Leveraging spatial memory for shortcuts through mid-air deictic pointing using Microsoft Kinect

Yoann Bourse
ENS, UPMC
Yoann.Bourse@ENS.fr

Summer internship in the VIA laboratory at Telecom Paristech directed by Eric Lecolinet

Abstract

The rise of computer-mediated living and smart televisions keeps adding numerous functions to the home media center. Efficient access and memorization of a wide number of function is therefore required. We leverage spatial memory to provide interactions enabling fast memorization of a big number of items.

We introduce two shortcut management systems designed to enable microinteraction in a couch-interaction setting. The first one is an adaptation of Marking Menus to in air directional interaction. The second one is a novel interaction relying on deictic pointing in which users assign the functions to objects in their environment, following a personal more or less symbolic mapping.

We first analyze the precision of our system based on Microsoft Kinect depth camera, and then study the memorization capabilities offered by those interactions. Our techniques push the limit of memorized items to 22 on average for only 3 presentation per item.

Contents

1 Introduction

Computer systems are spreading and covering more and more aspects of our daily lives. Smart televisions gained new functions through internet connection, while smartphones and tablets offer us more and more applications. Studies conducted by the Nielsen institute show that the number of application per phone has been increasing ever since the birth of the smartphone, to end up at 48 for the iPhone or 35 for Android phones [10].

Meanwhile, home automation keeps developing increasingly fast, and it is commonly agreed that the home media center is becoming the hub for computer-mediated living [5]. Its basic multimedia functions are growing in number and diversity: in addition to the numerous TV channels, it now handles pictures, music or video on demand. As technology progresses, it gains more and more functions, such as controlling the temperature, lights, locks or shutters of the house. Moreover, connected televisions offer on a new device the aforementioned multiple functions of smartphones and tablets: internet navigation, social functions like messaging or social networks [6], games, internet applications (Amazon, Youtube, sports, magazines, radios...).

The media centers will regroup all those functions, offering a tremendous amount of applications. It is therefore important to provide users with an easy access to as many of them as possible. Moreover, several of those functions will be repeated a lot, especially in regards of home automation (lights for instance). Hence the need for fast shortcuts that can be easily memorized.

It is all the more true as studies show that users currently don't use a lot of their tablet or smartphone applications frequently [26]. Although it is generally assumed that the cause of this is a characteristic user behaviour, we claim that the interaction technique is also partially responsible. That is to say users only use frequently a few functions of their devices because memorizing and accessing those functions is a hard task. Easing the learning process through better-designed interaction techniques may result in an increase of the number of functions used frequently. New interaction techniques are thus required to enable users to make the most of their digital experience and access more functions more frequently.

We answer this problem by proposing interaction techniques for media centers enabling fast memorization of a huge number of shortcuts by leveraging spatial memory.

1.1 Objective

We focus here on shortcut management. That is to say bypassing part of the tree structure representing the the system by allowing direct access to a function (turn on the lights, launch this channel...) or a subtree or folder (browse this gallery, access the comedy movie catalogue...).

This direct access is based on instant retrieval. It must not be confused with navigation within the system, which relies on continuous long interaction. On the contrary, shortcut retrievals are fast, rare and sporadic. This allows us to use in-air interaction without worrying about the physical fatigue which usually plagues a lot of in-air techniques.

As a consequence, we must provide fast micro-interactions. We also need the memorization process itself to be quick and easy. To that end, the transition between novice and expert mode must be seamless and well-designed in order to enhance the learning process. Finally, our objective is also to maximize the number of items memorized, and to enable a robust memorization (less recall errors). This will allow a huge interactional bandwidth at low cost for the user.

Inspired by well-known works in cognitive science and techniques such as the method of loci, which has been used through history to learn huge number of items by leveraging spatial memory [33], we propose a solution harnessing the power of spatial cognition and proprioception which consist of creating mappings between the shortcuts and either positions or directions in the user's environment.

1.2 Key points

The context of our work is the media-center of the home, that is to say, we focus on couch interaction. It is important for our techniques to be usable and pleasant in relaxed positions. We focus on in-air interaction because it seems practical and convenient for the user: no additional devices are required (like a remote controller which can be lost), allowing for direct interaction and multiple users.

Moreover, this means that the setting of our system is the living room, which is rich in visual cues (furniture, decoration). Those structure the space and provide familiar targets to use for memorization. The familiarity with the real case environment of our system is expected to boost the performances compared to our test in laboratories.

As input for the system, we are going to need at least the position of the user. We decided to retrieve it using a depth camera, and more precisely Microsoft Kinect

[18], because it seems likely that most households will be equipped with such a system in the future. In fact, a large number of them already are, making our work not only realistic but totally already applicable.

Moreover, this depth camera equips XBox, which is one of the most spread-out media center nowadays. The fact that this console is used more and more as a media center than for gaming [31] corresponds to an evolution of lifestyle which resonates with our work.

However, this choice comes with some constraints. First of all, Kinect, like any depth camera, is oblivious of the environment which is not in its field of view. Although it is possible to obtain a model of the environment by asking the user for a certain calibration, like sweeping the room with the camera [17], it is complicated and invasive for the user. We decided to tackle instead the interesting problem of reasoning in an inferred (mostly unknown) environment.

Kinect is also well known for its lack of precision. The price accessibility for the general public comes at the cost of low performances as a depth camera. This raises the question of maximizing the interactional bandwith with an imprecise and poor input.

1.3 Conceptual implications

On a broader perspective, our whole system relies on mapping between different "spaces". The heart of the memorization mecanism is the abstract mapping between the symbolic space inside the user's mind over which we have neither control nor knowledge towards the space corresponding to their perception of their environment. By their actions, users will create a link between their perceived space and the 3D real space around them. The last mapping happens between this real space and the restricted space that the depth camera perceives.

Each of those mappings is in fact a projection between two spaces, coming with a severe loss of information. Our challenge comes down to inferring the original information (in the user's mind) through a very imprecise, noisy and highly fractional projection of it.

The aforementioned symbolic space is characteristic of each person, and doesn't have a euclidean structure. It can be structured by semantic categories, but varies among users. Taking into account this diversity is one of the main guidelines of our work. We let users chose themselves the mappings between shortcuts and real world position or direction. It has been shown in the cognitive science litterature that choice-based processing enhances memory. Cloutier and Neil Macrae offer a well thought litterature review on the subject [7]. Respecting the particularities of each user's own representations should therefore result in a boost in memorization, as they get to pick and constitute themselves the abstract mappings binding the real world to the shortcuts, instead of learning an arbitrary one which can make no sense to them. This also accounts for the need of the system to be heavily customizable, as every user will have personal needs within the tremendous number of functions offered by the system.

From our technological choices rise challenging research questions that our system tackles.

2 Related work

2.1 Depth cameras

The release of Microsoft Kinect [18] as a low cost depth camera has created a rise in interest for depth cameras. Kinect relies on light coding technology, that is to say it projects an infrared pattern and deduce from its deformation a depth image of the scene. It is not to be confused with the time of flight technology relying on the measure of the travel time (through the phase shift) of an unseeable wave. The latter category is generaly more precise, but also more expensive [32].

Kinect offers audio channels (16 bits, 16kHz) unused in our work, as well as a 640x480 32 bit RGB camera and a 320x240 16 bit depth map (both 30 fps). Despite these poor specification, studies show that the camera is still relatively precise for its cost [14]. Using 7 Vicon camera as ground truth, Dutta measured the RMSE of Kinect detection in all directions to be 6.5cm on the left-right axis, 5.7cm on top-down and 10.9cm on depth. It is interesting to note that all those measures increase for big values of depth, particularly above 3 meters away from the camera. The measured field of view is 58.6 x 43.6, seeing from 0.47m to 3.6m away. A new version of Kinect called Kinect for Windows improves the low-range detection thanks to a "near mode".

Various solutions can be used to manipulate Kinect, the most spread out languages being C++, Java and C#. Microsoft provides an official Kinect SDK, but it currently only handles standing skeletons. In the context of couch interaction, we decided to use the open source OpenNI drivers and the NITE middleware which allow partial (sitting) skeleton tracking. Other options we chose not to use are the low level drivers LibFreeNet, and OpenCV, usually used for face detection or other vision algorithm. We used OpenGL (and C++) for the visual rendering of our program.

Numerous projects arise from the accessibility of this depth camera, mostly focusing on 3D in air gestures.

Large display distant manipulation is also an important field of use of depth camera. Surprisingly few studies of pointing have been made, the most notable one being the report of Daria Nitescu [22].

2.2 Deictic pointing

Pointing is rich field at the border between HCI and cognitive psychology. Delamare et al. [12] highlight the benefits of this technique in regards of computer-mediated living by focusing on the disambiguation of the targeted item in a home setting. As Kinect doesn't allow for such a precise selection, we will settle for a more straightforward and intuitive directional pointing.

However, this problem is richer than it seems, in particular in a 3D space. Cockburn et al. [9] compare different pointing techniques: selecting with a laser pointer in the hand, projecting the hand on a virtual 2D plan and use it like a mouse, or use the hand as a cursor in the 3D space (slow and inaccurate).

But the most intuitive type of pointing, used in daily life to show things to others, is none of the above. This natural pointing corresponds to a "What you point at is what you get" paradigm [24] that we want to follow. Nickel and Stielfelhagen [21] show that head-hand direction has a better precision to estimate the pointing direction than head orientation, finger direction, forearm orientation or shoulder-hand direction. They also come up with a hybrid HMM model taking those measures as input and outperforming them. But the best estimate of the natural pointing direction is learned by gaussian process regression [13]. However, considering the low precision of Kinect and the small performance differences between those methods, head-hand direction is a good enough estimate for the pointing direction in our system. More sophisticated methods would be wasted on such a low quality input.

Most pointing studies use hand-held devices and focus on user-centered frame of reference. It is the case of Virtual Shelves [20], a project studying the accuracy of pointing in various directions of space. The conclusion is that humans are significantly more precise in the vertical plan right in front of them (zero longitude), and that the targets below horizontal plane (negative latitude) are harder to reach.

However, we want to focus on an environmental frame of reference, in order for several users to manipulate our system. Moreover, we do not want any additional device, which raise the issue of in-air clicking. The in-air selection delimitor could be an action with the not-pointing hand, remaining in place for a given amount of time, a movement of the pointing hand, a voice command or a hand gesture. Daria Nitescu concludes that a relative movement of the pointing hand towards its target has the best index performance relative to Fitt's law [22], however, moving the hand might cause a loss of precision in the pointing direction. The work of Raheja et al. [27] gives us hope to circumvent the low precision of Kinect and to use the closing of the hand as a selection delimitor. This is the delimitor we chose as it seems discriminative enough not to cause any false positive in real context use.

