Context-Based Bounding Volume Morphing in Pointing Gesture Application

Andreas Braun[1], Arthur Fischer[2], Alexander Marinc[1],
Carsten Stocklöw[1], and Martin Majewski[2]

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
{andreas.braun,alexander.marinc,
carsten.stockloew}@igd.fraunhofer.de
[2] Technische Universität, Darmstadt, Germany
{arthur.fischer,martin.majewski}@stud.tu-darmstadt.de

Abstract. In the last few years the number of intelligent systems has been growing rapidly and classical interaction devices like mouse and keyboard are replaced in some use cases. Novel, goal-based interaction systems, e.g. based on gesture and speech allow a natural control of various devices. However, these are prone to misinterpretation of the user's intention. In this work we present a method for supporting goal-based interaction using multimodal interaction systems. Combining speech and gesture we are able to compensate the insecurities of both interaction methods, thus improving intention recognition. Using a p`rototypical system we have proven the usability of such a system in a qualitative evaluation.

Keywords: Multimodal Interaction, Speech Recognition, Goal-based Interaction, Gesture Recognition.

1 Introduction

Smart environments are often comprised of a plethora of networked and user controllable devices. Those are typically controlled by various remote controls or combined systems providing simplified graphical user interfaces. Pointing at devices for manipulation is a natural form of interaction that is often performed unconsciously when using traditional remotes. It is possible to realize this pointing manipulation by using a virtual representation of the physical environment in combination with gesture recognizing sensors [1]. The straightforward approach of finding devices is using an intersection between pointing ray and bounding volumes of devices in the virtual realm [2]. However, if the controllable devices are small or occluded selection might become difficult or even impossible. In this case means have to be provided to allow selecting the devices. Various options are available, such as conflict resolution strategies, e.g. via menu selection [3], the usage of visual indicators for aiding selection [4], or - as it is used in this work - using contextual information to infer the intention of the user of interacting with a specific device. This work will present the following contributions:

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 147–156, 2013.

- We propose a generic method to modify bounding volumes based on contextual information gathered by the environment or the interaction process
- We propose different methods of bounding volume morphing, such as static scaling, occlusion-based morphing and viewpoint-based space-filling methods [5].
- We test our method in a multimodal interaction scenario using a combination of speech and gesture

We use the contextual information generated by the smart environment to modify the selection process on a generic level by modifying the bounding volumes associated with the different devices, instead of modeling the uncertainty within the pointing process itself. By this generic approach we gain two distinct advantages, the contextual information allows to reduce the information required by other systems in multimodal interaction scenarios and the modification within the virtual representation allows other applications to directly use the modified bounding volumes. A particularly interesting application area for this method is multimodal interaction. Concerning gestural interaction a good candidate for an additional modality is speech. This allows interacting with devices by pointing at them and speaking out various commands. The intention as identified by Natural Language Processing applied to speech and the approximate can be considered context. E.g. if the user wants to make something “louder” this is unlikely to apply to lighting - if the user is pointing to the front he typically does not want to interact with devices behind him. Therefore if the devices are properly mapped to speech control it is possible to reduce the number of potential systems to interact with and use this information in the bounding volume modification. The overall process in this application scenario is following five steps; processing speech for interaction commands, modifying list of potential devices based on supported commands, modifying bounding volumes of candidate devices perform ray cast based on pointing direction and identify device and executing command on device.

2 Related Work

In the last few years novel interaction paradigms have seen a strong interest in the public eye. Particularly gesture interaction has seen considerable success; particularly in mobile applications with touch screens and gaming applications, with the Nintendo Wii and Microsoft Kinect.

There have been various research efforts to use gestural interaction in smart environments. Wilson et al have created the XWand, shown in Figure 1- left, a gesture interaction device based on accelerometers and infrared tracking of the device position [2]. The integrated sensors allow determining pointing direction and starting point, thus providing the ability to select modeled devices in a smart environment. The system also allows using speech commands to manipulate the selected devices. XWand models devices as Gaussian probability distribution, allowing for simple decision which device should be selected, however the method does not take into account ambiguous or occluded appliances. In our work we build upon a bounding

