A Situative Space Model for Mobile Mixed-Reality Computing

This article proposes a situative space model that links the physical and virtual realms and sets the stage for complex human-computer interaction defined by what a human agent can see, hear, and touch, at any given point in time.

Thomas Pederson
IT University of Copenhagen

Lars-Erik Janlert and Dipak Surie
Umeå University

As more computational devices are put into place, designing the user experience as an ensemble of interactions will become paramount. No single vendor should assume that their suite of devices and systems will become dominant. Consequently, it's important to begin planning and designing how real/virtual devices will cooperatively interoperate.

Such interoperation minimally requires communication protocol agreements and behavior standards for devices. But as a prerequisite for these, we need to begin understanding—as a community—what we want the overall experience of ubiquitously emplaced devices to be. We are designing more than just widgets; by creating many devices that communicate, we are creating entire environments.

The range of options is completely open to us: we can create a world where we live in a cacophony of many voices, each looking for a piece of user attention, or we can design a world of devices and systems that gently support our work and play, being continually responsive and sensitive to our preferred styles of interaction, announcement and intrusion.

—Dan Russell and Mark Weiser[1]

The design of mobile pervasive computing environments involves the adoption of fundamentally new computing hardware and software infrastructures, new communication models for human-computer interaction, and an expanded view of which resources the personal computing system should monitor and govern.

A fundamental idea in pervasive computing is that application logic (services, programs, and applications) should be device independent and roam between computing devices, enabling stationary devices to seamlessly take over from the slimmer mobile devices when more computing or interaction power is required.[2] Although much effort in the last decades has been devoted to the decoupling of computing hardware and software components,[3] considerably less attention has been given to the interaction design challenges presented by distributing personal computing in this way. One important issue is how to deal with the limited resource of human attention in mobile contexts. One option is to use the cumulated information from sensors and devices distributed in the environment to let the system draw conclusions and act autonomously without bothering the human agent. A complementary approach is to let the system choose a time and modality for interaction that minimize the overall attentional costs for the human agent. Unlike classical reactive HCI, such computing systems could initiate communication proactively, resulting in a mixed-initiative interaction dialog.[4] For this to work smoothly, the system must approach the level of human sensitivity in human-to-human communication and know when and how to pay attention and when and how to comment or raise an issue, all based on discreetly detectable cues such as the pose, body language, and actions performed by the dialog partner. To act (semi-)autonomously and engage in attention-economical mixed-initiative interaction, the system must be situation aware.

Mica Endsley developed her situation awareness model for describing the relationship between human agents and the environment when making decisions;[5] however, we find her distinction between different levels of abstraction (*situation awareness levels*) useful when discussing how computer systems could perceive and reason about the world. The situative space model (SSM) covers Endsley's level-1 situation awareness: *perception of the elements in the environment*. The model is unique in handling physical and virtual elements uniformly at this level of abstraction. We present results from bringing a personal computing system to level-2 situation awareness (*comprehension of the current situation*) by letting it analyze and compare sensor data classified according to the model over time. The model is used both for activity recognition and for situation-sensitive interaction device orchestration.

We also show how this model can be used to address a related problem in pervasive HCI—that of establishing and maintaining communication between the system and the human agent when the available means for input and output dynamically change and might include devices very different from traditional presentation and interaction devices. More specifically, we show how the model can be used to help computer systems choose the most effective communication channel (device and modality), thus paving the way for the creation of "plastic multimodal user interfaces."[6]

Getting Real-World Objects into the Loop

Human-computer interaction doesn't have to be explicit. It can occur in the background as a result of actions in the physical world or in any virtual environment accessed through devices. By monitoring the human agent's actions, useful implicit input can be generated at no extra cost to the human agent.[7] This possibility has inspired the development of context-aware systems and derivatives including the situation-aware systems targeted here. As the sensing of objects and events in the physical world improves in scope and detail, we foresee the use of ordinary physical objects for explicit input as well, allowing ad hoc commandeering of everyday physical objects for virtual functions, such as in graspable or tangible user interfaces.[8] That is a motive for incorporating not only dedicated interactive devices into the "device ensemble"[9] to be monitored but also smart objects with less computational ability, such as things labeled with passive RFID tags.

Pervasive personal computing systems need an interaction model different from the classical "window, icon, menu, pointing device" (WIMP) centered on applications offering tools for manipulating data files. Activity-based computing offers an interesting alternative by anchoring interaction to activities.[9] Ubiquitous instrumental interaction[10] breaks the concept of applications into toolboxes full of tools more freely applicable to domain objects than today's applications are, helped by *governor* system components that transform instrument manipulation of objects into something meaningful. Our model also breaks with application and device centricity at the interaction level. In addition it questions another long-standing habit in the design of interactive systems of treating physical objects as external to the computing system. Paul Milgram and Fumio Kishino define mixed reality as "anywhere between the extrema of the virtuality continuum" that extends from the completely real to the completely virtual environment.[11] Context-aware system design has gradually pushed computing models from the completely virtual extreme toward the center of this virtuality continuum by incorporating real-world phenomena. However, most context-awareness models are still biased toward interaction in the virtual world. Per their definition, real-world phenomena are treated as context to virtual-world processes. The models are based on what computer systems can sense today, and incrementally include more of the world as sensor technology improves. Our model has a different origin, starting from what humans can sense today (and tomorrow), taking human-centered computing[12] literally, with the intention of setting a solid long-term goal for engineering a useful mobile pervasive computing infrastructure.