2.3 Spatial cognition

Spatial cognition and HCI

Our goal is to use in-air pointing to leverage spatial memory, which has been known to play a major role in performance in user interfaces. Psychology literature on the benefits of spatial representation for learning is numerous [2]. Spatial learning is known to happen even without focused attention [1], and to strongly correlate with efficiency in computer-system manipulation. Eagan and Gomez [15] are one of the earliest such examples and show that spatial aptitudes are crucial to the manipulation of a software, here a document editor. Many more studies highlight the importance of spatial cognition in HCI performances.

Transparent novice-expert transition

Amongst the many benefits of spatial cognition, we find an intuitive, fast, transparent beginner to expert transition. Providing a fast interaction for experts is an important key to any human-computer interface, and easing the learning of this expert mode is desirable. The most notable technique emerging from these needs are the Marking Menus [19], multi-layered circular menus relying on a optional visual feedback for novices. By performing the same gesture for a given command, the transition to expert mode is smooth and transparent. The learning is implicit with repetition. We aim at putting such an emphasis on the novice to expert transition in our system.

Although many techniques derived from the Marking Menus exist, few attempts have been made at translating these Marking Menus to in-air interaction. Several of them use additional devices such as phones or wiimote [23]. Bailly et al. [3] achieve a good accuracy despite a relatively long manipulation time. Unfortunately, they don't offer an in-depth study of the memorization process. We focus on our work on the multi-stroke menus proposed by Zhao and Balakrishnan [34] in order to benefit from its accuracy to counterbalance the poor accuracy of Kinect.

As we focus on shortcut management, it seems important to focus our efforts on the learning process. Oc-

topocus [4] offers a perfect example of leveraging spatial memory to improve the memorization of expert mode commands in Marking Menus. We however wish to rely less on visual feedback, in order to give priority to the real-world environment.

Another work exploiting this aspect of spatial cognition is the CommandMaps [29] which leverages the already existing spatial knowledge of Microsoft Office's ribbons in order to provide a faster interaction means than ribbons and linear menus. Interestingly enough, they point out that resizing the window has a negative impact on spatial cognition, which we do not suffer from as one cannot resize their environment.

Huge memory capacity

Another crucial benefit we want to exploit is the huge capacity of spatial memory. Yates describe in The Art of Memory [33] mnemonic techniques, such as the method of loci, used through history. These methods harness the power of spatial cognition in order to memorize a huge number of items. They were used in particular before printing to learn important amount of data. For instance, ancient Greeks and Romans memorized law texts by associating each one of them to a stone in the layout of a familiar building. Such methods have been praised for their efficiency and widely studied by cognitive scientists.

However, only few interaction techniques attempt to leverage spatial cognition to provide the user with a lot of easily memorizable items. The most notable example of this is the Data Mountain, developed by Microsoft Research [28]. The subjects had to organize and then retrieve 100 Internet Explorer favorites, using both the classical hierarchical menus or a 3D plane on which they put thumbnails of the webpage. Spatial memory allowed for faster selection with less errors and failures in the latter system. Moreover, they also highlighted the durability of spatial memory, as the subjects came back 4 months later and showed no significant loss of performances [11].

The question of the benefit of spatial representation for memory has been studied by Cockburn and McKenzie [8] on one hand, who claim to see no significant improvement of memorization using a spatial 2D tree layout ; and Tavanti and Lind on the other hand [30] whose 2D isometric layout brought better performances. Out of their contributions, we can note the importance of the display layout. Letting the user organize their favorites theirselves probably played a great role in the Data Mountain success. We aim at benefiting from the same effects in our work.

However, all those work are limited to a 3D representation, using a 2D classical mouse interaction, creating a gap between manipulation and representation. To the best of our knowledge, no work attempted to leverage 3D space memory using the new 3D interaction means. It is our guess that the direct correspondence between interaction and representation will enhance the benefits of spatial cognition. Moreover, such an interaction relying on pointing can also benefit from the proprioceptive memory extending the benefits of spatialization [9].

Spatial mappings

Finally, spatial cognition might help drawing mappings between the real world environment whose knowledge we want to use and our virtual functions. Gustafson et al. [16] have obtained very good performances using this method by allowing users to press imaginary buttons on their hand as if it were their phone's homescreen. In the same way, our system can be seen as an imaginary interface (a shortcut map) superposed to the environment.

Our methods rely on abstract mappings between the user's environment and the functions of our system. Although symbolic mappings can seem like unreliable links, it has been shown that they perform very well, almost as good as straightforward semantic mappings [25]. It will be all the more so as the user will get to create those mappings themselve, since agentic and choice-based processing is known to enhance memory [7]. Thereby, we hope to achieve a fast learning of a big number of commands [28].

3 Pointing capabilities

The first part of our work consisted in designing a system which could infer the environment based only on Kinect's view of the world. It was important for us to provide a system working in the environment reference frame, and not the user, in order to allow multiple users and a use from any point of the environment. That is to say, we needed to know that a given object was pointed, not that the user was pointing towards left.

3.1 Pointing system

To answer this problem, we propose two paradigms to infer the room-based frame of reference from the content of Kinect's limited field of view. Both rely on considering a target point instead of the simple $[head, hand)$ pointing direction.

- In the **sphere** paradigm, we project the pointing ray on a virtual sphere around the field of view of Kinect. This sphere is centered at the middle of the field of view, and is designed to encompass the whole room (4 meter diameter). Being an abstract approximation,

this model is expected to have poor accuracy but to be easily transferable.

- In the **room** paradigm, we provide our system with a rough cuboid model of the room, with a preliminary calibration for instance. We then consider the point at the intersection of a face and the pointing ray. Note that this method would be very sensible to any movement of the Kinect and to the calibration process.

In order to be independant from the environment and the chosen paradigm, the aimed point is represented in our system and in this paper by the spherical coordinates (latitude θ, longitude ϕ) of its direction relative to the center of the camera's field of view.

3.2 Calibration

The aforementioned room paradigm requires the room dimensions and Kinect position in it. We tested a calibration mecanism based on asking the user to point the same point from two different positions. The estimation of the intersection of the pointing directions would represent the aimed point in our system. We could thereby obtain the position of the room's corners.

This calibration process was tested to estimate 6 corners of a room. Each corner was estimated 10 times. Our results were rather poor, obtaining an average distance between the estimated point and the measured ground truth of 3 meters. The standard deviation between the estimations for a same input point was 1.5 meters. It is noteworthy that augmenting the distance between the two pointing positions did not necessarily have a positive effect on the estimation.

Those results highlight the difficulty to easily obtain a decent model of the room. More complicated calibration process could be designed, relying on more pointing rays to improve the intersection estimate. A more accurate and less invasive calibration process could rely on continuous movement of the user. In the following experiment, we used a manually-inputted room model in order to evaluate if designing a precise calibration process was worth it. Therefore, the following evaluation does not suffer from any calibration bias.

3.3 Evaluation

Aware of the poor performances of Kinect, we proceeded to a technical evaluation of our system, in order to estimate its capabilities. In a testing room (6m ; 4.5m ; 2.7m), markers have been placed at 62 points corresponding to all possible latitude and longitude around the center of the camera's field of view considered with a $\frac{\pi}{6}$ step. A user then had to point at these markers and validate the pointing by clicking. We considered two positions for the user: in the middle of the camera's field of view and one big step (85cm) behind on the right. The user uses whichever hand is more convenient and practical so as not to create a biological bias.

Figure 1: Our manually-positioned markers representing the ground truth

We measure for each point the aimed point spherical coordinates θ (longitude) and ϕ (latitude), and their average deviation θ_d and ϕ_d. We summarize the measures in an average deviation score $d = \sqrt{\theta_d^2 + \phi_d^2}$ for clarity. An average of 30 measures was taken by point.

As we did not use high precision equipment to position the markers, we are fully aware that our manually-positioned markers are very imprecise (see Figure 1). Therefore, it comes as no surprise that the average deviation to the ground truth is way higher than the standard deviation (table 1) which can be seen as the deviation from the "average point" of our measures, the point which would represent the aimed point inside our system. Therefore, as the markes will not exist in real applications, this problem is not serious and we will consider in the following the standard deviation.

3.4 Observations

We study the influence of various parameters, that is to say the differences between latitude and longitude, the spatial variations, the influence of the user position and the model paradigm used. As a whole, we end up with very satisfactory results, and a precision which would enable any system based on these pointing techniques to discriminate between hundreds of positions (table 1)

Deviation	Sphere paradigm	Room paradigm
To marker	0.213 (43.3)	0.223 (45.4)
Standard	0.056 (11.2)	0.061 (12.2)

Table 1: Average deviation d (rad) and the corresponding size (cm) on a 2m away wall.

Deviation	Sphere paradigm	Room paradigm
To marker	0.317 (65.6)	0.265 (54.3)
Standard	0.047 (9.45)	0.040 (8.00)

Table 2: Same measures when the user is not centered.

3.4.1 Latitude versus longitude

Latitude and longitude being directly computed from the position of the head and the hand, their deviation has similar variation among space (fig 2). However, we observe a globaly better accuracy on ϕ (std = 0.031 rad) than on θ (std = 0.049 rad), because this latter heavily relies on the most imprecise coordinate of the Kinect z (close-far).

3.4.2 Spatial variations

Precision is globaly very good and spatially uniform (fig 2). Problems arise when the hand occlude the head or vice-versa (eclipse phenomenon). We notice therefore a very low precision behind the user ($\theta = \pi$) where the camera simply cannot see the arm. The same also stand at the very precise point of the camera $(0, 0)$, where the hand occludes the head, but considering how small this point is the inference of the direction from the previous positions is rather good. We observe a loss of precision for the more extreme values of ϕ: the points are more cluttered, and a small variation of the cartesian coordinates of the body translates into a huge variation of angles. There is also a very slight increase of the deviation around the sides ($\theta = \pm\pi/2$) as the angular resolution of the arcsine function decreases around these values.

