Cognitive-Aware Modality Allocation in Intelligent Multimodal Information Presentation

Yujia Cao, Mariët Theune, and Anton Nijholt

Abstract Intelligent multimodal presentation (IMMP) systems are able to generate multimodal presentations adaptively, based on the run-time requirements of user–computer interaction. Modality allocation in IMMP system needs to adapt the modality choice to changes in various relevant factors, such as the type of information to be conveyed, the presentation goal, the characteristics of the available modalities, the user profile, the condition of the environment, and the type of user task. In this study, we emphasize that modality allocation in IMMP systems should also take into account the cognitive impacts of modality on human information processing. We first describe several modality-related cognitive and neuropsychological findings. Then a user study is presented to demonstrate the effects of modality on performance, cognitive load and stress, using a high-load and time-critical user task. Finally, we show a possible way to integrate relevant cognitive theories into a computational model that can systematically predict the suitability of a modality choice for a given presentation task.

1 Introduction

The development of intelligent multimodal presentation (IMMP) systems has received much attention during the past two decades. The application domain of IMMP is very broad, including home entertainment [19], technical document generation [58], medical training [29], crisis management support [22], and much more. IMMP systems have been defined as knowledge-based systems, which exploit their

Y. Cao (✉) · M. Theune · A. Nijholt
Human Media Interaction, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands
e-mail: y.cao@utwente.nl

M. Theune
e-mail: m.theune@utwente.nl

A. Nijholt
e-mail: a.nijholt@utwente.nl

L. Shao et al. (eds.), *Multimedia Interaction and Intelligent User Interfaces,*
Advances in Pattern Recognition,
DOI 10.1007/978-1-84996-507-1_3,

knowledge base in order to dynamically *adapt* their design decisions to the run-time requirements of user–computer interaction, such as the user profile, task characteristics, nature of the information to be conveyed, etc. [8, 28]. They are *intelligent* in the sense that they are able to generate multimodal presentations *adaptively* at runtime. A key issue in this process is to automate modality allocation—a process that chooses one or more modalities to present a certain information content for achieving a certain presentation goal [8]. Modality allocation can also be considered as making the most suitable mappings between a set of information items and a set of modalities, constrained by certain factors [1]. The factors can be the type of information to be conveyed, the presentation goal, the characteristics of the available modalities, the user profile, the condition of the environment, the type of user task, or any other factors that are identified to be relevant to a specific application. In IMMP systems, modality needs to be allocated on the fly, adapting to changes in the selected factors.

In existing IMMP studies, modality allocation is commonly rule-based [2, 20, 28, 29, 33, 43, 44, 56, 57, 59]. Modality allocation rules typically associate factors with preferred modality choices. They are usually predefined and embedded in the knowledge base of the system. They are the core of the intelligence in the sense that they define what (factors) the system should adapt to and how it should adapt. To demonstrate modality allocation rules, several examples associated with various factors are listed as follows.

- *The type of information to be conveyed*: for location and physical attributes, use graphics; for abstract actions and relationships between actions (such as causality), use text; for compound actions, use both text and graphics (in [20] for technical document generation).
- *Presentation goal*: to inform the user about TV programs, use the text in a list (in [57] for home digital guide).
- *State of the environment*: if the noise level is greater than 80 Db, use visual or tactile modalities (in [44] for phone call reception announcement).
- *Application specific factor*: when the needle is outside the patient's body, use only sound to present the distance to the target; when the needle is inserted into the body, use both sound and color gauge; when the needle tip is very near the target point (< 10 mm), use only color gauge (in [29] for surgery training).

In order to be inferred by the system, modality allocation rules need to be translated into the representation language of the system, such as M3L used in [57] and MOXML used in [43]. For each presentation task, modalities can be allocated on the fly by searching the rule base for rules associated with the factor values at that specific point of presentation. Alternatively, some studies quantify the rules by translating them into numerical metrics of weights or appropriateness and then apply computational models for an overall optimization, such as the graph matching method used in [64] and the weighed additive utility model used in [28]. These computational methods were not often named as rule-based. However, the input metrics are still derived from rules. What differs is the way in which the rules are encoded and inferred by the system.

When viewing the modality allocation rules used in existing IMMP systems, it appears that most of them are disassociated from knowledge of human information processing. In other words, they do not seem to consider how information carried by different modalities is perceived and processed by the human cognitive system. Consequently, the efficiency of interaction might be affected due to the unnecessary cognitive load that the multimodal presentations impose on the user. As technology advances, computer systems are increasingly able to assist users in data-rich and time-critical applications, such as crisis management and stock monitoring. The cognitive compatibility issue could be particularly important in these applications, because users are very likely to work under high cognitive load and stress. The need to integrate relevant cognitive knowledge into IMMP has gained awareness in recent years and has been addressed in several articles providing design guidelines [41, 45, 54].

In this chapter, we first describe several findings from the field of cognitive psychology and neuropsychology on the relevance of modality to human information processing. These findings reveal the necessity of considering the cognitive impacts of modality and can serve as a theoretical foundation of cognitive-aware modality allocation. Then, we present a user study to further demonstrate the effects of modality on user performance and cognitive load, using a high-load and time-critical scenario. The experimental results are interpreted in the light of relevant cognitive theories. Based on the consistency of the results and the theories, we go one step further to construct a computational model for predicting the suitability of the modality variants that were not investigated in the experiment. This model also demonstrates a way to integrate relevant cognitive knowledge into the modality allocation for this specific presentation task. Lastly, several suggestions on adapting this model to other applications are given.

2 Modality and Human Information Processing

First, we present a conceptual model of human information processing proposed by Wickens [60]. This model provides a useful framework for further discussing the relation between modality and several stages of human information processing. The model, as shown in Fig. 1 represents human information processing as a series of stages. In the *sensory processing* stage, information from the environment is received by the brain as raw sensory data that can be processed by the brain. Then, attention is needed to select certain raw sensory data to be interpreted and given meaning in the *perception* stage. Afterwards, more complex cognitive operations (reasoning, comprehension, etc.) are conducted in the *working memory* stage. Working memory also has access to the long-term memory system. Based on the outcome of cognitive processing, decisions are reached on how to respond in the *response selection* stage. Finally, the selected response is executed. The feedback loop at the bottom of the model indicates that the human response to the environment can be observed again. This feedback loop makes it possible to keep adjusting