Accordingly, we base our model on two deliberate conceptual departures from classical HCI to allow the shift from device-centric to body-centric modeling, and to place the modeling focus in the middle of Milgram and Kishino's virtuality continuum.[11]

First, we use virtual objects and mediators instead of interactive devices. According to activity theory, there are good reasons for viewing input and output devices as *mediators* through which virtual objects are accessed.[13] If mediators reach a sufficient level of perceptual and cognitive transparency—that is, they don't hinder or distract the human agent from manipulating and monitoring the virtual objects they make present—the mediators disappear and only the virtual objects remain. This is what happens when expert users use digital devices, and we base our modeling framework on this effect. Taking such a stance lets us model physical and virtual objects as if they were situated in the same space.

Second, we use action and perception instead of input and output. The classical HCI concepts of input and output don't easily lend themselves to the description of physical object manipulation. We suggest substituting the concepts of (device) "input" and "output" with (object) "manipulation" and "perception." We acknowledge the inherent differences between physical and virtual objects[14]: the intention isn't to make them resemble each other in practice, but handle them uniformly on a high level of abstraction to better model mixed-reality situations. Object manipulation and perception can take place in any modality, including tactile, visual, and aural. Both of these conceptual standpoints are necessary to enable our body-centric and mixed-reality-neutral SSM.

To avoid taking on too big of a challenge all at once, we focus our modeling efforts on the interaction between a single human agent and the surrounding mixed-reality environment. In principle, we see no problem in expanding the framework presented here to include fellow human agents (to cope with cosituated human agents and more or less cooperative human activities) and intend to do so in future work.

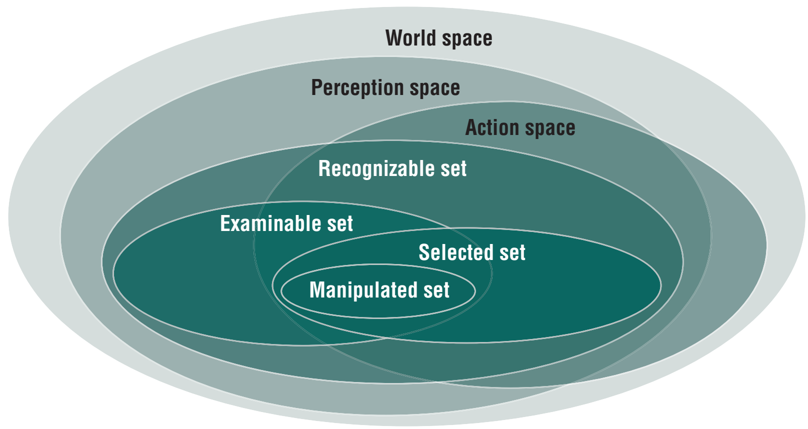

Figure 1. A situative space model (SSM). The spaces represent presence and approximate spatial relationship among physical and virtual objects with respect to what a specific human agent can perceive (*perception space*) and manipulate (*action space*) at a given moment in time. Whether objects are perceivable and manipulable depends on their relations to the human agent in all available interaction modalities (for example, vision, touch, and audio).

A Situative Space Model

The situative model is intended to capture what a specific human agent can perceive and not perceive, reach and not reach, at any given moment in time (see Figure 1). It is inspired by cognitive science theories related to context and situatedness.[15] This model is for the emerging "mobile-cum-pervasive" interaction paradigm (for the lack of a better label) what the virtual desktop is for the WIMP interaction paradigm: more or less everything of interest to a specific human agent is assumed to, and supposed to, happen here. Although spatial and topological relationships between objects within a particular space certainly are of interest, we have so far mainly taken into account whether an object is present in a space or set, or not. Applying the model in this simple way generates several objects for each space and set at any given time instant.

The following definitions are agent centered but not subjective. They're principally aimed at allowing objective determination and thus are suitable for automated tracking. The SSM might yet undergo some changes before it settles into its definitive form, and some of the definitions are more tentative than others.

The *world space* is a space containing the set of all physical and virtual objects that are part of a specific model.

The *perception space* is the part of the space around the agent that can be perceived at each moment. Like all the spaces and sets defined next, it's agent centered, varying continuously with the agent's movements of body and body parts. Perception space can be given a simple geometrical interpretation (like a cone, in the case of vision) as a rough approximation. Objects can occlude other objects and thus create (temporary) holes in the space.

Different senses have differently shaped perception spaces, with different operating requirements, range, and spatial and directional resolution with regard to the perceived sources of the sense data. Compare vision and hearing, for example. The perception space of vision requires light, is basically cone-shaped, with, in principle, infinite depth range if there are no obstructing objects, good angular resolution, and fairly good depth resolution at close range. The perception space of hearing requires air (or some similar medium), is basically ball-shaped, with quite limited range, good angular resolution for higher pitches, low resolution for low pitches, and rather poor depth resolution. You can't see what is behind your back, but you might hear it. On the other hand, many objects are silent (and, contrary to how vision works, offer little object-specific information by way of modulating sound from other sources at the scene) but can be seen. The different perception spaces of different senses complement each other.

