Tangible Displays for the Masses: Spatial Interaction with Handheld Displays by Using Consumer Depth Cameras

Martin Spindler[1], Wolfgang Büschel[2], Charlotte Winkler[1], Raimund Dachselt[2]

[1]University of Magdeburg, Germany [2]Technische Universität Dresden, Germany

spindler@isg.cs.ovgu.de, bueschel@acm.org, charlotte.winkler@isg.cs.ovgu.de, dachselt@acm.org

ABSTRACT

Spatially aware handheld displays are a promising approach to interact with complex information spaces in a more natural way by extending the interaction space from the 2D surface to the 3D physical space around them. This is achieved by utilizing their spatial position and orientation for interaction purposes. Technical solutions for spatially tracked displays already exist in research labs, e.g., embedded in a tabletop environment. Along with a large stationary screen, such multi-display systems provide a rich design space with a variety of benefits to users, e.g., the explicit support of co-located parallel work and collaboration. As we see a great future in the underlying interaction principles, the question is how the technology can be made accessible to the public. With our work, we want to address this issue. In the long term, we envision a low-cost tangible display ecosystem that is suitable for everyday usage and supports both active displays (e.g., the iPad) and passive projection media (e.g., paper screens and everyday objects such as a mug). The two major contributions of this article are a presentation of an exciting design space and a requirement analysis regarding its technical realization with special focus on a broad adoption by the public. In addition, we present a proof of concept system that addresses one technical aspect of this ecosystem: the spatial tracking of tangible displays with a consumer depth camera (Kinect).

KEYWORDS

Tangible Magic Lenses, PaperLens, tangible display ecosystem, spatial interaction, multi-display environment, interaction above the tabletop, Kinect.

Introduction

Combining mobile displays with each other or even with larger stationary displays, such as tabletops [39], offers exciting new possibilities not only in terms of an enlarged presentation space but also as an increased interaction space that is particularly useful for co-located parallel work and collaboration, e.g., see [10, 22, 29, 40]. Building and studying such multi-display environments is the subject of current research in modern HCI labs, e.g., [14, 26, 32, 35], where researchers can experiment with technically complex and costly hardware installations that are usually not suitable for the average office environment or living room. Often, a large interactive tabletop or wall-mounted display is central to these installations. They serve as a global display that can be shared by multiple users for simultaneous work. Besides investigating techniques for interacting **on** a tabletop, a recent research goal is to extend the interaction space to the physical space **above** its surface.

Our PaperLens project [30] is such a system. From a technical point of view, it provides a rather complex solution for projecting digital imagery onto lightweight handheld paper-based projection screens that are tracked in three-dimensional (3D) space with six degrees of freedom (6DOF). This requires an expensive tracking system, consisting of six or more infrared (IR) cameras (e.g., Optitrack FLEX:V100R2) and a short-throw projector that are attached to the ceiling (see Figure 1). Together with a self-tailored interactive tabletop this sums up to a price of more than $22.000.

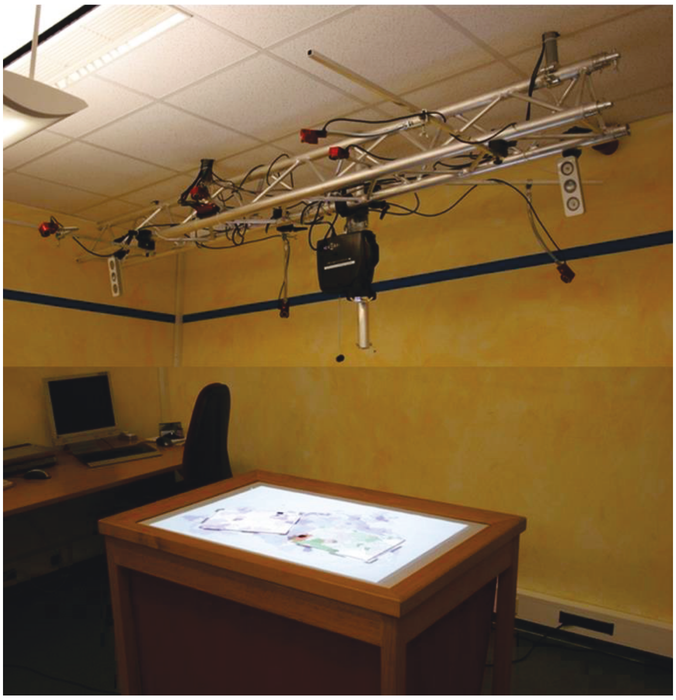

Fig. 1 The technical setup of the PaperLens system [30] as found in our lab is technically complex, expensive, difficult to setup and maintain, and too obtrusive in order to be suitable for the average office or living room

In terms of interaction, PaperLens utilizes the concept of spatially aware tangible displays (Tangible Magic Lenses) that users can interact with by grabbing and moving one or more personal handheld displays around in 3D space (spatial input). In this way, simultaneous exploration of complex information spaces is supported in a natural way [30, 32, 35]. Our experiences show that spatial input is a powerful input channel that integrates particularly well within multi-display multi-user environments and supports co-located parallel work and collaboration explicitly. This is also supported by many demo sessions and user studies, e.g., [30, 33], where we received very positive user feedback, not only from average users or children, but also from domain experts (e.g., biologists or radiologists). In this process, we have often been asked when and how these techniques could be available to them.

In this article, we address this question by proposing our idea of a low-cost tangible display ecosystem that is suitable for a broad audience, i.e., that integrates well with modern mobile display technology and is easy to setup and maintain, robust, affordable, extendable, and unobtrusive. In particular, we analyze important technical requirements of such ecosystems. We also present a prototypic implementation of one core aspect: the spatial tracking of handheld displays with a consumer depth camera (Kinect). We do this in the hope that it will bring spatial interaction with handheld displays a step further in becoming widely available beyond research labs.

The remainder of this article is organized as follows: First, we review related work including typical application domains. We then explore the fascinating design space of tangible displays, where we also assess two principle technical strategies for handheld displays. After this, we present a requirement analysis of a tangible display ecosystem suitable for the masses. We then continue with a description of a prototypic implementation that addresses one aspect of this ecosystem: the spatial tracking with a consumer depth camera. We finally conclude with an evaluation and discussion of the prototype and some final remarks.

Related Work

The question of how to make emerging technologies available to the masses has been investigated, among others, in the related field of ubiquitous computing, e.g., by Costanza et al. [7]. In this section, we will provide a brief overview of important Tangible Displays systems, outline possible fields of application, and review spatial tracking approaches with a special focus on depth cameras.

