Pers Ubiquit Comput (2007) 11:609–619
DOI 10.1007/s00779-006-0138-8

ORIGINAL ARTICLE

Towards a Multimodal Interaction Space: categorisation and applications

Bert Bongers · Gerrit C. van der Veer

Received: 27 November 2005 / Accepted: 23 August 2006 / Published online: 14 February 2007

Abstract Based on many experiences of developing interactive systems by the authors, a framework for the description and analysis of interaction has been developed. The dimensions of this multimodal interaction space have been identified as *sensory modalities*, *modes* and *levels of interaction*. To illustrate and validate this framework, development of multimodal interaction styles is carried out and interactions in the real world are studied, going from theory to practice and back again. The paper describes the framework and two recent projects, one in the field of interactive architecture and another in the field of multimodal HCI research. Both projects use multiple modalities for interaction, particularly movement based interaction styles.

Keywords Multimodal interaction · Interactive architecture · Framework

B. Bongers (✉)
HCI, Multimedia and Culture,
Department of Computer Science,
Vrije Universiteit, De Boelelaan,
1081a Amsterdam, The Netherlands
e-mail: bertbon@xs4all.nl
URL: www.bertbongers.com

B. Bongers
MAAS Interaction Lab, Cörversplein 1A,
Maastricht, The Netherlands

G. C. van der Veer
Human Computer Interaction, School of Computer Science,
Open University, Valkenburerweg 177, Heerlen,
The Netherlands
e-mail: gerrit@acm.org

1 Introduction and background

In this paper a theoretical framework to place and describe interactions is introduced, and illustrated by recent projects where it has been applied. Now that computers are becoming more omnipresent and ubiquitous, embedded in other technologies and increasingly networked, we find ourselves in an electronic ecology or what can be called the *e*-cology [1]. In order to interact with such a complex environment, interaction styles have been developed and applied which include multiple senses and modalities, various modes and layers. This research takes place in the field of human–computer interaction, where the 'computer' now is defined as a (possibly) networked and embedded, distributed system.

The multimodal interaction space described in this paper is based on a wide range of HCI literature, and informed by personal practical experience and experiments in numerous settings. These particularly include: electronic musical instruments [2, 3], interactive architecture [4], real time video performances [5], mobile projections, and multimedia home systems [6]. Further inspiration comes from interaction styles that are developed fields such as Augmented Reality [7] and collaborative design environments.

Going from theory informed by practice, to practical experiments and developments, back to theory enables us to work towards a framework to categorise interaction styles including gestures, writing, speech, etc. The dimensions of such a multimodal interaction space are becoming clear and are described in Sect. 2. This is then illustrated by two recent projects that one of the authors has been involved in that support the mixed reality of computer generated elements focussing on

the interaction (Sect. 4) and actual architecture in a more—though not quite—traditional sense (Sect. 3) (Fig. 1). This way a path is followed from theory to practice and back again. There are interactions found in the real world which are challenging the definition of a theoretical framework, potentially leading to improvements of the framework.

2 The multimodal interaction space

A *modality* is a communication channel, for instance related to the human senses or the form of expression. In HCI, a considerable amount of research has been done on combining multiple modalities in order to achieve a higher bandwidth of interaction between people and their technologies. The goal is not only to make the interaction more efficient or effective, but there can also be other objectives such as making the interaction more pleasurable or fun, or more natural.

There have been several projects to describe multimodal interaction between humans and technology [8–12]. In the final report of the MIAMI European project in the mid 1990s, a good overview is given and goes towards a categorization [13]. Other works look at how, through these media, communication takes place between people [14]. The World Wide Web Consortium works on a Multimodal Interaction Framework, the emphasis however is on the interpretation and semantic layer *inside* the system [15].

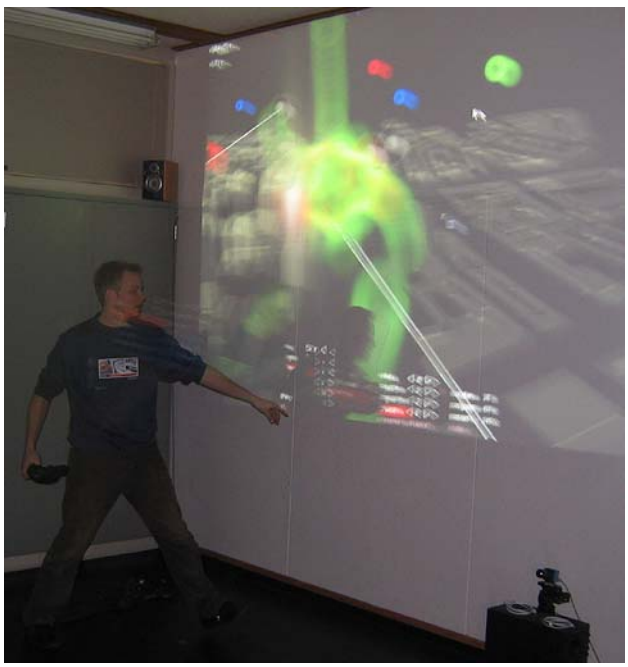

Fig. 1 Interaction in an architectural design environment

Our focus is on the *real space* where the interaction takes place, closer to humans, and we are developing a descriptive framework for interaction styles starting from the physical level.

The *Multimodal Interaction Space* (MIS) can be described in: *levels* (physical, syntactic, semantic, task, goal, etc.), *modes* (textual, continuous, non-verbal, subconscious, intentional, etc.) and sensory modalities (seeing, hearing, touching, etc.)