3.4.3 User position

We compare the condition "centered" in which the user stands at the center of Kinect's field of view, and a condition "moved". In the latter, the user was asked to take a step on his right and behind, arriving at a point 85cm away from the center of the field of view of Kinect, without changing the markers. The average deviation (table 2) to the ground truth is higher than the centered condition (table 1) for both the sphere and the room paradigm, however the standard deviation remains very good. The

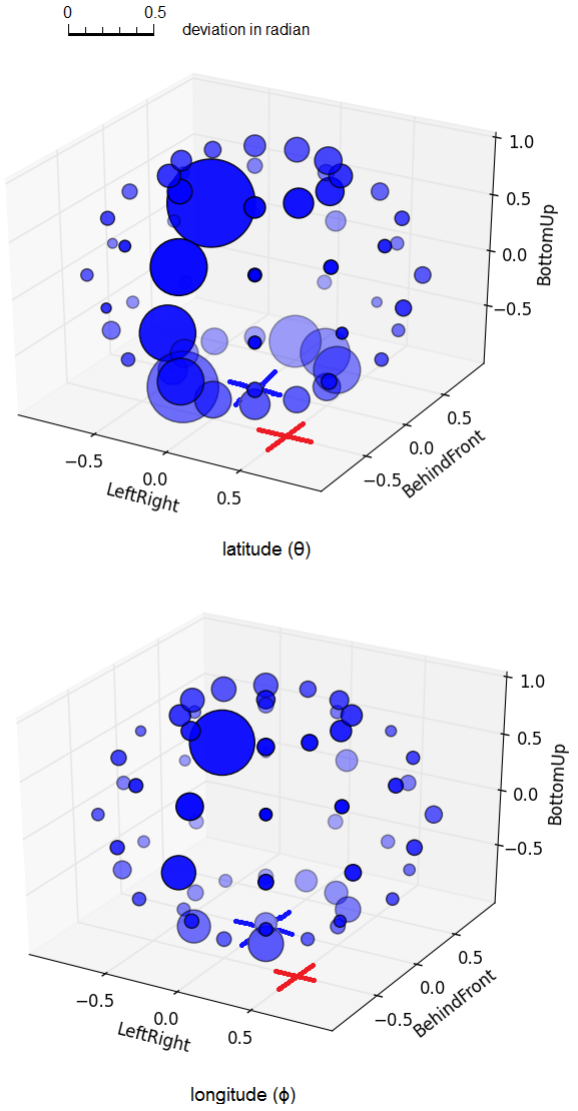

Figure 2: Spatial variation of the standard deviation of latitude and longitude when the user is at the center of the sphere/room (blue cross). The red cross correspond to position where the user will be in the "not-centered" condition.

points are therefore well determined by the system but differ even more from the marker's ground truth. This could be explained by the fact that from the new position, the markers are perceived as further away or more cluttered, which increases the impact of the camera's imprecision. On the whole, both our paradigms are therefore relatively robust to change of the user position and manage to detect fairly well the point aimed at in the environment regardless of the user position.

3.4.4 Modelization paradigm

The sphere and room paradigm end up having very similar variations. In particular, the sphere paradigm is surprisingly robust to the change of position (table 2) considering that this case should have seen a huge decrease of precision since for the same marker, the aimed point in the system is, strictly speaking, different. However, it seems that the two aimed points are close enough, because the drop in precision measured when the user is not at the center is relatively similar to the one of the room paradigm. The loss in precision due to the poor accuracy of the camera and the movement of the markers relative to the user outweight the loss of precision due to the sphere abstraction.

Moreover, we noticed during experimentation that the room paradigm was very sensible to the calibration and the orientation of Kinect: the slightest move of the camera will have a dramatic effect on the position of the "virtual" room. It stands also for the Sphere paradigm, but qualitative observation show that it is more robust to these kind of change, because the sphere, contrary to the cuboid, does not show any discontinuity. That brings us to the conclusion that our sphere paradigm brings fairly decent results and provides a robustness and an ease of use missing in the room paradigm. It doesn't require calibrating and allows user to move and point from different positions without too much loss of precision. Therefore, our Sphere paradigm is smart enough to enable us to fulfill our goal of inferring the environment from the partial and unprecise input data retrieved by Kinect.

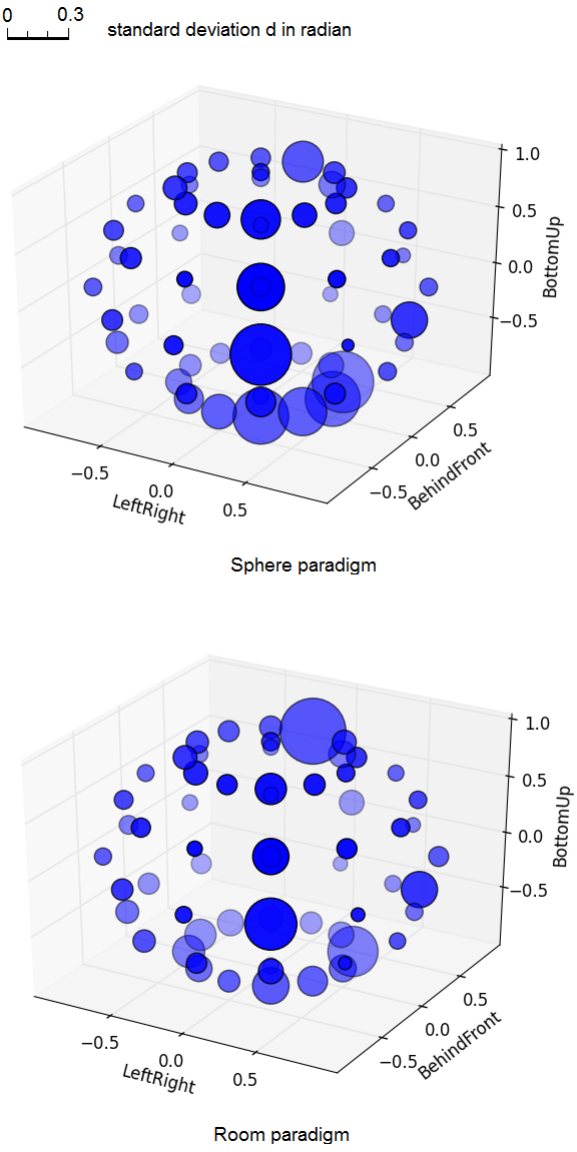

Figure 3: Spatial variation of the standard deviation d in the sphere and room paradigm, in the not centered position

4 Interaction techniques

We tried to take advantage of this pointing paradigm by designing interaction techniques which would allow us to leverage proprioceptive and spatial memory to ease the learning of many commands. We introduce two in-air microinteraction techniques: SMM (for Spatial Marking Menu), an in-air adaptation of the widely praised multi-stroke Marking Menus [34], and a novel interaction based

on deictic pointing called SPS (for Spatial Pointing Shortcuts).

In essence, SPS relies on pointing a given position, whereas SMM relies on a gesture combining several directions. Each therefore leverages a very different cognitive perception mechanism: positional versus directional. It is noteworthy that those two aspects can be easily combined to enrich each other into more advanced interaction techniques. One could for instance point a direction to select an item of their environment or a given program, and then chose a direction to select the function to be applied. This could be the theme of a further investigation, as it raises numerous new problems (such as the precise detection of the direction of a movement if its starting point varies).

Our two techniques can handle multiple users, and allow them to use any hand they want. The hand considered as manipulating is indeed automatically selected as being the furthest away from the body, that is to say the hand such that the pointing ray [head, hand) makes the most important angle with the vertical (latitude). This allows more freedom in the manipulation of our techniques.

Moreover, they both offer an expert mode which can be used in an eye-free situation, that is to say without even using any visual feedback. That prevents the need to look at the display, or even to turn it on (which could be convenient if the command to launch is not a multimedia application). But it also allows the execution of shortcuts without grabbing hold of the display, which may bother people in the room, especially if they are watching TV or a movie. In short, our techniques answer the needs of a couch-interaction in a computer-mediated home.

4.1 SMM: Spatial Marking Menus

Spatial Marking Menus (SMM) is an adaptation of the multi-stroke Marking Menus [34] to 3D in-air interaction. It is expected to bring the well-known learning properties of the Marking Menus to couch-interaction. It relies on the selection of two directions (two levels of hierarchy) among the 8 canonical ones.

The shortcuts are stored on a two-level Marking Menu of 8 branches by level (for a total of 64 shortcut storage capacity). Each item therefore corresponds to the selection of two branches, that is to say a gesture composed of two line segments, encoded in the system as a simple couple of integers.

To compensate for the poor precision of low cost depth cameras like Kinect, we chose to leverage the precision advantages of multi-stroke Marking Menus [34] by asking the user for clear delimiters of its selection, at every step on the way. A selection requires therefore three delimiters: starting the gesture, choosing the first level branch, and selecting the final object on the second level. Although other delimiters could be thought of, such as a snapping of the hand, we propose to use the brief closing of the hand, fast enough to be efficient, and distinctive enough in order not to create false hits in real-world situations. Moreover, it does not require any additional device.

4.1.1 Implementation :

The direction selected in a given level of the menu is computed from the position of the user hand in the 3D space. We retrieve its (x,y) coordinates, ignoring the depth, and compute the segment line created by the points where the hand is at at the time of the delimiters. A simple angle comparison allows us to select the closest branch. This mechanism is robust to changes and variations in the user gestures and allows a clear and precise selection.

In order to ease manipulation and make up for the flickering of the skeleton obtained through Kinect, if the user selects a direction corresponding to no item, we launch the command corresponding to the closest shortcut. Despite introducing a bias in the technique learning, since what the user get is not per say what they did, this makes the manipulation more pleasant.

Our solution is not perfect, as the human does not manipulate naturally on a vertical plane but on a sphere centered on him whose ray is the length of their arm. We end up with a slight difference between the desired direction and the direction measured by our system. Our tests show that this effect is neglectable most of the time compared to the flickering of Kinect's skeleton detection. It is our guess that smarter detection techniques could be devised. That being said, computing the directions from the analysis of the spherical coordinates (latitude, longitude) of the targeted point on the virtual sphere we introduced in our pointing paradigm (see 3.4.4) seemed to show relatively poor performances.

It is important to highlight that all those movements are relative to a starting point of the selection specified by the user, allowing for manipulation from anywhere in the room. It also accounts for the diversity of behavior in the user's manipulation of the system (size of gestures, personal offset...), allowing anyone to manipulate in their own way.

The creation of a shortcut requires an additional command, like a gesture of the hand, or more clearly a vocal command ("create SMM"). The user will then proceed to the gesture which will select this item (combination of two directions). If it is not already affected to a command, the binding is created. Reorganizing the short-

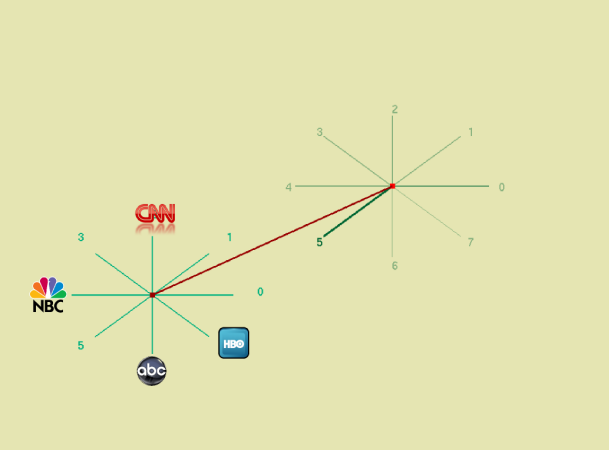

Figure 4: Visual feedback used for SMM

cuts would require an additional interface relying on vocal commands (”move SMM”, ”delete SMM”). Those shortcuts will later on be retrieved through the interaction described above.