Spatially Aware Tangible Displays

The approach presented in this paper continues our research into spatially aware tangible displays used in tabletop environments [30–35]. The underlying concept goes back to the vision of ubiquitous computing as defined by Weiser [41] that aims at merging the digital world with the physical (analog) world. This idea was adapted by Ishii’s and Ullmer’s tangible user interfaces (TUIs) [16], where interaction with digital information is provided through physical manipulation of real-world objects. Inspired by the notion of see-through interfaces, as introduced by Bier et al. [4], these objects can also be spatially aware handheld displays (e.g., mobile phones) that serve as tangible magic lenses into the virtual world. One of the first mobile displays for ubiquitous usage has been proposed by Fitzmaurice, who presented a spatially aware palmtop computer for exploring 3D-situated information spaces for revealing virtual information associated with arbitrary objects in an office environment [9]. Other prototypes include the Peephole Display by Yee [43] that combines the navigation in two-dimensional (2D) virtual workspaces with digital pen input. With the metaDESK project, Ullmer and Ishii [39] applied this idea to a tabletop environment, where users can freely navigate through polygonal 3D models by moving an arm-mounted LCD display through the space above the tabletop that is also responsible for providing contextual graphical information. More recently, Tsang et al. presented the Boom Chameleon [38], a tablet PC mounted to a mechanical arm that is tracked in 3D space. While not a spatially aware display in the narrower sense of the term, Siftables by Merrill et al. [27] laid important foundations for multi-display setups in which displays are aware of their mutual arrangement.

Application Domains

Spatially aware tangible displays have been applied to many application domains in fields like science, education and design, where they have demonstrated their support for co-located parallel work and collaboration. In the following, we present three of them as an example.

Exploration of Scientific 3D Datasets

One particular field of application for spatially aware handheld displays is the collaborative exploration and manipulation of large 3D data. Besides geological or biological data, prime examples are medical volume datasets acquired from MRI or CT imagery. In a collaborative interactive space consisting of a tabletop and multiple handheld displays, such datasets can be understood as residing in the physical space above the tabletop. When users move handheld displays through the interaction volume, arbitrary, user-defined cutting planes can be computed in real-time and displayed, as demonstrated by us in [30]. This form of direct interaction allows for a fast and flexible exploration of the whole dataset or specific structures within and also integrates well with touch- or pen-based input that can be used, e.g., for creating and managing annotations [31]. With Tangible Windows [35], we have shown how additional head input, e.g., used for personalized head-coupled perspective views on handheld displays, can help present 3D spatial relations even more realistically.

Video Sifting and Editing

As opposed to just passively consuming single videos, sifting and sense making are a crucial requirement for working with large amounts of unstructured video material, involving the exploration, analysis, rating, grouping, and editing of video snippets. More and more, these tasks are performed collaboratively. As a basic approach to map time to the vertical dimension we introduced the concept of temporal information spaces that we proposed to combine with a global video time line displayed on a tabletop [30]. Based on this idea and taking it a step further, Lissermann et al. demonstrated the PaperVideo project [26] that addresses many of these issues by using multiple spatially aware paper-like displays for working with multiple videos simultaneously. One of the key ideas of PaperVideo was to borrow from best practices of working with physical paper documents, namely the ability to spatially lay out, structure, and rearrange multiple documents in parallel, allowing users to grasp a higher amount of information in the periphery [21, 36]. Results from a recent user study [33] indicate that moving a handheld display up and down is more appropriate for time browsing (if considered as the primary interaction goal), while horizontal display movements should be reserved for secondary interaction goals, such as the selection of video snippets scattered on the tabletop.

Information Visualization

In many fields of business and science, large amounts of data have to be visualized and examined. Such complex datasets usually cannot be presented in a single image without the risk of cluttering. Filtering or presenting multiple views on the data can mitigate this problem. With Tangible Views [32], we have demonstrated how these views can be made physically tangible by using spatially aware handheld displays that not only provide additional screen space but also a new means of interaction, combined in single tangible objects. Tangible Views also support the focus and context concept, i.e., the tabletop serves as the contextual background view, while handheld displays provide local views into it. For interaction, the height of handheld displays is used, e.g., to encode zoom factors or specific levels of abstraction. Other examples include changing the parameters of a fisheye lens by rotating the display or exploring a 3D space-time cube by either slicing it vertically (overview of all time steps associated to a single location) or horizontally (selecting a single time step). In summary, the support for multiple handheld displays not only makes Tangible Views a suitable tool for the visual comparison of portions of the data, but also facilitates co-located collaboration between users.

Spatial Tracking of Objects in 3D Space

Reliable spatial tracking of objects in 3D space is a fundamental requirement for a spatially aware tangible display system. Various technical approaches have been utilized in the past. In the Shader Lamps project by Bandyopadhyay et al. [3], visual markers were deployed. Zhang et al. [44] use a Hough transform-based approach to visually track a panel without markers. In [23], a pattern is projected onto tangible displays. These are equipped with light sensors connected to a microcontroller that computes the position from the detected pattern. Infrared (IR) marker-based tracking is used in many projective tangible display systems, e.g., in [14, 26, 30].

Among the first works to utilize a depth camera for touch detection is [42] by Wilson, who uses a Kinect sensor to capture raw depth images and extracts touch events by thresholding operations based on a known model, e.g., the distance to a flat surface. In [13], Harrison, Benko and Wilson propose a wearable system for multitouch interaction, based on a depth-camera and a projector. Finger tips are detected using a gradient model. Surface reconstruction from Kinect depth data was done by Clark et al. [6] to provide real-time, user-configurable environments for an AR racing game. Extensive real-time mapping of scenes and simultaneous tracking of a single depth camera is presented by Newcombe et al. in their KinectFusion system [28]. In LightBeam [15], Huber et al. use a combination of a Kinect sensor and a regular webcam to track objects and surfaces. They aim at turning everyday objects into projection surfaces for nomadic pico projectors.

Design Space

We continue our discussion with a brief review of the design space of spatially aware tangible displays, e.g., as presented in [35]. In addition, we will examine important properties of two principle types of tangible displays: projective and active displays. We will compare their advantages and drawbacks, and discuss how these impact a tangible display system in terms of interaction and technical realization.

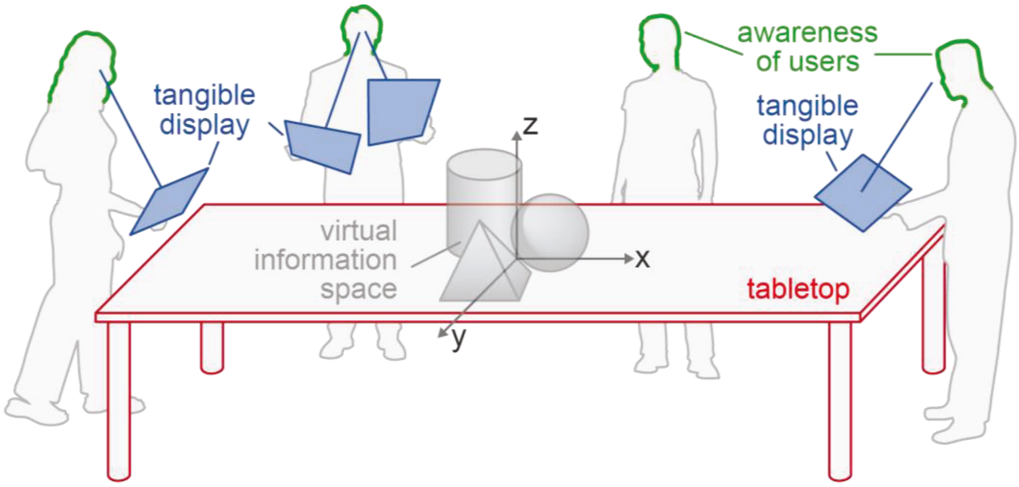

Fig. 2 A multi-display tabletop environment consisting of the following entities: one or more large global displays (e.g., a tabletop or wall display), multiple spatially aware handheld displays (e.g., mobile phones or pieces of paper), the ability of being aware of spatial positions of users (e.g., by head tracking), and digital content (e.g., volumetric datasets, geographical maps, or multimedia content). (Modified figure taken from [35])

