Multi-modal Multi-scale Attention Guidance in Cyber-Physical Environments

Guillermo Reyes
guillermo.reyes@dfki.de
German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus

Alexandra Alles
alexandra.alles@dfki.de
German Research Center for Artificial Intelligence (DFKI), Saarland University

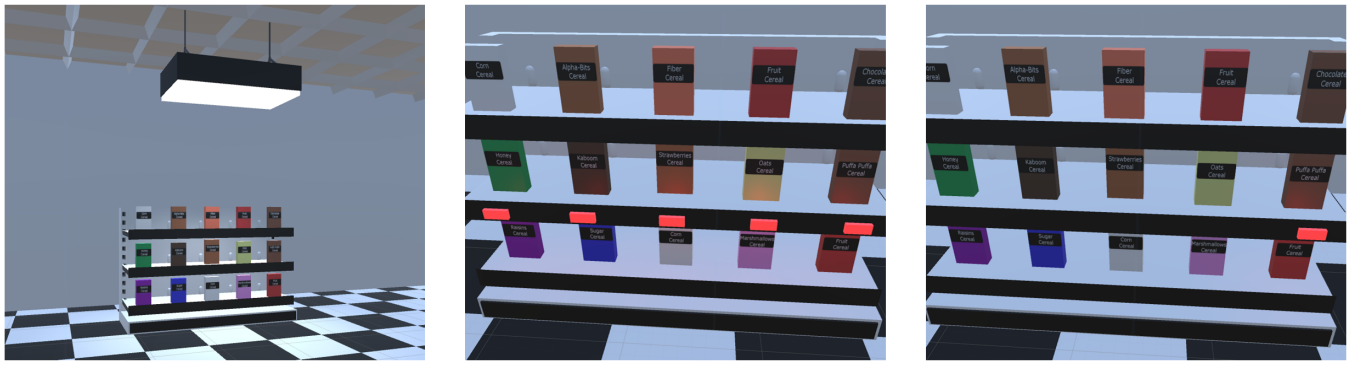


Figure 1: Attention guidance for three different scales.

ABSTRACT

This work proposes a new method for guiding a user's attention towards objects of interest in a cyber-physical environment (CPE). CPEs are environments that contain several computing systems that interact with each other and with the physical world. These environments contain several sensors (cameras, eye trackers, etc.) and output devices (lamps, screens, speakers, etc.). These devices can be used to first track the user's position, orientation, and focus of attention to then find the most suitable output device to guide the user's attention towards a target object. We argue that the most suitable device in this context is the one that attracts attention closest to the target and is salient enough to capture the user's attention. The method is implemented as a function which estimates the "closeness" and "salience" of each visual and auditive output device in the environment. Some parameters of this method are then evaluated through a user study in the context of a virtual reality supermarket. The results show that multi-modal guidance can lead to better guiding performance. However, this depends on the set parameters.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Ambient intelligence**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IUI '21, April 14–17, 2021, College Station, TX, USA
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450678

KEYWORDS

Attention Guidance, Attention, Cyber-Physical Environments, Intelligent Environments, Multi-modal, Multi-scale

ACM Reference Format:
Guillermo Reyes and Alexandra Alles. 2021. Multi-modal Multi-scale Attention Guidance in Cyber-Physical Environments. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397481.3450678

1 INTRODUCTION

Throughout the years, we have seen the evolution of computing give rise to cyber-physical systems (CPS). CPSs are physical and engineered systems whose operations are monitored, coordinated, controlled, and integrated by a computing and communication core [22]. These systems are capable of interacting with the physical world [24]. As CPSs become more and more common, certain environments will contain several of them. These systems can then be integrated to become a cyber-physical environment (CPE). A CPE consists of a large number of CPSs distributed in a local environment (factories, homes, cars, etc.). Part of the challenges these environments entail is that they must be dynamic and context-aware. In other words, they must monitor and react to changes in the environment [23]. This is done by reading information from the physical environment through a variety of sensors, identifying its current state, and inducing some change in the environment by the use of actuators or output devices.

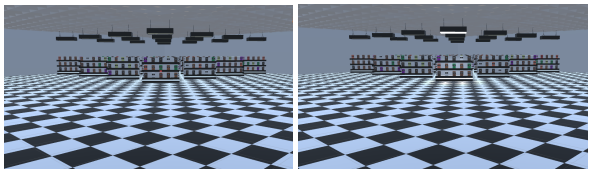
In certain situations, users in CPEs need to find or be aware of certain objects. For example, an apprentice in a car workshop who needs help finding a tool. In a smart home environment, the system might want to let the user know he left a window open before going to bed. A smart car could warn the driver of a pedestrian suddenly

attempting to cross the street. Whatever the case, the user needs to be made aware of the location of a certain object or area.

People communicate verbally and non-verbally, such as when pointing at things they are talking about [12]. Unfortunately, CPEs have had, so far, no way of pointing at the things they are referring to when interacting with humans. In order to breach this gap in communication between humans and CPEs, *attention guidance* is necessary.

In this context, attention guidance refers to the deliberate shifting of a user's focus of attention (FoA) towards an object or area of interest (target highlighting) or away from distracting ones (distraction avoidance). Guiding attention towards a target object could be accomplished by making the object more salient than its surroundings (see figure 1).

These target objects are not always located in the user's vicinity or range of perception. A small LED display may attract the user's attention when standing three meters away from it, but be barely visible at a distance of 50 meters. Nevertheless, there could be other devices that are more suitable to guide the user's attention at such a distance (see figure 2). In such situations it becomes necessary to use a *multi-scale* attention guidance method that determines the *most suitable device* to guide the user's attention towards the target at any given point. This method should take into account different parameters about the environment, the user, the target, and each of the devices' properties (position, effective ranges, etc.).



(a) Barely visible red display (b) clearly visible ceiling lamp

Figure 2: An example of how a different kind of device may be more suitable than another depending on the user-target distance. *Left*: an electronic label is not perceived from far away. *Right*: at the same distance a ceiling lamp succesfully attracts attention.

It is important that the system also takes the devices' modalities into account. Not only are certain modalities not suitable for some users (e.g. auditive devices for deaf users), but their usefulness could also vary with the situation. A visual device behind a user would not be as suitable as an auditive one to guide the user's attention.

Another challenge is that CPEs are very dynamic. Users are often moving and interacting with multiple CPSs at the same time. Thus, an attention guidance method should consider this and adapt the guidance accordingly.