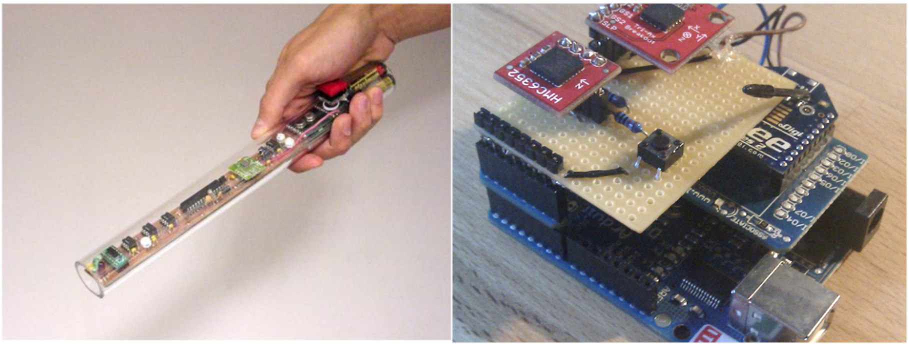

Fig. 1. Left - XWand gesture interaction device. Right - prototype interaction device.

volume approach previously presented [1] and introduce dynamically modified bounding boxes that change their shape based on the currently registered context, in this case speech and pointing direction. In contrast to the interaction device we have previously used (Figure 1- right) the new system is based on depth imaging.

Recognizing the intention of a person is a task typically performed subconsciously without rationalizing the motives of the conversation partner [6]. Even in simple conversations we evaluate the intentions continuously and use it as a supplement to our communication efforts to generate additional information that is important in the context of the conversation [7]. Heinze et al postulates that in inter-agent communication the recognition of intention is crucial if the transmission between the agents is flawed and ambiguous [6]. This is typically the case in Human-Machine-Interaction with natural input methods that mimic interpersonal communication [7].

3 Goal-Based Interaction in Context-Sensitive Smart Environments

The dynamic nature of an environment is making it difficult to distinguish between intentional interaction and random movements [8]. Goal-based interaction aims at abstracting explicit interaction from the user and instead of specific functions act

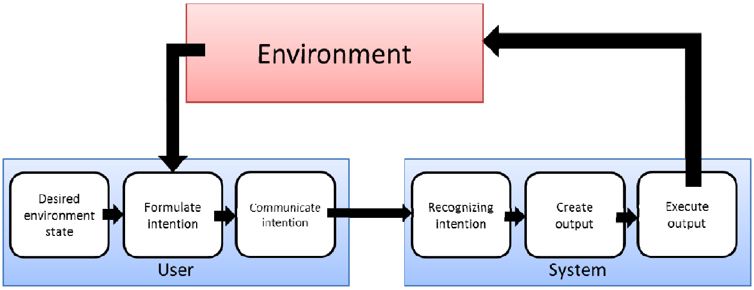

Fig. 2. Goal-based interaction without context support

based on the desired target of the interaction [9]. The general structure of a goal-based interaction system is displayed in Fig.2. A user is trying to achieve a desired environment state by formulating and communicating a specific intention. An interaction system is then trying to recognize this intention using the information communicated by the user. It will create the appropriate output and manipulate the environment accordingly.

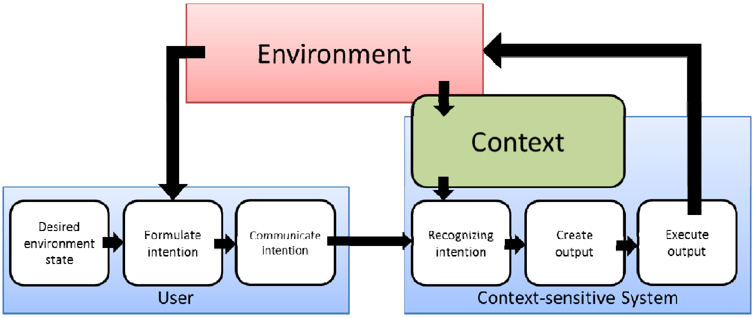

Fig. 3. Context-supported, goal-based interaction

This method however is not able to capture the implicit information. This is derived from interpersonal communication, wherein a considerable part of the information is exchanged implicitly within the current context; that is the situation surrounding the conversation and gives meaning to the specific interactions. In order to recognize this subtext it is necessary to monitor the user within the environment; analyzing the behavior and status to infer this information. The general structure of such a system is shown in Fig. 3 whereas the system has a second flow of information in order to recognize the intention using direct communication from the user and the context acquired in the environment. The latter method is particularly interesting concerning natural methods of interaction that abstract explicit functions from the user in order to allow interaction using the methods of interpersonal communication [10]. The question arises how we can use this concept in actual applications. A combination of speech and gesture is a common form of natural interaction that we are using to determine a suitable scenario for context-supported goal-based interaction. The direct channels of communication are the recognized gestures and the speech picked up by language processing. Combining these information channels with a modeled environment that is aware about its capabilities, those of the devices in the environment and activity information about the user we are able to create a scenario where we can improve the user experience by simplifying the interaction and making it more robust.

