Gesture-Based Configuration of Location Information in Smart Environments with Visual Feedback

Carsten Stocklöw(✉) and Martin Majewski

Fraunhofer Institute for Computer Graphics Research IGD,
Fraunhoferstr. 5, 64283 Darmstadt, Germany
{carsten.stockloew,martin.majewski}@igd.fraunhofer.de

Abstract. The location of objects and devices in a smart environment is a very important piece of information to enable advanced and sophisticated use cases for interaction and for supporting the user in daily activities and emergency situations. To acquire this information, we propose a semi-automatic approach to configure the location, size, and orientation of objects in the environment together with their semantic meaning. This configuration is typically done with graphical user interfaces showing either a list of objects or a representation of objects in form of 2D or 3D virtual representations.

However, there is a gap between the real physical world and the abstract virtual representation that needs to be bridged by the user himself. Therefore, we propose a visual feedback directly in the physical world using a robotic laser pointing system.

Keywords: Smart environments · Configuration · Personalization

1 Introduction

Technologies that realize the paradigm of Ambient Assisted Living (AAL) and Smart Environments have found an increasing interest in the scientific community and on the market. Most notably, sensors and actuators for Home Automation, smart entertainment systems like TV and Hifi sets as well as devices such as smartphones or tablet computers. Simple scenarios allow the user to directly interact with and to control the devices in the intelligent environment. More sophisticated applications try to analyze the context of the user, e.g. the location of the user and the location of objects in the surrounding. This can be used for a variety of scenarios. For example, a fall detection application could be enhanced to determine if the user has fallen down on the ground or is simply lying down on a sofa, if the location of the sofa is known; a burglar could be detected by a smart floor if activity is detected near a window and no activity was detected before from a user that could have gone inside the room to the window; a user interface is following the user from one room to the next, thus presenting the

N. Streitz and P. Markopoulos (Eds.): DAPI 2015, LNCS 9189, pp. 201–211, 2015.
DOI: 10.1007/978-3-319-20804-6_19

information on the device that is closest to the user; or for multimodal interaction (e.g. pointing at an object to control its parameters). Therefore, location of users and objects in the environment can be considered a very important piece of information for the system to enable high-level applications.

Consequently, there is a need to acquire location information of objects in the surrounding. This can be done online via camera-based systems that continuously monitor the area of interest. However, there may be multiple cameras needed to monitor a whole appartement, thus increasing the costs of such a system, and there are justified privacy concerns of constantly monitoring a private home. Another possibility would be to use active tags on each object and exploit the Received Signal Strength, but those mechanisms need the tags on each object. Some of those technologies have been investigated in the EvAAL competition [1].

Most of the bigger objects can be considered to be stationary. This includes some devices (like TV, refrigerator), furniture (like sofa, table), as well as built-in objects (like windows). Therefore, we use a one-time configuration that needs to be repeated if bigger changes in the environment occur. This configuration can make use of camera-based systems to gather the information needed without privacy concerns. However, the algorithms to detect and identify objects from a camera image are not yet fully reliable, and objects that are not in the visible range cannot be recognized. Thus, we propose a semi-automatic approach to detect objects that can be enhanced with detailled information from a user, e.g. by selecting a specific model for the detected TV. In this case, the user would be a technician that performs the one-time configuration.

This configuration is typically supported by the system with visual feedback in form of graphical user interfaces showing either a list of objects or a representation of each object in form of 2D or 3D virtual representations. However, there is a gap between the real physical world and the abstract virtual representation that *needs to be bridged by the user himself*. Therefore, we provide a visual feedback directly in the physical world using a robotic laser pointing system that is able to show where the user is pointing to. This can be used to reliably select elements in the environment as it has been shown that a direct visual feedback is an important aspect to increase the accuracy of pointing gestures.