The contributions of this work are twofold: first, we describe a novel attention guidance method that quantitatively ranks the suitability of visual and auditive devices for a given user and target. Second, we discuss the results of the first user study using this method, which investigates some of the parameters of this method.

The next section will describe some of the background and related work in the area of attention guidance. Section 3 will disclose the developed method, explaining each of the parameters of the main functions. Then, sections 4 and 4.4 will describe an experiment that validates some of these parameters and discuss the main findings. Finally, section 5 summarizes this work and proposes possibilities for future research directions.

2 BACKGROUND AND RELATED WORK

To understand how we are able to find an object in our surroundings, cognitive mechanisms like selective attention are necessary. Selective attention is the process of selecting relevant and deselecting irrelevant information. Two major factors influence the focus of attention: top-down and bottom-up. While top-down processes are goal-driven and regulated by our intentions and expectations (endogenous), bottom-up processes are object-driven and attentional shifts occur involuntarily (exogenous) [10, 19, 21]. An important factor is the salience of an object. As stated in Wolfe [29] "Salience is the signal to noise ratio". Therefore, when searching for an object, its salience represents the differences between objects in each feature dimension (e.g. color) [20, 29]. In terms of guided attention, visual and/or auditive cues play an important role. Cues don't always attract attention to a single point in space, in fact, they can attract attention to an area of certain dimensions. There is evidence that objects that are located inside the focus of attention (FoA) can be detected more easily [9, 21]. This area is adjustable and the wider it is, the harder it is to locate targets within it [9]. Contrary to location-based selective attention (FoA is shifted to an area), there exists as well object-based attention (FoA is shifted to an object) [6]. There is evidence that both processes work at the same time and parallel to each other [7, 13]. Other research has focused on Cross-Modal Cueing i.e., cues presented in another sensory modality than the target (e.g. auditive cue for visual target). Several other studies have shown that auditive cueing can influence the visual FoA and if valid enhance the detection performance [2, 5, 25].

Selective attention has been of great importance for wayfinding (knowing where you are and where you want to go and how to get there) [4]. For instance, in videogames, visual and auditive cues lead the players attention to find places and objects. A comparison of five types of cues found that, while some cues are obvious and easy to follow, others are not very precise [27]. However, a key aspect of designing wayfinding systems for games is that they are not meant to be as clear as directions found in the real world [18]. By using several game parameters such as objects in the scene, objective, camera position, etc., the ALVA (Adaptive Lighting for Visual Attention) system was able to determine which game objects were important and dynamically highlight them by using in-game lighting [8]. An advantage of games and virtual worlds is that the designer has full knowledge and is in complete control of the environment. Cues to guide attention can be placed anywhere in the environment. In the real world, however, knowledge of the environment is limited to whatever information sensors provide, and control over it will depend on the actuators and output devices found therein. In the real world, attention manipulation is often done manually. However, computer-based approaches have also been implemented in the real world. Booth et. al. [3] used a technique called Subtle Gaze Direction [1]. A limitation of their approach was the simplicity of the scene. Environments such as CPEs can be very complex. Project

REAL [26] tried to assist users in instrumented environments with navigation and shopping. Two kinds of navigation were described: macro- and micro-navigation. Macro-navigation was used when the goal was outside the user's perception. Directions were given in either a PDA or public displays. Micro-navigation tried to guide the user's FoA to a spot within their range of perception using a steerable projector. Similarly, the IRL SmartCart [15] also adopted these kinds of navigation. However, macro-navigation was handled with an on-screen map that displayed the route from a user's current position to their target destination. Both of these works required specialized hardware that might not be available or even applicable to many CPEs. A better alternative would be a guiding method that can be adapted to different kinds of environments and output devices and that automatically chooses the most suitable device to guide the user.

3 CONCEPT

Although related work on the topic of attention guidance exists, corresponding research for CPEs is still lacking. The reviewed methods require specialized hardware or are simply not applicable to some environments. This section describes a multi-modal, multi-scale, attention guidance method applicable to all CPEs that have controllable audio and visual devices, and where the position and orientation of the user relative to these devices is known. The next few sections describe the suitability function of an attention-management system in charge of selecting the devices with which to guide the users' attention. We define this function with primary and secondary functions. The range of all primary and secondary functions is kept to be within $[0, 1]$. Primary functions (sec. 3.1) are the main idea of this attention guidance method. Secondary functions (sec. 3.2) expand the primary functions and show how the method might work in general. An important feature of this distinction is that secondary functions could be easily substituted by other functions as long as the range of the function is kept the same. These functions could then be better suited for a specific CPE. Secondary functions for the auditive modality are then discussed in section 3.3 and additional details of the suitability function are given in section 3.4.

3.1 Primary functions

The role of an attention management system is to choose the *most suitable device* d^* to guide the user's attention towards a target area or object. For this, it is necessary that it takes into account user properties u like their position, orientation or FoA and the target properties t. If one could represent the suitability of a device as a function $\mathcal{F}$, then one could simply select the device that maximizes this function:

$$d^* = \underset{d \in \mathbb{D}}{\operatorname{argmax}}\, \mathcal{F}(d, u, t; \mathbb{C}) \tag{1}$$

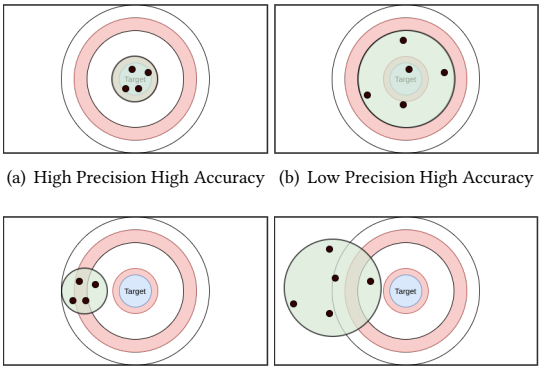
where $\mathbb{D}$ is the set of all output devices in the CPE and $\mathbb{C}$ is a set of context parameters about the environment. Unfortunately, even if such a function did exist, it would still be difficult to envision what it could look like in detail. We argue that such a function would consist of at least two things: the salience of the device and its closeness to the target. If the device is not salient enough, the user will not perceive it. If the device is not close to the target, cueing the user will distract or lead the user away from the target. One could further imagine a kind of trade-off between these two aspects: when a user is far away from the target, saliency might be more important and how close a cue is to the target may be less relevant. However, as the user approaches the target, the discrimination of these devices' closeness to the target becomes much more relevant, as otherwise the user's attention might be guided in the wrong direction. If one could quantify how salient a device's cue is to a user and how close it leads the user to the target through two functions $\mathcal{S}$ and C respectively, one could then define a function $\mathcal{F}$ as follows:

$$\begin{aligned}\mathcal{F}(d, u, t; \mathbb{C}) =& (1 - w_i)\Big(w_s \cdot \mathcal{S}(d, u; \mathbb{C}) + (1 - w_s) \cdot C(d, u, t; \mathbb{C})\Big) \\ &+ w_i \cdot \Big(\mathcal{S}(d, u; \mathbb{C}) \cdot C(d, u, t; \mathbb{C})\Big)\end{aligned} \tag{2}$$

where $+w_i$ is the *interaction weight*.