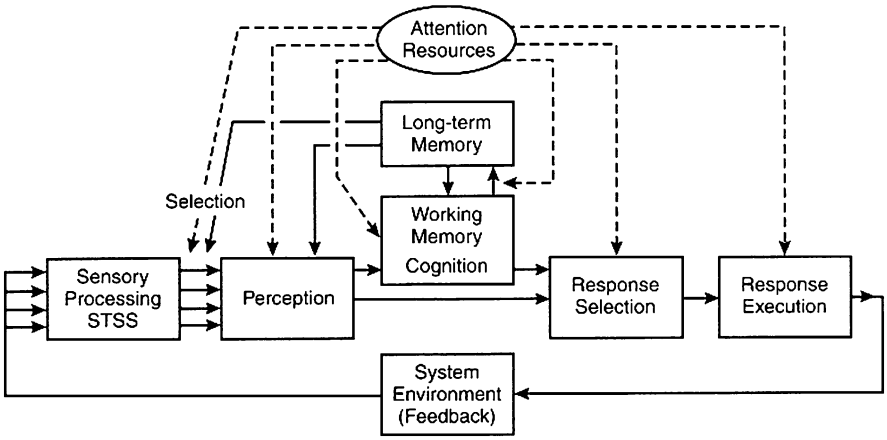

Fig. 1 A model of human information precessing stages proposed by C.D. Wickens (reproduced from [60], p. 11)

the response to reach a certain goal. This is important for many real-world tasks, such as walking and driving.

From the perspective of a system (in the bottom block), output modalities[1] mostly influence three stages of information processing: sensory processing, perception, and working memory. The response selection stage has not been explicitly related to the use of modalities in the literature, because the response to an event is mostly based on the output of cognition rather than the modality of information presentation. However, in multimodal interaction, users might choose to respond to the system with modalities that are consistent with the output modalities, such as speech response to speech outputs. The response execution stage is modality-specific because different modalities are generated by different parts of the body, such as hands for tactile response and vocal organs for speech response. These modalities are input modalities[2] from the perspective of a system, thus are outside the focus of multimodal information presentation. In the remainder of this section, we describe the role of modality in sensory processing, perception (selective attention from sensory processing to perception), and working memory (cognition).

2.1 Modality and Sensory Processing

At the very first sensory processing stage, the distinction of modalities is physically determined, because the five human senses are realized by different sensory receptors. The receptors for visual, auditory, tactile, olfactory, and gustatory signals are

[1] Output modalities refer to modalities a system uses to present information to users.

[2] Input modalities refer to modalities users use to interact with a system.

found in the eyes, ears, skin, nose, and tongue, respectively. Each sensory receptor is sensitive to only one form of energy. The function of these receptors is to transduce the physical energy into electrochemical energy that can be processed by the brain.

2.2 Modality and Perception

Sensed stimuli do not have to be consciously attended to and actively interpreted. Instead, attention is needed to select certain raw sensory data to be perceived, given meaning and further processed by the brain [42, 60]. This selection process is referred to as "selective attention" [27]. Modality plays a role in selective attention, because different modalities vary in their abilities to attract attention, mostly based on their sensory properties. Here, we focus on visual and auditory modalities.

2.2.1 Visual Attention

Visual attention guides what we are looking at. The visual field is divided into foveal and peripheral fields. Only foveal vision is able to observe details of objects, but it has a very limited angle of only about two degrees. Therefore, without foveal visual attention, people often have surprising difficulty in detecting large changes in visual scenes—a phenomenon known as "change blindness" [46, 47]. Peripheral vision is sensitive to motion and luminance changes. Visual attention can be directed in a top-down manner or a bottom-up manner [14]. The top-down manner means that visual attention is consciously directed by top-down knowledge, such as task-dependent goals [36], contextual cues [12, 40], current items in the working memory [15, 25], and expectations of what to see [53]. In contrast, the bottom-up manner is saliency driven, meaning that the visual stimuli which win the competition for saliency will *automatically* be attended. When an object in a visual field contains some unique features, this object seems to "pop out" and captures the attention [26]. Through the bottom-up mechanism, attention shifts can be influenced by how the visual information is presented. Items that have higher priority should be presented with a unique (compared with surrounding) color, shape, intensity, orientation, depth, size, or curvature [39, 63].

2.2.2 Auditory Attention

The auditory modalities are different from the visual ones in three aspects regarding attention attraction. First, auditory modalities are more salient than visual modalities. Usually, attention is promptly directed to an auditory signal upon the onset of its presentation [53]. This feature makes auditory modalities a preferred choice to present information with high priorities, such as warnings and alerts [55]. The risk of using auditory modalities is that they might interrupt an ongoing task by pulling

full attention away from it, referred to as "auditory preemption" [62]. Second, unlike visual information which needs to be in the visual field in order to be attended to, auditory information can grab attention no matter which direction it comes from, and its direction can be distinguished if perceived by both ears. This feature makes it possible to assist visual search by providing location cues via auditory modalities. For example, it was demonstrated in [6] that 3D audio information could indeed assist pilots to locate outside-the-window visual targets faster. Third, auditory information is transient if no repeat mechanism is added to it. Therefore, it is force-paced, meaning that in order to be fully perceived, attention needs to be held on to an auditory stream during its presentation. In contrast, static visual information tends to be more continuously available and thus offers more freedom of perception in terms of time [60].

2.2.3 Cross-Modal Attention

In real-life situations, attention often must be simultaneously coordinated between different senses—a fact that motivated the development of a relatively new research topic, crossmodal attention [52]. It has been proved that a shift of attention in one modality toward a certain spatial location tends to be accompanied by corresponding shifts in other modalities toward the same location [16, 21]. Such crossmodal links can operate in a reflexive (automatic) manner or a voluntary (controlled) manner. The reflexive manner means that an irrelevant but salient event in one modality tends to attract attention toward it in other modalities as well. Such reflexive links have been found for many modality combinations. For example, a salient auditory event (e.g., a loud bang) can generate rapid shifts of visual attention towards its direction; a tactile event on one hand (e.g., being touched) can generate shifts of visual and auditory attention toward the location of the touch. Crossmodal links can also direct attention voluntarily. When a person strongly expects an event in one modality at a particular location, his/her sensory sensitivity improves at that location not only for the expected modality but also for other modalities, even if there is no motivation to expect events from other modalities to occur at that location [51]. The crossmodal attention shifts have been supported by electrophysiological evidences from event-related brain potential (ERP) studies [17, 18]. There might be a single crossmodal attentional system that operates independently of sensory modality and controls shifts of spatial attention for all senses. In summary, spatial attention toward a location typically spreads across modalities, and this finding has implications for multimodal information presentation to better support attention management in complex and data-rich interface applications.

2.3 Modality and Working Memory

The working memory stage following the perception stage also works in a modality-specific manner. Two theories about this are discussed below.