Any interaction style can be placed in this space. Interaction usually (ideally!) takes places using many possible combinations of modalities, sequentially and/or in parallel. An interaction style is therefore not a place in the Interaction Space but a trajectory through it, particularly described in the levels (getting from the goal to the action, and back again analyzing the results of the action).

The framework is human centred, i.e., it is not concerned with machine input and output modalities. The physical level of human interfaces with technology is described in the Physical Interface Design Space [16].

2.1 Levels of interaction

An interaction can be described in several layers, taking the user from a goal and intention, formulating a task and subtasks, carrying out these actions whilst receiving feedback on the physical level, and evaluating the result.

An action is usually initiated in order to achieve some higher order goal or intention, which has to be prepared and verbalised, and finally presented and articulated through physical actions and utterances. The presentation and feedback by the computer passes through several stages as well, before it can be displayed, possibly in various modalities including the haptic, in order to be perceived by the user. The actual interaction takes place at the physical level. In the standard literature, often three levels are discerned: semantic, syntactic, and lexical [17], but for more specific cases more levels can to be described. Jakob Nielsen's virtual protocol model [18] is example of this, specifying a task and a goal level above the semantic level, and an alphabetical and physical level below the lexical level. It is interesting to note that a hierarchical task analysis (HTA) often reflects these levels. The levels not only particularly describe well the spoken or written language, but can also be applied on direct manipulation interface paradigms [19]. Donald Norman makes a useful explicit discrimination between input and output flows of information in *stages* in his Theory of Action [20]. Users have to accomplish their

goals through the physical system's action through two processes, having to bridge a Gulf of Execution and a Gulf of Evaluation by the flows of actions in various stages which emphasises the asymmetry in the interaction.

The Layered Protocol is an example of a more complex model [21], particularly to describe the dialogue using the speech modality, but also applied to general user interface issues [22].

When more sensory modalities are included in the interaction, models often have to be refined. Applying the Layered Protocol in the interaction which includes active haptic feedback, introduces the idea of (higher level) E-Feedback which has to do with *expectations* of the system of the user, and the I-Feedback which communicates the lower level *interpretations* of the user's actions by the system [23].

It can be said that virtual messages are exchanged between higher levels between user and system (still through translations to the physical level though), and that various messages are multiplexed into others and vice versa [24].

Garett's Elements of User Experience is an example of a more recent model, developed to include approaches from design and engineering particularly of web site architectures [25].

The articulatory feedback (or interpretation feedback) on gestural control which is studied in the research described below in Sect. 4, takes place at the physical level but can be extended to include the semantic levels.

Summary of levels: goal, task, semantic, syntactic, lexical, alphabetical, physical

2.2 Human input modalities (senses)

An interaction can be based on addressing all possible sensory modalities such as the visual and the auditory. There are more than the traditional five senses (seeing, hearing, smelling, tasting and feeling), lumped together under the fifth sense of feeling (or the bodily senses) are in fact a number of senses. One can feel pain (nociception), motion, gravity, acceleration, equilibrium, pressure, and so on, which are all very relevant in the context of the physical interface.

Our sense of touch, the tactual sense, has three sources: the signals from the mechanoreceptors in the skin (our cutaneous sensitivity) informing our *tactile* sense, the mechanoreceptors in the muscles and joints (our proprioceptors) inform our *kinaesthetic* awareness of the location, orientation and movement of body parts, and the efferent copy signal that occurs when a person is actively moving by sending signals from the brain to the muscles. *Haptic* perception involves all three channels, which is usually the case when a person manipulates an object or interacts with a physical interface [26].

Furthermore, there is the issue of self-perception or *proprioception*. When interacting, an individual is inherently active, and therefore aware of it. There are internal feedback loops that guide the control of the act, for instance when focussing the eye, articulating speech, moving around and guiding manipulation. It makes a difference if a stimulus is imposed or obtained (as in the difference between tactile and haptic). The internal feedback often goes together with feedback perceived externally, which in the case of technology has to be provided by the system and explicitly designed, built in or programmed.

In summary: visual, auditory, tactual, olfactory, gustatory, tactual, temperature, nociception, vestibulary (almost all of these senses have an outside as well as an inside—proprioceptive—element).

In order to establish a better match the human senses need to be studied in more detail, as has been done in the field of psychology of human perception. However, the majority of this research is based on stimulus–response paradigms in fixed laboratory conditions. In the context of HCI research, we need to take into account the whole loop, and preferably study them in more complex situations. Generally, in real life, perception and action are closely linked. Therefore the work of J. J. Gibson is useful in the study of human–technology interaction, because of his emphasis on active perception and the role of the context or ecology that the interaction is part of. This is described in his third book [27], including the notion of affordances as later applied in HCI in the work of Donald Norman and Bill Gaver. In Gibson's second book he already proposes to 'consider the senses as perceptual systems', in five categories (leaving the proverbial sixth sense intact) of *systems*: Basic Orientation, Auditory, Haptic, Taste–Smell, and Visual. He emphasises the *activity* in each system, e.g., looking, listening, and touching rather than seeing, hearing and feeling [28].

2.3 Modes

Interactions can take place in several modes, for instance a text modality or a manipulation modality (here called 'continuous', described as 'analog' in Niels Ole Bernsen's Modality Theory [29]). Furthermore, human utterances can be unconscious and in some cases also involuntary.