4.1.2 Novice mode and feedback :

The transparent transition from novice to expert mode in the Marking Menu paradigm is enabled by an optional display of a visual help for novices. Traditional Marking Menus display this visual feedback after a short period of inactivity without moving. This could be applied directly to in-air interaction, but it has been reported by our test subject to be a little annoying and painful to wait without moving during an in-air interaction. Therefore, we propose to activate the visual help by a voluntary command, such as an audio order (”display SMM”), at any point during the manipulation. A gesture from the non-pointing hand could also be used and may be preferable in certain contexts.

The visual help displayed is fairly similar to the mouse Marking Menus (see figure 4). It evolves during the manipulation, displaying the first level of menus and then the second level when selected. Although this feedback has the advantage to be interactive, it does not provide a full view of the system (only the selected second level, and not all of them). Novices need to rely on exploration to find what they look for if they don’t know where to look, hence the crucial importance of organizing the items with such an interaction technique. They need therefore to navigate within the menus and cancel their last action, which can be done by a vocal command (”back”) without wasting any interaction possibility.

We chose not to display the current position of the hand of the user. Continuous feedback seemed indeed to steal the focus of the attention of the user from the performed gesture to the display screen. As our goal was to train users to become experts as efficiently as possible, we wanted them to focus on the actual action to realize in order to learn it better. Our tests showed that despite being a little harder to come to grips with, users quickly got used to it, and stopped thinking of the task as a pointer manipulation but as a gesture instead. We thereby managed to twist user conception of the technique so that they conceive it in terms of directions, in order to leverage spatial memory.

By transposing Marking Menus to an in-air manipulation context, with the required heavy modifications, Spatial Marking Menus leverage spatial perception focusing on directions. It provides a well structured environment, hierarchical by design, allowing for easy organization within the shortcuts. This comes at the cost of a complexified manipulation action (two direction choices) and the impossibility to have a full view of the system. It is also not flexible and has a limited storage capacity. It is also important to stress out that it is mostly oblivious to the real-world environment.

4.2 SPS: Spatial Pointing Shortcuts

Spatial Pointing Shortcuts (SPS) is a novel microinteraction technique allowing a very direct shortcut selection in the context of couch interaction. It relies on direct deictic pointing of the elements of the user’s environment, which allows the user to create an abstract mapping between their representation of their real-world environment and the symbolic space of the shortcuts.

4.2.1 Implementation :

To create a shortcut, the user only needs to point the item on which they want to create a shortcut. Although a gesture from the other hand could be used, we propose to use Kinect’s microphone to allow for an audio command ”create shortcut”. If the position is not already taken by another command, the shortcut is stored. A simple drag and drop interface can be designed to manage the stored shortcuts (change location) and remove them if needed.

In our system, pointing this position will correspond to a coordinate couple (latitude, longitude) in respect to the center of Kinect’s field of view, directly obtained by projecting the [head, hand) ray onto our model of the room, but from the user point of view, it will seem that the command is associated to the object in the environment. This seemless association is the heart of SPS memorization mechanism.

Since our tests show that our paradigm is robust to the change of position of the user (see 3.4.3), the retrieval can take place from anywhere in Kinect's field of view, and could even be made by another person. The retrieval is straightforward and happens by simply pointing to the desired object and closing the hand without moving. This selection delimiter has been selected to be fast and easy to realize, without requiring additional devices. It is also important that the delimiter is unnatural enough not to be triggered by accident, especially in a couch setting where people might move and stretch. We hope that this delimiter will not lead to false hits, but it could be changed if it were the case, to an audio command for instance.

If the user points to a place where no item is stored, we decided not to ignore this and to trigger the command corresponding to the closest shortcut. We believe our delimiter significant enough that we must not ignore it, especially since the empty selection comes most likely from an imprecision in Kinect's skeleton detection. In order not to select anything and cancel the selection, the user can simply unstretch their arm without doing the delimiter.

4.2.2 Novice mode and double-level feedback :

The heart of the learning process for this technique, much like in the case of the Marking Menus [19], is that the novice user does the same action for a given command, guided by an optional help. This allows for a transparent learning and an eye-free expert mode. Our test have shown that using a timer in order to activate help automatically after an inactivity period is either painful and annoying for the user if the delay is too long, or always activates the help at the moment of selection (as the user stops moving and tends to be inactive for a small moment), which can be a bother for an expert user.

To circumvent this issue, we propose a double feedback mechanism. When the user stretches an arm, as they are probably on the verge of making a selection, we trigger a non-invasive audio feedback giving the name of the command currently pointed (hovered). This situation is thought to be rare enough, and the pointing directions precise enough in order not to be annoying. Moreover, our tests show a lot of users, even experts, mixing up two different items (confusing their respective positions). Even though it might seem annoying and unnecessary, such an immediate audio feedback for experts avoids all those confusions. However, we suggest leaving the possibility for this audio feedback to be used only when the visual feedback is activated, in order not to bother users with unwanted audio feedback.

Figure 5: Visual feedback used for SPS

4.2.3 Visual feedback :

We propose another modality of feedback to answer the case where the user has completely forgotten where the correct shortcut might be. By saying a distinctive audio command ("SPS map"), they will trigger the display of a sketch of the room with all the memorized items on it on the nearest monitor.

However, it is important to highlight that the walls of the room are unlikely to be in the field of view of Kinect. Our system therefore has no precise information about the environment of the user, which makes displaying a map of it a very hard task. Promising works hint the possibility to draw a 3D model of the room by sweeping it with the depth camera [17], which could greatly enhance our visual feedback.

We however settled for simplicity, portability and ease of configuration for the user by sketching only a rough geometrical representation of the room, displayed as a blue cuboid seen from the inside (see figure 5). This allows us to be compatible with our environment-oblivious pointing mechanism (see 3.4.4), at the cost of being severely imprecise. Although this raised the problem of representing a 3D object on a 2D screen, the system turned out to be usable. The main drawback was the impossibility to print the wall behind the user, which is not a real problem since we achieve very poor performances when the user points behind them, as Kinect doesn't manage to detect their skeleton anymore (see 3.4.2). This limitation of the display therefore acts as a mean to dissuade the user from placing shortcuts behind them.

Our tests showed that users have no problem understanding those limitations, made clear by the simplicity of the display, and rely on our visual feedback to find the approximate position of shortcuts, or their position relative

to each other (which is on the other hand accurate since it relies only on information present in the system).

The question of displaying on this map the user position (a silhouette representing the user) and the pointing direction has arisen. As we wanted this feedback to be an occasional on demand help for learning, we chose to put the focus on the real world environment by displaying neither. Indeed, our tests have shown that face with a representation of themselves, and in particular of the pointing ray, people shift the focus of their attention from the room to the screen and tend to let themselves guide by the continuous visual feedback (which is an imprecise map). This results in shortcuts placed at random positions, without regards for the environment whatsoever. In addition to poor memorization results, the ultimate consequence was that users had no idea where their shortcuts actually were. Removing the continuous aspect of the feedback by offering only a static map forced the user to focus on the target of their pointing.

This feedback mechanism on two level answers in a non-invasive way different needs. The visual feedback on demand can be seen as a "macro-feedback", allowing users to point out the approximate location of the shortcut they are looking for within the big space around them. The audio feedback is on the other hand a "micro-feedback", which will allow for precise localization and precise discrimination between neighboring items. By combining the strengths of those two modalities, we offer a robust and complete feedback mechanism.

4.2.4 Additional perspectives on visual feedback :

Variations of the visual feedback mechanism have been studied. In particular, we considered a hierarchical feedback, based on displaying only one category of shortcuts at a time in order to lighten the display. This idea has been discarded because it forces the user to organize the shortcuts in an explicit hierarchy and slows down the interaction by forcing the choice of a category. That being said, users are still free to create hierarchical structure by organizing their shortcuts however they want in the 3D space (clusters...).

Mechanisms such as fish eye around the targeted point or a separate zoom display could also be useful. We did not chose to use them as the display did not seem particularly overloaded, so we did not see a reason to break the straightforward direct mapping between the euclidean 3D real world and the display. It seemed that it would add a lot of unnecessary complexity to the user's perception.

Finally, the evolution of hardware gives us hope that in a near future houses could be equipped with a multi-directional projector (or several projectors), or interactive walls. In this context, we could use as visual feedback mechanism a direct projection of the shortcut icon at the real-world position where it is stored (on top of the actual item), eliminating the need for a display. This would bypass the representation issue (3D space displayed on a 2D screen) and may improve the overall performances of the techniques.

4.2.5 Parameters :

Several outside parameters depending on the environment are thought to have a great impact on the performances of this technique. SPS relies indeed on the items of the real world. Psychology literature [2] leads us to believe that our technique would be sensible to the density of the visual cues in the space surrounding the users, their nature or to their organization (chaotic or logical). Our guess is that their number, but also the brightness of their colors, the emotional involvement of the user and so on could improve the memorization performance of our system. However, such factors are deeply dependent of real world situations, and it seems hard to reproduce in laboratory.

To sum up, Spatial Pointing Shortcuts is a shortcut technique based on positional pointing. It relies on the combination of a straightforward in-aid deictic pointing and the smart use of spatial memory. It offers a direct access by only a simple action to any shortcut stored in the system. One of the main features of this technique is a huge storage capacity. Our precision study (see 3.4) indicates indeed that our system could potentially discriminate bewteen several hundred positions. It also provides the user with a highly customizable experience and handles a huge variability in the shortcut positioning, allowing the users to have a very personal organization scheme. This makes up for the lack of hierarchy, as all the items in the system are considered on the same level.

5 Memorization evaluation

To evaluate our techniques, we designed a testing experiment to measure their memorization capability. Numerous pre-tests have lead us to take significant decision about the experimentation.

5.1 Pre-testing

The heart of a memorization experiment is to expose the subject to a stimulus several times and to assess whether they remembered it or not. Therefore, the number of exposure events has to be large enough to allow for memorization. Moreover, we want to highlight the big stor-

age capacity of spatial memory, so we need a consequent number of items to be memorized. However, our micro-interaction technique is not designed to be done a lot of times in a row. The experiment was therefore very tiring. Plus we needed the experiment not to be too long in order to stay pleasant for the subject. This highlights a trade-off between the number of items and exposure events that the experimenter want to maximize on the one hand, and the fatigue of the user and the length of the experiment that we need to minimize for the subject's comfort on the other hand. We ended up decreasing the number of items, exposure events and memorization evaluation as much as possible.