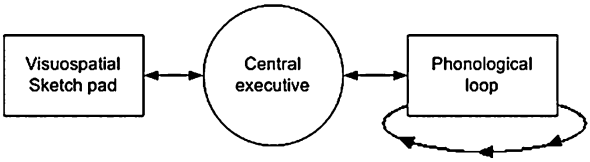

Fig. 2 The working memory model from Baddeley and Hitch [5]

2.3.1 Working Memory Theory

In 1974, Baddeley and Hitch proposed a three-component model of working memory, which has been well supported by scientific evidence from cognitive psychology, neuroimaging, and anatomy [4, 5]. According to this model, working memory contains a central executive system aided by two subsidiary systems, a visual-spatial sketch pad and a phonological loop (Fig. 2). The phonological loop has a phonological store for temporarily storing auditory information. It also includes a rehearsal system. Auditory traces within the store are assumed to decay over a period of about two seconds unless being refreshed by the rehearsal system. Particularly, the rehearsal system relies on speech coding to maintain the memory trace, meaning that information is usually rehearsed in the mind via subvocal speech [3]. The visual-spatial sketch pad is assumed to temporarily maintain visual information and to form a relation between visual and spatial information. The information stored in the two subsidiary systems is retrieved by the central executive system, which is assumed to be an attentional system whose role extends beyond memory functions. As the name indicates, it is believed to be a processing and control system which is involved in attention management, learning, comprehension, decision making, reasoning, judgement, and planning. Neuroimaging and anatomical studies have indicated that these three components of working memory are localized in different brain regions. There is clear evidence of the phonological loop being on the left temporoparietal region. The visual-spatial pad is identified to be primarily localized in the right hemisphere [34, 50]. There is the least agreement among research findings on the anatomical location of the center executive. It seems possible that different executive processes are implemented by different brain components. It can be inferred from this theory that the visual and auditory channels consume separated perceptional resources. Therefore, two perception tasks can be better performed in parallel when they make use of different channels, compared to when they compete for resources in the same channel [61].

2.3.2 Dual Coding Theory

At about the same time when the working memory theory was proposed, Paivio proposed a dual coding theory which addresses another modality-specific feature of human cognition [37]. This theory assumes that cognition is served by two separate symbolic systems, one specialized for dealing with verbal information and the other with nonverbal information (Fig. 3). The two systems are presumed to be interconnected but capable of functioning independently. The verbal system processes

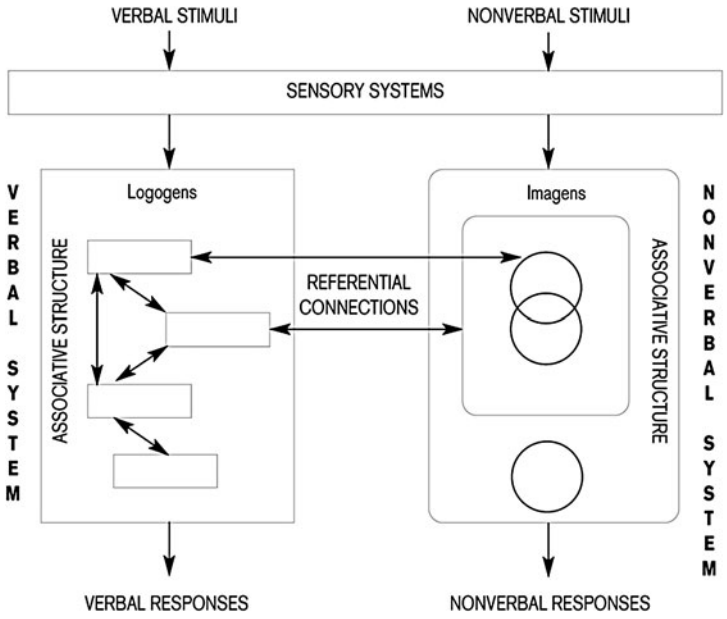

Fig. 3 Verbal and nonverbal symbolic systems of Dual Coding Theory [38]. Logogens and imagens refer to verbal and nonverbal representational units, respectively

visual, auditory, and other modality-specific verbal codes. The nonverbal system processes images, environmental sounds, actions, and other nonverbal objects and events. The two systems are linked into a dynamic network through referential connections. The referential connections convert information between two systems and join corresponding verbal and nonverbal codes into knowledge that can be acted upon, stored, and retrieved for subsequent use. It has been demonstrated that the referential connections play major roles in various educational domains, such as knowledge comprehension, memorization, the learning of motor skills, etc. [13]. Neuroimaging studies have provided support for the dual coding theory by showing that different parts of the brain are responsible for the passive storage and active maintenance of verbal, spatial, and object information [48, 49].

2.3.3 Relating the Two Theories

The aforementioned two theories have not been explicitly related to each other by their founders. However, they are complementary instead of contradictory. It seems reasonable to assume that the center executive selectively retrieves information from modality-specific mental systems, integrates them into a unified percept, and then implements executive processes (reasoning, decision making, etc.). The center executive may also be responsible for the transfer of information between modalities. Since the rehearsal of information in the working memory is based on subvocal

speech [3], rehearsing written materials during reading is an example of modality transfer from visual to auditory system. Moreover, mental imagination of the appearance of an object upon hearing its name is an example of modality transfer from verbal to nonverbal system.

In the educational psychology domain, multimedia learning studies have applied both theories to understand the impacts of various learning material designs on the learning performance. Regarding the dual coding theory, it was found that it was more beneficial to present knowledge both verbally and nonverbally than only verbally or only nonverbally [30, 31]. This is because the mental processes of associating related verbal and nonverbal information can help deepen the understanding of the knowledge and thus lead to better problem-solving transformation. Regarding the dual coding theory, it was found that when nonverbal information (illustration, animation, diagram, etc.) was provided visually (on paper or on screen), the associated verbal explanation was better presented in speech than in text [9, 32, 35]. When all information was presented visually, perceptional resources in the visual channel had to be divided for verbal and nonverbal items, causing a so-called split-attention effect. By replacing text with well-synchronized narration, related verbal and nonverbal units could be concurrently perceived via two channels. As a result, more cognitive resources are available for further processing of the knowledge.

In the study presented here, we intended to apply these two cognitive theories together with findings on attention (Sect. 2.2) to a high-load and time-critical user task rather than learning. Our goal was twofold. The first was to investigate/validate the modality effects on user performance, cognitive load and stress, with our high-load and time-critical task setting. Second, we intended to interpret the experimental results in association with relevant cognitive findings. By doing so, we could also investigate whether these theories could be used to predict how suitable a certain modality choice is for this presentation task.