The description of modes reflects primarily the human output modalities with which it influences its

environment and communicates with other people (possibly mediated through technology). The modes are: symbolic (e.g., text, speech, Braille), iconic (mimicking), para-linguistic (or non-verbal, e.g., accompanying symbolic mode), involuntary (not under conscious control) and even subconscious. A manipulative output or action is called continuous.

Note that these modes often depend on the context: when typing on a keyboard the movements of the fingers (gestures) have a different meaning than when playing on the piano, tapping on the table, etc.

2.4 Human output modalities (action)

Classifying human output modalities is not as straightforward as the input modalities. Every modality has the goal to establish communication, and therefore aims to be perceived. Whether it is a human output modality or computer system output (display), a way of describing often found in the literature is by sensory modality. However, in some cases an utterance, for instance a gesture which is intended to be perceived visually (by another person or an electronic system) can also be perceived haptically if the other person is touched. This influences the action, because it becomes an interaction. Another example is the Tadoma method, where a deaf and blind perceiver puts the hand on/against the face of the speaker, perceiving the spoken utterances through a combination of the vibrotactile and haptic senses.

The same is true for the meaning of the action (at the semantic level, see above). A gesture in free space may mean nothing, until it encounters for instance the light switch. This means that in the case of a more complex interactive system the person making the gesture must be aware of whether the motion is tracked with a camera system, and what results the actions (might) have. This is getting frightfully close to the philosophical question about the sound that a tree makes when falling in the middle of the woods where no one can hear it.

Human output modalities are usually involving our muscles, such as for manipulating things, locomotion, and the fine motor control involved in producing speech. Not only are our perceptions often multimodal, most utterances or actions are too. For instance, when speaking not only information is conveyed through the meaning of the words, but also the tone of the voice (pitch, prosody) and the accompanying gestures and body language.

There are many other human output modalities not involving muscles. There are several somatic (bodily) modalities such as blood pressure, temperature (blushing), excretion (sweating, crying), heartbeat, some of which are not under conscious control and may be unintentional. It therefore makes a difference whether the actor cries in a movie, or a person cries for a genuine reason. People still communicate through smell (not as much as animals do, or our ancestors), either involuntary by body odors or intentionally by putting on perfumes. Some output modalities can only be applied by involving an interface, in the case of some of the somatic modalities as described above, and particularly in the case of bio-electricity.

Communications can be asynchronous, for instance sending a letter by pigeon or e-mail, or preparing a meal through which the cook will address the taste–smell system of the perceiver.

2.5 Summary of the MIS framework

The point of this section is to illustrate how many more possibilities there are to increase the bandwidth of the interaction between humans and their technological environment. Through the use of technology, from a pen or paintbrush, a musical instrument, to new media, humans can express and act in a far bigger scale and with more variety than ever before, and this is still increasing. This has implications for the way the interactions are organised. The framework is not just a classification of existing interaction styles but tries to take into account what *would* be possible.

In this section the interaction between humans and technological environment has been analyzed in its parts and brought together in the descriptive model of the *Multimodal Interaction Space*. The dimensions of this design space are:

- levels (physical, syntactic, semantic, task, goal, etc.)
- modes (textual, continuous, non-verbal, subconscious, intentional, etc.)
- senses/modalities (seeing, hearing, touching, etc.)

To complete the human–system interaction loop, it is good to include the processing levels (cognition and memory) at both sides. Further experimenting is going on, as well as studies of real-world interactions. A visual representation has to be developed which shows clearly and quickly the various interaction possibilities.

In the next sections some example projects are described which served as test cases for the developing multimodal interaction architectures. One is actually an architectural project, and the other is about the development of an interaction style for mixed reality situations. The terms as described in this section on the multimodal interaction framework are used in the next section which illustrates the practical applications. The terms are indicated by underlining.

3 Multimodal interaction in protospace

At the architecture department of the Technical University of Delft, a new interactive space has been set up in the last years called Protospace[1], by the Hyperbody research group of Professor Kas Oosterhuis [30] in collaboration with the MaasLab. Rather than the virtual reality (VR) approach with its emphasis on the world inside the computer and projecting this outwards in order to involve human interaction [31], the emphasis in this research is on merging the virtual with the real, leading to a mixed or augmented reality, including a dynamic architecture [32, 33].

The aim of Protospace is that through multiple, full field of view and eventually 3D projections (using polarised light), teams of designers can work collaboratively on the creation of structures and environments. The parametric nature of these kinds of architectural designs is particularly well suited for interactivating, that is, actively being interacted with by the users through sensor systems. For Protospace, a system has been developed consisting of a combination of on-body and in-space sensing techniques, to control the virtual worlds and elements.

The group often uses games as a metaphor for the collaborative design activity, so below we often use terms such as player instead of user.

3.1 Design approach and process

The design and development of such a complex and new interactive system preferably takes place in an iterative way, in a combination of *bottom up* (technology driven, engineering) and *top down* (visionary, intuitive) approaches. We made an overview of system functions, controls, and feedback. These are organised in *palettes*. The elements of the palettes are linked through experimentally established *mappings*. In this section the focus is on the palette of interfaces or *interaction appliances*. Every interaction appliance is linked to a certain (set of) functions in the design environment, and combined with the appropriate feedback. This is in fact similar to a traditional workshop, such as a mechanical workshop, a dentist, or an instrument builder's atelier. What one will see here is a set of tools, arranged in a spatial manner supporting overview and availability. A professional developer or designer in a traditional workshop has a tool at hand for any (set of) tasks or operations to be carried out, rather than having one general purpose tool (Swiss army knife, Leatherman). Bill Buxton has made this comparison, between the ''strong specific'' and the ''weak general'' [34]. With the computer, the standard paradigm is a one-interface-fits-all, general purpose interaction style. In experimental interactive environments such as Protospace it is possible to apply many different interaction styles.