Principle Setup

The principle setup of a tangible magic lens system is illustrated in Figure 2. The main components are one or more large stationary displays (in our case a tabletop) and one or multiple handheld displays that provide independent views into a virtual information space. Optionally, the system can be aware of the locations and view directions of one or multiple users, e.g., by spatially tracking their heads, such as demonstrated in [35]. The combination of stationary and handheld tangible displays provides several benefits. As a multi-display environment, it enables the simultaneous use of global and local views and thus facilitates co-located collaboration. When employing active displays, e.g., tablets or smart phones, users can also take their data with them, seamlessly alternating between mobile usage and a fixed interactive space.

Users can interact with such a system through three fundamental types of input that together constitute the interaction space of spatially aware tangible displays [35], namely surface input, spatial input, and head input. **Surface input** refers to interactions that are performed directly **on** the surface of displays, e.g., by touch and digital pen input [32]. It provides a good means for performing direct selection or manipulation tasks of objects that are visible on the screen, e.g., by pointing at a particular data item with a finger or by annotating/drawing a picture with a pen. **Spatial input** relies on the system's knowledge about spatial positions of handheld displays that are being tracked in physical space with six degrees of freedom (6DOF). This allows users to directly interact **with** a handheld display by moving or rotating it through the physical space. In this way, a rich set of interaction techniques becomes available that exploit the spatial interplay of displays. As a first but non-exhaustive attempt to categorize such techniques, see the interaction vocabulary described in [32]. **Head input** refers to the optional awareness of head movements that can help to distinguish between users in terms of user ID and position. This is valuable information for systems supporting user collaboration. It can also be employed for secondary interaction tasks, e.g., by providing users with the right perspective during 3D navigation [35] or by presenting different levels of detail depending on the distance of head and display [12], which can help make the interaction more natural.

In conclusion, two major features of this design space are (1) the support for co-located parallel work and collaboration by combining multiple personal handheld displays with one or more shared global displays and (2) the simultaneous support of different input modalities that are close to what users are familiar with in everyday life. This includes the pointing and drawing with fingers and pens, the spatial arrangement and manipulation of objects, and the use of the head for changing the focus to a particular object. Certainly, some of these input types address particular interaction goals better than others. By cleverly assigning them to particular tasks, such as spatial input to navigation and surface input to selection, we are able to get a step closer to the overall goal of making the interaction more natural. The use of orthogonal input strategies can decrease the need of mode switches, which are often perceived as being distractive. This, in turn, may help take away mental load from users, thus potentially freeing intellectual capacity for more productivity and creativity.

Projective and Active Displays

We distinguish between two principle types of spatially aware handheld displays: **projective** and **active displays**.

Projective Displays

Traditional display solutions, such as LCD panels, do not always provide suitable form factors for a seamless integration into a tangible display system. This is because these displays are often too heavy, thick, big, rigid, and sometimes even too expensive if many devices are needed. This lack of technology motivated researchers to come up with a variety of lightweight handheld display solutions that can easily be customized in terms of shape and size. Most of these solutions use a projective approach, where digital image content is dynamically projected onto spatially tracked (non-instrumented) projection mediums that are made of paper, cardboard, acrylic glass, porcelain, cloth, etc. Prominent examples for this are the paper-like screens from PaperWindows [14]. Projective displays also include everyday objects, such as mugs, playing cards, or the surface of a table [18].

One particular advantage of projective displays is their flexibility in terms of form factors. They can be made very thin and lightweight (e.g., by using cardboard or foam board), do not feature annoying display frame borders, can show image content on their front and back (e.g., useful for flipping [32]), allow for arbitrary shapes (e.g., discs [30]), can be extended into the third dimension (e.g., as cylinders or cubes [2]), are inexpensive, and usually they are easy to reproduce. The projective approach also allows for advanced form factors that further enrich the interaction space, such as changing a display's shape and size. Examples for this are rollable displays, e.g., Xpaaand by Khalilbeigi et al. [19], and foldable displays, such as presented by Lee et al. [24] and Khalilbeigi et al. [20].

Projective handheld display technology has been integrated into tabletop environments, e.g., in PaperLens [30]. With their SecondLight system, Izadi et al. [17] presented a more self-contained approach to allow rear-projection on both a tabletop and handheld projection screens above it. It supports Frustrated Total Internal Reflection (FTIR) based touch input [11] on both tabletop and mobile displays. Unfortunately, it is technically more complex in that it is based on electronically switchable diffusers and thus does not support large table displays so well.

As a downside, projective displays exhibit a rather limited mobility. This is because they only work within technically complex environments that are usually stationary. These installations are necessary to precisely determine the position and orientation of the mobile projection screens in 3D space and to provide the infrastructure for projecting images onto them. Often, projective displays also suffer from poor image quality in terms of resolution and noticeable shifts between object and projection space that are caused by imprecise tracking and projection. Also, curved surfaces or materials with poor reflective properties may limit the projection quality. Beyond that, occlusion (i.e., shadows) can be a problem, e.g., for projection screens positioned on top of each other.

Active Displays

Using active displays, e.g., smart phones and tablets, can solve many disadvantages of projective displays. They feature high-quality displays (e.g., the iPad's Retina display) and thus do not require complicated projector setups. This also implies that device tracking is solely used for spatial interaction and therefore can be less accurate. This is an important benefit and allows for the application of less obtrusive tracking technology, e.g., marker-less approaches. Another advantage of many active displays is that they provide precise multi-touch capabilities out of the box. Beyond that, they are often instrumented with a variety of useful sensors, e.g., accelerometers, near field communication (NFC), and compasses, which add further degrees of freedom to the interaction. In this way, active displays address two technical challenges of a tangible display system: They provide a built-in display solution and a multi-touch interface.

Despite all these advantages, active displays are less flexible than projective displays in terms of form factors. They usually are heavier and thicker, have noticeable display frames, are less variable in shape (if at all), and support only a front display. Technical progress might change this in the future. For example, organic light-emitting diode (OLED) technology or bendable e-Ink displays (e.g., as used in PaperTab [37]) could be applied. Nevertheless, a seamless integration of everyday objects that serve as tangible displays, such as mugs or playing cards, is only realistic with the projective approach. We therefore believe that a fully functional tangible display system should support both active and projective displays.