The perception space in our definition can be interpreted as the complex superposition of the different senses' perception spaces, or it can be interpreted as dealing with each perception space in isolation. Here we want to specifically focus on visual perception space, and some of the definitions might need revision when considering other perception spaces.

Within perception space, an object might be too far away to recognize and identify. As the agent and the object come closer to each other (either by object movement, agent movement, or both) at some point, at some distance, the agent will be able to identify it as *X*, where *X* is a certain type of object or a unique individual.

An object can be of several different types—for example, at different levels of abstraction (my car, a Toyota, a car, a moving object, and so on)—but for a particular type *X* the distance at which it can be perceived as *X* can approximately be related to attributes of *X* such as size and presence of distinguishing perceptible features. For vision, viewing angle might also be important; many objects are difficult to recognize from certain angles. In the dynamics of a certain situation, the agent will be able to compensate by changing the viewing angle (by head movements, locomotion, waiting for the object to turn, or actively turning the object). Increasing the distance again, or changing the viewing angle, the perception of the object, although not sufficient to recognize the object *ab initio* as of type *X*, might still be able to serve as a token—a perceptual reminder of its type. For each type *X*, the predicate "perceptible-as-*X*" will cut out a sector of perception space, the distance to the farthest part of which will be called *recognition distance*.

A *recognizable set* is the set of objects currently within the perception space that are within their recognition distances. The kind of object types we're particularly interested in here are those that can be directly associated with the agent's activities—ongoing activities and activities potentially interesting to start up—what in folk-taxonomy studies is known as the basic level.[16] This is the level of a hierarchical taxonomy at which within-category similarities are maximized and between-category similarities are minimized. Objects belonging to the same category at the basic level are typically similar and distinctive in appearance and can also be associated with similar and distinctive motor operations. In a sense, the basic level represents the basic operative level of human activities. For example, when we think of activities involving hand tools, relevant basic-level object types include hammer, saw, and screwdriver—each type easily recognized from its distinctive visual appearance.

To perceive a designed object's status with regard to its operationally relevant (perceivable) states (operations and functions as defined by the artifact's designer), the object will often have to be closer to the agent than its recognition distance. The outer limit is the *examination distance*.

The *examinable set* is the set of objects currently within the perception space that are within their examination distances. Normally, the examinable set will be a proper subset of the recognizable set.

The *action space* is the part of the space around the agent that is currently accessible to the agent's physical actions. Objects within this space can be directly acted upon. The outer range limit is less dependent on object type than the perception space, recognizable set, and examinable set, and is basically determined by the agent's physical reach, but obviously depends qualitatively also on the action type and the physical properties of objects involved. (For example, an object might be too heavy to handle with outstretched arms.) Because many actions require perception to be efficient or even effective at all, the action space is qualitatively affected also by the perception space's current shape.

From the viewpoint of what at this stage can be relatively easily automatically tracked on a finer time scale, it will be useful to introduce a couple of narrowly focused and highly dynamic sets within the action space (real and mediated).

The *selected set* is the set of objects currently being physically or virtually handled (touched, gripped, or selected in the virtual sense) by the agent.

The *manipulated set* is the set of objects whose states (external and internal) are currently being changed by the agent. Normally, the manipulated set will be a subset of the selected set.

For all of these spaces and sets, geometrically defined sectors and object-type-dependent memberships are, in principal, computable. This computability, together with current sensor technology, make it possible to automatically track their contents without requiring an elaborate real-time model of the agent's cognitive processes. Clearly, that an object is known to be within the visual perception space, for example, is still no guarantee that it actually has been perceived or that it will be. All these spaces and sets, with the obvious exception of the selected and manipulated sets, primarily provide data on what is potentially involved in the agent's current activities. They're still quite useful in creating a first rough approximation of what is going on—good enough to make usable detections and predictions of ongoing and upcoming actions and activities, as we'll see in the prototyping experiments reported later.

Using the Situative Space Model

Like many interaction models, the SSM is intended to be used as a design tool to inform the system design, and to be implemented as a system component to improve and guide behavior in mixed-initiative dialog. We provide examples of both. Exactly how the model can be integrated in existing design processes remains an open issue, just as it is for many other post-WIMP interaction models that challenge concepts deeply embedded in classical HCI approaches (for example, an input/output device). With respect to implementation, the model will continue to be an engineering challenge for years to come, but this is intentional: the SSM provides a long-term sensing and modeling goal based on human characteristics, independent of current technology.

Representing Existing Mixed-Reality Situations

Say a human agent *O* sits down at the kitchen table to have breakfast. The kitchen table is fitted with a visual display in the center of the tabletop. The agent has a cellular phone in his pocket and a wireless headset on his right ear. A wall calendar two meters away has an embedded touch screen. Various software applications are running on a server ready to interact with *O* through these mediators. Figure 2 illustrates this scene with the mediators and a few objects highlighted. Figure 3 shows the SSM applied to the same situation.