As can be seen from equation (2), as the weight assigned to saliency (w_s) decreases, the weight assigned to closeness ($(1 - w_s)$) increases. In the same way, the interaction term *interaction weight* controls how important it is that a device is both salient and close at the same time. By decreasing this weight more importance is given to salience and closeness as individual components. We now describe what is meant by these two terms in more detail, starting with the visual modality.

One could understand closeness as an estimate of how well a cue might lead to the target. We argue two things are important in this regard: accuracy and precision (see eq. 3 and figure 3).



(a) High Precision High Accuracy (b) Low Precision High Accuracy

(c) High Precision Low Accuracy (d) Low Precision Low Accuracy

Figure 3: Depiction of accuracy and precision. The area to which the cue attracts attention is denoted by a green circle. The target is denoted by a blue circle. 3(a): A perfectly close cue. It guides a user's attention to the target and nothing else. 3(b): A potentially acceptable cue, as it attracts attention to the target, though it also attracts attention to other nearby-objects. It could be good to attract attention to a general area 3(c): A very specific cue targets an area where the target is not located. Depending on the distance to the target, this cue might still be relatively good. 3(d): A cue that is neither precise nor accurate. This cue will probably guide the user's attention away from the target.

We assume that, depending on how well the cued area fits the target and how similar the cued area is to the target's size, the cue will guide the user's attention to the target differently. Consider the target shooting analogy in figure 3 where one could interpret each of the black dots as a saccade or the gaze of a user looking for a target within a certain period of time. If the cue is perfectly accurate and precise, the user's attention is guided directly to the target. However, if the cue lacks precision, the user's attention gets guided to a larger area near the target, but the user still has to scan this area to find the target. On the other hand, if the cue is precise but lacks accuracy, the user's attention gets guided towards a specific object or area that is not the target. It could then take some time for the user to realize that the guidance is not accurate and start looking around outside the cued area for the target. The time it takes the user to find the target would depend on whether or not the user knows what they're looking for and how inaccurate the cue was. Finally, if the cue is neither precise nor accurate, the user will be looking for the target in a completely wrong area. One could, thus, define the closeness function as follows:

$$\mathcal{C}(d, u, t; \mathbb{C}) = w_{\mathrm{p}} \cdot \mathcal{P}(d, u, t; \mathbb{C}) + (1 - w_{\mathrm{p}}) \cdot \mathcal{A}(d, u, t; \mathbb{C}) \quad (3)$$

where $\mathcal{P}$ and $\mathcal{A}$ are functions that determine how precise and accurate a cue is respectively and w_{p} is the weight assigned to precision. These functions are further explained in section 3.2.

The second part of equation 2 is salience. Salience could be understood in this context as the "probability" that a cue will be detected by the user. It is certainly difficult to quantify users' knowledge, goals, and expectations such that one can estimate whether or not a user will see a cue or not. Salience, on the other hand, could be manipulated. Stimuli can be different across several feature dimensions like luminance, size, color, orientation, etc., but also in others like loudness and frequency in the auditive channel. Although it could be possible to create an accurate computational model that predicts how salient a cue will be to a user, it is difficult to do so outside a controlled environment and in particular in a CPE. This is mainly because of the lack of information in a very complex and dynamic environment. This work does not claim to objectively predict exactly how salient a certain cue might be for a user. As it stands, the saliency function simply outputs a confidence that the user will detect the cue produced by a certain device based on some heuristics. Saliency could be decomposed into different conspicuity factors such as intensity, color, orientation, etc. If not enough is known about the environment one could then make some assumptions about it and what kind of features will be salient in it. For some features, it's difficult to make assumptions about what the environment might look like regarding e.g. color or orientation. Here, we make the assumption that brighter and larger cues attract more attention. Luminance is the first type of information extracted by our visual systems and luminance contrast seems to be the primary variable on which visual salience computation is based [28]. Also, increasing the size of a non-salient target decreases the time it takes to locate it, compensating for the lower saliency. Not only that, but it seems as if the function relating increase of salience to feature value ratios is similar for the size and salience dimensions [14]. Having selected these two parameters, one could define the saliency function to be:

$$\mathcal{S}(d, u; \mathbb{C}) = w_{\mathrm{z}} \cdot \mathcal{Z}(d, u; \mathbb{C}) + (1 - w_{\mathrm{z}}) \cdot \mathcal{B}(d, u; \mathbb{C}) \quad (4)$$

where $\mathcal{Z}$ and $\mathcal{B}$ are functions that determine how salient a cue will be to the user depending on its size and brightness respectively.

If more is known about the environment or if at some point a new method could predict how salient a cue will be for a user, or how likely a user is to pay attention to the cue either by top-down or bottom-up factors it could be plugged into this method and improve it without having to change any other function. For this, the range of the function should be consistent with the other functions and be set to $[0, 1]$.

3.2 Secondary functions

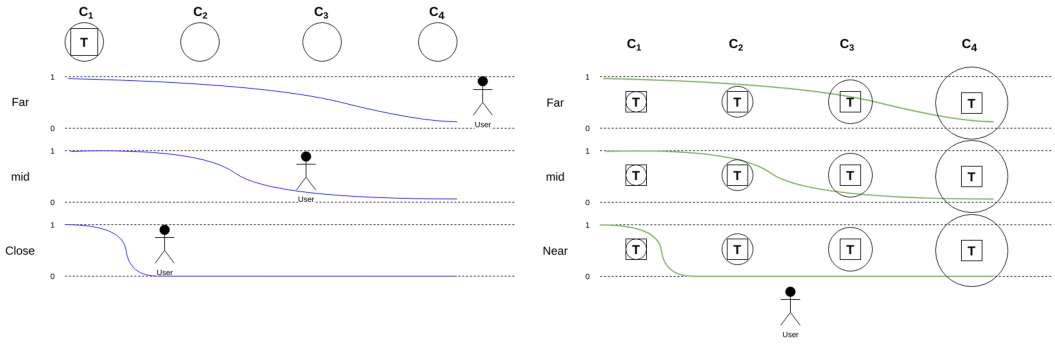
As previously mentioned, secondary functions expand on and further explain primary functions. However, the exact mathematical definition of these secondary functions is outside the scope of this work. Instead, we abstract from the details and provide an intuition on how these functions work. We start by explaining the closeness function (eq. 3).