3 Experiment on Modality Effects in High-Load HCI

A user study was conducted, using an earthquake rescue scenario, where the locations of wounded and dead people are continuously reported to the crisis response center and displayed on a computer screen. Based on these reports, a crisis manager directs a doctor to reach all wounded people and save their lives. In this experiment, the subject plays the role of the crisis manager, and his/her task is to save as many wounded victims as possible. Note that it was not our goal to make the crisis scenario realistic, and subjects were not required to have any experience in crisis management. The choice of scenario was made to better motivate a high-load and time-critical user task.

3.1 Presentation Material

For each victim report, two types of information could be provided: basic information and additional aid. The basic information included the type of the victims (wounded or dead) and their location. The additional aid reduced the searching area by indicating which half of the screen (left or right) contained the victim.

To convey these two types of information, four modalities were selected based on their visual/auditory and verbal/nonverbal properties, namely text (visual, verbal), image (visual, nonverbal), speech (auditory, verbal), and sound (auditory, nonverbal). The basic information could be efficiently conveyed by locating a visual object on a map. Therefore, we selected text and image to present the victim type (Fig. 4, left), and the location on a grid-based map indicated the location of the victim (Fig. 4, right). Three modalities were selected to present the additional aid. They were image (a large-size left arrow or right arrow right below the map area), speech ("left" or "right"), and sound (an ambulance sound coming from the left or the right speaker).

Finally, five experimental conditions were chosen, two without additional aids and three with aids (see Table 1). We predicted that image would be better than text for presenting victim types, because it has been found that the categorization and understanding of concrete objects are faster when they are presented by image

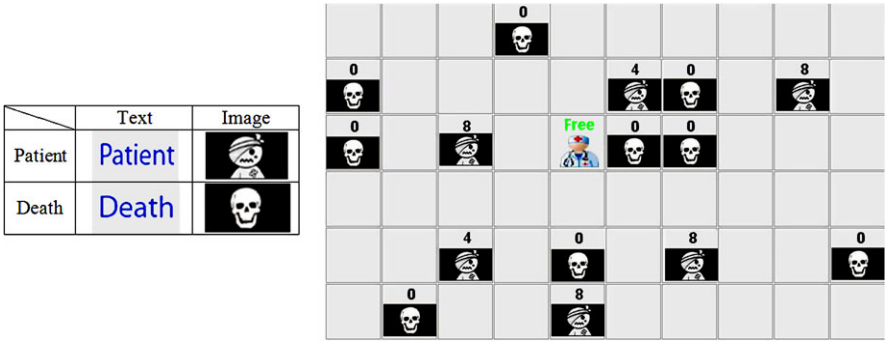

Fig. 4 Presentations used in the experiment. *Left*: text and image presentations of the victim types. Wounded and dead victims are named "patient" and "death," respectively. *Right*: a part of the grid-based map (the full size is 20 grids by 13 grids)

Table 1 Five experimental presentation conditions

Index	Basic Information	Additional Aid	Modality Properties
1	Text	None	Visual, verbal
2	Image	None	Visual, nonverbal
3	Text	Image	Visual + visual, verbal + nonverbal
4	Text	Speech	Visual + auditory, verbal + verbal
5	Text	Sound	Visual + auditory, verbal + nonverbal

than by text [7]. Therefore, in order to better observe the benefit of the additional aid, basic information was always presented by text when additional aids were provided.

3.2 Task and Procedure

The subject played the role of the crisis manager, whose task was to send the doctor to each patient by mouse-clicking on the presentation (text or image). New patients appeared at random intervals of 2 to 5 seconds, usually at the same time as one or more dead victims. A patient had a lifetime of 10 seconds and would turn into a dead victim without a timely treatment. A number above the presentation of a patient indicated his remaining lifetime. When timely treated, patients disappeared from the screen. In each trial, 100 patients were presented in about 5 minutes. Dead victims served as distracters that required no reaction.

The difficulty of the task could be regulated by the number of distracters (dead victims). At the beginning of a trial, there were no any objects on the grid map, and the task was relatively easy. As the number of dead victims grew, it became more and more difficult to identify a patient in the crowded surroundings. The task difficulty reached the maximum (about 40% of the cells contained objects) after about 150 seconds and remained unchanged for the rest of the trial.

Twenty university students (bachelor, master, or Ph.D.) volunteered to participate in this experiment. A participant first received an introduction to the experiment and then performed a training session in order to get familiar with the task and presentation conditions. Afterwards, the participant performed all five experimental trials with a counterbalanced order. Short breaks were placed between trials, during which the questionnaires were filled in. At the end of the experiment, an informal interview was carried out to obtain additional feedback from the participant. The whole experimental procedure lasted for about 80 minutes.

3.3 Measurements

The performance was assessed by three measurements. *Reaction time* (RT) measured the time interval between the moment when a patient was presented and the moment when the doctor was sent (in seconds). *Number of patients died* (ND) referred to the number of patients that were not treated within 10 seconds and died. *Time of the first patient death* (TF) measured the time interval between the start of a trial and the moment when the first patient died in the trial (in seconds). Since the number of distracters increased gradually in the first half of a trial, TF actually reflected how tolerant the performance was against the increase of task difficulty.

Besides performance, we also obtained subjective assessments on cognitive load (SCL) and stress (SS). Based on the Task Load Index from NASA [23], the rating scale was designed to have 20 levels, from 1 (very low) to 20 (very high).

3.4 Hypotheses

We constructed the following four hypotheses.

1. The image (nonverbal) condition is superior to the text (verbal) condition in terms of better performance, lower cognitive load, and lower stress, because image is better than text for presenting concrete objects.
2. The auditory (speech and sound) aids are superior to the visual (image) aid, because they can be better time-shared with the visual rescue task.
3. The nonverbal (image and sound) aids are superior to the verbal (speech) aid, because the location information is nonverbal in nature, so that verbal presentations require additional mental resources to be converted.
4. Additional aids lead to benefits in terms of performance, cognitive load and stress, because they carry useful information and are meant to assist the user.

4 Results on Performance, Cognitive Load and Stress

Due to the within-subject design, we applied repeated-measure one-way ANOVAs on the five dependent measurements, using modality as the independent factor. Results are presented in this section.

4.1 Performance

RT. The average reaction time of all trials is shown in Fig. 5 (left). On average, it took subjects between 1.9 seconds and 3.1 seconds to react to a patient. The reactions were the fastest in the "text + speech aid" condition and the slowest in the text condition.

ANOVA results revealed a significant modality effect on reaction time, $F(4, 16) = 12.76$, $p < 0.001$. Post-hoc tests (Bonferroni tests) were then conducted for pair-wise comparisons. Significant differences in reaction time were found between the five condition pairs. The reaction was faster in the "text + speech aid" condition than in the text, "text + image aid," and "text + sound" conditions. The reaction was faster in the image condition than in the text and "text + image aid" conditions.