We quickly noticed the need for clear and interpretable measures, in particular when it comes to recall and feedback use. This lead to the decision of not allowing mistakes in selection despite Kinect's poor precision. To make up for that, in case of selection error, the experimenter will ask the subject for their intended target. This will allow us to determine if the error comes from memorization or is Kinect-related.

It is noteworthy that it is impossible to distinguish between cases where the error comes from an imprecision or a drop in the skeleton tracking we used and cases where the error comes from the user who didn't actually perform the same gesture they intended. It is actually not unrare to see people pointing in the right direction but with a small offset for SPS, like pointing another part of the object they intended, which could be closer to another target. People sometimes also have trouble to perform a precise direction in SMM, and end up doing a diagonal line instead of horizontal, instead of a horizontal one for instance. Our guess is that it is due to the fact that the human hand moves on a sphere centered on the user. Determining whether an interpretation error comes from a flickering in the skeleton tracking of Kinect or from a user manipulation mistake would require a permanent arbitrary labeling of Kinect's input video feed. Therefore, we classify both those kind of errors as "Imprecision errors" (as they rely on depth camera manipulation or performances). These are the errors our validation by the experimenter will circumvent in order to measure the memory recall. They correspond to cases where the user knows where their target is but did not manage to reach it.

Another challenge linked to the low precision depth camera was the trade-off between smoothing, which made up for the flickering of the skeleton tracking, and precision. Smoothing also causes a delay in the skeleton tracking, which results in the skeleton being always a little behind the actual user silhouette. To circumvent this issue, we added an arbitrary delay between the moment where the selection is ordered by the user and the moment it is treated by the system, to give time to the smoothed skeleton tracking to find the right position. The delay has been set up to offer maximum performances without being perceptible by humans. The smoothing algorithm has also been changed to be less active for huge movement. That way, the skeleton tracking followed the user efficiently without delay when he moved from one position to another, but still got rid of the flickering of the detection when precision is required.

In the manipulation of SPS, we first let our users chose any position they wanted in the 3D space surrounding them. This resulted in interesting behavior, such as one user creating an artificial grid to place the items. However, we noticed poor performances coming from such behavior. In particular, since there was no physical referent to direct the aim of the pointing, the distance to the target tended to increase over time. Those users didn't learn precise position, but rather tended to forget them, as they got blurred over time. They reported that "it is very hard to point when there is no physical reference". As a consequence, we decided to order the users to aim only at precise objects. Furthermore, we asked the users not to store shortcuts in the area right behind them, which causes problems in Kinect's detection (see 3.4.2). Most of them told us that they did not intend to anyway, showing that this constraint does not impair much their freedom of use of our techniques.

Finally, we noticed that selecting the closest item in case of empty selection could create an artifact in memorization. In particular, people ended up learning that a shortcut was placed on a given item, when they had formerly placed it on the item right next to it. This was however not a real problem as every selection on this item, even though it was not the originally intended item, ended up on the right shortcut selection. We therefore decided to keep this mechanism, as it was very efficient at making up for the poor precision of Kinect.

5.2 Experimental protocol

For this experiment, we wanted to measure the memorization performance of our interaction technique with as much precision as possible, without being tied to a technological system which could evolve in the future. In particular, in order not to introduce a noise coming from the performance of vocal analysis or closing-hand detection, we used a mouse to emulate those delimiters with a perfect accuracy. For SPS, the user therefore had to point to a direction and then click. In the case of SMM, this meant for the user to specify the starting point of their gesture, the

Figure 6: Room used for the experiment

first level selection and the second level selection through clicks (two segment lines, three clicks). We also asked the subjects to perform standing to improve skeleton tracking.

In order to maximize the comfort of the subject, we decided that they could manipulate with whichever hand they felt like using. We considered the hand furthest away from the body as the active hand (biggest angle between the arm and the vertical).

Since we wanted to have a clear measure of the influence of help feedback, we created two artificial conditions common to the two techniques. Subjects would by default enter the expert mode, where manipulation happened without any feedback whatsoever. A right click would trigger the novice mode, that is to say enable all audio and video feedback mechanisms once and for all. In order to compare the two techniques on an equal footing, we added an audio feedback to SMM saying the name of the hoovered item. The novice mode of SMM also allows for exploration: a right click replaces the audio command "back" and cancels the last selection and provides users with a way to switch category and find an item.

Our interaction techniques are designed for couch-interaction in a home environment. However, our laboratory does not dispose of a believable living room for testing. Therefore, additional visual cues have been added to one of our testing rooms to emulate the decor of a living room. Those consisted of pictures of decorum elements (lamps, vases, plants...) hung on the walls (see figure 6). In SPS, the users could point indistinctly between real world items or those artificial visual cues.

We used in the experiment a neutral vocabulary to represent the shortcuts. It consists of 5 categories (animals, leisure activities, colors, fruits and clothing items). We used 5 items per category, adding up to a total of 25 items. The items were different between categories, but not between users. Each item was represented by a small icon with its name written under it. Since we wanted to test the maximal capacity of memory, we decided not to consider the items following a Zipf law but presenting every item the same number of times (see discussion in 1).

The experiment was taken by a total of 12 subjects, aged from 15 to 30, average 23. 3 of them were women. Most of them had no previous Kinect experience. They all tested the two techniques. They were asked to memorize as many items as possible, and to select them as quickly and precisely as possible. The order of the techniques was balanced among participants following a latin square. A two-factor ANOVA on starting phases and techniques with repeated measures on the technique factor (two by subjects) later showed that the order of phases had no significant effect neither on time nor on memory performances, validating this experimental protocol.

5.2.1 Procedure

Each technique test began with a small example phase in order for the subject to get familiar and to understand the manipulation involved. Each user was then asked to chose the position for each item either in the 3D space surrounding them (SPS) or on the two layered directional menu (SMM). Much like in real use, the user therefore gets to pick the position of the items, enhancing memorization (see 2.3). Moreover, to mimic this real-world use, the subject does not know beforehand what items are going to come in the future, creating a big constraint on their organization scheme.

Retrieval phases then took place, where the subject was asked by a visual and audio stimulus to retrieve the stored items. An audio feedback lets the user know if they were right or not. The order of the retrievals was randomized every time. Each phase consisted of one retrieval per item. Each technique was tested on 4 retrieval phases, for an overall total of 200 selections by subject. In the fourth phase of each technique, access to the novice mode was disabled, in order to evaluate what had been memorized (after positioning and 3 exposure to stimuli) without any feedback whatsoever. In the other phases, for every item, the user had the possibility to trigger the novice mode if he judged it necessary.

At the end of the experiment, a survey was given to the subject in order to measure their personal opinion. A small discussion aimed at highlighting the organization strategies and memorization techniques they used, in order to proceed to a user study.

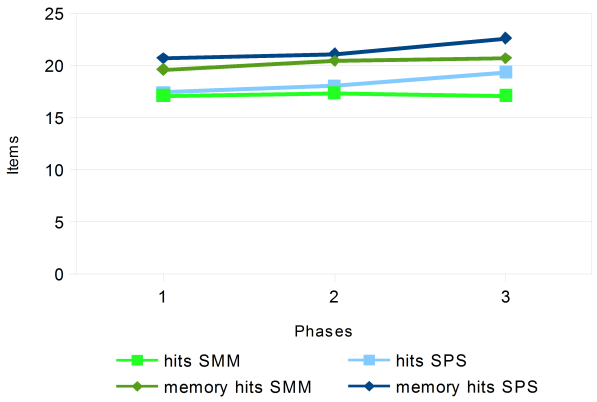

Figure 7: Global performance of the system

5.3 Aggregated results

We first studied the aggregated data of all our subjects to draw general conclusions on our techniques.

5.3.1 Global performances

To better understand the global performances of our techniques as a whole (including the optional helping feedback), we distinguish two hit scores. The basic "hits" score is the number of good selections measured by the system, whereas the "memory hits" score correspond to the number of correct intended selections, measured by the experimenter as discussed in 5.1, that is to say the number of cases where the subject knew where the item he wanted was. The difference between those two figures corresponds to the aforementioned "Imprecision errors" (either coming from a bad movement differing from the intent from the user or a poor detection from the depth camera).

Figure 7 show the evolution of performance of the full techniques during the phases where the feedback can be activated. It highlights how well users perform in general with our techniques, with the optional help if needed.

Performance is quite high from the start, which leaves little room for increase over time. SMM evolves very little, whereas SPS gets more and more efficient. Familiarity with the technique and the layout of the items plays therefore a bigger role in SPS. The two techniques reach relatively similar performances in this setting, SPS ending up better by 1.91 memory hits. A one-way ANOVA shows that this different is however not significant.

SMM imprecision errors	14%
SPS imprecision errors	11%

Table 3: Proportion of imprecision errors in the total of selections

5.3.2 Imprecision errors

These measures also allow us to evaluate a "imprecision errors" score, linked to the current state of technology, evaluating the detection of Kinect as well as the ease of the user to manipulate it. It corresponds to the cases where the selection is a "memory hit" corrected by the experimenter but not detected correctly by the system. As discussed in 5.1, they can come from a poor skeleton tracking or from a imprecise manipulation from the user. We compute a "imprecision errors" score as the proportion of cases where such imprecision errors happened in the total number of selections.

Table 3 show that the imprecision errors are fewer for SPS, probably because there are less errors coming from the user. However, a one-way ANOVA shows that this difference is not significant. We achieve overall very few imprecision errors, with a few outliers dragging the means down (for instance people having a hard time performing a correctly-detected horizontal movement).

5.3.3 Success rates

The remaining items to reach 25 in addition to the "memory hits" discussed before correspond to true errors where the subject did not know where the item was. They were either to attempts of selection without help who ended up failed (in large majority), that is to say cases where the user believed the item to be somewhere but was mistaken; or to cases where the help was not enough to perform the right selection. For instance, some subjects got from the help the approximate position of an item in SPS but still failed the selection because they were too much in a hurry to look for the audio feedback. Others ended up selecting a wrong item in SMM by lack of attention.

To investigate the origin of these errors, we analyzed the evolution of the success rate (percentage of hits in the total number of selections) over our whole dataset, with and without help used (figure 8). The number of cases in which help wasn't enough to reach success is indeed low, as the success rate with help reaches 0.89 for SPS and 0.95 for SMM. The success rate without help show a nice progression for the two techniques, showcasing learning with our techniques.

