Multitouch-Tischen. *Mensch & Computer 2011: 11. fachübergreifende Konferenz für interaktive und kooperative Medien. überMEDIEN-ÜBERmorgen*., 305.

Mann, S., Huang, J., Janzen, R., Lo, R., Rampersad, V., Chen, A. & Doha, T. (2011). Blind navigation with a wearable range camera and vibrotactile helmet. *Proceedings of the 19th ACM international conference on Multimedia*., 1325–1328.

Ramakers, R., Vanacken, D., Luyten, K., Coninx, K. & Schöning, J. (2012). Carpus: a non-intrusive user identification technique for interactive surfaces. *Proceedings of the 25th annual ACM symposium on User interface software and technology*., 35–44.

Rekimoto, J. (1997). Pick-and-drop: a direct manipulation technique for multiple computer environments. *Proceedings of the 10th annual ACM symposium on User interface software and technology*., 31–39. New York, NY, USA: ACM.

Schmidt, D., Chong, M.K. & Gellersen, H. (2010). HandsDown: hand-contour-based user identification for interactive surfaces. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*., 432–441. New York, NY, USA: ACM.

Schönauer, C. & Kaufmann, H. (2011). Wide Area Motion Tracking Using Consumer Hardware.

Schöning, J., Rohs, M. & Krüger, A. (2008). Using mobile phones to spontaneously authenticate and interact with multi-touch surfaces. *Workshop on designing multitouch interaction techniques for coupled public and private displays*., 41–45.

Stellmach, S., Jüttner, M., Nywelt, C., Schneider, J. & Dachselt, R. (2012). Investigating Freehand Pan and Zoom. *Mensch & Computer 2012: interaktiv informiert–allgegenwärtig und allumfassend!?*.

Tong, J., Zhou, J., Liu, L., Pan, Z. & Yan, H. (2012). Scanning 3d full human bodies using kinects. *Visualization and Computer Graphics, IEEE Transactions on*. *18*(4), 643–650.

Turnwald, M., Nolte, A. & Ksoll, M. (2012). Easy collaboration on interactive wall-size displays in a user distinction environment. *Workshop "Designing Collaborative Interactive Spaces for e-Creativity, e-Science and e-Learning."*

Xu, D., Cai, J., Cham, T.J., Fu, P. & Zhang, J. (2012). Kinect-Based easy 3D object reconstruction. *Proceedings of the 13th Pacific-Rim conference on Advances in Multimedia Information Processing*., 476–483. Berlin, Heidelberg: Springer-Verlag.