2 Related Work

Bridging the gap between the physical world and a virtual representation of it was subject of many scientific publications and areas. The term *Mixed Reality* is often used to describe technologies that "involve the merging of real and virtual worlds somewhere along the 'virtuality continuum' which connects completely real environments to completely virtual ones" [11]. As part of this continuum, Augmented Reality "refers to all cases in which the display of an otherwise real environment is augmented by means of virtual (computer graphic) objects". This is often realized by smartphones or tablet computers that need to be directed towards the real object, making it cumbersome for devices with large displays or

impose restrictions related to the interaction in case of devices with small displays. Another class of devices are specialized glasses or head-mounted displays that are out of the scope of this work.

The MirageTable [2] uses a curved screen; an image is shown on that screen with a stereoscopic projector. Interaction with hand gestures can be done directly on that screen, but the interaction area is restricted to the screen.

Hossain et al. [5] present a system that combines the home automation system of the real world with the virtual world of Second Life[1]. Events from one representation is reflected in the other. The system has been found to be "appealing and useful to the user". However, there is no mixture of the two representations; the user interacts either in the real or in the virtual world.

The XWand, part of the WorldCursor [15] project by Microsoft Research in 2003 is a signal-processing pointing device, that can be seen as an indirect predecessor of the very popular Nintendo Wii Remote controller shipped with the same-named Wii video games console. The XWand is equipped with a three-axis magnetoresistive permalloy magnetometer that measures its yaw angle relating to the Earths magnetic field and a two-axis accelerometer to sense the acceleration relative to the gravity vector. Its purpose is to determine the pitch and roll angle of the device so it can be used as a spatial pointing device. The Nintendo Wii Remote[2], released three years after Microsoft's publication, uses a more advanced technology featuring a single three-axis accelerometer chipset. The Sensor Bar, a kind of visual homing system, completes it. The Sensor Bar includes a couple of spatial separated infrared LEDs that are tracked by the Wii Remote's image sensor providing additional orientation information.

The second part of the WorldCurser project is a small laser robot arm that projects a small red light-spot into the pointing direction of the XWand device. This way a direct environmental feedback system is created that tries to close the gab between the users and the systems interpretation of the performed action. There are two main limitations within the WorldCursor project. First, it is mandatory for the user to carry a pointing device to perform the action and is therefore not meeting the paradigm of unobtrusiveness in terms of Ambient Intelligence. Second, the system is not aware of any location-based information, like the spatial distribution and dimensions of walls and furniture. Therefore just mimicking the movement of the XWand, either absolute or relative, results in the feedback projection.

Majewski et al. proposed a more advanced solution to solve these two concerns with the Visual support system for selecting reactive elements in intelligent environments [7] that is also part of the technical realization of this paper. The Environmental Aware Gesture Leading Equipment recognizes marker free pointing gestures performed by the user with the Kinect RGB-D camera, calculates the intersection point of the pointing direction with the internal virtual representation of the environment and provides a direct absolute cue-projection onto that physical location. Majewski et al. [8] presented an extension to this system

[1] http://secondlife.com.
[2] http://us.wii.com/ (visited on 05/03/2015).

that combines different location technologies, like the CapFloor [4] system, to dynamically determine forbidden areas where projection could cause harm and should therefore not be performed at all, e.g. when projected directly onto the eye-area of another person.

The Beamatron project of Wilson et al. [14] shows an even more advanced marker free interaction approach in terms of information projection using several Kinect cameras, a microphone array setup, as well as a high definition projector mounted on a stage-light robot arm. While the usage of the mentioned sensing system provides direct user-to-machine interaction, the projector enables a high detail information feedback of the performed interaction directly into the users environment such as menus and 3D graphics. This bidirectional interaction paradigm makes it also possible to use the projected information for interaction like grabbing a projected document and dragging it to different locations.

Stahl et al. [12] introduced the concept of'synchronized realities' and created a 3D virtual representation of a real living lab. Actions in either representation are synchronized and reflected in the other one. However, this system is used only to control elements of the environment, not for configuration or to change the location of an element.

3 System Overview

This chapter describes the main components that are needed to realize our interaction concept - a vision based reconstruction of the environment, a visual feedback system with a robotic laser pointing system, and a semantic model to classify and describe the objects and their properties. We use Microsoft Kinect as RGB-D sensor for the reconstruction, to track the user, and to recognize free-air gestures from the skeleton data.