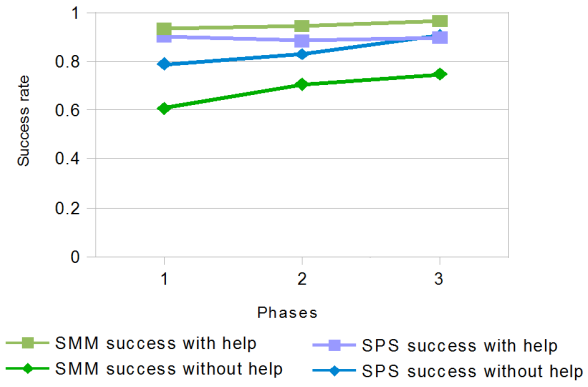

Figure 8: Success rate with or without help

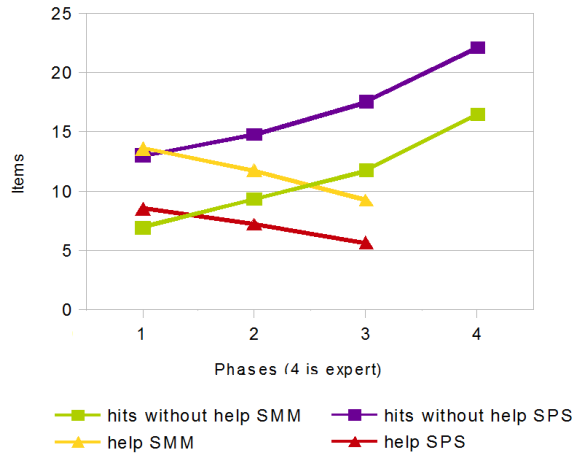

Figure 9: Usage of help

5.3.4 Use of help

The behavior in the use of help were rather diverse, which we will discuss later on (see 5.4). However, we can still draw conclusions from the averaged behavior. Figure 9 shows the evolution of help usage and good selections without help during the experiment.

For both techniques, help use decreases as the user learns their shortcut layout. Despite the fact that SPS clearly requires less help, a two-factor ANOVA with repeated measures on both factors applied to the factors technique and phase show that our sample is not big enough to draw a significant conclusion ($p = 0.061$). The decrease over time is on the other hand statistically very significant ($p = 0.00014$, $F = 13.6$).

In the same way, the number of hits (we here consider memory hits, in order for our analyze not to depend on

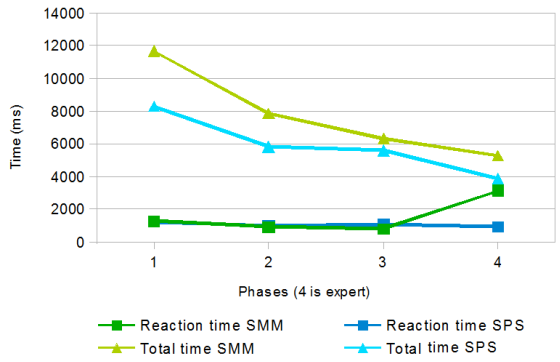

Figure 10: Evolution of reaction and total selection time

current technology) without help use is smoothly increasing, at the same rate for the two techniques. This tends to imply that the learning process of the two techniques might be similar, and although SPS offers better memorization performance from the start, a longer training could maybe improve SMM performances by making up for the initial handicap of SMM.

The two techniques present a smooth and efficient novice to expert transition, which was one of our main concerns. Moreover, this transition is decently fast, for an average of 3.1 items learned by phase, which results in high memorization scores for very few exposure events.

5.3.5 Time

Another measure showcasing the transition from novice to expert is the manipulation time. Our system proceeds to two different measures: the total selection time between the apparition of the stimulus and the user's retrieval, and the reaction time between the apparition of the stimulus and the moment when the system records a significant movement. Their evolution is plotted on figure 10. The space between the two curves corresponds to the time taken by the actual selection movement.

This distinction offers us a very interesting observation. In the expert phase (4), in the SMM condition, a relatively big amount of the time is taken by the "reaction time". This highlights a significant hesitation in expert mode using SMM not observed in SPS. However, once this hesitation is settled, the selection movement is very fast. SPS seems to bring more confidence in the learned items than SMM. It is however possible that in SPS, the hesitation takes place after moving, if the user has only a vague idea where the item actually is. Our subjective guess from observation of the experiments is that this happens sometimes but not significantly. Another experiment

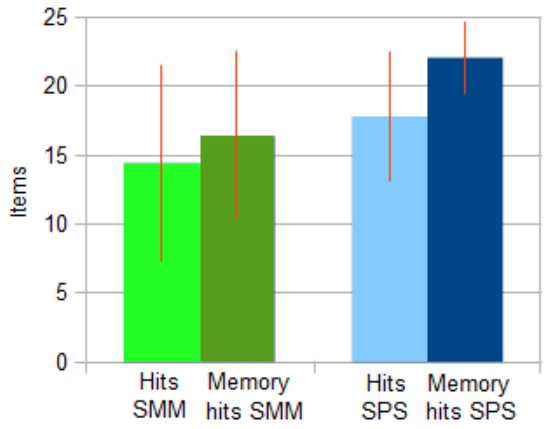

Figure 11: Means and standard deviations of recall performances in expert phase (4)

would have to be designed to settle the matter.

The lack of huge difference between SPS time scores in phase 2 and 3 could imply that we are getting closer to the maximal action speed offered by the technique (not yet reached however).

On the whole, SPS is always faster than SMM. This can be explained by the fact that SPS consist of one single selection, when SMM needs a two step action. However, SPS requires more ample movements, as the user could be lead to point everywhere in the room, whereas SMM could be performed with very small movements right in front of the user. Our guess is that with a longer training, experts in SMM could perform faster than with SPS.

However, in our study, a one-way ANOVA on the total time in phase 4 showed that the factor "technique" has a clear and significant effect ($p = 0.006$, $F = 11.3$). SPS is therefore significantly faster than SMM, ending up at 3888ms against 5267ms for SMM. Although those measures could seem long for micro-interaction, they are fairly decent for in-air interaction. They will not be perceived as annoying by the user since our techniques are used for rare sporadic actions once in a while.

5.3.6 Memorization

Measure of phase 4 (without any feedback) allows us to assess raw memorization after only 3 exposure by item. Our guess is that incident learning has also taken place, but there is however no clear way to measure it in our setting.

Figure 11 showcase the recall performances of both techniques (and the recall as perceived by the system).

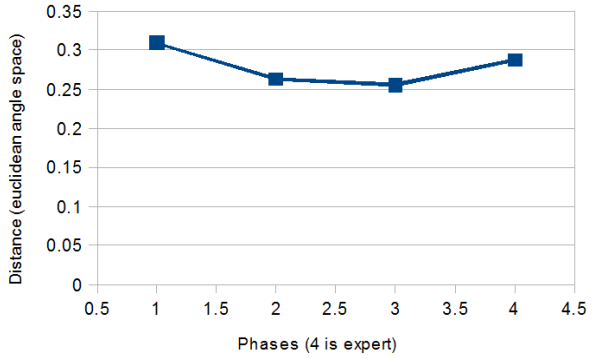

Figure 12: Evolution of distance between target point and real aim for SPS

We manage to reach with so few exposure event a recall score of 16.4 items for SMM and 22.1 for SPS. A one-way ANOVA for correlated samples shows that the factor "technique" has a very significant effect on the memory hits, that is to say the overall number of memorized items ($p = 0.0055$, $F = 11.8$). Therefore, SPS clearly outperforms SMM when it comes to memorization.

5.3.7 Distance to target in SPS

It is also interesting to look at the average distance between the actual target and the goal of the user in order to evaluate how SPS is mastered. Figure 12 show a progress curve as expected: as the user gains mastery over the system, they become more and more precise. However, without the guiding feedback, some of this benefit is lost. This will probably not happen in real-world use since the audio feedback will still be present.

5.4 Diversity of users

The aggregation of all users in average data does not do justice to the diversity of behaviors we observed in the experiment. However, it is very hard to come up with clear measures to describe them. Some conclusions can nonetheless be drawn from qualitative observation.

The diversity in the use of help can be a problem for analysis of memorization. Indeed, people can be more or less prone to risk or uncertain, blurring our measures of memorization during phases where help is accessible. Fortunately, we begin to distinguish two main trends of help usage. People unsure of themselves tend to use help very often, sometimes showing only a very small decrease in help usage. We call this behavior "timid". Others are

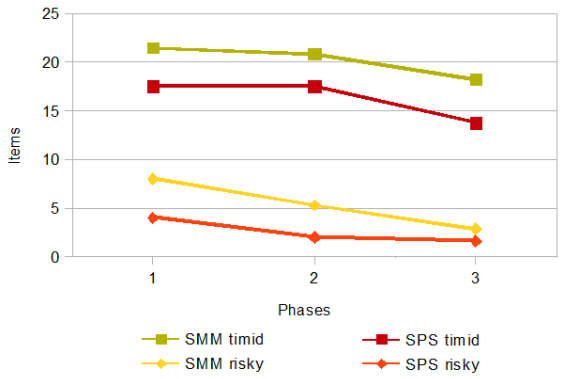

Figure 13: Different profiles for help usage (average per profile)

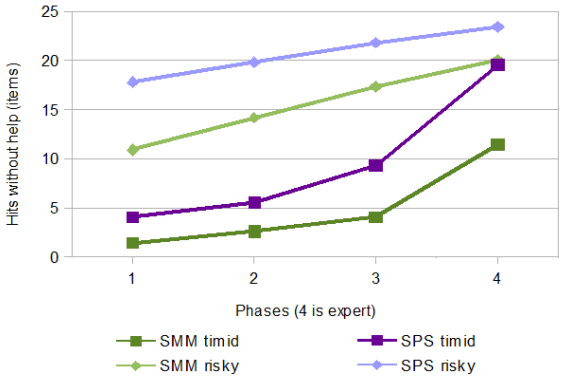

Figure 14: Hits without help performances for the different help usage profiles

	SMM	SPS
risky	7	8
timid	5	4

Table 4: Help profiles distribution

either more risky or feel more comfortable with the techniques right away, resulting in the "risky" behavior.

We tried to validate this distinction by performing two (separating the techniques) two-factor (help usage, phase) ANOVA with repeated measures on the factor phase, applied to our measures of help usage. The results show a very significant effect of our "help profile" distinction on the measure of help usage : $p < 0.0001$ for both techniques, $F = 68.2$ for SMM and $F = 106.2$ for SPS. This tends to legitimize the profiles we defined, although the number of subjects we had seem to small to draw any significant conclusion.

The average behavior for the two techniques and the different help usage profiles is plotted in figure 13. Interestingly enough, it seems that the learning could be taking place later with the "timid" behaviour than with the "risky" one.