Requirement Analysis

Spatially aware tangible displays are not necessarily restricted to a fixed interaction space, such as an office or living room, but are suitable for public places, too. For example, in a shopping mall, several users could simultaneously navigate through the mall's map that is displayed on one of many large public displays by holding and moving their mobile phones on and above the big screen. In the same way, two of these people could hold their phones close to each other in order to discuss a private matter – now with a single display twice as large. Also, one of the users could zoom in and out of a map that is displayed on her phone by simply moving the device up and down relative to her head; the latter would also work in the wild, e.g., by solely relying on a phone's internal cameras and motion sensors.

These examples illustrate different aspects of a simple idea: using the spatial relationships of displays and their interplay with each other for making the interaction more natural. Of course, this only works provided that all involved components cooperate proactively. We believe that as of today the technical instruments are already available to implement such scenarios. What is still missing is a platform that brings together the different components into a unifying system. In context of the introduction of the iPhone/iPad and its associated development platform that dramatically propelled the research and spread of multi-touch-based interaction design, we think that a similar effect could happen to spatial-based interaction principles. In fact, we expect a second revolution of mobile computing that in the future will rely much more on spatial input and the interplay between displays.

In this context, our previous research addressed the foundations of tangible display interaction mostly using the example of a tabletop environment. In particular, we presented basic interaction patterns, design guidelines and case studies on the basis of working prototypes and their evaluation. However, our experiences also show that the development of functional applications with real benefit to users is still too difficult, if possible at all. We identified three major areas that need to be addressed to tackle this problem: (a) provide reliable input and output technology, (b) support easy application development, and (c) bring technology to users.

Technical Requirements

In order to overcome these obstacles, a tangible display system should be based on affordable and broadly available consumer hardware. It should also be modular, interoperable and easy to setup and maintain, as well as provide open and easy access for public application development. On the technical side, the system should address a catalog of features including:

R1 **Reliable spatial tracking** of active displays, passive projection media, and optionally users, e.g., their heads

R2 **Reliable surface input**, e.g., touch- or pen-based input on mobile phones and everyday objects

R3 **Open protocols for inter-device communication**, e.g., for streaming device states and digital image content, posting system notifications, and exchanging application-specific data between various displays

R4 **Support of projective displays** by projecting image content onto spatially tracked objects

R5 **Automatic calibration** of tracking and projection equipment

R6 A **system-wide component catalog** that provides detailed information about available equipment. Besides a unique identification (ID) for each device, this should also include, e.g., the 3D shape of a coffee mug, the precision of a tracking sensor, or the sensor functionality that a tablet is willing to share with the environment.

R7 Automatic or manual **de/registration of components**, e.g., a new smart phone joining the workspace or a tracking sensor that is being removed from the environment.

R8 **Expansion of the programming model** used for application development, e.g., by supporting spatial- and pen-based input events

R9 **Unification of application development** ("write once, run anywhere") that not only supports all popular mobile platforms (e.g., iOS and Android), but also projective displays.

R10 **Support of state-of-the-art software frameworks**, such as used in medical or scientific computing (e.g., VTK and Qt). This is in particular an issue for projective displays, where changing projection angles introduce perspective distortions[1].

A Cooperative Environment

Addressing these requirements is a complex endeavor that demands cooperation between all participating components. This includes: (a) active and/or projective **tangible displays**, (b) **users** and (c) an **augmented workspace**, i.e., a spatial environment that provides and/or connects various sensors, computational power, public displays, and possibly one or more digital projectors. While the workspace will often exist in a pre-installed form, ad-hoc augmentation might play just as vital a role, e.g., by using high-bandwidth Internet cellphone links that tether various smart devices, such as mobile phones, portable sensors and projectors.

In the tangible display ecosystem, cooperation and handshaking usually rely on technical and social protocols that all components should adhere to. For example, when new tablets arrive at a workspace a registration process is triggered. This informs the workspace that there are new devices requesting to be tracked and possibly to share a public display. The workspace then provides the tablets with these resources, but in turn, might also need some help, e.g., information about their shape and size or access to internal sensors (sensor fusion). During spatial tracking, the workspace could eventually lose some displays or confuse them with each other, e.g., because they overlap. In such cases, the tablets could show visual hints, indicating where users should move them, e.g., away from each other, in order to let the tracking

[1] While the correction of such distortions is usually trivial, the tricky part is to grab the graphical output of a rendering framework before it is sent to the framebuffer of the projector.

system find them again. Such visual clues could even encode individual calibration patterns that help the tracker to distinguish between the displays. When leaving the workspace, devices should actively deregister themselves from the workspace. This can be done automatically, e.g., when the distance to the workspace exceeds a certain threshold, or manually, e.g., when a user switches a device off. Deregistration of devices helps keeping the system-wide component catalog up-to-date (see requirement R6). This catalog stores valuable information that can simplify the spatial tracking of objects. For example, if the system knows that there are only two tablets to be tracked, false candidates can be rejected easier.

Envisioned Setup

In order to implement such cooperative tangible display environments, we propose a principle setup that extends the idea of LuminAR Bulb [25]. LuminAR Bulb combines a Pico-projector and a camera in a single device with a compact form factor. It can be screwed into standard light sockets everywhere. This allows for a simple and unobtrusive way of setting up a tangible display workspace, e.g., as illustrated in Figure 3. We envision that these bulbs do not need to provide all functionality, but can rather be specialized to a specific task. A Tracking Bulb, for example, could solely address the spatial tracking of active displays. In contrast, a Projection Bulb could specialize on projecting digital content onto arbitrary surfaces possibly using the information provided by a Tracking Bulb. In this way, using multiple bulbs that wirelessly communicate with each other can extend a workspace to better fit the demands of users.

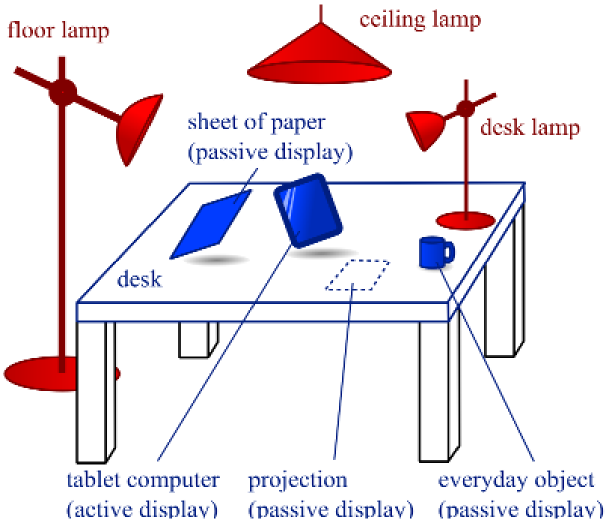

Fig. 3 We propose an enhanced technical concept for Tangible Displays that is easier to setup and maintain, less expensive, robust, modular, and unobtrusive

Prototype

While implementing a tangible display ecosystem is the long-term goal of our research, in this work we only address the short-term goal of building a prototypic (yet affordable) technical solution for the spatial tracking of displays, in our case handheld projection screens made of cardboard and the iPad. To achieve this goal, we use a cost-efficient consumer depth sensor: the Microsoft Kinect. The use of one iPad and a single Kinect enables us to build a simple tangible display system for less than $900, though not including the optional tabletop display or further iPads. The Kinect also supports our design goal of concealing as many of the technical aspects as possible from the user, e.g., there is no need to glue markers on the iPad.