Activity Recognition

To assess the SSM's utility before tackling the engineering challenge of prototyping the necessary sensor infrastructure, we simulated the capture of some of the model's spaces and sets using immersive virtual reality (VR), as Figure 4a illustrates. The assessment of multimodal interaction potential proved to be challenging in this setup, but we acquired interesting results with respect to activity recognition. The project aimed to develop a prototype for a wearable "cognitive prosthesis" to be carried by individuals suffering early dementia, which could help them successfully perform activities of daily living. Activity recognition naturally plays an important part in such an application as the system needs to monitor the carrier of the cognitive prosthesis in detail. Information captured in the simulated home environment consisted of the presence or absence of everyday objects in the perception space, the presence or absence of everyday objects in the action space, and events generated from object selection

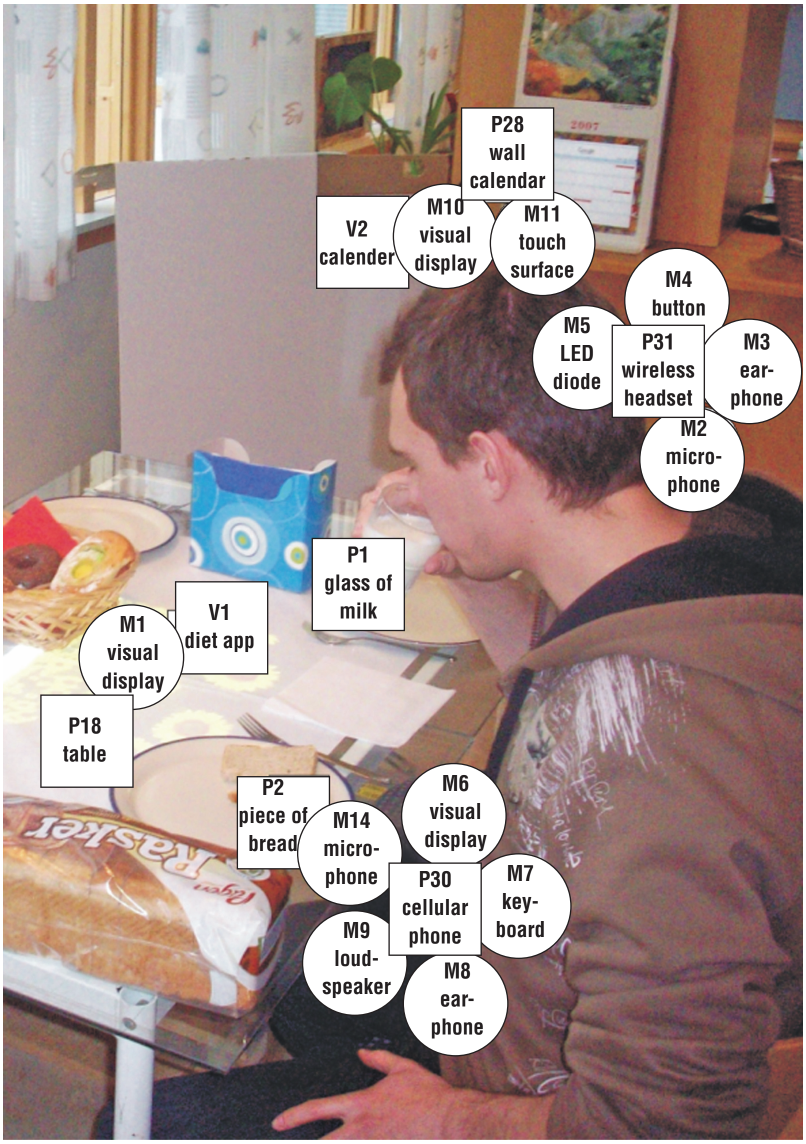

Figure 2. Human agent *O* having breakfast. The table (P18) has a visual display (M1) currently showing information from a diet application (V1). The wall calendar (P28) has a visual display (M10) and a touch-sensitive surface (M11) currently showing and providing access to a calendar application (V2). The wireless headset (P31) in *O*'s right ear contains a microphone, earphone, button, and LED (M2–M5). The cellular phone (P30) in *O*'s trouser pocket includes a visual display, keyboard, earphone, loudspeaker, and microphone (M6–M9, M14). All mediators offer means for explicit interaction between *O* and virtual objects. Also highlighted in the figure are a glass of milk (P1) currently being manipulated by *O*, and a piece of bread (P2). (Other physical objects lack labels to keep the figure simple.)