In general, the intuition of whether two objects are close to each other or not depends on what the distance between both objects is being compared to. The distance to the moon might be considered close when compared with the distance to the sun. While closeness could in general be considered a target-cue relation, it is also important to compare it to something. One could compare it to the distance from the user to the target (see figure 4(a)). If the user is far away from the target, the attention guidance system should be less "picky" with what makes up a close cue. As the user gets closer to the target, it becomes more and more important, that the system outputs more accurate cues that guide the user in the right direction.

A sigmoid function could be adapted between the target and the user. When the user is far away, the sigmoid could look almost linear, resulting in far away cues being still relatively accurate, but still discriminating between cues that are farther away from those that are closer. As the user gets closer to the target, the sigmoid is adapted in such a way, that farther cues' accuracies are dropped to nearly zero. One could also adapt what is considered precise as the user moves in the environment in a similar way to the accuracy function.

An important distinction is that, while the accuracy function evaluates the distance from a cue to the target in meters, the precision function evaluates the relative error in the size of the area highlighted by the cue. The relative error is the difference in size between the highlighted area and the target proportional to the target. This could again be implemented with a sigmoid function that is adapted to the distance from the user to the target (see figure 4(b)). When the user is near the target, only the cues with the most similar size will be considered precise. But if the user is far away, even cues that highlight areas are much larger than the target are considered somewhat precise.

Having both the accuracy and the precision functions, one can then calculate how close all cues are to the target in terms of accuracy and precision and in relation to the user. Now we focus on the saliency function (eq. 4). The size of the cue is not just relevant for precision, but also for the saliency function. However, the focus



(a) Distance-dependent Accuracy

(b) Distance-dependent Precision

Figure 4: 4(a): Accuracy is adapted depending on the distance from the user to the target (marked by a 'T'). Three scales are presented. As the user approaches the target, the accuracy of the cues (noted by circles c_1 to c_4) farther away decreases to zero. 4(b): Shows how the perception of what cues are precise becomes more relaxed the farther away the user is from the target.

here is on the user-cue relation in contrast to the target-cue relation when talking about precision. In this case, what matters is how big cues are perceived by the user. A smaller cue might be perceived by the user as the same size as a larger cue that is farther away. To account for this, one could consider the angular size of the cue, instead of the actual size. The result could then be again rescaled with the sigmoid function to ensure that the range of the function remains between 0 and 1.

Regarding brightness, we should distinguish between luminosity and brightness. Whereas the first is the constant amount of light emitted by a source, the second is the amount of light perceived at a certain distance. To calculate this, one could apply the inverse square law. The inverse square law implies that as the observer gets closer to the source of light, the perceived brightness will tend towards infinity, so in order to ensure that the range remains between 0 and 1 as in the other functions, the output of the function is again rescaled.

Finally, one should also consider that, in case of visual cues, they are only visible if they are within the user's field of view. One could then use an activation function that outputs 1 if this is the case or 0 if not and multiply the output by equation 4. This would result in the visual device having 0 salience whenever its produced cue lies outside the user's FoV. Visibility is of course not a problem for auditive cues. In the next section, the secondary functions are revisited with a focus on the auditive modality.

3.3 Auditive Modality

Visual devices are not the only ones capable of attracting human attention. A great advantage of auditive cues over visual ones is that they are not limited by peripheral vision, but can still attract the user's attention and orient the user in the right direction. The user could then be guided further by the same auditive cue, or by a visual one which was previously not within the user's field of view. While visual and auditive cues are inherently different, they do have certain properties in common and, in particular, they can be closer or farther away from the target and they can be salient or not for the user. In principle, this means their suitability could be represented in the same equations as for the visual modality, with slight adaptations.

In regards to closeness, if the position of an auditive cue is considered to be the device's position, one could directly calculate accuracy by applying the same equation as in the visual modality. For precision, on the other hand, it is not entirely clear what the dimensions of the focus area should be. However, since sound sources don't need to lie within the user's field of view, they could be used to attract the user's attention in roughly the right direction and letting visual cues take over the guidance as soon as this happens. When talking about auditive cues, it might be enough for the user to know the general direction of the target and to define precision by how precisely a cue points the user in that direction. The angle between the cue and the target could then be scaled with a sigmoid function just like the other functions (see figure 5).

The other component would be the saliency of the device. Here again, the definition of visual cues does not apply to auditive ones. However, just like with brightness for visual cues, we make the assumption that louder cues attract more attention. Instead of calculating brightness, we calculate the sound pressure level (SPL) which describes how loud a sound is perceived at a certain distance, which follows the inverse square law, similar to brightness. To scale the result to be in a range between 0 and 1, we first define a constant in decibels that is considered maximally salient and then scale it again using the sigmoid function. This is because very loud noises can cause discomfort or even harm humans.

3.4 Multi-scale transition

Up to this point, we have discussed a function that selects the most suitable device to guide the user's attention towards a target in the environment. An important question would be how the transition from one device to the other takes place as the user gets closer to the target. If the system always selects the most suitable device, space would be partitioned into regions with hard borders such that crossing a region's border would switch the selected device. This could make the system very volatile. Due to imprecision in the system, standing still at one of these hard borders could switch

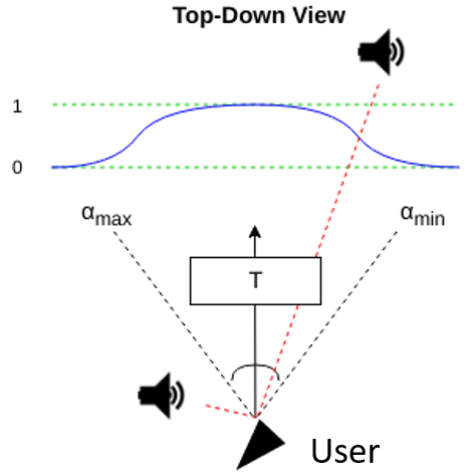


Figure 5: A graphical representation of the precision function for the auditive modality. A user (black triangle) is presented with two auditive cues (speaker icons) for the target (T). The user is depicted as if looking away from the target in the direction of the triangle. As the angle difference between the target and the device increases, the precision of the cue decreases in a sigmoid-like fashion. The left speaker shows a cue that is relatively accurate but not precise and the right speaker a cue that is relatively precise, but not accurate.