ND. On average, the number of dead patients in each condition was between 2 and 12 (see Fig. 5, right). As 100 patients were presented in each trial, the percentage of saved patients was between 88% and 98%. The most patients were saved in the "text + speech aid" condition, and the least were saved in the text condition.

ANOVA results indicated that there was a significant modality effect on the number of dead patients, $F(4, 16) = 16.81$, $p < 0.001$. Pairwise comparisons showed five significant effects. More patients died in the text condition than in the image,

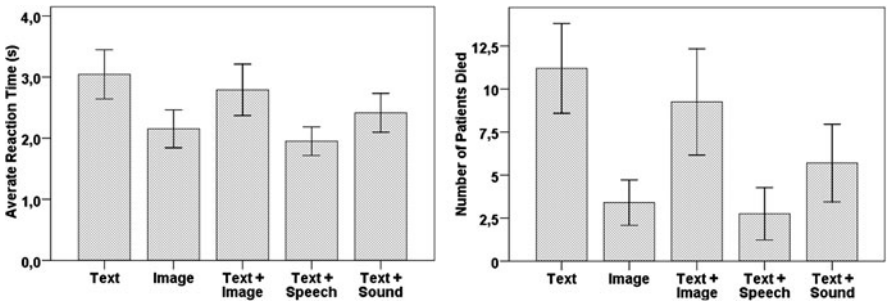

Fig. 5 Average reaction time (*left*) and number of patients that died (*right*) in five modality conditions. *Error bars* represent standard errors

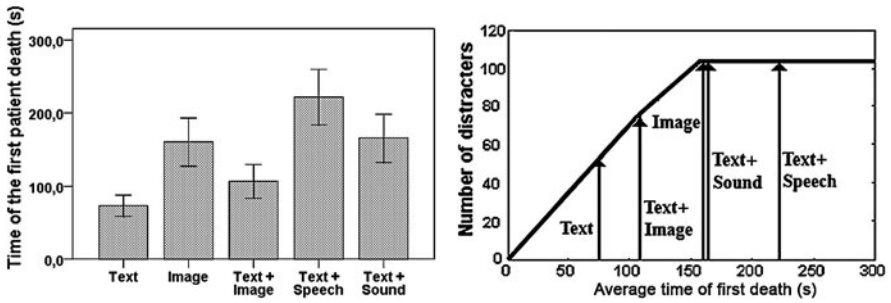

Fig. 6 Time of the first patient death. *Left*: average TF in all modality conditions. *Error bars* represent standard errors. *Right*: average TF shown on the curve of task difficulty over time

"text + speech aid" and "text + sound aid" conditions. More patients died in the "text + image aid" condition than in the image and "text + speech aid" conditions.

TF. As Fig. 6 shows, the first dead patient occurred the earliest in the text condition (at the 73th second on average), and the latest in the "text + speech aid" condition (at the 221th second on average). Again, ANOVA revealed a significant modality effect on this measurement, $F(4, 15) = 17.71,\ p < 0.001$. According to post-hoc tests, the first patient death occurred significantly earlier in the text condition than in the image, "text + speech aid" and "text + sound aid" condition. The first patient death also occurred significantly earlier in the "text + image aid" condition than in the "text + speech aid" condition.

The effects found from this measurement actually indicate that the use of modality significantly affected how tolerant the performance was against the increase of task difficulty. As Fig. 6 (right) shows, in the text condition, the performance dropped when the task difficulty increased to about half of the maximum. In contrast, in the "text + speech aid" condition, the good performance was maintained for more than 50 seconds after the task difficulty reached the maximum.

4.2 Cognitive Load and Stress

SCL. The average rating scores on subjective cognitive load mostly fell in the higher half (10–20) of the rating scale (see Fig. 7, left). Subjects considered the text condition as the most difficult one and the "text + speech aid" condition as the easiest. The cognitive load ratings were significantly affected by the use of modality, $F(4, 16) = 17.06$, $p < 0.001$. Generally, two groups could be identified among the five modality conditions. The image and the "text + speech aid" conditions formed a group of higher ratings. The remaining three conditions formed a group of lower ratings. Results of post-hoc tests showed that there were significant differences in rating scores between any two conditions taken from different groups (six condition pairs in total).

SS. As shown in Fig. 7 (right), the text condition was rated the most stressful, and the "text + speech aid" condition was rated the least stressful. ANOVA results show a significant subjective stress ($F(4, 16) = 9.379$, $p < 0.001$). According to post-hoc tests, the stress level was significantly higher in the text condition than in the image, "text + speech aid" and "text + sound aid" conditions. The "text + image aid" condition was also rated significantly more stressful than the "text + speech aid" condition.

A very similar pattern can be seen when comparing the two graphs in Fig. 7. Indeed, there is a strong positive correlation between ratings on cognitive load and stress (Corr. = 0.855), suggesting that subjects felt more stressed when they devoted more cognitive efforts to the task. Moreover, the subjective measurements were also found to be positively correlated with the performance measurements RT and ND. There are positive correlations at the 0.01 confidence level between ND-SCL, ND-SS, RT-SCL, and RT-SS. In combination, these correlations indicate that when the task was more difficult (due to a suboptimal use of modalities), subjects devoted more cognitive effort, felt more stressed, and performed worse.

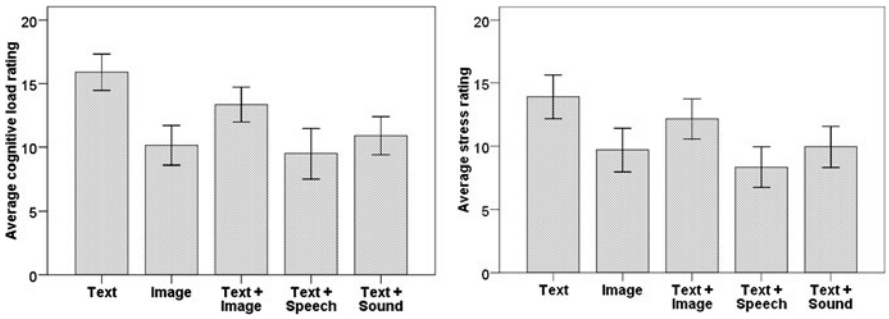

Fig. 7 Average subjective rating scores on cognitive load (*left*) and stress (*right*) in five modality conditions. *Error bars* represent standard errors

5 Discussion

The experimental results clearly showed that the use of modality affected the performance of the task, as well as the experienced cognitive load and stress. In this section, the experimental results are discussed in association with the hypotheses and the cognitive theories.