Background and Previous Situation

The tracking approach presented in this article continues our work on PaperLens [30] that already addresses many of the technical requirements discussed in the previous section. PaperLens is a fully functional projective tangible display system that supports the reliable spatial tracking of paper-like projection screens (requirement R1) as well as the projection of digital content onto them (requirement R4). It also provides basic surface input on paper displays (requirement R2) that in our case is based on Anoto digital pen technology and Arduino pressure-sensitive buttons [35].

We successfully integrated Qt 4.8 (a cross-platform application/GUI framework) into PaperLens. This enables us to project arbitrary Qt-widgets onto tracked paper displays, which opens the door to a rich world of software frameworks (requirement R10). We also expanded the Qt-event model by new event types, e.g., concerning spatial input. We channel relevant events, such as coming from digital pens or movements of displays, to the proper GUI-elements, e.g., the ones that are associated with a particular paper display (requirement R8). While this simplifies the application development for projective displays considerably, the world of active displays is still a separate one. With the announcement of Qt 5 that promises the support of iOS and Android, we have high hopes of bringing these two worlds together in the near future (requirement R9).

PaperLens includes a modular architecture for inter-device communication that allows us to stream standardized events between devices (requirement R3), such as 2D positions of digital pens, 6DOF positions of displays and heads, and button states. This enables us to decouple tracking technology from application development, thus providing a simple way to seamlessly switch between various tracking approaches (and to add further ones). This includes tethered magnetic tracking (Polhemus Fastrak), infrared marker-based tracking (OptiTrack IR-cameras), and a mouse-based GUI-controlled tracking that we use, e.g., when coding and testing outside the lab.

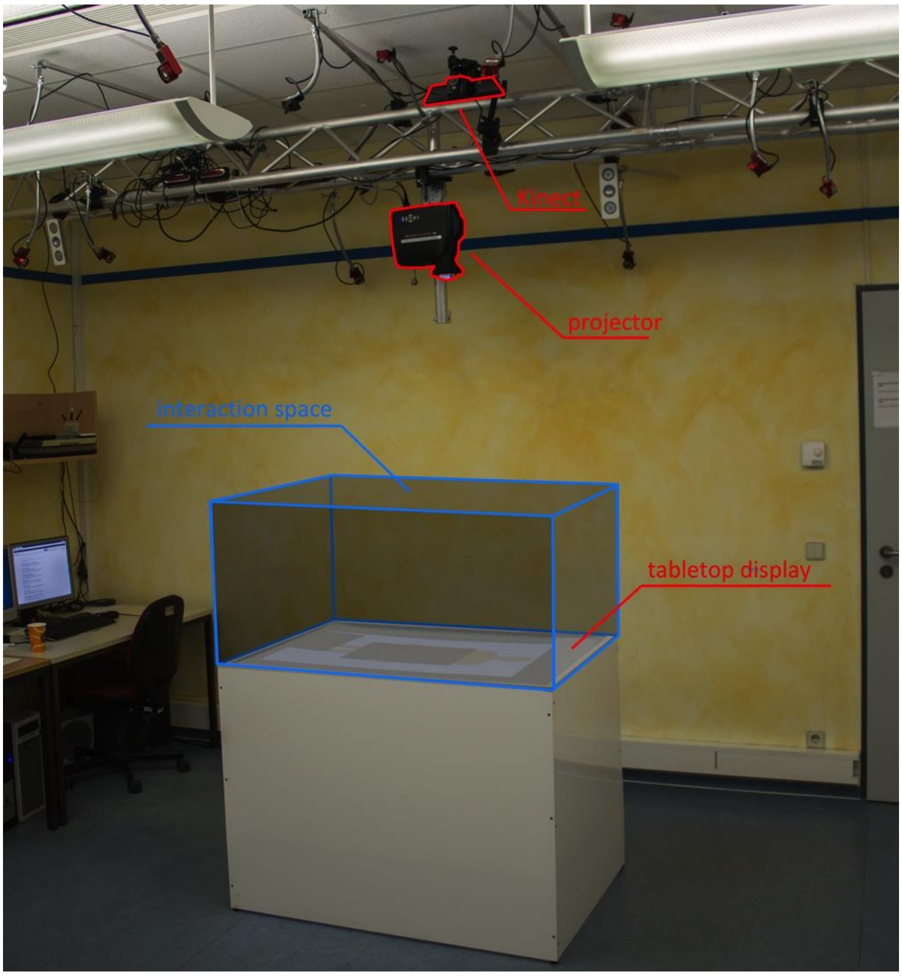

Fig. 4 Setup in our lab with a Kinect at the ceiling that replaces the IR cameras previously used

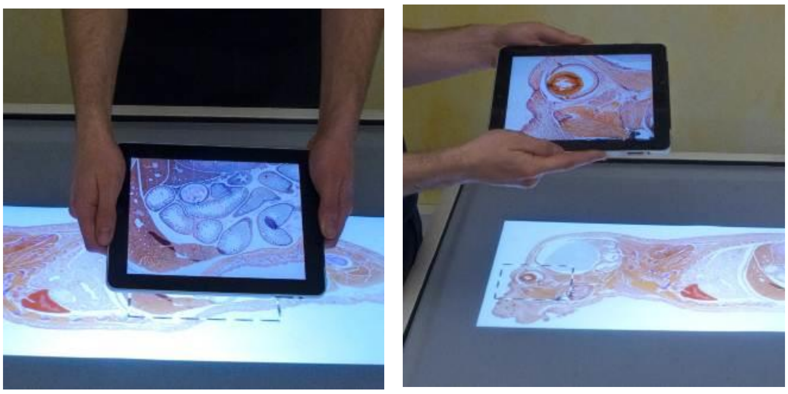

Fig. 5 An iPad is used to explore a high-resolution image of a cut through a rat embryo. Lifting the device up or down controls the zoom factor. Horizontal movements allow for panning. Due to the marker-less tracking provided by the Kinect, iPads are readily functional, i.e., no markers need to be attached to them.

Implemented Tracking Algorithm

We attached a Kinect to a crossbeam on the ceiling approximately 1.60 m above the table surface (see Figure 4). Following a convention from PaperLens, we use a world coordinate system that has its origin in the middle of the tabletop with the Z-axis pointing up. This world coordinate system is used for the system-wide exchange of positions and orientations of all tracked displays and heads. For this purpose, we extended the system with the capability to stream device information between iOS devices in real-time and use this as input for the interaction with them (see Figure 5). Our goal was to spatially track multiple tangible displays simultaneously by solely relying on the depth images provided by the Kinect that we feed to a designated server (PC). One of our design decisions was not to make use of the color (RGB) channel. This is because we consider the visible spectrum as an unreliable source of information, due to unpredictable changes of screen content on tracked displays. However, future iterations of the system might use this additional information, for example, to allow tangible displays to optically send feedback back to the system, such as visual patterns that encode display IDs, whenever the system is requesting such information. Our spatial tracking approach assumes flat displays that we model as 2D planes in 3D space. The algorithm consists of the following steps:

(1) Camera calibration: This is done only once during setup, as long as the positions of the tabletop and Kinect are not changed. For this purpose, we find the transformation between the Kinect's image coordinate system and its own local coordinate systems by internal calibration. We then compute the mapping between local space and world space using the known size and geometry of the tabletop.