Previously the interaction styles developed were based on wireless game controllers and various sensors. The game controllers are quite versatile, they contain a number of buttons for mode switching or other actions, and two small analogue joysticks used for navigation and manipulation. The mapping between these input elements and system parameters is not the most intuitive, but it works and multiple devices can be used. Various sensors are placed in the space, such as switch mats in the playing field and photocells and motion detectors which enable a spatial control of switching on/off parameters, combined with proximity sensors that would allow continuous changes in the space. This set up is described in several research papers [35,36]. The earlier choice of input modes was felt to be too limited.

In the recent phase of the project, a team of people have been working on researching gesture tracking and speech recognition. First a thorough investigation and overview was made of existing speech recognition and video tracking, for the latter the particular focus was on systems developed for the performance arts such as music and dance. Most systems were tested out and worked with in both labs, including comparing latency issues of various hardware and software elements (different cameras, connections, drivers and applications).

We have developed and implemented a system which is highly flexible and scalable. Our aim was not to solve one particular problem (which would be, after all, inherently unknown) but to create a platform, a set of tools to work with, an expandable palette of interaction possibilities. The investigations and choices have been driven by the practice of a collaborative architectural design environment.

3.2 System overview

The system developed consists of a separate 'interaction computer' for real-time data, audio and video manipulation, communicating with one or more 'parametric architectural model computer(s)'. The interaction computer is an Apple PowerMac G5, running the Max/MSP/Jitter graphical programming environment. Max is particularly developed for handling real time data, MSP is the sound processing extension while Jitter has many objects for real-time

[1] see http://www.protospace.bk.tudelft.nl

video processing[2]. It can receive all performance data from all the sensors, the various input devices, cameras and microphones, so that it can *interpret* and relate all data to individual player's actions. It then passes on semantic data to the system that generates and manipulates the real-time parametric architectural models on PC computers running a real-time rendering environment called Virtools (originally intended for game development[3]), which projects the parametric architectural models (potentially in 3D) in Protospace. The interaction computer generates the direct feedback on the player's actions, to facilitate articulation and guiding. An overview of the system can be seen in Fig. 2.

The reason for introducing a separate computer to handle the interaction with the people in order to ensured that all timing requirements are met. It is believed that in order for the user experience to be convincing, and to make the interaction optimal, at least the articulatory feedback has to be presented within the time accuracy of the various human perceptual systems. For instance, for a trained musician the time discrimination lies in the order of tens of milliseconds in the most extreme cases, and the haptic system operates optimally under similar conditions. The Max environment has a precision of 1 ms, and will not be interrupted by any other task carried out by the operating system (Mac OSX).

The communication between the interaction computer and Virtools computer is done using OSC, Open Sound Control. OSC works over Ethernet (also wireless) and is suitable for the transmission of high bandwidth real time data. We are using an OSC 'building block' for Virtools [37].

The screen shot in Fig. 3 shows a part of the Virtools graphical editing environment, with its building blocks at the bottom and the final image produced at the top left. The picture in Fig. 4 shows an early stage of the Max/MSP/Jitter 'patch', with the objects and images grabbed from the camera input.

The goal of the Virtools program is to render real-time images, and is very much based on frame rate. The Max environment was chosen for the interaction computer because it is very suitable for manipulating real-time data.

3.3 Gesture tracking

The video tracking is done with industrial zero-latency Firewire (IEEE 1384) cameras (Fig. 5) interpreted in software. Before developing the semantic layer(s) in the gestural interaction, we first concentrated on making an optimal continuous mode of interaction for direct manipulation and navigation. To make the gesture tracking more precise and responsive, optical *beacons* are used (Fig. 6). These are small tracking objects with lights, gentle glowing coloured jumbo-LEDs and infrared LEDs. They are combined with a small battery, or mounted on the game devices and powered from the internal battery. Using these beacons, the system can be used under realistic conditions, i.e. not disturbed by other movements, against any background and under various lighting conditions. Experiments have been carried out with coloured or reflective material but this didn't work so well. Some of the cameras are equipped with filter material which blocks all visible light, enabling more accurate tracking of infrared beacons undisturbed by other light sources or conditions. Tracking speed and latency are important issues at this level of interaction. The common technique of analyzing the difference between two successive frames is done by a Jitter object 'FindBounds'. At a frame rate of 25 fps (at a resolution of 320 × 240 of each camera) each frame is already 40 ms long, and with a minimal amount of processing time a response time of below 100 ms should be obtainable, which is an acceptable value for continuous control and feedback. The interaction computer generates real-time auditory articulatory feedback generated by MSP, and passes the data on in real-time to the Virtools environment which visually represents the changes in the architectural model's parameters as well as generating visual articulatory feedback on the screen.