the selected device back and forth. Moving quickly through the environment could make devices switch immediately after being selected, constantly attracting the user's attention in different directions. This could be distracting and annoying for the user but also confusing and lead to difficulties locating the target. To avoid this, one could take inspiration from the way thermostats work: the heating is turned on after the temperature drops to a certain level and is kept on until it is hot enough. Once this happens, the temperature starts decreasing, but the heating is not immediately turned on until a certain threshold is reached. For multi-scale attention guidance, what could be done is adding a margin λ. To be selected, a device would have to be at least more suitable than the currently selected device plus this margin. By doing this, the current guiding device is kept until a significantly better device is found:

$$d_t^s = \begin{cases} d & \text{if } \mathcal{F}(d,u,t;\mathbb{C}) > \mathcal{F}(d_{t-1}^s,u,t;\mathbb{C}) + \lambda \\ d_{t-1}^s & \text{otherwise} \end{cases} \quad (5)$$

where d_t^s is the device to be selected at time t and d_{t-1}^s is the currently selected device. This would of course mean that once a device has a suitability larger than $1-\lambda$ no other device would be selected. Since it would be desirable that at some point the best guiding device does get selected, selected devices gain momentum for a short time. The margin λ is active only while the device has momentum and is slowly decreased as momentum is lost. This ensures that after momentum is lost, a more suitable device is selected if such a device exists.

4 STUDY

The multi-modal and multi-scale attention guidance method described above was evaluated in a virtual reality (VR) supermarket environment. For these purposes, we conducted an experiment in which participants were asked to find a target cereal box. Users had to walk in the virtual supermarket until they found the shelf and the target cereal, which was cued by the system according to a certain combination of conditions.

4.1 Hypothesis

The goal of this study was to examine the effectiveness of the multi-modal multi-scale guiding method, as well as the effects of the margin and interaction weight. This led us to the following hypotheses:

- A mixed guiding method of visual and auditive cueing leads to better guiding performance compared to a guiding method with only using visual cues.
- No interaction weight (w_i= 0) leads to a worse guiding performance compared to conditions with w_i of 0.5 and 1.
- No margin (λ=0) leads to a worse guiding performance compared to the condition with a λ of 0.2.
- No margin makes a more volatile system compared to conditions with a margin of 0.2.

4.2 Method

4.2.1 Demographics. A group of 46 participants took part in the study. One participant could not complete the experiment due to dizziness and one third of the data of two participants was lost due to an error during the evaluation. These cases were excluded from the analysis of the data. The remaining participants (22 female, 20 male, 1 not specified) had a mean age of 24.95 years old (*SD* = 4.11). The genders showed equal distributions across the conditions. Visual impairments[1] were reported by 39.5% of the participants. The majority reported seldom to no previous interaction with VR (n=39). The average dizziness level was 2.84 (*SD* = 1.49) in a 1 to 7 rating scale (1=not at all dizzy, 7= very dizzy). Participants reported only *little* (n= 17) or *no* (n= 26) difficulties seeing clearly in VR.

4.2.2 Setting. The participants started in the center of the environment, surrounded by four groupings of 4x4 supermarket shelves with a rotation of either 0° or 45°. To ensure a standard starting orientation, participants had to fixate a red ball for two seconds at the beginning of each trial. Each shelf contained 15 different labeled cereal boxes (see fig. 6).

Before the start of each trial, all the products on the shelves were randomized and a randomly selected target was placed on one of them. The visual cues could be either a lamp located over the shelf (area light), a row of red lights indicating the level of the shelf (group light), or a single light (point light) directly below the cereal box (see fig. 1). In addition, auditive cues in the form of speakers behind the cereal boxes were used to guide the participant. The salience weight w_s, was set to 0.5, as it is not clear whether or not

[1] All participants performed a sight test before the experiment and were able to read the boxes' labels in the virtual environment. Participants with corrected vision were allowed to wear glasses or contact lenses while performing the sight test and the experiment.

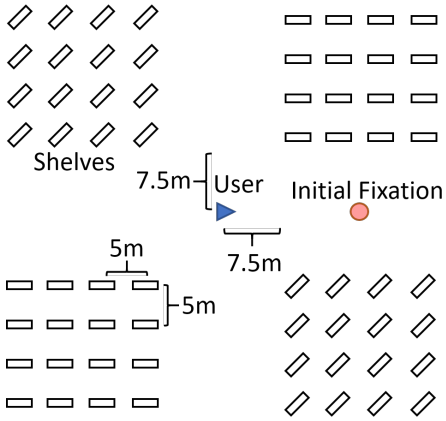


Figure 6: Top-down view of the evaluation environment. The user (blue triangle at the center of the map) looks at a red ball for 2 seconds to start the experiment. The positions of the shelves containing the four kinds of devices are indicated by the rectangles.

salience or closeness should be given more importance. For the same reason w_p, and w_z for the visual modality were also set to 0.5.

4.2.3 Study Design. For the experiment, we used a between-within fractional factorial design, in which the modality (between factor), as well as the interaction weight w_i and the margin λ (within factors), was varied. The interaction weight was set at either 0, 0.5, or 1 while the margin was set at 0 or 0.2. Due to the chosen experimental design, we decided to vary interaction weight when the margin was set at 0.2 and vary the margin when interaction weight was set to 0.5. This resulted in four conditions: one in which the margin was 0 and interaction weight was 0, 5, and three conditions where the margin was 0.2, and interaction weight was either 0, 0.5, or 1. Participants were randomly assigned to a cueing modality: In the visual-only modality, participants were only guided by visual cues (area light, group light, and point light). Likewise, participants in the mixed modality received additional guiding information through auditive cues.

4.2.4 Procedure. Participants were asked to imagine that they were employees in a supermarket and that their task was to collect products for a delivery service. Thus, they should work as swiftly as possible. A random permutation of the experiment condition was generated for each participant. After the explanation of the task and the controls, the participants had to perform four test trials (one for each of the conditions). The main experiment consisted of 20 trials (5 of each condition).