(2) Detection of candidate regions: During operation, each incoming depth image (see Figure 6a) is masked so that only blobs with a distinct height above the table remain visible (see Figure 6b). Since these candidate regions are (almost) free from sudden changes of depth, we consider each of them to contain not more than one of the tracked displays.

(3) Rejection of false positives: As users hold the displays, candidate regions usually include parts of the hands and arms. However, some of these regions may not even contain a display at all. We discard such false positives by ignoring the entire candidate region if it is too small or thin. For this purpose, we apply a distance transform and thus get a rough estimate of the maximum extension of the thickest region within a candidate measured in pixels (see Figure 6c). If this value is above a certain threshold, we can be quite sure to have found the center of a display and not a part of the hand or arm. As the thresholds depend on how close a display (or arm) is to the depth sensor, we dynamically change them depending on the corresponding value in the depth buffer.

(4) Determination of spatial positions: For each positive candidate region (see Figure 6d), we then compute a robust mass center by considering multiple depth pixels that lie in the vicinity of the display center as found in step (3). We then transform the average depth of these pixels along with the 2D mass center into world space and use the result as the spatial position of the display.

(5) **Determination of spatial orientations:** In order to determine the display's spatial orientation, we compute its normal vector by utilizing a RANSAC-based [8] plane fitter that we feed with random depth samples of the candidate region. We also calculate the eigenvectors of the covariance matrix that represent the principal components (orientation) of a candidate (see Figure 6e). Due to symmetric display shapes, there are ambiguities with respect to front vs. back and top vs. bottom orientations. We address this issue by applying constraints, e.g., a display's orientation cannot change more than 10 degrees between two frames.

(6) **Assignment of IDs:** To distinguish between displays and assign IDs to them, we use their most prominent visual properties: their shape and size. Sampling the distance map along the principles axes allows us to determine the approximate width and height of a candidate. This information is sufficient to distinguish between a rectangular, quadratic and circular paper display as well as the iPad (see Figure 6f). Although we cannot reliably distinguish between two iPads, we maintain consistency by reusing the last known ID at a particular position.

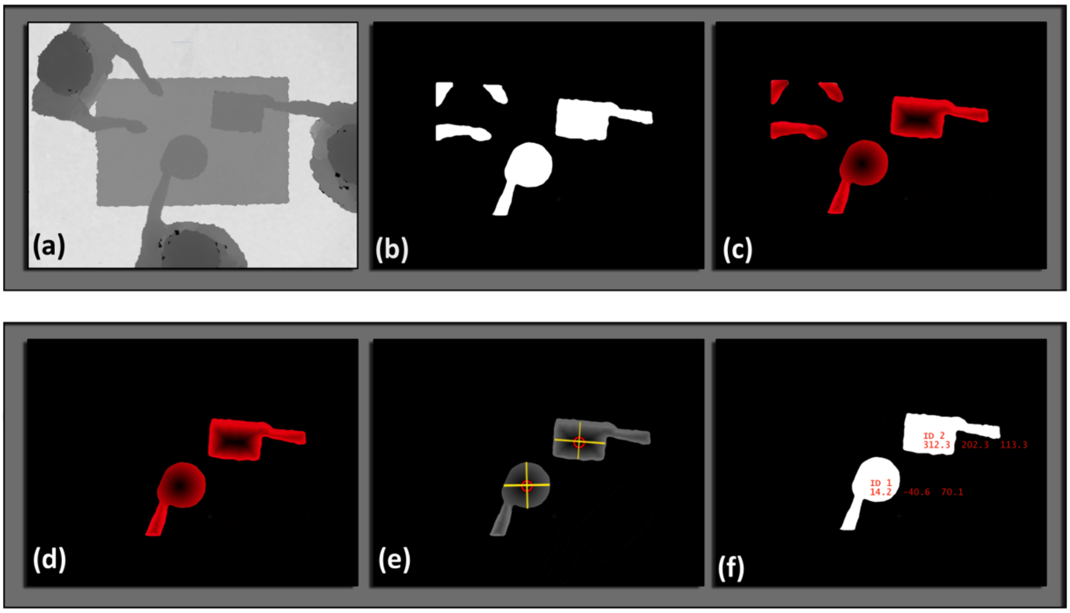

Fig. 6 The tracking pipeline: (a) incoming depth image, (b) candidate regions, (c) distance transforms, (d) rejection of false positives, (e) mass centers and principle axes, (f) final IDs and positions/orientations in 3D world space

Evaluation and Discussion

We have tested our prototype with respect to performance and spatial precision. In the following, we will present some of the results gathered and also discuss further issues of our approach.

Performance

All tracking computations were performed on a designated server (PC, Intel Quad-Core i5 CPU, 2.67 GHz with 8GB RAM) running a 64-bit Windows operating system that was connected to a single Kinect. Depth images were grabbed with a rate of 30 frames per second. In our tests, the system achieved a tracking rate of 18 to 20 Hz (i.e., 56 to 50ms per frame) for a single display. This is relatively close to the input frame rate given by the Kinect and certainly provides an interactive user experience. However, there is still some room for improvement, e.g., by accelerating computations via graphics hardware. Each additionally tracked display consumes 2 to 3ms extra per frame. Thus, the performance of the algorithm scales well with a rising number of displays.

Spatial Precision

To evaluate the accuracy of our algorithm, we compared the precision of the Kinect-based tracking with our OptiTrack system. For this purpose, we moved a display through the interaction volume above the table surface and tracked it with both systems simultaneously for about 15 seconds. Based on (N = 355) sampled positions, we detected an average difference of 4 mm (X-axis, SD = 3.0mm), 5 mm (Y-axis, SD = 3.6mm), and 9 mm (Z-axis, SD = 7.0mm) between OptiTrack and our system. In comparison, at a distance of 2m, the Kinect's nominal spatial (X,Y) resolution is 3mm and its depth (Z) resolutions is 10 mm [1]. When used with our projective system, our experiences visually support these findings.

Display IDs

Currently, our system can discriminate between the iPad (241mm x 186mm) and three paper-based displays: rectangle (295mm x 210mm), square (210mm x 210mm), and circle (203mm diameter). Since this process depends entirely on the shape and size of the displays, the system can get confused occasionally, such as when two displays of the same type (e.g., the iPad) get too close together. One notable effect is that the content of both displays is swapped occasionally. Despite these little disruptions, our experiences indicate that displays are generally recognized well, especially when users do not hold them too steep, so that they remain fairly parallel to the tabletop (and thus the Kinect), which is the most common use case. When the current frame does not provide sufficient information to determine a particular display, we use knowledge from previous frames.