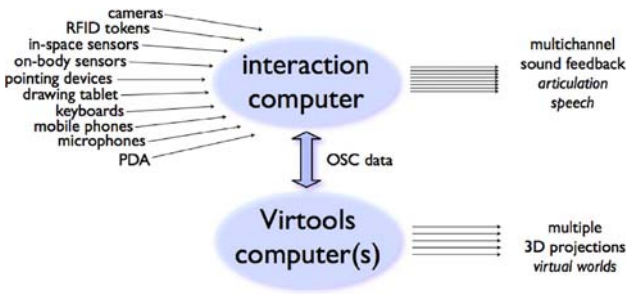

Fig. 2 Overview of the Protospace interaction system

The auditory feedback uses the parameters of pitch, volume, timbre and panning related to the movements and identity of the individual players. The overall sound level was kept low, giving peripheral rather than overly explicit feedback.

3.4 Speech recognition

For the speech recognition a Max object called 'Listen' was used. This Max object communicates with the Mac

[2] http://www.cycling74.com.

[3] http://www.virtools.com

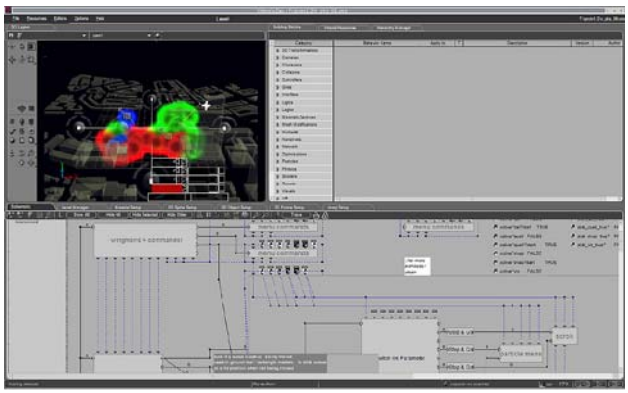

Fig. 3 The Virtools programming environment and urban deisgn section

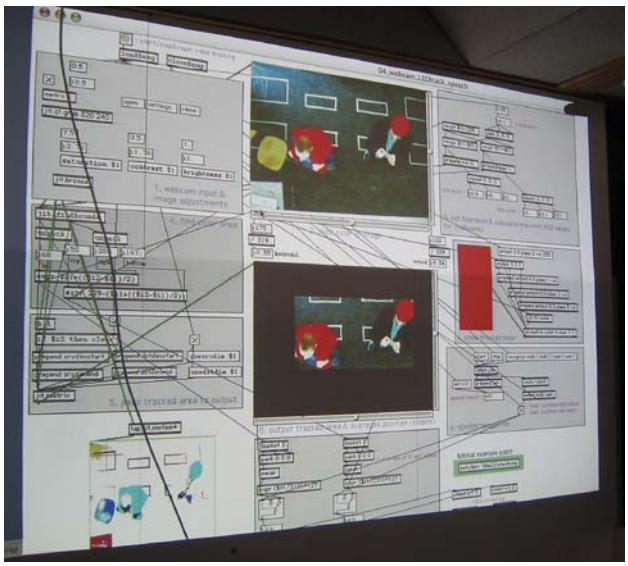

Fig. 4 Max patch for experiments with camera tracking

Fig. 5 Industrial Firewire camera

Fig. 6 Game controller with LED beacon

OSX built in recognition system. This system is particularly well suited for this application as it doesn't require training so that it can be used by different players. Apple's speech recognition is speaker independent and adaptive. At the current stage in the Protospace project it is only used for recognizing single words as commands, the symbolic mode of interaction. The words to be recognised are defined in Max. Using head mounted wireless microphones makes the speech recognition more reliable, and less obtrusive for the user. Feedback is generated by the Mac OS recognition system, but could be generated from within Max as well. There are Max objects that can analyse (voice) sounds and determine through Fourier analysis the frequency distribution and amplitudes. For instance, the 'fiddle' object [38] was used to extract basic pitch and map that to a parameter of the system. This way, the voice is used in continuous control interaction mode. It was found however that architects seem to be reluctant to use their voice in this mode. The command mode works very well though, with different speakers.

The high-quality audio system for feedback consists of six channels (mid and high frequencies) and one channel of low frequency sound (subwoofer). This enables us to place articulatory feedback sounds in 3D space. The speakers can also used for synthesized speech feedback for the symbolic interaction mode.

3.5 Demo and results

A demo has been developed in Virtools showing the interactions and behaviours of the models. In the demo a team with different roles of designer/architects, a project leader, an economist, a client, etc., all with their

own ways of interacting with the system and work on a collaborative architectural design.

The approach was to think and develop in *palettes*. There is a palette of interaction modalities, a palette of feedback modes, and a palette of parameters in the modeling environment. Between these palettes *mapping* is worked out, finding the optimal interaction style for each task. In a next phase, we not only want to add more to the palette but particularly further develop the application of the tools in a practical and realistic environment. This way a large part of the Interaction Space is covered.

With this stage of the project a basic system has been implemented and convincingly demonstrated. From this further developments are possible, particularly to further involve larger design teams to carry out work in this environment.

The research direction from here will stick to the path of merging the real and the virtual worlds, the mixed reality.

3.6 Current work

At present we are extending the interaction palettes of Protospace even further. We are improving the existing technologies, investigating a great number of promising techniques. A number of generic operations have been identified (such as pointing, navigating, manipulating, selecting, storing and moving) and we are developing interaction appliances that fit these operations. For instance, in many cases the participants need to quickly adjust a number of parameters. For this the 'menu rollator' was developed, an assembly of two wheels (rotary encoders), to be manipulated with the thumb of the right hand (Fig. 7). One wheel, with the up/down movement (rotary DoF around the Y-axis) is used to select the parameters which are arranged in a vertical list on the screen. The other wheel, with a left/right movement (rotary DoF around the Z-axis), sets the value. The wheel movements and the effects are enhanced with auditory feedback, little clicks that support the motion.