5.1 Text vs. Image

Comparing the two conditions without performance aid, the five measurements all suggested that image had advantage over text in this scenario. Thus the first hypothesis has been clearly confirmed. Image, as a nonverbal and analogue modality, is better for presenting concrete concepts [7, 24], such as wounded and dead victims in this experiment. For this task, image made it easier to distinguish between the two types of objects and thus led to faster and better performance, lower cognitive load, and lower stress. In contrast, text, as a verbal modality, is known to be less suitable for presenting concrete information but more suitable for abstract concepts, logic, quantitative values, relations [7, 24]. In this study, as the two words in text had the same font, size, and color, the two icon images were also designed to have similar shapes and colors. We believe that the advantage of image over text would become even more notable if the two images showed larger contrasts in color, shape, and size. These findings stand in line with the dual-coding theory, because they show that verbal and nonverbal presentations of the same information indeed have different impacts on how well the information can be processed. This in turn suggests that the verbal/nonverbal property needs to be taken as one dimension of modality selection in IMMP system design.

5.2 Visual Aid vs. Auditory Aid

Here, we compare the "text + image aid" condition to the "text + speech aid" and the "text + sound aid" conditions. The results from all five measurements consistently showed that the speech aid was significantly more appropriate than the image aid. In terms of average values, the sound aid was also superior to the image aid in all five measurements. However, this advantage only reached a statistical significance in the cognitive load measurement. Overall, we could conclude that the auditory aids were more beneficial than the visual aid in this experiment. The second hypothesis is confirmed.

The explanation of this finding is twofold. First, auditory signals are more able to attract attention than visual signals, especially when the eyes are occupied with another task (Sect. 2.2.2). Therefore, while busy searching for patients, visual aids displayed at the bottom of the display were more likely to be missed than speech

aids. Besides conveying the search area for a patient, a performance aid also indicated that a patient was newly added onto the map. If a visual aid was missed, the arrival of that patient could be missed as well. Therefore, in the "text + image aid" condition, subjects were likely to lose track of the number of patients remaining unattended.

Second, even when being attended to, visual aids still have drawbacks due to cognitive resource competition. According to the working memory theory from Baddeley (Sect. 2.3.1), separated perceptional resources are used for visual and auditory information. Therefore, auditory aids could be perceived in parallel with the ongoing rescue task. In contrast, the perception of visual aids cannot be time-shared with the rescue task. Limited visual perceptional resources needed to be shared between the rescue map and the aids. When the rescue task was already demanding, visual aids were more likely to cause overload than to be of help. Not surprisingly, many subjects mentioned during the interview at the end that they sometimes had to consciously ignore the image aids in order to concentrate on the rescue task.

5.3 Verbal Aid vs. Nonverbal Aid

Although the image aid is nonverbal, it has been identified as inappropriate for this task (Sect. 5.2). Therefore, we focus the comparison on the speech aid and the sound aid. In terms of average values, all five measurements showed an advantage of the speech aid over the sound aid. The difference in reaction time was significant. When asked to compare these two conditions, the majority of subjects preferred the speech aid. These results clearly contradict the third hypothesis. The understanding of words "left" and "right" is highly automatic for most people. So the additional load associated with it (if any) was probably too little to harm the task performance. Then, why were speech aids better than sound aids? Subjects provided two explanations. First, it was commonly mentioned that speech aids made it easier to maintain a short queue of newly reported patients ("left"s and "right"s) in mind, while solving a current one. It was however harder to do the same with the sound aids. Baddeley's working memory theory states that the working memory usually relies on subvocal speech to maintain a memory trace (Sect. 2.3.1). That is to say speech aid "left" and "right" could be directly rehearsed, but the direction of a sound, as a nonverbal information, had to be converted into a verbal form in order to be maintained. This conversion (via referential connections) consumed additional cognitive resources, and this was probably why subjects found it harder to maintain a queue of untreated patients with sound aids than with speech aids. Second, a few subjects disliked the ambulance sound. They found it disturbing when used at a high frequency, and they could not concentrate well on the rescue task.

Interestingly, the dual coding theory (Sect. 2.3.2) leads to a different suggestion for our task than for learning material design. A learning task requires comprehension and long-term memorization of presented knowledge. The combined use of verbal and nonverbal presentation invokes referential connections which have been

shown to be essential to a deeper understanding and a better memorization [13]. In contrast, our task required short-term memorization and did not involve comprehension of complex information. In this case, the additional cognitive effort spent on building referential connections was less useful and more harmful.

5.4 Additional Aid vs. No Aid

Of the five conditions, the text condition was the worst one, shown by all measurements. However, when text was combined with speech aid, the condition became the best of the five. This comparison seems to suggest that providing additional aids is beneficial compared to not providing them. However, the benefit of additional aid was only conditional, because it could be influenced by the modality used to present the aid.

When comparing the image condition with the "text + image aid" condition, one can see that the former led to shorter reaction times (RT), better rescue performance (ND), and lower cognitive load (SCL) than the latter. Considering average values, time of the first patient death (TF) and subjective stress (SS) also showed an advantage of the image condition over the "text + image aid" condition. However, the differences did not reach statistical significance. This comparison shows that presenting less information using an appropriate single modality (image) could be more beneficial than presenting more information using an inappropriate modality combination (text + image aid). Therefore, the additional aids can be of real help only when they are presented via an appropriate modality. The fourth hypothesis is only partially confirmed.

5.5 Low Load vs. High Load

We further investigated whether the modality effects reported above would also occur without the high-load condition. At the beginning of each trial, no objects were on the grid map, and thus the rescue task was relatively easy. As more and more objects were presented, it got more and more difficult to identify a patient in the crowded surroundings. According to the data from the TF measurement, the first patient death occurred after 60 seconds in all trials of all subjects. Therefore, we considered the first 60 seconds as a relatively low-load period. The average reaction time was recalculated with this period (see Fig. 8). Comparing Fig. 8 to Fig. 5 (left), a similar up-and-down trend can be recognized, suggesting that the relative difference in task difficulty between conditions remain unchanged. However, the differences in reaction time between conditions were much smaller during the first 60 seconds. On average, reactions in the fastest condition ("text + speech aid") was about 0.15 seconds faster than in the slowest condition (text)—a difference that was only about 14% of the value calculated from the whole trial (1.09 s, Fig. 5, left).