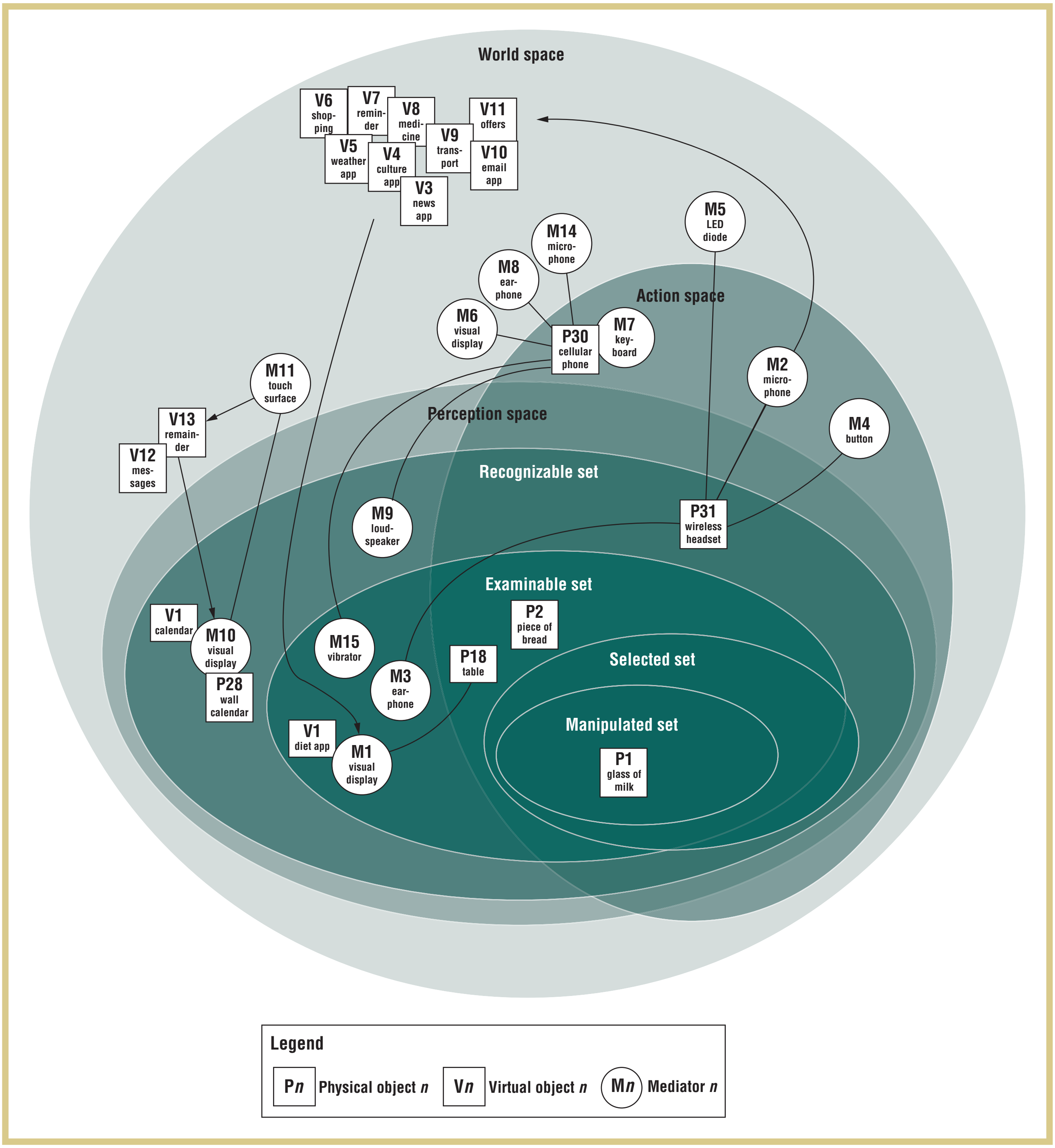

Figure 3. The breakfast scenario of Figure 2 as viewed through the situative space model (SSM). Some virtual objects (V3–V13) not visible in Figure 2 are shown here in the world space, ready to be made accessible to agent *O* through mediators in the perception and action spaces. Arrows indicate flows of interaction—specifically, manipulation of virtual objects and perception of the results. Lines without arrowheads indicate more static relationships among objects.

(specifically, object grab and object release). A learning and recognition design based on hidden Markov models, consisting of separate information channels for each of the two spaces and the object selection, resulted in an activity-recognition precision of 89 percent and recall accuracy of 98 percent across 10 household activities when all

(a) (b)

Figure 4. (a) The simulated home as viewed from the human agent wearing a head-mounted display and six degrees-of-freedom (DOF) gloves. (b) Illustrations of how the perception space (left) and action space (right) were approximated.[17]

Figure 5. The instrumented apartment with stationary mediators (green ovals) and wearable mediators (orange ovals), from left to right: display embedded in bathroom mirror; accelerometer embedded in wristband; wireless headset; display embedded in cutting board; wall-projected display; display embedded in bookshelf; display embedded in photo frame; display embedded in wall calendar; display embedded in tabletop.

three information channels were combined. When fed by the channels independently, the system achieved the best results when the perception space was used as source for identifying the ongoing activity (P 81, R 98) compared to action space (P 77, R 99) and object selection (P 79, R 95).[17] Figure 4b shows how the perception and action spaces were operationalized.

Interaction Management

We instrumented an apartment with a set of stationary and wearable mediators (see Figure 5) and a set of tracking systems for sensing some of the spaces and sets in the SSM and to experiment with multimodal interaction. To decide where, when, and how system-initiated communication between system and human agent should occur, our system's interaction manager component (corresponding to Eric Horvitz and his colleagues' "Notification Platform"[18]) considers several parameters completely or partially derived from a continuously updated representation of the SSM, such as the human agent's situation, activity, and focus of attention. In our system, however, decisions are grounded on a high-level model of human perception and action, not on statically defined device-centric parameters.[18] The interaction manager also considers other information, such as predefined object-specific properties of virtual objects (for example, privacy restrictions for public displays and modality rendering options), as well as the importance level specified by applications maintaining the virtual object to be rendered. The interaction manager

TABLE 1
Means by which the tracking systems maintain the situative space model (SSM) for a given human agent and enable multimodal interaction.