At the beginning of each trial, participants were given the label of the target and had to fixate a red ball. This indicated the starting position and orientation. Both, the ball and the target label disappeared after 2 seconds of fixating the ball. Participants were then able to move in the environment using an HTC Vive controller at a maximum speed of $3m/s$. Once the participants were close enough to the target, they were allowed to move physically in an area of 2.2m × 2.2m and grab the product with the controller. As soon as the participants grabbed the correct product the trial ended and they were transported back to the middle of the map. Visual feedback was given if the wrong product was chosen. Participants were also able to re-display the target label by pressing a button on the controller at any point during the experiment run.The participants were told they could take a break after the first half of the experiment as well as any time they felt dizzy or tired. After the experiment, the subjects were asked to answer a short questionnaire. Once they were done, they thanked and rewarded with 5€ for their participation.

4.2.5 Measurements. We hypothesize that a multi-modal multi-scale guiding methods leads to a more effective guiding performance. Effective performance was defined as fast task-solving and less walking distance. We observed, that a lot of participants, had only little experience with VR and that many reported a feeling of discomfort and dizziness in the experimental situation. This might have impacted the time a participant needs to find a target. Taking into account that the walking distance has a high influence on the time it takes to solve a task, we decided to measure guiding performance as the distance that participants traveled to a target in meters (*distance traveled*). For our calculations, we included the mean of the trials for each condition. The initial travel distance to the target was randomized and should have no influence on the results. This is supported by the fact that *extra travel distance* (*distance traveled − initial distance* to the target) and *distance traveled* are highly correlated (r=.98, p<.001).

We assumed that in the no-margin condition more cues occur due to the subject's movement and that the system becomes more volatile as a result. Hence, as an indicator of a more volatile system, we measured the number of cues presented in a trial (*number of cues*). Similar to guiding performance, we calculated a mean for the *number of cues* per condition. All dependent variables were normally distributed.

4.2.6 Apparatus. The test environment was developed using Unity 3D version 5.6. As user input, Unity's standard first-person controller was taken as it provides all required movements out of the box. An obvious shortcoming of this is that the user's gaze direction cannot be tracked, so it was assumed to be the camera's directional vector. The input and output of the application were integrated through a VR headset. The VR headset used in this study was the HTC Vive which has a refresh rate of 90Hz, an FoV of about $110^\circ \times 110^\circ$ and supports spatial audio. It is capable of tracking motion from the headset and the controllers in a room with the help of two base stations located at opposite corners of the room. The experiments were conducted with an MSI GT73VR 6RF Titan Pro with an NVIDIA GeForce GTX 1080 graphics card, an Intel Core i7-6820HK CPU with a clock speed of 2.70GHz, and 32GB of RAM running Windows 10 64-bit.

4.3 Results

The reported effects were statistically significant associated with a *p value* below .05. In average the participants traveled a distance of 35.29m (*SD*=4.86). The values for distance traveled per condition are shown on table 1.

To estimate the effects of interaction weight and modality we calculated a MANOVA to compare the differences in *distance traveled*

Table 1: Descriptive statistics

condition	Mixed M(SD)	Visual-only M(SD)	total M(SD)
$\lambda(0.0) \times w_{\mathrm{i}}(0.5)$	35.32 (7.95)	34.51 (5.30)	34.90 (6.62)
$\lambda(0.2) \times w_{\mathrm{i}}(0.0)$	33.74 (4.40)	36.81 (7.99)	35.31 (6.60)
$\lambda(0.2) \times w_{\mathrm{i}}(0.5)$	33.12 (3.38)	39.02 (9.79)	36.21 (7.95)
$\lambda(0.2) \times w_{\mathrm{i}}(1.0)$	33.63 (4.53)	36.05 (5.37)	34.87 (5.07)

Mean and standard deviation of distance traveled for each condition in mixed and visual-only modality and in total for all participants. M = Mean, SD = Standard Deviation, λ = margin, w_{i} = interaction weight.

for the three interaction weight levels in each modality under a constant margin of λ= 0.2. We found a significant effect between the two modalities (F(1,41)=5.97, p=.019, η_p^2=.127). Participants walked less, when guided with auditive and visual cues compared to only visual guiding (see fig. 7(a)). The interaction weight had no significant effect on distance traveled (*Pillai-trace*=.03, F(2,40)=0.70,p=.503, η_p^2=.034). There was no significant interaction between interaction weight and modality (*Pillai-trace*=.09, F(2,40)=1.20,p=.149, η_p^2=.091).

A second MANOVA was calculated to examine the effects of the margin in *distance traveled*, while interaction weight was on w_{i}=0.5. There was no significant effect neither for margin (*Pillai-trace*=.02, F(1,40)=0.80, p=.376, η_p^2=.020) nor modality (F(1,40)=2.07, p=.158, η_p^2=.049). Instead a significant interaction between margin and modality was found (*Pillai-trace*=.15, F(1,40)=6.79, p=.013, η_p^2=.145). Figure 7(b) shows that when the margin was λ=0.2 participants walked less in the mixed modality (M=33.12, SD=3.38) compared to the visual modality (M=39.02, SD=9.79). However, when the margin was set to λ=0 the difference between the modalities decreased and in average visual-only guidance (M=34.51, SD=5.30) was more effective than mixed guidance (M=35.32, SD=7.95).

A third MANOVA was calculated to analyse the effect of the margin on the *number of cues*. The margin showed a significant effect on these variables (*Pillai-trace*=.83, F(1,41)=194.97, p<.001,η_p^2=.826). The no-margin condition was assosciated to significantly more cues (M=25.16, SD=4.71) compared to the 0.2-margin condition (M=14.05, SD=4.61).

The effect of the number of cues on *distance traveled* were held constant in an explorative MANCOVA with modality and margin for *distance traveled*. When controlling for *number of cues*, the significant interaction of margin and modality remained significant (*Pillai-trace*=.12, F(1,39)=5.11, p=.029, η_p^2=.116).