3.1 Vision-Based Reconstruction

The reconstruction of the environment is performed using a RGB-D camera that provides a 2D camera image with additional depth information (Microsoft Kinect One). We first analyse the scene as a whole by detecting the floor, ceiling, and walls with a plane recognition method. With this information the recognition of objects can be enhanced using this contextual information. Similar approaches have been reported by Koppula et al. [6] who used visual appearance, local shape and geometry, and geometrical context (e.g. a monitor is always on-top-of a table); the micro precision in home scenes increased by 16,88 % with this additional reasoning on geometrical context. Xiong et al. [16] used Conditional Random Field to model geometrical relationships (orthogonal, parallel, adjacent, and coplanar) between planar patches to classify walls, floors, ceilings, and clutter. The recognition rate increased from 84 % to 90 % when using certain contextual information in combination with local features.

As the reconstruction is not the main part of this contribution, we used a simple approach as shown in Fig. 1. We first use RANSAC to detect planes in the

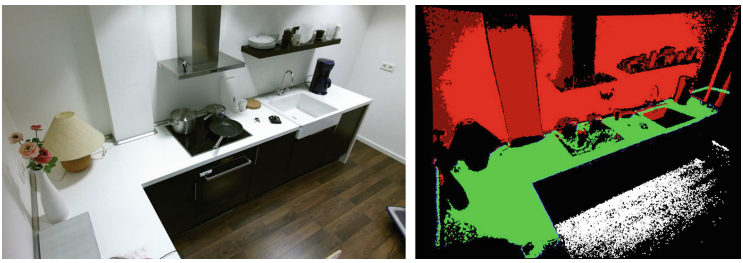

Fig. 1. Reconstruction of the environment: camera image (left), detected objects (right) (Color figure online)

3D point cloud. The normal vector of these planes then determines whether the plane belongs to a wall (vertical plane, shown in red in the figure). The lowest horizontal plane is considered the floor (gray), and planes parallel to the floor plane are further evaluated according to their distance to the floor plane, their shape and size. That way, we can recognize, for example, a table (green) or a sofa in the environment.

3.2 Visual Feedback System

Visual feedback in the physical world is provided by the E.A.G.L.E. Eye, a robotic laser pointing system, introduced by Majewski et al. [7]. The system is shown in Fig. 2. The E.A.G.L.E. Eye is based on an Arduino microcontroller board that is operating a laser mounted on two servo motors that allow free and precise positioning of a laser dot in the room.

Fig. 2. Visual Feedback Robot (left) and mounted on living room ceiling (right)

By showing a laser dot in the room it is possible to provide feedback about where the user is pointing at. The framework tracks the skeleton data of the user and can calculate the intersection of the pointing ray with the environment. The

laser pointing system can show the laser dot at exactly this intersection point, thereby following the user's pointing direction and giving direkt feedback in the physical world.

If the location, size, and orientation of an object is known, it can also be used to highlight a selected object by showing the laser dot at the center of that object. Optionally, the laser is able to go into a blinking mode, indicating that a selection has been completed. Additionally, the laser can also indicate the location, size, and orientation of an object by moving along the silhouette of that object, as described in Sect. 4.

3.3 Semantic Model

The semantic model to describe objects and their properties is realized as a set of ontologies with each ontology representing a certain application domain. The ontologies are taken from universAAL[3].

A basic ontology models physical things and their location. This ontology is described in detail by Marinc et al. [10]; an excerpt is depicted in Fig. 3. Each physical thing has a location that can be a room with a certain function. Additionally, each PhysicalThing has a shape and each location can be contained in or adjacent to another location (not shown here).