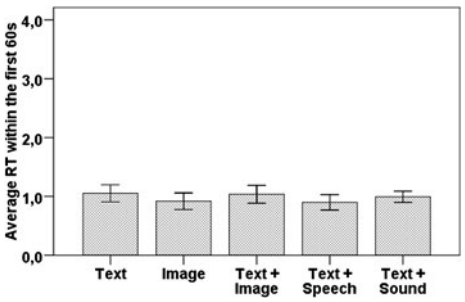

Fig. 8 Average reaction time from the first 60 seconds in five modality conditions. *Error bars* represent standard errors

Furthermore, ANOVA analysis did not reveal any modality effect on the reaction time during the first 60 seconds ($F(4, 16) = 1.61$, n.s.). These results suggest that in the low-load period, the use of modality influenced the task performance to a smaller extent, compared to in a high-load condition in which this influence became significant. Therefore, it is particularly important for IMMP systems with high-load applications to integrate modality-related cognitive principles into the modality allocation processes.

6 A Modality Suitability Prediction Model

The discussion of experimental results showed that cognitive theories of working memory and attention, together with the expressive feature of modalities, accounted for variations in user performance and experienced cognitive load and stress. In this section, we demonstrate a possible way of integrating these theoretical foundations into a model that can systematically predict the suitability of a certain modality choice for this presentation task. Several suggestions on adapting this model to other applications are also given.

Again, we assume that the set of available modalities consists of text, image, speech, and sound. Regarding the two basic information elements, all four modalities are suitable to present the victim types, but only text and image are suitable to present the victim locations. Speech can refer to a location by a row index and a column index, or a zone index. Sound can use variations in tone, pitch, or direction to convey location. However, since the grid-map used for this task contains 260 location units (grids), using only auditory modalities without any visual hint actually locating a point on the map would be much too inefficient to convey the locations. It would be particularly hard or even impossible for users to distinguish between 260 sound variations. Therefore, only text and image are chosen as candidates for presenting basic information. The additional aid, if provided, can be presented by all four modality candidates. A total of 10 possible modality choices are identified to be evaluated (Table 2).

A weighted additive utility model (Eq. 1) has been constructed which takes modalities as inputs and outputs a numerical value describing the level of suitability

Table 2 Predicted suitability of 10 possible modality usages

Index	Modality for basic info.	Modality for additional aid	B (0.5)	P (0.3)	M (0.2)	Suitability score
1[a]	text	none	1	0	0	0.5
2	text	text	1	−1	2	0.6
3[a]	text	image	1	−1	1	0.4
4[a]	text	speech	1	1	2	1.2
5[a]	text	sound	1	1	1	1.0
6[a]	image	none	2	0	0	1.0
7	image	text	2	−1	2	1.1
8	image	image	2	−1	1	0.9
9	image	speech	2	1	2	**1.7**
10	image	sound	2	1	1	1.5

[a]Experimental conditions

of the input modality choice. The higher the output value is, the more suitable the input modality choice is:

$$Suitability = f_b \times B + f_p \times P + f_m \times M. \quad (1)$$

The model contains three attributes. For each of them, suitability values are assigned to all modality candidates, based on predictions from relevant theories.

1. **B:** the expressive feature of the modality that presents the basic information. Modality candidates are image and text. Image is more suitable than text to present concrete objects such as victim types (see Sect. 5.1), and thus a 2 is assigned to image and a 1 to text.
2. **P:** the perception property of the modality that presents the additional aid. Four modality candidates are either visual or auditory. Based on the attention and working memory theory, visual aids harm the rescue task and auditory aids benefit the task (see Sect. 5.2). Therefore, a –1 is assigned to the visual modalities and a 1 to the auditory modalities.
3. **M:** the verbal/nonverbal property of the modality that presents the additional aid. Two modality candidates are verbal, and two are nonverbal. According to the working memory theory and the dual-coding theory, verbal aids are more beneficial than nonverbal aids (see Sect. 5.3). Thus, a 2 is assigned to verbal modalities and a 1 to nonverbal modalities.

Furthermore, a weight f is assigned to each attribute, determining how much the attribute contributes to the final suitability score. The summary of the three weights is 1. The basic information and the additional aid are considered equally important, and therefore attribute B gets a weight of 0.5, and P and M get 0.5 in total. Comparing P and M, our experimental results suggest that the difference between visual and auditory aids was more notable than the difference between verbal and nonver-

bal aids, which in turn suggests that P may have a larger influence on the suitability evaluation than M. Therefore, f_p is set to 0.3, and f_m is set to 0.2.

Finally, the suitability predictions for 10 possible modality choices are shown in Table 2. The outcomes for the five investigated conditions are consistent with the experimental results, indicating the validity of this model. The "image + speech aid" combination is predicted to be the best modality choice for this specific presentation task.

This suitability prediction model demonstrates the possibility of quantitatively evaluating the cognitive effects of modalities and systematically selecting the best modality usage for a specific presentation task. To adapt this model to other applications, the following aspects need to be reconsidered: (1) the input: what are the available modalities and possible allocation choices; (2) the output: how to define suitability based on the presentation goal (performance, cognitive load and stress in our case); (3) the attributes: which factors have an influence on the suitability assessment and which criteria can be used to predict the influence; (4) the weights: how large is the relative influence of each attribute.

7 Conclusions

In this study, we emphasized that modality allocation in IMMP system needs to consider the cognitive impact of modalities, especially for high-load and time-critical applications. A user experiment was conducted, the results of which confirmed that the use of modality significantly affected the performance and experienced cognitive load and stress. The experimental findings were well explained by relevant cognitive theories and the expressive features of modalities. Furthermore, a suitability prediction model was constructed to predict the suitability of other uninvestigated modality choices for this specific task. This model demonstrated a possible way of integrating cognitive theories into the modality allocation process in IMMP systems. Further work is needed to evaluate and extend this model for more complex user tasks and a larger set of modalities.

Acknowledgements The user experiment and results presented in this paper have been published previously in [10] and [11].

References

1. André, E.: The generation of multimedia presentations. In: Handbook of Natural Language Processing, pp. 305–327 (2000)
2. Arens, Y., Hovy, E., Vossers, M.: On the knowledge underlying multimedia presentations. In: Intelligent Multimedia Interfaces, pp. 280–306 (1993)
3. Baddeley, A.D.: Essentials of Human Memory. Taylor & Francis, London (1999)
4. Baddeley, A.D.: Working memory: Looking back and looking forward. Nat. Rev., Neurosci. **4**, 829–839 (2003)