	Vision		Audio		Touch		Gesture	
	Physical objects	Virtual objects	Physical objects	Virtual objects	Physical objects	Virtual objects	Physical objects	Virtual objects
Perception space								
Perception space	proxyTrack	orientTrack + proxyTrack		proxyTrack				
Recognizable set	proxyTrack	orientTrack + proxyTrack		proxyTrack				
Examinable set	proxyTrack	orientTrack + proxyTrack		proxyTrack				
Action space								
Action space				proxyTrack/ speechTrack	proxyTrack/ idTrackChest	proxyTrack/ idTrackChest		gestureTrack
Selected set				speechTrack	stateTrack/ idTrackHand	stateTrack/ idTrackHand		gestureTrack
Manipulated set				speechTrack	stateTrack	stateTrack		gestureTrack

uses a priority queue, partially derived from preferences expressed by the human agent, to handle the traffic in case many software applications simultaneously call for the human agent's attention. More details on how the SSM provides essential information to our system's interaction management component are available elsewhere.[19]

Sensing the Spaces and Sets, and the Possibilities for Multimodal Interaction

As mentioned earlier, the accurate tracking of objects and mediators needed for real-time application of the SSM will probably remain a challenge for years to come. Here, we briefly present our efforts in acquiring the necessary real-world data for some of the SSM spaces and sets as well as information useful for multimodal interaction. The design has been guided by our experience in obtaining the data in the VR simulated environment presented earlier and by experimenting with available sensor technology. We list six tracking systems, in various stages of implementation and evaluation, and categorize them according to their roles in maintaining the SSM and the kind of modality they operate in (see Table 1).

ProxyTrack is a WLAN signal-strength-based proximity-tracking system for determining the set of physical objects and mediators around a human agent's body. It provides important information for determining the structure of both perception space and action space pictured in Figure 1. ProxyTrack uses an off-the-shelf wireless access point (WRT54GL) worn on the agent's chest, its antenna shielded with aluminum to make it directional, connected to a wearable computer (a notebook PC in the initial experiments). Tracked physical objects and mediators have embedded ASUS WL-167g WLAN adapters connected to thin-client boards for calculating WLAN signal strength (we used the Java Wireless API developed at Luleå University of Technology, Sweden) with reference to the wearable access point. Because indoor walls play an important role in determining the perception space's boundaries, they're covered with aluminum foil to dampen the signal from objects behind them.

For each object, we determined signal strength thresholds corresponding to the situative spaces from an empirical study conducted within the instrumented apartment shown in Figure 5: Two subjects positioned themselves in 32 different locations in the apartment. They spent 6 to 8 minutes at each location, constantly changing their proximity and orientation with reference to individual physical objects and mediators. Every 5 to 10 seconds, the subjects told the experimenters in what set or space specific physical objects or mediators were present, according to their own perception and action possibilities. We used this information to define the signal strength thresholds at the borders of the spaces and sets by identifying the time-stamped values recorded in logfiles generated by the WLAN adapters in the corresponding situations.

We then asked the same subjects to perform everyday activities in the instrumented apartment, reporting objects perceived as being inside the spaces and sets every 10 to 20 seconds. Using that information as ground truth, we ascertained the system's accuracy based on the signal-strength threshold values determined in the previous study, as Table 2 shows. The study's small scale

TABLE 2
Precision and recall values of the different spaces and sets within the situative space model (SSM).[20]

Space	True positives	False positives	False negatives	Precision (%)	Recall (%)
Perception space (visual domain)	743	3	16	99.6	97.9
Perception space (audio domain)	717	410	86	63.6	89.3
Recognizable set (visual domain)	450	1	31	99.8	93.6
Action space (touch domain)	164	0	134	100.0	55.0

doesn't allow us to draw any certain conclusions, but the results encourage further research in this direction.

The system performs well in determining perception space and the recognizable set in the visual domain (P 99.6 percent and P 99.8 percent, respectively), while performing significantly worse for the recognizable set in the audio domain (P 63.6 percent) due to background noise and the fact that subjects weren't very confident in distinguishing audio information within the situative spaces. For the action space, the precision reached 100 percent but with a recall value of 55 percent, meaning the system couldn't make a decision 45 percent of the time, which is unacceptably high and must be reduced in the future. Further information about proxyTrack is available elsewhere.[20]

OrientTrack (still under development) is an infrared signal-strength-based orientation-tracking system for determining the set of physical objects and mediators that are visually perceivable by a human agent based on their orientation relative to the agent's head. In other words, this tracking system is intended to help determine the perception space, recognizable set, and examinable set shown in Figure 1.