Spatial Orientation

With a maximum error of about 5°, the computation of display normals (local Z-axes of displays) is less accurate than the tracking of spatial positions. However, it is precise enough to project image content in correct perspective onto displays in the majority of situations. One reason for this is that displays are usually held horizontally – a default case for which the algorithm delivers the most accurate orientations (maximum error less than 1°).

The rotation within the display plane (i.e., the rotation around the display's Z-axis) is calculated depending on clearly distinguishable principal axes in candidate regions (see Figure 6d/e). If such features are not present, the algorithm cannot determine this angle properly. This applies in particular to circular displays, where we leave the rotation angle around the display’s Z-axis to zero. To further improve the accuracy of orientations, we experimented with integrating internal sensors of active displays (iPads), in particular accelerometers and compasses. Although still in an early stage of implementation, we are confident that this sensor fusion along with the information acquired from the depth camera will considerably improve the precision and robustness of the overall tracking process, in particular for spatial orientations. This can also help solve the problem of ambiguities due to symmetric display shapes.

Overlapping Displays

The tracking of (partially) overlapping displays is required for a variety of interaction techniques, such as filtering by physically stacking displays. Our general strategy for such cases is to only consider the candidate closest to the Kinect, while all candidates below are masked out in the current frame. This means that for all occluded displays we simply maintain the last known positions. Though our support for overlapping displays can only be considered experimental, preliminary results show that the overall strategy is promising, especially when differences in heights are significant (more than 5cm).

Limitations

Due to the limited depth resolution of the Kinect (10mm error at 2m distance), we currently cannot reliably detect displays close to or on the table surface, which unfortunately is a frequent use case. However, we can soften this problem by relying on object permanence, i.e., once a tracked display is close to the table and thus it becomes invisible in the depth sensor, chances are it is still there. This enables us to handle one of the most common situations: putting a display down on the table and picking it up again after a while. Other problems include occlusion (e.g., a user or other displays occlude a tracked display) and unfavorable viewing angles that let the perceived appearance of a tracked display collapse to a line in the worst case. By using two or more depth sensors at different view angles, not only could these problems be mitigated, but they would also allow for an increased tracking stability as well as the coverage of a larger interactive space. Unfortunately, sensors operating with the structured light approach (like the Kinect) are prone to increased noise when the patterns of two or more Kinects are projected onto the same surface. This can eventually lead to a complete failure of tracking and limits the use of multiple sensors to setups with little or no overlapping of the covered area. We hope that future iterations of the Kinect (drivers) will add better support for simultaneous use of two or more devices, e.g., by slightly vibrating the Kinect units so that each depth sensor sees its own projected pattern sharply, but a blurred version of the patterns of the other sensors [5].

Conclusion and Future Work

In this article, we presented the exciting design space of tangible displays in a tabletop environment. In this context, we analyzed and compared two principle classes of spatially aware handheld displays: the active and projective approach. As our long-term research agenda is to make the underlying interaction principles available to a broader audience, we analyzed technical requirements for a low-cost tangible display ecosystem that uses affordable consumer hardware and integrates well with established software frameworks. Our vision is that such ecosystems will seamlessly support both principle display approaches in the future.

The technical implementation of all requirements is clearly beyond the scope of a single article. We therefore focused on one aspect only: the marker-less spatial tracking of projective and active handheld displays by an off-the-shelf consumer depth sensor: the Microsoft Kinect. We designed and developed a tracking prototype so it integrates well with the PaperLens system [30, 32, 35]. PaperLens is the technical cornerstone of our research agenda that already addresses many of the requirements discussed in this article. The Kinect-based tracking approach allows for considerably easier setups with fewer components. With lower costs, it can be made available for new application areas outside of specialized laboratories. Furthermore, by using mobile active displays (e.g., the iPad), users can collaborate on complex interaction tasks and then access the data beyond the boundaries of a fixed environment. While the tracking prototype may be relatively simple, we see one important contribution in the integration into a complex tangible display framework, thus addressing one crucial aspect of our research towards a low-cost tangible display ecosystem for the masses.

For future work, one of our short-term goals is to take the next step from a functional prototypic to a robust tracking system by addressing issues such as: sensor fusion, automatic registration of active displays, usage of a visual back channel via displays, and a better integration of cooperation strategies. In particular, we plan on employing multiple depth sensors to improve the treatment of singularities and overlapping displays. The reliable recognition of multi-touch finger input on paper displays is another important issue that we want to address in future iterations of our framework. Although not the focus of this article, we are particularly enthusiastic about addressing the seamless application development of projective and active display within a unifying software framework (requirement R9).

In conclusion, our vision of a tangible display system for the masses is an interdisciplinary endeavor of great scope that requires the cooperation of experts from different fields, such as interaction design, image processing, hard- and software engineering, as well as application development. In this context, we see this article as a call to arms and hope to inspire others in joining this journey.

Acknowledgements
This article is a revised and extended version of a paper that was presented at the DCIS 2012 workshop on Designing Collaborative Interactive Spaces for e-Creativity, e-Science, and e-Learning held at AVI 2012 [34]. Our work was funded by the German Ministry of Education and Science (BMBF) project ViERforES-II (01IM1000 2B) and the German Research Foundation (DFG) project IPAR (DA 1319/2-1).

References

1. OpenKinect – Imaging Information
http://openkinect.org/wiki/Imaging_Information

2. Akaoka E, Ginn T, Vertegaal R (2010) DisplayObjects: prototyping functional physical interfaces on 3d styrofoam, paper or cardboard models. In: Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction (TEI), ACM, pp 49-56

3. Bandyopadhyay D, Raskar R, Fuchs H (2001) Dynamic Shader Lamps: Painting on Movable Objects. In: Proceedings of the International Symposium on Augmented Reality (ISAR), IEEE, pp 207-216

4. Bier EA, Stone MC, Pier K, Buxton W, DeRose TD (1993) Toolglass and Magic Lenses: The See-Through Interface. In: Proceedings of SIGGRAPH, ACM, pp 445-446

5. Butler DA, Izadi S, Hilliges O, Molyneaux D, Hodges S, Kim D (2012) Shake'n'sense: reducing interference for overlapping structured light depth cameras Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI), ACM, pp 1933-1936

6. Clark A, Piumsomboon T (2011) A Realistic Augmented Reality Racing Game using a Depth-sensing Camera. In: Proceedings of the International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI), ACM, pp 499-502

7. Costanza E, Giaccone M, Kueng O, Shelley S, Huang J (2010) Ubicomp to the masses: a large-scale study of two tangible interfaces for download. In: Proceedings of the ACM international conference on Ubiquitous computing (Ubicomp), ACM (2010), pp 173-182

8. Fischler MA, Bolles RC (1981) Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm. of the ACM, Vol 24, ACM, pp 381-395.