Although there is a large core of users adopting a "risky" behavior for the two techniques, it is not rare to see people having a "risky" attitude in a technique and a "timid" attitude in the other. The total distribution is given by table 4.

We proceeded to two (separating the two techniques) one-factor ANOVA test with "help usage profile" on the final memory hit counts. The results showed that this usage of help had a significant impact on the memorization performance ($p = 0.012$ and $F = 9.3$ for SMM, $p = 0.011$ and $F = 9.8$ for SPS).

This distinction in help usage profiles allows us a more meaningful look at the phases where help was available. Indeed, in the previous section, Figure 9 aggregated the results of people having an important use of the help with those of people who barely used it. In those two different cases, the value "hits without help" has however a different scale: "timid" profiles will have a low score on this variable simply because there are fewer items where help hasn't been used. We are now able to plot these evolution curves on two different scales (see Figure 14).

The relatively big gap between the scores measured on phase 3 and 4 for the "timid" condition tends to imply that the users know more items that they give themselves credit for, and that they use help more often than they actually need. This is especially true for SPS.

In order to further investigate the influence of help usage profiles in the learning process, we applied two (one per technique) two-factor ANOVA on the factors phase (with repeated measure) and help usage profiles, for the value of the slope of the "hits without help" curve during the "novice" phases, which represents here the memorization taking place at a given time. The results tend to imply that the help usage profile has no significant impact on the learning process ($p = 0.064$ for SMM and $p = 0.63$ for SPS).

5.5 User perception

Our experiment, in particular due to the fact that users were free to chose the positioning of their shortcuts, allowed us to study the diversity of the personal behavior of the subjects in respect to our techniques.However, due to the variability even within the choices of the same user, it

	SPS	SMM
I liked the technique	4 *(0.85)*	3.25 *(1.48)*
This technique is easy to get to grasp with	4.25 *(0.97)*	3.08 *(1.51)*
This technique is fun	4.33 *(0.89)*	3.08 *(1.31)*
This technique is efficient	3.92 *(0.9)*	3.17 *(1.27)*
This technique is fast	4 *(1.13)*	3.75 *(1.22)*
I found what I was looking for easily	3.75 *(1.14)*	3.25 *(1.48)*
I memorized the position of items easily	3.75 *(1.14)*	2.83 *(1.34)*
I was able to memorize a lot of positions	4.17 *(0.94)*	3 *(1.41)*
I'm satisfied with how I got to organize my items	3.58 *(1.56)*	3.42 *(1.38)*
This technique is tiring	2.67 *(1.07)*	3.92 *(1.38)*

Table 5: Means and standard deviation of subjective answers to the user survey (1=strongly disagree, 5=strongly agree). Statistically significant differences are highlighted in bold.

is hard to draw precise measures out of this. A survey and a discussion allowed us to gather their personal opinions and commentaries, thereby enriching our observations.

5.5.1 User survey

Table 5 sums up the results of the subjective survey distributed to the users at the end of the experiment in order to evaluate their personal perceptions of the techniques. All the values have been tested for statistical significance of the mean differences by a paired difference t-Test for correlated samples, to assess their significance as best as possible considering the small size of our dataset. The conclusion was that SPS was significantly more fun, and was perceived as easier to get to grasp with and allowing easy memorization of items. This shows that SPS successful leverages the benefits of spatial memory for easy transition from novice mode to expert mode. SPS is also perceived as better than SMM to memorize a big number of items. Even though the difference is not statistically significant for such a small sample, it is still relatively big. Our guess is that it might turn out to be significant on a larger sample of the population. This makes sense as SPS memorization mechanism is closed to the method loci. Both indeed rely on associating items to objects in a familiar environment to enhance memorization.

Other results show that SPS is on the whole prefered to SMM and considered more efficient. SPS is also perceived as less tiring, which makes sense as it requires only one action per selection against the two-level gesture required in SMM. However, we can notice that there is no big difference in the evaluation of the organization scheme. The freedom in organization offered by SPS was not leveraged to create better organization schemes. This tends to show that constraints may help people to organize themselves.

5.5.2 Additional user comments

Unsurprisingly, user comments highlight that "the efficiency depends a lot on the initial placement", and that it is "hard to organize the items without knowing them beforehand". In real-world use, the performances of our techniques might therefore be better as the users will have a better idea of what they want to do with them.

SPS has been reported as "more pleasant", and "requiring less reflection", but also as "more suited to everyday life". Indeed, "it is easier to command the fridge by pointing on it than with an arbitrary gesture". Those comments bring high hope for our novel interaction technique, in particular in the context of computer-mediated living. It can indeed handle various levels of symbolic abstraction in the mappings involved.

SMM on the other hand was praised for its organizational capacity. It is "practical to regroup concepts, but not to memorize them". On the whole, people seem to have liked the constraints of SMM which obligated them to use a decent organization scheme. Comments like "it is easier to be lost for a big number of items" suggest that this technique might be better suited for small amounts of commands.

5.6 Qualitative observation

Observation during the experiences allowed us to notice some recurring trends. For instance, we saw people making the right selection against their will (i.e. a hasty selection which they regretted but turned out to be right) or without being aware of it, in particular for SPS. This tend to show some kind of proprioceptive memory, where the body knows the right gesture to perform but the subject is not conscious of it.

Some users retraced their steps during the positioning phase in order to remember where they had put a given item based on the logical decision they believed they had taken. Several of them were reluctant to use the help and preferred to think it over and to observe their environment for inspiration. This may be characteristic of the aforementioned risky help usage profile (see 5.4).

In SPS, we observed that spatial cognition was very strong and often provided an approximate idea of the desired position ("I know I put it somewhere in this area..."). This could however can lead to some drawbacks, especially in our experiment where the audio feedback is by default deactivated. We saw several times users learn a bad position for an shortcut, by targeting for instance a neighboring item of the real world. Even when they have been notified to be wrong, some of those errors persist all along the experiment.

Almost every subject confused two items positions in SPS, mistaking one for the other. This highlights that the users have no difficulty remembering on which item of the real world they put a command, but rather what command they put where. We have good hope that real-world use with more meaningful items will put an end to this kind of behavior. Moreover, the real-case use of SPS features an immediate audio feedback which should get rid of all the aforementioned observed artifacts.

5.6.1 Organization and memorization strategies

There is a great diversity in the positioning scheme of users, even within the different items for the same user. It was sometimes overlooked and not thought about. It is common to see users trying to come up with a decent organization and memorization techniques and "give up after a while because there were too many items", that is to say not manage to scale those successfully. Some subjects applied a partial organization scheme which they had to adapt for an unforeseen item. It is therefore impossible in those conditions to draw clear objective measures on the impact of items organization and memorization strategies on user performances.

However, we still can draw interesting conclusions from rough observations. People who apply a clear organization scheme tend to perform better at memorization, plus it eases greatly exploration for unknown items during the novice phases. Items placed without any particular reason or consideration tend to lead to poor performances, as their learning is forced and doesn't benefit from any investment from the subject (see 2.3). Any kind of mnemonic device was useful, in particular to distinguish between the members of a given category (low-level selection).

It is not rare to see users ignore or tweak the existing category scheme to create their own. 5x5 items was indeed not optimal for SMM in particular. We saw for instance some subdivisions inside the proposed categories (animals split into flying or not, leisure activities split into sports or not...).

Outliers, such as items placed far away from the rest of their category, seemed to be better memorized as a whole, but some of them stay persistent errors. Similarly, one user reported that it was "easier to memorize the items which he had to replace" because the initially chosen position was already taken. This forced him to think of a second choice and this anomaly behavior seemed to enhance memorization.

5.6.2 SMM

For SMM, the large majority of subjects used the categories suggested by our item taxonomy for their organization, using the category as first level choice (first direction). For the second level, the default behavior was to place the items in incoming order, without any particular reason. As a consequence, they had no trouble finding the right first category, but the second level selection was hard, and sometimes random.

In this forced learning situation, as well as in general, some movement combinations were easier to remember: twice the same directions (up up), or opposite directions (right left). The importance of this symbolic perception of directions is highlighted by frequent confusions between opposite directions (such as confusing a "left right" movement with a "right left" movement for instance). A particular user even used this fact to store related items in opposite directions, bypassing any category organization.

Fortunately, some users still came up with decent organization scheme within categories. Most of the time, the organization mechanisms was different between categories. It could be straightforward, such as organizing the colors from brightest to darkest. It could rely on objective semantic link between the objects (putting raspberry and strawberry closed to each other, or the bird animals...), or subjective links (place the panda bear up because panda bears climb trees). Some relied on arbitrary subjective sentiment assignation: for instance, the "worst" item could be placed "up", because "up" is bad since it is hard to reach. This assertion could also rely on the combination of directions (left then right is bad because coming back on ones footstep is perceived as a failure). These subjective assignations could also come from memories ("I put the dog up because I remember my dog jumping very high when I was a child").

We observed two subjects coming up with very interesting memorization techniques for SMM that they applied at least partially. The first one leveraged elements of the decor to remember the position of the items, much like in SPS. For instance, they associated up-right to pineapple because in the upper right corner of the wall was a plant similar to a palm tree.

The other one tried to use their environment in a different way, using the real-world items as starting point for their gesture. This could be very efficient if the menus in SMM changed with the starting position. In our condition, they encountered frequent conflicts, as "up up" is the same item whatever the starting position is. The same subject tried to mimic either the shape of the object or the shape of the first letter of the object with the two directions, going away from direction selection and more into shape recognition. This could be an incentive for further research in this direction.

5.6.3 SPS

For SPS, some subjects did not seem to follow any particular category for their organization scheme. Some of them later came to regret this choice. The other ones generally used regions of space, forming a cluster with items of the same category. They sometimes used walls and ceiling as categories, despite the fact that there are very few visual cues on the ceiling (such an organization scheme limited them to only 4 categories). A subject told us that "maybe a semantic link between the items of a category would have been better, such as putting all the leisure activities on real-world books".

The most common strategy used to place items in the environment was to try to find some kind of semantic mapping, however personal and arbitrary it may be. It includes for instance putting banana on a plant that reminds the subject of a banana tree, putting the dolphin on the picture of a boat because of their connection to the see, or putting the deck of cards in the library because that's where the user would place actual cards back home. The link could be more straightforward and rely on physical resemblance, like putting the banana on the yellow plate, or the tie on a flower whose shape is vaguely similar to a tie. It is fairly common to see people leveraging their own personal life memory: putting the dog on the vase because their dog broke a vase, the gloves on the stepladder because of real life habits, or associating grapes and a pen because of the story of a hungry professor. Funnily enough, it was sometimes the memory of the positioning phase who played a major role ("I remember this item because when I was placing it a thought entered my mind and it made me laugh"). Finally, for some rare cases, the body position was the determining factor ("I remember this item because to point at it I had to assume a funny stance").