4.4 Discussion

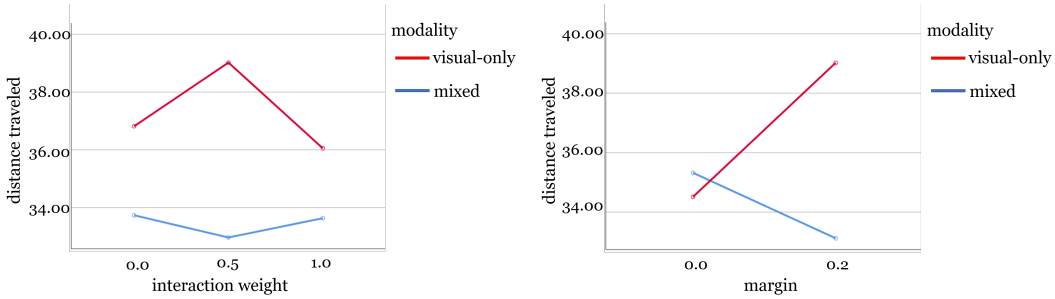
In this user study, we aimed to evaluate a multi-modal, multi-scale attention guiding method, while varying the margin and interaction weight for each participant, as well as the modality between participants. The results suggest that when a margin is present (λ=0.2), a better guiding performance can be found in the visual-auditive guiding method compared to a guiding method using visual cues only, independently of the interaction weight. This partially confirms our first hypothesis, which stated both visual and auditive cues leads to better guiding performance, than using visual-only cues. This is in accordance with the literature [2, 5, 25]. It is surprising to see that, this benefit was lost in case there was no margin (λ=0). Performance in the mixed condition and visual-only condition was comparable when λ=0. Instead, in the presence of a margin (λ=0.2) and an interaction weight $w_i(0)$=0.5, the performance in the mixed modality was slightly improved, while in the visual-only modality it declined. The reason for this interaction effect is not yet completely explained.

According to our hypothesis, the absence of a margin should lead to a more volatile system. This hypothesis was confirmed. In the no-margin condition, there was a significantly higher number of cues compared to a 0.2-margin condition. We further examined the influence of the number of cues on distance traveled. The interaction of margin and modality remained significant, while the effect size of the interaction between margin and modality was lower when controlling for number cues. This might indicate that the present effect could be partly explained by the number of cues.

For the *visual-only* condition, the reduced number of cues under the presence of a margin could imply less regularly updated guiding information. This may have contributed to a longer traveling distance. We assume that a user might have been guided to an inaccurate cue for a longer period of time. In the *mixed modality*, the auditive cues could perhaps compensate for the decreased amount of visual cueing by providing better guidance. It may have been that whenever the participant was guided towards an inaccurate cue in the 0.2-margin condition, the auditive cue could have redirected the participant towards the target location. Moreover, sporadic cueing might have also resulted in slightly better guidance when visual and auditive cues were combined since many cues of multiple modalities might have confused the participant in the no-margin condition. This seems to be the case in the context of warnings [11, 17]

Contrary to our hypothesis, the main effect of the interaction weight failed to reach significance. This may suggest that the parameter had no strong influence on the guiding performance. Another possible explanation might be that the design of the virtual supermarket simply did not have enough cues that were either only salient or only close, such that the interaction weight makes a difference in the guiding performance. An alternative design of the environment could still prove the usefulness of this parameter. This could be assessed in a future study.

When looking at the results, however, one has to consider the limitations of the study. First, the study was conducted in a virtual environment instead of a real-world environment. While VR-environments have advantages, such as complete control of the setting, there are also downsides. Most users had little or no previous experience with VR which lead to difficulties in navigation and the feeling of dizziness. In turn, this could have been contributed to noise in the data. Second, as there are conflicting results about how the number of cues might impact the guidance performance, further research is necessary. Comparing the developed method to a system that highlights the target with a single cue could better explain how the quality and the number of cues impact the guiding performance and usability. Third, is that the current design is a fractional factorial design: margin levels were varied only under specific parameter settings of the interaction weight and vice versa. In that case, the effects of interaction weight and margin are



(a) Mean distance traveled for interaction weight and modality

(b) Mean distance traveled for margin and modality

Figure 7: Mean distance traveled (see also table 1) under various conditions of interaction weight, margin, and modality. In 7(a) mean distance traveled as a function of interaction weight and modality while margin is set at 0.2. In 7(b) mean distance traveled as a function of margin and modality while interaction weight is set at 0.5.

not completely independent of interacting effects between these parameters.

5 CONCLUSION AND FUTURE WORK

Attention guidance in CPEs is important because it allows the environment to highlight areas and objects of importance to the user. However, selecting the right devices or cues to guide the user's attention is not a trivial decision. In this work, we address this problem by introducing a novel method for multi-scale attention guidance for CPEs. This consists of a flexible function that can be adapted by a designer to a specific CPE by modifying a few parameters and weights. The function quantifies the suitability of a device to lead a user's attention to a specific target object or area. We argue that a suitable device is one that is salient enough to attract the user's attention and is close to the target such that the user is guided in the right direction. The salience and closeness functions were then defined with the help of some secondary functions that attempt to quantify salience and closeness individually for both the visual and the auditive modalities. One of the advantages of defining salience and closeness this way is that the secondary functions could easily be replaced by better alternatives should they arise. Another important feature was the transitioning between devices. To avoid an overly volatile system, a numerical advantage was given to the currently selected device for a short period of time. Continuous device switching could then be avoided.

We tested our guiding method in a user study in terms of guiding performance for visual-only or mixed visual and auditive guiding information while varying specific parameters (margin and interaction weight). The results suggest that multi-modal guidance can have a supportive effect on the distance a user travels. However, these results seemed to be dependent on distinct parameter settings, in our case the margin. Future work could provide a more detailed insight on how different parameter settings could enhance the guiding performance in general as well in specific multi-modal multi-scale attention guidance. In the same way, it would be interesting to examine how changes in the properties of the cues may affect the performance while interacting with the system. For example changing the frequency or the intensity of the visual and auditive cues.

This work opens up many new opportunities and research directions. While the attention guidance method evaluated in the user study managed to guide users to the target objects in VR, an important next step would be to test the system in a real-world environment. Doing this would present the system with new challenges like sensor imprecision and lack of availability of output devices. Coming up with better alternative secondary functions that can more accurately estimate the salience and closeness of a device would also be of great interest. In addition, other devices or even modalities might improve the guidance performance of the user. Haptic feedback, for instance, has already been successfully used to guide users towards a goal[16]. Finally, this method has, thus far, focused solely on how to guide a single user's attention in a CPE. It would be interesting to see how this method could be adapted to a multi-user setting, where cues from one device may be perceived by more than one user.

ACKNOWLEDGMENTS

This work is partially supported by the German Federal Ministry of Education and Research (01IW17004, 01IW20008).