New and more flexible sensor converters are used, including the extension through wireless sensor networks [39], extending the approach of the 'interaction computer' with a distributed computing layer.

Linking real world objects to the system through RFID tags and other ways of recognition (also through the cameras), enables us to link to the underlying data structures with a token-based interaction. When a player wants to interact with an appliance on a particular screen, the appliance (which contains an RFID tag) is held near the reader below the screen. The system then links the interaction appliance to the chosen screen.

Fig. 7 Assembly of rotary encoder wheels to manipulate menus and variables

Environmental parameters such as light, humidity, and temperature are sensed and represented in the virtual worlds in an implicit way. A number of standard input devices are added to the palette, including a drawing tablet (for the continuous mode of interaction, and also for the symbolic mode through character recognition). Bluetooth devices such as GSM phones can be used as control devices as well now, which enables the incidental visitor to participate in the design process with limited functionality.

A handheld miniature computer (PDA) is used for interaction, to send commands from the touch screen to the system and receive visual feedback on the internal screen, extending the interaction space.

4 LaserTouch pointer

Another example which explores multimodal interaction styles is the LaserTouch project at the Vrije Universiteit Amsterdam. Here a gestural controller is used based on a laser pointer and a camera tracking system, with added tactual feedback and with the explicit aim to interact with both the virtual as well as the real world.

4.1 Laser pointer tracking

To use video tracking of the dot of a laser pointer is a well known technique [40]. A recent paper by Brad Myers et al. [41] gives a good overview of such systems, and reports critically on the low accuracy of the laser pointer based interaction due to the 'magnification',

the leverage of hand instability when operating over a larger distance. The paper further investigates the influence in the shape of the pointer on the accuracy, and compares the laser pointer technique with other input modes.

Others have developed real world exploratory applications of this idea, for instance to apply as an aid for blind people [42] currently with auditory feedback and potentially with tactual feedback too. However it proves difficult to replace the traditional 'cane' with all its richness and various modes of interaction, as can be read on an Internet forum discussing this research[4]. This gives considerable insight into the actual issues involved (including social) in using such a cane, including the safety issues related to waving around laser beams (by both blind or sighted people).

As often, it seems that by focusing on overcoming the limitations of existing technologies or interaction styles, as those certainly present in the case of the 'cane' (for instance its limited length), the inherent strengths may disappear too. In the case described below, we therefore first approach the interactions not possible before and from there hope to include the established layers.

4.2 Remote touch and ubiquity

The reason for us to use a laser pointer is to explore the possibility to point at *both real and virtual objects*. The virtual objects are projected by a video projector, and real objects such as light switches and loudspeakers are present in the space. If the camera tracking system knows the coordinates of these elements in space, appropriate responses can be generated. In the ubiquitous computing paradigm after all, the parameters of various systems would all be controllable through one interface. The parameters of these objects, whether real or virtual, can then be manipulated.

As it has been found in other researches including our own, presenting *active tactual feedback* to the user helps the articulation process [43]. We therefore included a small vibrotactile actuator in the device, enabling a kind of *remote touch*, feeling the pixels on the screen as in the ''Palpable Pixels'' [44], as well as other objects in space. This research is an extension of the earlier work with a mouse with active tactual feedback, now in a situation of unguided gestural control where only the kinesthetic awareness is informed by the internal signals in the human body (by the proprioceptors and efferent copy). It is expected that under these circumstances the added feedback will play a great role in the improvement of the articulation and steadiness of the control function.

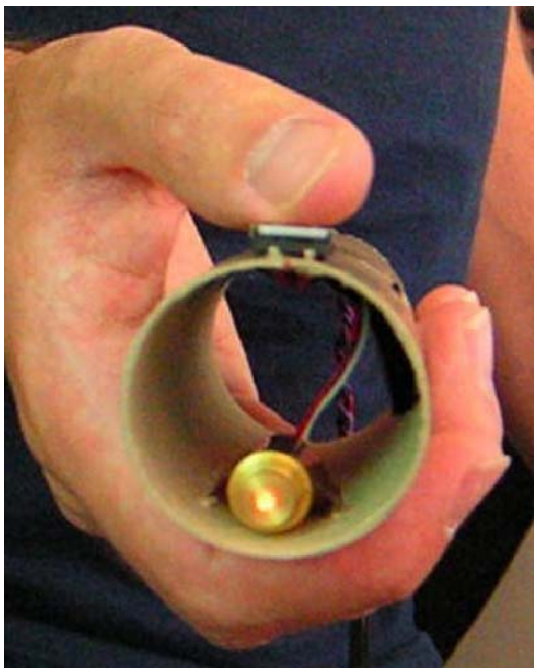

Fig. 8 The LaserTouch pointer prototype

4.3 System set-up

Again in this project we use a 'patch', a program written in the Max/MSP/Jitter software, for the video tracking (Jitter) and handling other sensor input (Max) through a Teleo USB module see [5]. The tactile feedback is generated as low frequency sound by MSP, and linked to the textures projected by the computer. For the gesture tracking, a Firewire camera is used, in this case an Apple iSight. The camera has a filter to block environmental light, and is precisely tuned to the wavelength of the laser pointer so that only the dot appears in the system.