It was also fairly common to use associations between the shortcut items. This mostly translated into using physical proximity to represent semantic proximity, be it straightforward (cat close to dog, sea animals together, birds together...) or more far fetched ("let's put those two things together because I find them both cute"). The spatial relative position of the items was also leveraged in some cases ("the fish is below the cat").

Even the subjects who had no inspiration at first for where to put the items ended up "inventing stories" or finding mnemonic devices to enhance memorization. This tends to imply that forced learning is a bit easier with SPS, which could be measured by another experiment where the user does not get to chose the position of the items.

Several users reported using their visual memory, in particular for the relative position of items such as colors. In this case, it was helped by visual homogeneity in the items. However, visual memory was also leveraged for totally fictional constructs ("I remembered the fish on the trashcan because I clearly visualized a fish in the trashcan"). Spatial cognition obviously played an important role, even in the most abstract way ("the cat was left"). Some users even memorized the order in which the items were placed and used it to order their shortcuts (from left to right for instance). Audio memory was also reported to be used ("I easily remembered that the coat was on the picture of a boat because coat rhymes with boat"). This technique therefore leveraged many more memory modalities than expected, enhancing all the more the memorization performances.

In conclusion, SPS was efficient at direct item retrieval, and suffered from poor hierarchical organization. SMM showed the exact opposite, being very efficient at providing the user with an organization scheme but performing poorly at distinguishing among the items within a given category. Those two techniques could well be combined to make up for each other's mistakes.

6 Conclusion

In order to answer the rise of capability in our home media centers, we aimed at designing device-free in-air interactions which would leverage the well known properties of spatial memory and offer a easy memorization of a huge number of items much like in the method of loci [33]. After designing a pointing paradigm which could detect the targeted object without knowing the environment and which was robust to changes in the user's position, we proposed two micro-interaction techniques relying on detection by a depth camera. Spatial Marking Menus (SMM) is a mid-air adaptation of the single-stroke marking menus [34], relying on the selection of two directions of space. Spatial Pointing Shortcuts (SPS) is a novel interaction technique relying on deictic pointing to create mappings between the real-world environment and

the symbolic space of the shortcuts, following a "what you point is what you get" intuitive paradigm. SPS can even handle various levels of symbolic abstraction, ranging from totally abstract to semantically straightforward (pointing on an object to issue orders to it).

Our pointing paradigm relies on considering the intersection of the $[head, hand)$ ray and an imaginary sphere encompassing the field of view of the camera. Despite the poor performances of Kinect, this system is precise enough to distinguish between squares of edge lengths 20cm on a wall 2 meters away, from any position in the field of view. Even with poor precision equipment, this abstraction allows us to reach fairly decent performance, answering our goal to efficiently infer not only the surrounding real-world environment but also the symbolic representation that the user has of it from a blurry imprecise and very partial input. Thereby, we maximize the interactional bandwith with a relatively poor input.

Our two techniques have then been tested in our lab by a panel of users. Both performed fairly well, enabling the memorization of 16.4 items for SMM and 22.1 for SPS after only 3 exposure to each stimulus. SPS is also faster than SMM, and overall preferred by a subjective opinion survey. Both technique show the desired smooth, easy and fast transition from novice to expert mode, their learning rate do not significantly differ.

Our dataset seems to show two different trends in the use of help feedback: some users use it a lot, whereas the others only rarely. This raises the question of self-confidence in a human-computer interface and its impact on performance. Although it seems that the "confident" group, adopting a "risky" behavior, obtains better performances, additional research is needed to come to a significant conclusion. These trends relative to help use may even be more general than the context of our two techniques.

Many parameters play an important role on spatial cognition, and by extension on our techniques. Now that they have been introduced, further studies could investigate the influence of the number of visual cues in the environment, of their organization, of their nature (color, emotional link, etc...), of the type of feedback received by the user, of the freedom offered during the positioning of the items, or even if this positioning is left to the user to leverage their agentic memory (see 2.3) or not.

We noticed that SPS was particularly efficient at direct retrieval, whereas SMM was great at providing the constraints required to create an efficient organization scheme. It is noteworthy that those techniques, although designed for shortcuts retrieval at the scale of the whole home-center level, can be used within applications to store another level of shortcuts, conditioned by the applicative context. In the same way, those basic interactions can be extended into more advanced interaction mechanisms to provide multi-level control mechanisms (for instance pointing on an object and then moving in one direction to select the action to apply to this object). This could be a way to combine their strength and make up for their weaknesses. Those enrichment solutions raise a lot of other questions (such as the visual feedback) and could be the object of further research.

The evolution of technology is bound to improve the performances of our techniques with better low-cost depth cameras. In addition to improving the user's overall comfort, novel feedback mechanism, such as projection of the shortcuts on real-world objects, could contribute to easing even more the learning process.

On a broader note, we hope to revive interest for spatial memory, and particularly for the very powerful method of loci [33], used through history to learn huge number of items. Our work indeed extends the conclusions of the Data Mountain [28] to a 3D input interaction. We showed that the democratization of 3D in-air interaction created new means to leverage existing human powerful capabilities to enhance human-computer interaction, and we hope to lead the way for further research in that direction.

References

[1] Andrade, J., and Meudell, P. Short report: is spatial information encoded automatically in memory? *The Quarterly Journal of Experimental Psychology 46*, 2 (1993), 365–375.

[2] Baddeley, A. *Human memory: Theory and practice*. Psychology Pr, 1997.

[3] Bailly, G., Walter, R., Müller, J., Ning, T., and Lecolinet, E. Comparing free hand menu techniques for distant displays using linear, marking and finger-count menus. *Human-Computer Interaction–INTERACT 2011* (2011), 248–262.

[4] Bau, O., and Mackay, W. Octopocus: a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, ACM (2008), 37–46.

[5] Cabrer, M., Redondo, R., Vilas, A., Arias, J., and Duque, J. Controlling the smart home from tv. *Consumer Electronics, IEEE Transactions on 52*, 2 (2006), 421–429.

[6] Cesar, P., and Geerts, D. Past, present, and future of social tv: A categorization. In *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, IEEE (2011), 347–351.

[7] Cloutier, J., and Neil Macrae, C. The feeling of choosing: Self-involvement and the cognitive status of things past. *Consciousness and cognition 17*, 1 (2008), 125–135.

[8] Cockburn, A., and McKenzie, B. Evaluating the effectiveness of spatial memory in 2d and 3d physical and virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, ACM (2002), 203–210.

[9] Cockburn, A., Quinn, P., Gutwin, C., Ramos, G., and Looser, J. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. *International Journal of Human-Computer Studies* (2011).

[10] Company, N. The state of mobile apps, 2011.

[11] Czerwinski, M., Van Dantzich, M., Robertson, G., and Hoffman, H. The contribution of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3d. In *Proc. Interact*, vol. 99 (1999), 163–170.

[12] Delamare, W., Coutrix, C., and Nigay, L. Pointing in the physical world for light source selection.

[13] Droeschel, D., Stückler, J., and Behnke, S. Learning to interpret pointing gestures with a time-of-flight camera. In *Proceedings of the 6th international conference on Human-robot interaction*, ACM (2011), 481–488.

[14] Dutta, T. Evaluation of the kinect sensor for 3-d kinematic measurement in the workplace. *Applied Ergonomics* (2011).

[15] Egan, D., and Gomez, L. Assaying, isolating, and accommodating individual differences in learning a complex skill. *Individual differences in cognition 2* (1985), 173–217.

[16] Gustafson, S., Holz, C., and Baudisch, P. Imaginary phone: learning imaginary interfaces by transferring spatial memory from a familiar device. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, ACM (2011), 283–292.

[17] Izadi, S., Newcombe, R., Kim, D., Hilliges, O., Molyneaux, D., Hodges, S., Kohli, P., Shotton, J., Davison, A., and Fitzgibbon, A. Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In *ACM SIGGRAPH 2011 Talks*, ACM (2011), 23.

[18] Kinect, M. http://www.xbox.com/fr-fr/kinect.

[19] Kurtenbach, G., and Buxton, W. Issues in combining marking and direct manipulation techniques. In *Proceedings of the 4th annual ACM symposium on User interface software and technology*, ACM (1991), 137–144.

[20] Li, F., Dearman, D., and Truong, K. Virtual shelves: interactions with orientation aware devices. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, ACM (2009), 125–128.

[21] Nickel, K., and Stiefelhagen, R. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing 25*, 12 (2007), 1875–1884.

[22] NITESCU, D., Lalanne, D., and Schwaller, M. Evaluation of pointing strategies for microsoft kinect sensor device.

[23] Oakley, I., and Park, J. A motion-based marking menu system. In *CHI'07 extended abstracts on Human factors in computing systems*, ACM (2007), 2597–2602.

[24] Pan, G., Ren, H., Hua, W., Zheng, Q., and Li, S. Easypointer: what you pointing at is what you get. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, ACM (2011), 499–502.

[25] Poage, M., and Poage, E. Is one picture worth one thousand words? *The Arithmetic Teacher 24*, 5 (1977), 408–414.

[26] Purcell, K. Half of adult cell phone owners have apps on their phones. *Pew Research Center's Internet & American Life Project. Accessed January 9* (2011), 2012.

[27] Raheja, J., Chaudhary, A., and Singal, K. Tracking of fingertips and centers of palm using kinect. In *Computational Intelligence, Modelling and Simulation (CIMSiM), 2011 Third International Conference on*, IEEE (2011), 248–252.

[28] Robertson, G., Czerwinski, M., Larson, K., Robbins, D., Thiel, D., and Van Dantzich, M. Data mountain: using spatial memory for document management. In *Proceedings of the 11th annual ACM symposium on User interface software and technology*, ACM (1998), 153–162.

[29] Scarr, J., Cockburn, A., Gutwin, C., and Bunt, A. Improving command selection with commandmaps. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, ACM (2012), 257–266.

[30] Tavanti, M., and Lind, M. 2d vs 3d, implications on spatial memory. In *Symposium on Information Visualization* (2001), 139–145.

[31] Times, L. Xbox now used more for online entertainment than online gaming, March 26, 2012.

[32] Xiang, X., Pan, Z., and Tong, J. Depth camera in computer vision and computer graphics: an overview. *Jisuanji Kexue yu Tansuo 5*, 6 (2011), 481–492.

[33] Yates, F. *The art of memory*, vol. 64. Random House UK, 1992.

[34] Zhao, S., and Balakrishnan, R. Simple vs. compound mark hierarchical marking menus. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*, ACM (2004), 33–42.