IdTrack (also under development) is intended to identify physical objects and mediators selected by the human agent (idTrackHand)—that is, members of the selected set pictured in Figure 1, as well as those in the immediate vicinity of the agent's body (idTrackChest) based on RFID readers worn on the agent's wrists and chest. IdTrackChest complements proxyTrack in determining the action space by sensing RFID-tagged objects in front of the human agent, whereas proxyTrack only senses objects fitted with WLAN transmitters.

StateTrack is a ZigBee-based wireless sensor network that helps determine the selected set and manipulated set (see Figure 1). Physical objects and mediators in an environment are equipped with simple state-change sensors (such as touch sensors, temperature sensors, and light sensors) that determine objects' state changes caused by a human agent's manipulation. Sensor motes were designed using a Maxstream XBee 802.15.4 transceiver and Atmel ATMEGA88-20PU microcontrollers. The system uses a star networking topology with sensor motes transmitting state/state-change information to the wearable computer.[21]

SpeechTrack is a speech-recognition system based on Microsoft Speech SDK 5.1 API for speech recognition and a BTH-8 Bluetooth microphone worn by the human agent or microphones embedded in the environment. The system determines the action space, selected set, and manipulated set (see Figure 1) in the audio modality of our user interface implementation by filtering out speech commands from the real-world sound stream, allowing for selection and manipulation of virtual objects. Evaluation results are omitted here due to space limitations.[19]

GestureTrack is a gesture-recognition system based on Phidgets 1059 three-axis accelerometers worn on the human agent's hands. Analogously to speechTrack, but in the gesture modality instead of audio, gestureTrack determines the action space, selected set, and manipulated set (see Figure 1) by filtering out explicit gesture commands from the flow of everyday physical actions. Currently, gestureTrack works with a predefined set of gestures, but we intend to include user-defined gestures in the future. Evaluation results are omitted due to space limitations.[19]

Perception Space

Our work on determining the perception space has so far focused on the visual modality, with the exception of audio as a modality for perceiving the presence and state of virtual objects, hence the empty cells in the touch × perception area of Table 1.

Because proxyTrack senses the presence of both physical objects and mediators, it provides information about the perceptibility of physical as well as virtual objects in the visual domain. Determining the perceptibility of virtual objects requires more information. A mediator's rotation (critical for 2D displays) and proximity relative to the human agent's chest is encoded in the WLAN signal strength as measured by the mediator and filtered using a signal-strength threshold determined empirically for each visual display. A virtual object's presence in perception space is then determined by its visual perception threshold value, its size on the display, and the size of the display itself—all known to the system. The mediator-orientation information provided by proxyTrack is to be complemented with more accurate information provided by orientTrack.

ProxyTrack also plays an important role in determining the perception space in the audio modality. Our current implementation considers the relative proximity of loudspeakers (determined by proxyTrack), the sound volume they can generate, and the contextual noise level, to determine whether a virtual object represented in the audio modality is within the perception space. The aurally recognizable and examinable sets are formed by applying empirically determined proximity thresholds.

Action Space

For the action space, our choice of gesture and speech as input modalities in our interaction infrastructure gives the speechTrack and gestureTrack systems important roles in enabling the manipulation of virtual objects.

The speech action space outlines the range within which objects can be manipulated through speech. The presence of a suitable mediator (microphone) together with the state of the interactive system (whether it's in speech input mode) can be used to determine the presence of virtual objects in the speech action space. The gestural action space and its subsets are by and large defined analogously to the speech action space.

The touch action space outlines the range within which physical and virtual objects can be manipulated using touch. ProxyTrack is currently used to approximate object tracking in this space, soon to be complemented by idTrackChest. For the selected set, stateTrack is currently used but is limited to objects capable of communicating state changes to the interactive system running on the wearable computer. To include less-advanced physical objects, we'll introduce idTrackHand in the near future. IdTrackHand will allow any object tagged with a passive RFID tag to be accurately included in the selected set once grabbed by the human agent. Determining whether the selected object has actually been manipulated and has changed its internal state, however, still requires stateTrack. The process for including virtual objects in the touch action space follows the same path as for physical objects, but they depend on selection and manipulation of a mediator first, associating it with a virtual object. We're only in the initial phases of exploring the structure of this interaction design space, which includes the use of any tagged everyday object as mediator or physical token in the manner of tangible user interfaces, provided by the combination of gestureTrack and idTrackHand.

Because the SSM is body-centric instead of device-centric, it reveals how emerging wearable interaction technology such as gesture, posture, gaze, and brain–computer interfaces can be useful in pervasive computing settings that also include more traditional computing devices and everyday physical objects. Apart from guiding prototype design using current sensor technology (of which our own efforts have been briefly reported here), the SSM also suggests a direction to a research community often accused of lacking one. Fundamental human perception and action capabilities, which form the basis of the SSM, do not change.

Our immediate future work is to finish implementing and fine-tuning the idTrack and orientTrack systems, which include modeling multiple human agents. We will also investigate how we can incorporate more details about object manipulation and object state changes into the model, and how the perception and action spaces translate into the virtual environments offered by mediators. Finally, we plan to develop more advanced heuristics for multimodal interaction management based on the model.