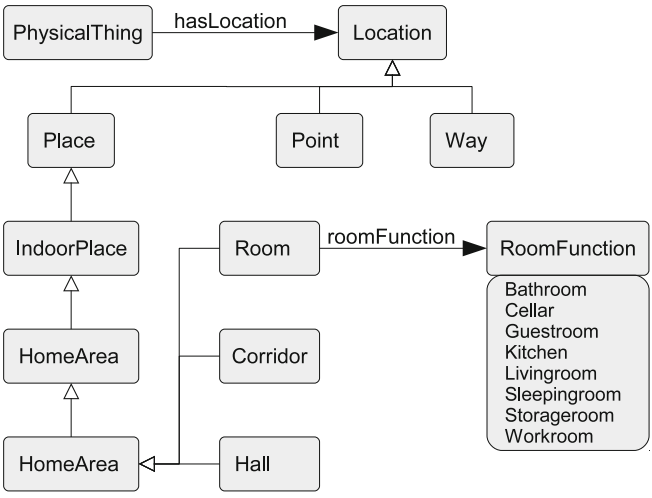

Fig. 3. Location ontology (excerpt)

Other ontologies model different types of devices, e.g. a LightSource (from device ontology), a TV, or a Stereoset (from multimedia ontology), as shown in Fig. 4. Each device has appropriate properties, e.g. a light source has a brightness

[3] http://depot.universaal.org.

value. As subclasses of the class PhysicalThing they inherit also the location property. Other devices are, for example, blinds, curtain, heater, humidity sensor, smoke sensor etc.

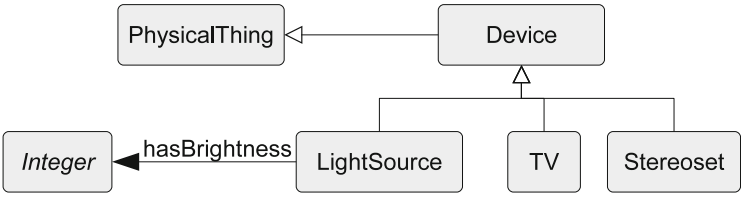

Fig. 4. Device ontology (excerpt) and multimedia ontology

The furniture ontology represents different kinds of furniture. These ontology classes have no additional properties except the ones that are inherited from the super class, PhysicalThing (Fig. 5).

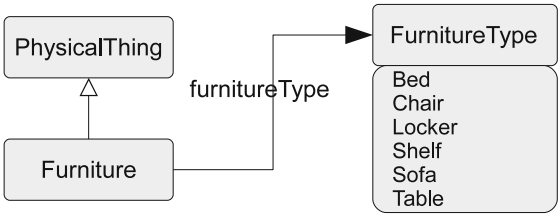

Fig. 5. Furniture ontology

4 Interaction Between Physical and Virtual World

In this section we describe our concept of bridging between the physical and the virtual world. Basically, we can distinguish between the two directions of (1) interacting in the physical world with reactions in the virtual and (2) interacting with the virtual representation with reactions in the physical world.

We assume that the reconstruction of the environment has already segmented and annotated different objects. However, this annotation can only be done according to a certain recognition rate and may be incorrect or not precise enough. For example, a TV could be recognized by the system automatically, but to actually use this device in a smart environment we may need the specific model to interoperate with the right protocol. Hence, it is needed to modify existing annotations. Additionally, the location, size, and orientation of objects may be incorrect and needs to be adjusted. Therefore, the system should be

able to select and to modify information in a unified way that synchronizes the virtual representation with the physical world.

Figure 6 shows an example of a reconstructed environment. The objects are shown with their bounding box, which is highlighted when the object is selected.