9. Fitzmaurice GW (1993) Situated Information Spaces and Spatially Aware Palmtop Computers. In Comm. of the ACM – Special Issue on Computer Augmented Environments: Back to the Real World, 36(7):39–49

10. Haller M, Leitner J, Seifried T, Wallace JR, Scott SD, Richter C, Brandl P, Gokcezade A, Hunter S (2010) The NiCE Discussion Room: Integrating Paper and Digital Media to Support Co-Located Group Meetings. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 609-618

11. Han JY (2005) Low-cost multi-touch sensing through frustrated total internal reflection. In: Proceedings of the 18th annual ACM symposium on User interface software and technology (UIST), ACM, pp 115-118

12. Harrison C, Dey AK (2008) Lean and Zoom: Proximity-aware User Interface and Content Magnification. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 507–510

13. Harrison C, Benko H, Wilson AD (2011) OmniTouch: Wearable Multitouch Interaction Everywhere. In: Proceedings of the International ACM Symposium on User interface Software and Technology (UIST), ACM, pp 441–450

14. Holman D, Vertegaal R, Altosaar M, Troje N, Johns D (2005) Paper Windows: Interaction Techniques for Digital Paper. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 591-599

15. Huber J, Steimle J, Liao C, Liu Q, Mühlhäuser M (2012) LightBeam: Interacting with Augmented Real-world Objects in Pico Projections. In: Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM). ACM, Article 16.

16. Ishii H, Ullmer B (1997) Tangible Bits. Towards Seamless Interfaces between People, Bits and Atoms. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 234-241

17. Izadi S, Hodges S, Taylor S, Rosenfeld D, Villar N, Butler A, Westhues J (2008) Going beyond the display: a surface technology with an electronically switchable diffuser. In: Proceedings of the 21st annual ACM symposium on User interface software and technology (UIST), ACM, pp 269-278

18. Junuzovic S, Inkpen Quinn K, Blank T, Gupta A (2012) IllumiShare: Sharing Any Surface. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 1919-1928

19. Khalilbeigi M, Lissermann R, Mühlhäuser M, Steimle J (2011) Xpaaand: Interaction Techniques for Rollable Displays, In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 2729-2732

20. Khalilbeigi M, Lissermann R, Kleine W, Steimle J (2012) FoldMe: Interacting with double-sided foldable displays. In: Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction (TEI), ACM, pp 33-40

21. Kirsh D (1995) The intelligent use of space. Artificial Intelligence - Special volume on computational research on interaction and agency, part 2, 73(1-2):31-68

22. Kray C, Rohs M, Hook J, Kratz S (2008) Group coordination and negotiation through spatial proximity regions around mobile devices on augmented tabletops. In: Proceedings of the 3rd International Workshop on Horizontal Interactive Human Computer Systems (TABLETOP), IEEE, pp 1-8

23. Lee JC, Hudson SE, Summet JW, and Dietz PH (2005) Moveable Interactive Projected Displays Using Projector Based Tracking. In: Proceedings of the International ACM Symposium on User interface Software and Technology (UIST), ACM, pp 63-72.

24. Lee JC, Hudson SE, and Tse E (2008) Foldable Interactive Displays. In: Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), ACM, pp 287-290

25. Linder N, Maes P (2010) LuminAR: Portable Robotic Augmented Reality Interface Design and Prototype. In: Adjunct Proceedings of the International ACM Symposium on User Interface Software and Technology (UIST), ACM, pp 395-396

26. Lissermann R, Olberding S, Petry B, Mühlhäuser M, Steimle J (2012) PaperVideo: Interacting with Videos on Multiple Paper-like Displays. In: Proceedings of the ACM International Conference on Multimedia (MM), ACM, pp 129-138

27. Merrill D, Kalanithi J, Maes P (2007) Siftables: towards sensor network user interfaces. In: Proceedings of the 1st international conference on Tangible and embedded interaction (TEI), ACM, pp 75-78

28. Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, Kohli P, Shotton J, Hodges S, Fitzgibbon A (2011) KinectFusion: Real-time Dense Surface Mapping and Tracking. In: Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, pp 127-136

29. Rekimoto J (1998) A multiple device approach for supporting whiteboard-based interactions. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 344-351

30. Spindler M, Stellmach S, Dachselt R (2009). PaperLens: Advanced magic lens interaction above the tabletop. In: Proceedings of the International ACM Conference on Interactive Tabletops and Surfaces (ITS), ACM, pp 69-76

31. Spindler M, Dachselt R (2009) Towards Pen-based Annotation Techniques for Tangible Magic Lenses Above a Tabletop. In: Accompanying DVD of the ACM International Conference on Interactive Tabletops and Surfaces (ITS), ACM

32. Spindler M, Tominski C, Schumann H, Dachselt R (2010) Tangible views for information visualization. In: Proceedings of the International ACM Conference on Interactive Tabletops and Surfaces (ITS), ACM, pp 157-166

33. Spindler M, Martsch M, Dachselt R (2012) Going Beyond the Surface: Studying Multi-Layer Interaction Above the Tabletop. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 1277-1286

34. Spindler M, Büschel W, Dachselt R (2012). Towards Spatially Aware Tangible Displays for the Masses. In: Proceedings of Workshop on Designing Collaborative Interactive Spaces for e-Creativity, e-Science and e-Learning (DCIS) at AVI

35. Spindler M, Büschel W, Dachselt R (2012) Use Your Head: Tangible Windows for 3D Information Spaces in a Tabletop Environment. In: Proceedings of the International ACM Conference on Interactive Tabletops and Surfaces (ITS), ACM, pp 245-254

36. Sellen AJ, Harper RHR (2001) The Myth of the Paperless Office, The MIT Press

37. Tarun AP, Wang P, Girouard A, Strohmeier P, Reilly D, Vertegaal R (2013) PaperTab: An electronic paper computer with multiple large flexible electrophoretic displays. In: CHI Extended Abstracts on Human Factors in Computing Systems (CHI EA), ACM, pp 3131-3134

38. Tsang M, Fitzmaurice GW, Kurtenbach G, Khan A, Buxton B (2002) Boom chameleon: simultaneous capture of 3D viewpoint, voice and gesture annotations on a spatially-aware display. In: Proceedings of the International ACM Symposium on User Interface Software and Technology (UIST), ACM, pp 111-120

39. Ullmer B, Ishii H (1997) The metaDESK: Models and Prototypes for Tangible User Interfaces. In: Proceedings of the International ACM Symposium on User Interface Software and Technology (UIST), ACM, pp 223-232

40. Wallace JR, Scott SD, Stutz T, Enns T, Inkpen K (2009) Investigating teamwork and taskwork in single- and multi-display groupware systems. Personal and Ubiquitous Computing 13(8):569-581

41. Weiser M (1991) The Computer for the 21st Century. Scientific American. 265(3):66-75

42. Wilson D (2010) Using a depth camera as a touch sensor. In: Proceedings of the International ACM Conference on Interactive Tabletops and Surfaces (ITS), ACM, pp 69-72

43. Yee K (2003) Peephole Displays: Pen Interaction on Spatially Aware Handheld Computers. In: Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI), ACM, pp 1-8

44. Zhang Z, Wu Y, Shan Y, Shafer S (2001) Visual Panel: Virtual Mouse, Keyboard and 3D Controller with an Ordinary Piece of Paper. In: Proceedings of the Workshop on Perceptive User Interfaces (PUI), ACM, pp 1-8