REFERENCES

1. D. Russell and M. Weiser, "The Future of Integrated Design of Ubiquitous Computing in Combined Real and Virtual Worlds," *Proc. ACM Human Factors in Computing Systems* (CHI 98), ACM Press, 1998, pp. 275–276.
2. M. Satyanarayanan, "Pervasive Computing: Vision and Challenges," *IEEE Personal Comm.*, vol. 8, no. 4, 2001, pp. 10–17.
3. B.N. Schilit and U. Sengupta, "Device Ensembles," *Computer*, vol. 37, no. 12, 2004, pp. 56–64.
4. E. Horvitz, "Principles of Mixed-Initiative User Interfaces," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems* (CHI 99), ACM Press, 1999, pp. 159–166.
5. M.R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors*, vol. 37, no. 1, 1995, pp. 32–64.
6. J. Vanderdonckt et al., "Multimodality for Plastic User Interfaces: Models, Methods, and Principles," *Multimodal User Interfaces: Signals and Communication Technology*, D. Tzovaras, ed., Springer, 2008, pp. 61–84, doi:10.1007/978-3-540-78345-98.
7. A. Schmidt, "Implicit Human Computer Interaction Through Context," *Personal Technologies*, vol. 4, nos. 2–3, 2000, pp. 191–199.
8. G. Fitzmaurice, H. Ishii, and W. Buxton, "Bricks: Laying the Foundations for Graspable User Interfaces," *Proc. ACM Conf. Human Factors in Computing Systems* (CHI 95), ACM Press, 1995, pp. 432–449.
9. J.E. Bardram, "Activity-Based Computing for Medical Work in Hospitals," *ACM Trans. Computer-Human Interaction*, vol. 16, no. 2, 2009, pp. 1–36, doi: 10.1145/1534903.1534907.
10. C.N. Klokmose and M. Beaudouin-Lafon, "VIGO: Instrumental Interaction in Multi-surface Environments," *Proc. 27th Int'l Conf. Human Factors in Computing Systems* (CHI 09), ACM Press, 2009, pp. 869–878.

the AUTHORS

Thomas Pederson is an associate professor in the Innovative Communication Group at the IT University of Copenhagen. His research interests include mobile mixed-reality interaction, tangible user interfaces, and pervasive computing. Pederson has a PhD in computing science from Umeå University. He's a member of IEEE and the ACM. Contact him at tped@itu.dk.

Lars-Erik Janlert is a professor of computing and cognitive science at Umeå University. His research ranges from knowledge representation, information space theory, and philosophy of information to interaction design, design theory, and the use and impact of information technology in a wider perspective. Janlert has a PhD in computing science from Umeå University. He's a member of the ACM. Contact him at lej@cs.umu.se.

Dipak Surie is a PhD student in the Department of Computing Science at Umeå University. His research interests include ambient intelligence, smart environments, and context-aware computing. Surie has an MSc in computing science from Umeå University. Contact him at dipak@cs.umu.se.

11. P. Milgram and A.F. Kishino, "Taxonomy of Mixed Reality Visual Displays," *IEICE Trans. Information and Systems*, vol. E77-D, no. 12, 1994, pp. 1321–1329.
12. A. Jaimes et al., "Human-Centered Computing: Toward a Human Revolution," *Computer*, vol. 40, no. 6, 2007, pp. 30–34.
13. V. Kaptelinin, "Computer-Mediated Activity: Functional Organs in Social and Developmental Contexts," *Context and Consciousness*, B. Nardi, ed., MIT Press, 1996, pp. 45–68.
14. T. Pederson, *From Conceptual Links to Causal Relations—Physical-Virtual Artefacts in Mixed-Reality Space*, PhD thesis, Dept. of Computing Science, Umeå Univ., report UMINF-03.14, 2003.
15. B. Nardi, "Studying Context: A Comparison of Activity Theory, Situated Action Models, and Distributed Cognition," *Context and Consciousness*, B. Nardi, ed., MIT Press, 1996, pp. 69–102.
16. E. Rosch, "Principles of Categorization," *Cognition and Categorization*, E. Rosch and B.B. Lloyd, eds., Erlbaum, 1978, pp. 27–48.
17. D. Surie et al., "Activity Recognition Using an Egocentric Perspective of Everyday Objects," *Proc. IFIP Conf. Ubiquitous and Intelligent Computing*, LNCS 4611, Springer, 2007, pp. 246–257.
18. E. Horvitz et al., "Models of Attention in Computing and Communication: From Principles to Applications," *Comm. ACM*, vol. 46, no. 3, 2003, pp. 52–59.
19. D. Surie et al., "Exploring Egocentric Interaction Paradigm for Ambient Ecologies," *J. Pervasive and Mobile Computing*, special issue on ambient ecologies, 2011, submitted.
20. D. Surie et al., "Situative Space Tracking within Smart Environments," *Proc. 6th Int'l Conf. Intelligent Environments*, IEEE Computer Society, 2010, pp. 152–157.
21. D. Surie, O. Laguionie, and T. Pederson, "Wireless Sensor Networking of Everyday Objects in a Smart Home Environment," *Proc. 4th Int'l Conf. Intelligent Sensors, Sensor Networks and Information Processing* (ISSNIP 08), IEEE Press, 2008, pp. 189–194.