A quick assembly was made on a carton pipe with laser pointer, selection switch, and a small loudspeaker as tactile element. In the picture (Fig. 8) it can be seen that yes, it is actually a toilet roll, it was made by students and in the VU HCI-Lab we often deliberately try to work with low tech materials whenever possible.

Currently this contraption is wired, but could quite easily be made wireless in a next phase using Bluetooth. We have investigated this, and using the recently introduced HiFi headset profile sound quality would be good enough to accurately display vibrotactile cues (the standard headset profile is proved to be not good enough).

[4] http://www.engadget.com/entry/1234000690023779

[5] http://www.makingthings.com

4.4 Demo and experiences

To try out this combination of modalities a demo was created, with which first experiences have been gathered.

Compared to the Gyropoint gyroscopic 'air mouse' that the first author uses frequently for presentations, the laser pointing technique seems to work fine, with the added benefit of extending the operating range outside the projected image. The tactile feedback seems helpful, although the speed needs to be improved in order to create a convincing experience.

5 Discussion and conclusions

In this article we have laid out a descriptive model of the interaction space, which is multimodal and multilayered. This *Multimodal Interaction Space* (MIS) is a design space to describe interactions in:

- levels (physical, syntactic, semantic, task, goal, etc.)
- modes (textual, continuous, non-verbal, subconscious, intentional, etc.)
- senses/modalities (seeing, hearing, touching, etc.)

This MIS is illustrated by two example projects, to further explain and apply the framework. For the development of interaction styles we think in *palettes*, of modalities and system parameters, which have to be mapped onto each other. This is depending on the 'task' or application and context, which varies over time and therefore a flexible, scalable and configurable system is being developed.

The interaction system in Protospace is developed at a proof-of-concept level, based on the anticipated needs of a design team that the developers are part of. It was found that having a common language to describe the interactions by the terms of MIS, discussions between the team members was improved. Creating an overview of the interactions and the mappings between human modalities and system parameters was possible through the application of the MIS framework. The framework also facilitated identifying missing interactions styles, directing research into these opportunities. In the next phase more thorough work sessions will be conducted in Protospace with multidisciplinary teams of designers. This will inform the development of the suitable mappings, guided by the interaction framework.

The system is scalable enough to be further expanded to include more interaction styles in the future, without performance degradation that would influence the interaction. It was found very useful to have a separate 'interaction computer', that handles all interactions (input and feedback) in real time. The communication between the interaction computer and the computers running Virtools for the generation of the models can now be done using OSC.

The new interaction style of the LaserTouch pointer, combining gestural spatial control with vibrotactile feedback, looks promising but has to be further improved before the necessary user tests can be carried out.

These and other developments of multimodal interactions will continue to inform the development of the Multimodal Interaction Space.

With the solutions as presented in this paper, a coupling of people and electronic environment is established through interaction appliances. These interaction appliances link human modalities to system functions in an intuitive and flexible way. The modular approach offers solutions to the users, who can select and manipulate the appropriate interaction appliances for any task on hand.

Acknowledgments The authors like to thank the student interns who worked on the LaserTouch Pointer in the summer of 2004 at the Vrije Universiteit (VU), Bart Gloudemans and Sylvain Vriens. We would express out thanks to Eric-Jan van Duijn from the Laser Centre of the Atomic Physics Group at the VU for his help with the laser light filter. The Multimodal Interaction in Protospace team consist of Prof. Kas Oosterhuis, Hans Hubers, Dieter Vandoren, Tomasz Jaskiewicz and Christian Friedrich of the Hyperbody research group at the TU Delft, and Yolande Harris of the MaasLab in Maastricht. We thank the reviewers for the insightful and constructive criticism on the draft of this paper.

References

1. Bongers AJ (2004) Interaction with our electronic environment; an *e*-cological approach to physical interface design. Cahier Book series, Hogeschool van Utrecht
2. Bongers AJ (2000) Physical interaction in the electronic arts, interaction theory and interfacing techniques for real-time performance. In: Wanderley MM, Battier M (eds) Trends in gestural control of music. IRCAM, pp. 41–70
3. Paradiso J (1997) New ways to play: electronic music interfaces. IEEE Spectr 34/12:18–30
4. Bongers AJ (2002) Interactivated spaces. In: Proceedings of the symposium on systems research in the arts, Baden-Baden, Germany, August 2002
5. Bongers AJ, Harris YC (2002) A structured instrument design approach: the video-organ. In: Proceedings of the conference on new instruments for musical expression, Dublin, May 2002
6. Sluis R, van de Eggen JH, Jansen J, Kohar H (2001) User interface for an in-home environment. In: Proceedings of the interact conference, Tokyo 2001
7. Bowman DA, Kruijff E, LaViola JJ, Poupyrev I (2004) 3D User interfaces, theory and practice. Addison Wesley, Reading