4 Bounding Volume Morphing and Multimodal Interaction

The combination of speech and gesture is a common form of multimodality [11, 12]. We use it in natural interaction, e.g. by pointing at a specific item, creating the implicit information that all subsequent information in this dialogue is centered on this item, without explicitly mentioning it every time. We can exploit this in a similar fashion for Human-Machine-Interaction. In this work we present a system supporting multimodal control of devices in smart environments. The supported method is the selection and manipulation of systems that are arbitrarily placed in the room. If the number of controllable devices is high it may be difficult to interact, e.g. considering small devices that have to be pointed at with gestural control, or numerous similarly named systems with speech control. If we combine both modalities we can create a model that supports and simplifies both methods of interaction by reducing the required inputs and increasing reliability. Based on this premise we have created a model that modifies the gestural selection process based on speech input and vice versa.

An overview of this process is given in Fig. 4. The user is communicating in a multimodal fashion using speech and gesture. The system is picking up this information and is additionally holding a model of the environment that is storing data about the different appliances, their capabilities and location. Both environment model and speech recognition influence the gesture recognizer while the final manipulation of the environment is depending on both speech and gesture recognition.

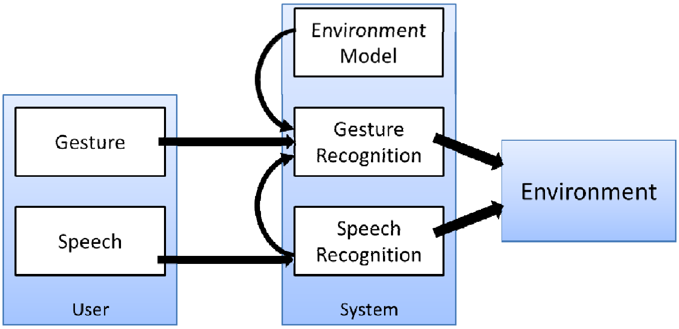

Fig. 4. Environment manipulation using speech and gesture

We are explaining this process by example of a user that is trying to control a lamp in a living room. He is pointing at the lamp he wants to turn brighter, however in the same region there are various other devices that make identification difficult for the gesture recognizer. Yet the system is aware of the device capabilities. The user now utters the words “brighter” indicating that he wants to control a device that is capable of changing lighting intensity. This information is going back to the gesture recognizer that discards devices that do not possess this ability, e.g. stereo or heating. The probability that the user is intending to select those devices can be lowered accordingly. One method to realize such a change in probability with regard to gesture recognition is modifying the bounding volumes of appliances, increasing or decreasing their

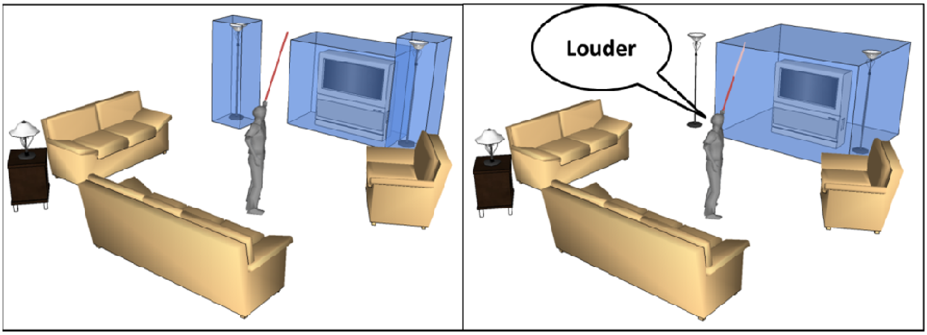

Fig. 5. Intersecting with a modified bounding volume after appropriate speech input

spatial representation in the environment model and thus adjusting the chances of intersecting this specific volume. To give an example, if there are three controllable devices, two lamps and a TV, and the user gives the command "louder", the lamps can't be affected lacking the capability. This behavior is shown in Fig. 5. If the lamps are discarded the bounding volume of the TV can be enlarged increasing the chances to be intersected.