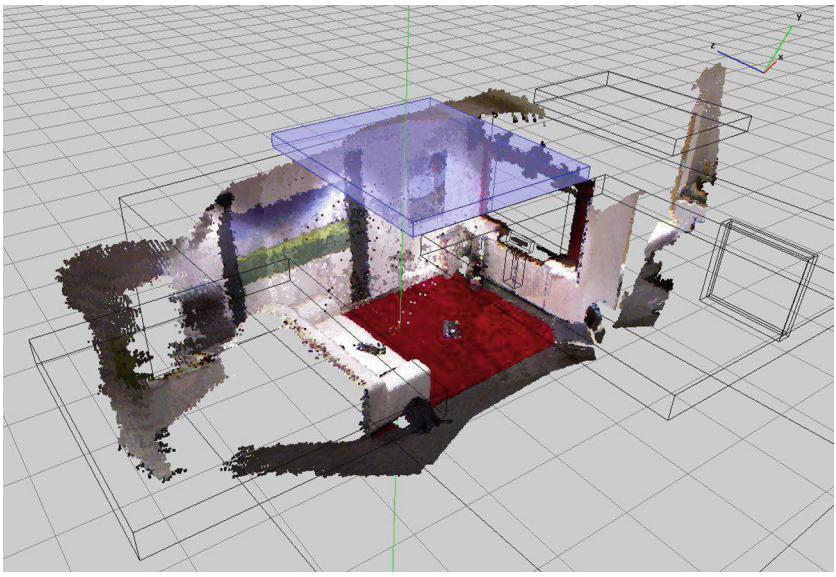

Fig. 6. Reconstructed environment with segmented objects; the ceiling lamp is highlighted

One way of interaction is the selection of elements. Using a graphical user interface (GUI), this method is well-known utilizing a mouse or a touchpad. For the real world, existing work often uses pointing gestures, e.g. by exploiting the skeleton data coming from RGB-D sensors. When synchronizing the two worlds, this can lead to an additional selection in the GUI. However, if an object is selected in the GUI, there is normally no visual feedback in the real world. This is enabled in our system with the robotic laser pointing system that can visually support the selection by showing where the user is pointing to and by pointing at the center of an object if that object is selected. To verify the location, size, and orientation of that object the laser pointing system can also move along its projected bounding box, or along the shape, if the shape could be reconstructed. For solid objects, the bounding box may need to be shrinked so that the laser pointer is always visible on the object. For non-solid objects, the bounding box may need to be expanded. For example, for a window the bounding box is expanded so that the laser pointer is moving around the object and is always visible on the wall surrounding the window.

When an object is selected the metadata, e.g. the type of the object, can be modified. This can also be done with gestures, for example by selecting a type

from a list of elements as shown by Stocklöw et al. [13]. However, as there may be many options to choose from this interaction method may be too cumbersome. Therefore, the usual method of selection in a GUI should be preferred. The location, size, and orientation could also be modified in the real world. A possible method is descibed by Marinc et al. [9]

5 Individual Evaluations

We evaluated our concept with different prototypes for the different tasks of the configuration process.

The robotic laser pointing system was evaluated by Majewski et al. [7] with 20 subjects between 22 and 65 years and a median age of 27 years. Each subject had to aim and select a sequence of eight different targets of different size that were placed in a room. It was shown that the visual feedback system has significantly improved the pointing accuracy.

The method ef modifying the bounding box of existing objects was evaluated by Marinc et al. [9] with eleven users, measuring the time they needed to create three boxes for Couch, TV, and Light. It was concluded that this method is well suited to allow common users to build up a virtual environment.

Braun et al. [3] have evaluated the selection and control of objects in the environment with multimodal input using gesture and speech. The test was performed by nine subjects between 21 and 29 years. The subjects considered the interaction to be intuitive and easy to master and particularly liked how pointing can simplify the complexity of speech commands. There was a noticeable learning effect from the first to the last tasks, reducing the number of wrong attempts and increasing the interaction time.

6 Conclusion

Bridging the gap between the virtual and the physical world is a task that often needs to be performed by the user. Several approaches exist for various use cases and technologies to support the user regarding this task. In this work we have proposed a method to use free-air hand gestures for interaction and a robotic laser pointing system for visual feedback in the real world; synchronizing the information in the two realms. This method was applied to the domain of configuration of location information in smart environments as this is an important piece of information to enable advanced and sophisticated use cases for interaction and for supporting the user in daily activities and emergency situations.

Future work includes improvement of the vision-based reconstruction and the integration into an overall system.

Acknowledgements. This work is partially financed by the European Commission under the FP7-ICT-Project Miraculous Life (grant agreement no. 611421).