REFERENCES

[1] Reynold J. Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. 2009. Subtle gaze direction. *ACM Trans. Graph.* 28, 4 (2009), 100:1–100:14.
[2] Carryl L. Baldwin, Charles Spence, James P. Bliss, J. Christopher Brill, Michael S. Wogalter, Christopher B. Mayhorn, and Thomas K. Ferris. 2012. Multimodal Cueing: The Relative Benefits of the Auditory, Visual, and Tactile Channels in Complex Environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1, 1431–1435.
[3] Thomas Booth, Srinivas Sridharan, Ann McNamara, Cindy Grimm, and Reynold J. Bailey. 2013. Guiding attention in controlled real-world environments. In *ACM Symposium on Applied Perception 2013, SAP' 13, Dublin, Ireland, August 22-23, 2013*, Ludovic Hoyet, Betsy Williams Sanders, Joe Geigel, and Jeanine Stefanucci (Eds.). ACM, 75–82.
[4] Rudolph P. Darken and Barry Peterson. 2014. Spatial Orientation, Wayfinding, and Representation. In *Handbook of Virtual Environments - Design, Implementation, and Applications, Second Edition*, Kelly S. Hale and Kay M. Stanney (Eds.). CRC Press, 467–491.
[5] Jon Driver and Charles Spence. 1998. Attention and the crossmodal construction of space. *Trends in Cognitive Sciences* 2, 7 (July 1998), 254–262.
[6] John Duncan. 1984. Selective attention and the organization of visual information. *Journal of Experimental Psychology: General* 113, 4 (1984), 501–517.
[7] Robert Egly, Jon Driver, and Robert D. Rafal. 1994. Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General* 123, 2 (1994), 161–177.

[8] Magy Seif El-Nasr, Athanasios V. Vasilakos, Chinmay Rao, and Joseph A. Zupko. 2009. Dynamic Intelligent Lighting for Directing Visual Attention in Interactive 3-D Scenes. *IEEE Trans. Comput. Intell. AI Games* 1, 2 (2009), 145–153.

[9] Charles W. Eriksen and James D. St. James. 1986. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics* 40, 4 (July 1986), 225–240.

[10] Michael W. Eysenck and Mark T. Keane. 2010. *Cognitive psychology : a students handbook* (6 ed.). Psychology Press, Chapter 1: Approaches to Human Cognition, 1–32.

[11] Gregory M. Fitch, Jonathan M. Hankey, Brian M. Kleiner, and Thomas A. Dingus. 2011. Driver comprehension of multiple haptic seat alerts intended for use in an integrated collision avoidance system. *Transportation Research Part F: Traffic Psychology and Behaviour* 14, 4 (July 2011), 278–290.

[12] Hede Helfrich. 2019. *Sprache und Kommunikation* (2. aufl. 2019 ed.). Springer Berlin Heidelberg, Chapter 9: Sprache und Kommunikation, 121–136.

[13] Andrew Hollingworth, Ashleigh M. Maxcey-Richard, and Shaun P. Vecera. 2012. The Spatial Distribution of Attention within and across Objects. *Journal of Experimental Psychology: Human Perception and Performance* 38, 1 (Feb. 2012), 135–151.

[14] Liqiang Huang and Harold Pashler. 2005. Quantifying object salience by equating distractor effects. *Vision Research* 45, 14 (June 2005), 1909–1920.

[15] Gerrit Kahl, Lübomira Spassova, Johannes Schöning, Sven Gehring, and Antonio Krüger. 2011. IRL SmartCart - a user-adaptive context-aware interface for shopping assistance. In *Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI 2011, Palo Alto, CA, USA, February 13-16, 2011*, Pearl Pu, Michael J. Pazzani, Elisabeth André, and Doug Riecken (Eds.). ACM, 359–362.

[16] James R. Marston, Jack M. Loomis, Roberta L. Klatzky, and Reginald G. Golledge. 2007. Nonvisual Route following with Guidance from a Simple Haptic or Auditory Display. *Journal of Visual Impairment & Blindness* 101, 4 (2007), 203–211.

[17] Christina Meredith and Judy Edworthy. 1995. Are there too many alarms in the intensive care unit? An overview of the problems. *Journal of Advanced Nursing* 21, 1 (1995), 15–20.

[18] Dinara Moura and Lyn Bartram. 2014. Investigating players' responses to wayfinding cues in 3D video games. In *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 - May 01, 2014, Extended Abstracts*, Matt Jones, Philippe A. Palanque, Albrecht Schmidt, and Tovi Grossman (Eds.). ACM, 1513–1518.

[19] Hermann J. Müller, Joseph Krummenacher, and Torsten Schubert. 2015. *Aufmerksamkeit und Handlungssteuerung*. Springer Berlin Heidelberg, Chapter 3: Selective Aufmerksamkeit, 19–38.

[20] Hermann J. Müller, Joseph Krummenacher, and Torsten Schubert. 2015. *Aufmerksamkeit und Handlungssteuerung*. Springer Berlin Heidelberg, Chapter 5: Visuelle Suche, 45–56.

[21] Michael I. Posner. 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology* 32, 1 (1980), 3–25.

[22] Ragunathan Rajkumar, Insup Lee, Lui Sha, and John A. Stankovic. 2010. Cyber-physical systems: the next computing revolution. In *Proceedings of the 47th Design Automation Conference, DAC 2010, Anaheim, California, USA, July 13-18, 2010*, Sachin S. Sapatnekar (Ed.). ACM, 731–736.

[23] Bill N. Schilit and Marvin M. Theimer. 1994. Disseminating Active Map Information to Mobile Hosts. *IEEE Network: The Magazine of Global Internetworking* 8, 5 (Sept. 1994), 22–32.

[24] Vivek K. Singh and Ramesh Jain. 2009. Situation based control for cyber-physical environments. In *MILCOM 2009 - 2009 IEEE Military Communications Conference*. 1–7.

[25] Charles Spence and Jon Driver. 1997. Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics* 59, 1 (Jan. 1997), 1–22.

[26] Christoph Stahl, Jörg Baus, Boris Brandherm, Michael Schmitz, and Tim Schwartz. 2005. Navigational- and Shopping Assistance on the Basis of User Interactions in Intelligent Environments. In *Proceedings of the IEE International Workshop on Intelligent Environments (IE 2005)*. University of Essex, Colchester, UK, 182–191.

[27] Karl E. Steiner and Lavanya Voruganti. 2004. A comparison of guidance cues in desktop virtual environments. *Virtual Real.* 7, 3-4 (2004), 140–147.

[28] Rufin VanRullen. 2003. Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris* 97, 2 (March 2003), 365–377.

[29] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. 1989. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance* 15, 3 (Aug. 1989), 419–433.