The result is a two-step process, where first unsuitable appliances are discarded based on their capabilities and the results of the speech recognition and secondly the bounding volumes of all remaining devices are modified to increase the reliability of the gesture recognition.

Only modifying bounding volumes allows for generic application of various different methods. A first example is space-filling, whereas the bounding volumes are extended until they fill the available room; that is until they intersect the space boundaries or intersect with other bounding volumes. A second method is normalization, whereas the bounding volumes are extended to a fixed size, giving all objects the same probability of being intersected. Another example is uniform extension, leading to all bounding volumes being increased in size by a fixed ratio. All three methods are shown in Fig. 6 in a simple two-dimensional case.

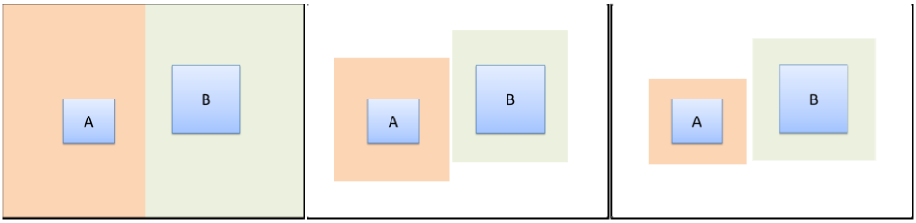

Fig. 6. Left space-filling method - middle, normalization method - right, fixed ratio method

When considering which method too chose it is crucial to think about the potential drawbacks of bounding volume based methods. We can distinguish two different types of errors. An occurring Type I error means that we are targeting at the actual device but there is a bounding volume mismatch that does not allow us to properly select the system; Type II error means that an overly large bounding volume of another device is preventing us from being able to intersect the intended device [4]. Therefor it is crucial to select a method that is reducing both types of errors by creating optimal bounding volumes.

5 Prototype System

Based on the process described on the previous pages we have created a prototype system and installed it in our Living Lab. The devices in the lab are interfaced using a KNX bus system, that allows setting and manipulating various appliances within the premises, e.g. lighting, TV, windows and blinds. We have decided to use the Microsoft Kinect as gesture recognizing sensor using the OpenNI[1] framework. For speech recognition a dedicated microphone is used and interfaced with the CMU Sphinx framework[2] that allows recognizing speech commands using a combination of natural language processing with Hidden Markov Models. The virtual representation of the environment is based on X3D files, with the bounding volumes stored separately and modified accordingly. A software module combines the sensor input with the virtual representation and implements the device recognition using the bounding volume modification methods presented previously. Afterwards this module sends the control signals to the KNX-networked devices. The overall structure of this prototype system is shown in Fig. 7. Given a set of possible devices and commands the system will combine them to determine the most probable device and execute the action intended with a command. For this purpose several cases in terms of the size of the sets have to be considered: In the trivial case one of the sets is empty and the system will just drop the current recognition process. In case there is only one possible device it will be assumed to be the final desired one and from all commands this device is capable of the most current command will be chosen. Finally, if the set contains multiple devices the most likely pair of device and command will be determined in four steps:

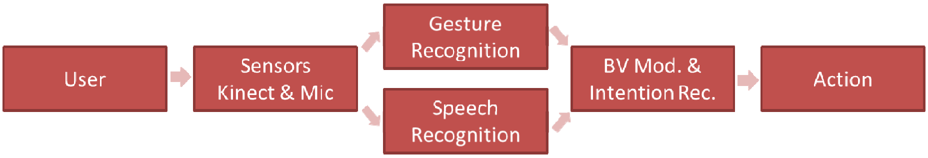

Fig. 7. Functional structure of the prototype system

[1] http://www.openni.org/

[2] http://cmusphinx.sourceforge.net/

1. Remove all commands which are not part of the capability of any device
2. Remove all devices which are not capable of any of the remaining commands
3. Take the most recent command and increase the bounding volumes of all devices capable of it
4. Recalculate the intersection point of the pointing gesture and the environment.

The device the user is pointing on now is considered as being the users intended choice. Afterwards the final device-command pair will be forwarded and executed. In this procedure the third step defines that only the last command is a valid one in case of still existing uncertainty. This is due to the time frames around a detected pointing gesture. One or more commands arriving within on frame are expected to be corrections of the previous command. Changing step three to a sequential processing of all speech commands can be alternatively used. According to that corrections by the user would be realized by undoing previous commands instead of skipping the allegedly wrong commands.