References

1. Barsocchi, P., Potort, F., Furfari, F., Gil, A.: Comparing AAL indoor localization systems. In: Chessa, S., Knauth, S. (eds.) EvAAL 2011. CCIS, vol. 309, pp. 1–13. Springer, Heidelberg (2012)
2. Benko, H., Jota, R., Wilson, A.: Miragetable: freehand interaction on a projected augmented reality tabletop. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 199–208. ACM, New York (2012)
3. Braun, A., Fischer, A., Marinc, A., Stocklöw, C., Majewski, M.: Context-based bounding volume morphing in pointing gesture application. In: Kurosu, M. (ed.) HCII/HCI 2013, Part IV. LNCS, vol. 8007, pp. 147–156. Springer, Heidelberg (2013)
4. Braun, A., Heggen, H., Wichert, R.: CapFloor – a flexible capacitive indoor localization system. In: Chessa, S., Knauth, S. (eds.) EvAAL 2011. CCIS, vol. 309, pp. 26–35. Springer, Heidelberg (2012)
5. Hossain, S., Rahman, A., El Saddik, A.: Bridging the gap between virtual and real with second life client in a virtual home automation system. In: 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 001212–001217, May 2011
6. Koppula, H., Anand, A., Joachims, T., Saxena, A.: Semantic labeling of 3D point clouds for indoor scenes. In: NIPS (2011)
7. Majewski, M., Braun, A., Marinc, A., Kuijper, A.: Visual support system for selecting reactive elements in intelligent environments. In: Proceedings of 2012 International Conference on Cyberworlds, Fraunhofer-Institut für Graphische Datenverarbeitung (IGD) and Technische Universität Darmstadt (TUD) and European Association for Computer Graphics (Eurographics) and IFIP Working Group 5.10 on Computer Graphics and Virtual Worlds, pp. 251–255. IEEE Computer Society Conference Publishing Services (CPS), Los Alamitos (2012)
8. Majewski, M., Dutz, T., Wichert, R.: An optical guiding system for gesture based interactions in smart environments. In: Streitz, N., Markopoulos, P. (eds.) DAPI 2014. LNCS, vol. 8530, pp. 154–163. Springer, Heidelberg (2014)
9. Marinc, A., Stocklöw, C., Braun, A.: Building up virtual environments using gestures. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2013, Part III. LNCS, vol. 8011, pp. 70–78. Springer, Heidelberg (2013)
10. Marinc, A., Stocklöw, C., Tazari, S.: 3D interaction in AAL environments based on ontologies. In: Wichert, R., Eberhardt, B. (eds.) Ambient Assisted Living. ATSC, vol. 2, pp. 289–302. Springer, Heidelberg (2012)
11. Milgram, P., Kishino, F.: A taxonomy of mixed reality visual displays. IEICE Trans. Inf. Syst. **E77–D**(12), 1321–1329 (1994)
12. Stahl, C., Frey, J., Alexandersson, J., Brandherm, B.: Synchronized realities. J. Ambient Intell. Smart Environ. (JAISE) **3**(1), 13–25 (2011)
13. Stocklöw, C., Wichert, R.: Gesture based semantic service invocation for human environment interaction. In: Patern, F., de Ruyter, B., Markopoulos, P., Santoro, C., van Loenen, E., Luyten, K. (eds.) AmI 2012. LNCS, vol. 7683, pp. 304–311. Springer, Heidelberg (2012)
14. Wilson, A., Benko, H., Izadi, S., Hilliges, O.: Steerable augmented reality with the beamatron. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, UIST 2012, pp. 413–422. ACM, New York (2012)

15. Wilson, A., Pham, H.: Pointing in intelligent environments with the world cursor. In: Proceedings of Interact 2003 (2003)
16. Xiong, X., Huber, D.: Using context to create semantic 3D models of indoor environments. In: Proceedings of the British Machine Vision Conference, pp. 45.1–45.11. BMVA Press (2010). doi:10.5244/C.24.45