8. Nigay L, Coutaz J (1993) A design space for multimodal systems: concurrent processing and data fusion. In: Proceedings of the InterCHI, 1993
9. Dragicevic P, Navarre D, Palanque P, Schyn A, Bastide R (2004) Very-high-fidelity prototyping for both presentation and dialogue parts of multimodal interactive systems. EHCI-DSVIS In: Preproceedings 61–88, Hamburg, 2004
10. Schaefer R, Bleul S and Müller, W (2004) A novel dialog model for the design of multimodal user interfaces. EHCI-DSVIS In: Preproceedings, Hamburg, 2004, pp. 390–391
11. Bongers AJ, Eggen JH, Keyson DV, Pauws SC (1998) Multimodal interaction styles. HCI Letters J 1/1:3–5
12. Camurri A, Mazzarino B, Ricchetti M, Timmers R, Volpe G (2003) Multimodal analysis of expressive gesture in music and dance performance. In: Camurri A, Volpe G (eds) Gesture based communication in human–computer interaction, LNAI2915.Springer, Berlin Heidelberg New York, pp 20–39
13. Schomaker L, Münch S, Hartung K, (eds) 1995, A taxonomy of multimodal interaction in the human information processing system. Report of the ESPRIT project 8579: MIAMI
14. Kress G, Leeuwen T van (2001) Multimodal discourse, the modes and media of contemporary communication. Oxford University Press, Oxford
15. Larson JA, Raman TV, Raggett D (eds) (2003) Multimodal interaction framework. W3C Note, on line at http://www.w3.org/TR/mmi-framework
16. Harris YC and Bongers AJ (2002) Approaches to creating interactivated spaces, from intimate to inhabited interfaces. J Organised Sound, Cambridge University Press, Special issue on Interactivity 7/3
17. Dix A, Finlay J, Abowd G, Beale R (1993) Human–computer interaction. Prentice Hall, Englewoods Cliffs
18. Nielsen J (1986) A virtual protocol model for computer–human interaction. Int J Man Mach Stud 24:301–312
19. Nielsen J (1992) A layered interaction analysis of direct manipulation. http://www.useit.com/papers
20. Norman DA (1986) Cognitive engineering. In: Norman DA, Draper SW (eds) User centered system design, chpt. 3. Lawrence Erlbaum Ass., Hillsdale
21. Taylor MM (1988) Layered protocol for computer-human dialogue. I: Principles. Int J Man Mach Stud 28:175–218
22. Eggen JH, Haakma. R, Westerink JHDM (1996) Layered protocols: hands-on experience. Int J Man Mach Stud 44:45–72
23. Engel. FL, Goossens P, Haakma R (1994) Improved efficiency through I- and E-feedback: a trackball with contextual force feedback. Int J Man Mach Stud 41:949–974
24. Taylor MM and Waugh DA (1991) Multiplexing, diviplexing, and the control of Multimodal Dialogue. In: Taylor MM, Néel F, Bouwhuis DG (eds) The structure of multimodal dialogue II. Acquafredda di Maratea, Italy
25. Garrett JJ (2003) The elements of user experience. New Riders Publishing, Indianapolis
26. Loomis JM, Leederman SJ (1986) Tactual perception. In: Handbook of perception and human performance, Chp. 31
27. Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin, Boston
28. Gibson JJ (1966) The senses considered as perceptual systems. Houghton Miffling, Boston
29. Bernsen NO (1993) Modality theory: supporting multimodal interface design. In: Proceedings from the ERCIM workshop on multimodal human–computer interaction, Nancy, pp. 13–23, November
30. Oosterhuis K (2002) Architecture goes wild. 010 Publishers, Rotterdam
31. Rosenbloom A (2004) Interactive immersion in 3D computer graphics. In: Communications of the ACM 47/8:28-31
32. Zellner P (1999) Hybrid space—new forms in digital architecture, Thames & Hudson, London
33. Jormakka K (2002) Flying dutchmen, motion in architecture. Birkhäuser
34. Buxton WAS (2002) Less is More (More or Less): Uncommon sense and the design of computers. In: Denning PJ (ed) The invisible future, McGraw-Hill, pp 145–179
35. Hubers JC (2005) Parametric design in protospace 1.1. In: Martens B, Brown A (eds) Learning from the pasta foundation for the future. Osterreichischer Kunst- und Kulturverlag, Vienna
36. Hubers JC (2005) Design environment protospace 1.1. In. Sariyildiz S, Tunçer B (eds) Innovation in architecture, engineering and construction. Delft university of technology chair technical design & informatics, The Netherlands
37. Rodet X, Lambert JP, Cahen R, Gaudy T, Gosselin F, Mobuchon P (2005) Study of haptic and visual interaction for sound and music control in the phase project. In: Proceedings of the international conference on new interfaces for musical expression (NIME) Vancouver, May 2005
38. Puckette MS, Apel T, Zicarelli DD (1998) Real-time audio analysis tools for Pd and MSP. In: Proceedings of the international computer music conference (ICMC)
39. Culler D, Estrin D, Srivastava M (2004) Overview of sensor networks. Guest editors introduction. IEEE Comput Spec Issue Sensor Netw 37/8:41–49
40. Olsen Jr, DR and Nielsen T (2001) Laser pointer interaction, In: ACM CHI Conference proceedings: human factors in computing systems. Seattle, pp. 17–22
41. Myers BA, Bhatnagar R, Nichols J, Peck CH, Kong D, Miller R, Long AC (2002) Interacting at a distance: measuring the performance of laser pointers and other devices. In: Proceedings of the CHI conference, April 2002
42. Yuan D and Manduchi R (2004) A tool for range sensing and environment discovery for the blind. In: Proceedings of the IEEE workshop on real-time 3-D sensors and their use
43. Bongers AJ and Veer GC van der (2006) Tactual articulatory feedback and gestural input. J Hum Comput Stud (submitted)
44. Bongers AJ (2002) Palpable pixels, a method for the development of virtual textures. In: Ballesteros S, Heller M (eds) Touch, blindness and neuroscience. UNED Ediciones, Madrid