6 Evaluation

We have performed a usability study in which the subjects had to perform simple tasks by using speech commands and pointing at the device to be controlled. The test was performed by nine users, aged between 21 and 29. Most had previous experience with gesture recognition systems, while most had little experience with speech recognition. The users had to perform a set of 11 different task controlling different devices in the environment, e.g. turning off lighting in the living room area. The devices were intentionally positioned to test cases that are relevant for context-based bounding volume adaptation, i.e. using small devices far away from the users and devices standing beside each other. The results were compared to a time-based selection, where interaction was enabled by holding a selection gesture for a certain amount of time. In this initial study we were mostly interested in getting an idea about the feasibility of our system and get on how users like the idea of using this multimodal interaction to control their smart environments. All subjects were able to perform all of the tasks with a noticeable learning effect from the first to the last tasks, reducing the number of wrong attempts and increasing the interaction time.

In a following interview the test persons considered the combination of speech and gesture to be preferable to gesture or speech alone. The subjects considered the interaction to be intuitive and easy to master and particularly liked how pointing can simplify the complexity of speech commands. However only one candidate could imagine using such a system right now to control devices and there were concerns about the performance of speech recognition, which can be attributed to the fact that the training had to be performed unspecified.

7 Conclusion and Future Work

We have presented a method that combines speech and gesture recognition to simplify interaction in smart environments. Using a virtual presentation of the environments we are able to control the gesture recognition using bounding volume modification. A test system based on the Microsoft Kinect and CMU Sphinx speech recognition was set up and tested with nine different subjects. The system compared favorably to time-based selection methods and all users were able to complete the defined set of tasks. Combining speech and gesture to control smart environments offers a huge potential. We can use the combined information to simplify interaction of the different modes. Using bounding volumes to realize this multimodal combination allows a direct integration in virtual representations of the smart environment and the possibility for modeling other aspects such as uncertainty or give an importance measure for the different devices, e.g. by changing the scaling factors based on confidence and a user-assigned weight. The initial results make us confident that the combination of speech and gesture to select and control devices is an approach that should be followed further.

We intend to upgrade our prototype system to a more capable speech recognition that does not require the user to hold a microphone, e.g. by using on-line speech recognition and microphone arrays. The gesture recognition performed favorably but can be improved using different feedback methods and a more precise skeleton tracker. In terms of bounding volumes we want to compare the results of different modification methods both quantitative in terms of how they fill the space and acquire a qualitative result on how they influence user experience. Another idea is to provide a measure how well-suited a given environment is for this kind of interaction based on size, capabilities and position of the included devices.

References

1. Braun, A., Kamieth, F.: Passive identification and control of arbitrary devices in smart environments. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part III, HCII 2011. LNCS, vol. 6763, pp. 147–154. Springer, Heidelberg (2011)
2. Wilson, A., Shafer, S.: XWand: UI for intelligent spaces. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, pp. 545–552. ACM (2003)
3. Cao, X., Balakrishnan, R.: VisionWand. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology - UIST 2003, pp. 173–182. ACM Press, New York (2003)
4. Majewski, M., Braun, A., Marinc, A., Kuijper, A.: Visual Support System for Selecting Reactive Elements in Intelligent Environments. In: International Conference on Cyberworlds, pp. 251–255 (2012)
5. Shneiderman, B.: Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on Graphics 11, 92–99 (1992)
6. Heinze, C.: Modelling intention recognition for intelligent agent systems (2004)
7. Tahboub, K.A.: Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. Journal of Intelligent and Robotic Systems 45, 31–52 (2006)

8. Yamamoto, Y., Yoda, I., Sakaue, K.: Arm-pointing gesture interface using surrounded stereo cameras system. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, pp. 965–970. IEEE (2004)
9. Heider, T., Kirste, T.: Supporting goal-based interaction with dynamic intelligent environments. In: ECAI, pp. 596–602 (2002)
10. Valli, A.: The design of natural interaction. Multimedia Tools and Applications 38, 295–305 (2008)
11. Oviatt, S., Cohen, P., Wu, L., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., Ferro, D.: Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions. In: Human-Computer Interaction, vol. 15, pp. 263–322 (2000)
12. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. ACM Transactions on Computer-Human Interaction 9, 171–193 (2002)