5. Baddeley, A.D., Hitch, G.J.: Working memory. Psychol. Learn. Motiv. Adv. Res. Theory **8**, 47–89 (1974)
6. Begault, D.R.: Head-up auditory displays for traffic collision avoidance system advisories: a preliminary investigation. Hum. Factors **35**(4), 707–717 (1993)
7. Bernsen, N.O.: Multimodality in language and speech systems—from theory to design support tool. In: Multimodality in Language and Speech Systems, pp. 93–148 (2002)
8. Bordegoni, M., Faconti, G., Feiner, S., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias, P., Wilson, M.: A standard reference model for intelligent multimedia presentation systems. Comput. Stand. Interfaces **18**(6), 477–496 (1997)
9. Brünken, R., Steinbacher, S., Plass, J.L., Leutner, D.: Assessment of cognitive load in multimedia learning with dual-task methodology: auditory load and modality effect. Instr. Sci. **32**, 115–132 (2004)
10. Cao, Y., Theune, M., Nijholt, A.: Modality effects on cognitive load and performance in high-load information presentation. In: Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'09), pp. 335–344. ACM, New York (2009)
11. Cao, Y., Theune, M., Nijholt, A.: Towards cognitive-aware multimodal presentation: the modality effects in high-load HCI. In: Proceedings of the 8th International Conference on Engineering Psychology and Cognitive Ergonomics: Held as Part of HCI International 2009, pp. 12–21. Springer, Berlin (2009)
12. Chun, M.M., Jiang, Y.: Top-down attentional guidance based on implicit learning of visual covariation. Psychol. Sci. **10**(4), 360–365 (1999)
13. Clark, J.M., Paivio, A.: Dual coding theory and education. Educ. Psychol. Rev. **3**(3), 149–210 (1991)
14. Connor, C.E., Egeth, H.E., Yantis, S.: Visual attention: bottom-up versus top-down. Curr. Biol. **14**(19), 850–852 (2004)
15. Downing, P.E.: Interactions between visual working memory and selective attention. Psychol. Sci. **11**(6), 467–473 (2000)
16. Driver, J., Spence, C.: Cross-modal links in spatial attention. Philos. Trans. R. Soc. Lond. B, Biol. Sci. **353**(1373), 1319–1331 (1998)
17. Eimer, M.: Can attention be directed to opposite locations in different modalities? An ERP study. Clin. Neurophysiol. **110**(7), 1252–1259 (1999)
18. Eimer, M., van Velzen, J., Forster, B., Driver, J.: Shifts of attention in light and in darkness: an ERP study of supramodal attentional control and crossmodal links in spatial attention. Cogn. Brain Res. **15**(3), 308–323 (2003)
19. Elting, C., Michelitsch, G.: A multimodal presentation planner for a home entertainment environment. In: Proceedings of the Perceptive User Interfaces (PUI'01), pp. 1–5. ACM, New York (2001)
20. Feiner, S.K., McKeown, K.R.: Automating the generation of coordinated multimedia explanations. Computer **24**(10), 33–41 (1991)
21. Ferris, T.K., Sarter, N.B.: Cross-modal links among vision, audition, and touch in complex environments. Hum. Factors **50**(1), 17–26 (2008)
22. Fitrianie, S., Poppe, R., Bui, T.H., Chitu, A.G., Datcu, D., Hofs, D.H.W., Willems, D.J.M., Poel, M., Rothkrantz, L.J.M., Vuurpijl, L.G., Zwiers, J.: Multimodal human–computer interaction in crisis environments. In: The 4th International ISCRAM Conference, Delft, The Netherlands (2007)
23. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Hum. Ment. Workload **1**, 139–183 (1988)
24. Heller, R.S., Martin, C.D., Haneef, N., Gievska-Krliu, S.: Using a theoretical multimedia taxonomy framework. J. Educ. Resour. Comput. **1**, 1–22 (2001)
25. Huang, L., Pashler, H.: Working memory and the guidance of visual attention: consonance-driven orienting. Psychon. Bull. Rev. **14**(1), 148–153 (2007)
26. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vis. Res. **40**, 1489–1506 (2000)
27. Johnston, W.A., Dark, V.J.: Selective attention. Annu. Rev. Psychol. **37**, 43–75 (1986)

28. Karagiannidis, C., Koumpis, A., Stephanidis, C.: Adaptation in IMMPS as a decision making process. Comput. Stand. Interfaces **18**(6), 509–514 (1997)
29. Mansoux, B., Nigay, L., Troccaz, J.: Output multimodal interaction: the case of augmented surgery. People Comput. **20**, 177–192 (2007)
30. Mayer, R.E., Anderson, R.B.: The instructive animation: helping students build connections between words and pictures in multimedia learning. J. Educ. Psychol. **84**(4), 444–452 (1992)
31. Mayer, R.E., Gallini, J.K.: When is an illustration worth ten thousand words. J. Educ. Psychol. **82**(4), 715–726 (1990)
32. Mayer, R.E., Moreno, R.: A split-attention effect in multimedia learning: evidence for dual processing systems in working memory. J. Educ. Psychol. **90**, 312–320 (1998)
33. McRoy, S.W., Channarukul, S., Ali, S.S.: Multimodal content adaptations for heterogeneous devices. J. Digit. Inf. **7**(1), 1–34 (2006)
34. Mishkin, M., Ungerleider, L.G., Macko, K.A.: Object vision and spatial vision: two cortical pathways. In: Philosophy and the Neurosciences: A Reader, pp. 199–208 (2001)
35. Moreno, R., Mayer, R.E.: Learning science in virtual reality multimedia environments: role of methods and media. J. Educ. Psychol. **94**(3), 598–610 (2002)
36. Navalpakkam, V., Itti, L.: A goal oriented attention guidance model. Lect. Notes Comput. Sci. **2525**, 453–461 (2002)
37. Paivio, A.: Coding distinctions and repetition effects in memory. Psychol. Learn. Motiv. Adv. Res. Theory **9**, 179–211 (1975)
38. Paivio, A.: Mental Representations: A Dual Coding Approach. Oxford University Press, Oxford (1986)
39. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vis. Res. **42**, 107–123 (2002)
40. Peterson, M.S., Kramer, A.F.: Attentional guidance of the eyes by contextual information and abrupt onsets. Percept. Psychophys. **63**(7), 1239–1249 (2001)
41. Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.C., McTear, M., Raman, T.V., Stanney, K.M., Su, H., Wang, Q.Y.: Guidelines for multimodal user interface design. Commun. ACM **47**(1), 57–69 (2004)
42. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To see or not to see: the need for attention to perceive changes in scenes. In: Psychological Science, pp. 368–373 (1997)
43. Rousseau, C., Bellik, Y., Vernier, F., Bazalgette, D.: Architecture framework for output multimodal systems design. In: Proceedings of OZCHI'04 (2004)
44. Rousseau, C., Bellik, Y., Vernier, F., Bazalgette, D.: A framework for the intelligent multimodal presentation of information. Signal Process. **86**(12), 3696–3713 (2006)
45. Sarter, N.B.: Multimodal information presentation: design guidance and research challenges. Int. J. Ind. Ergon. **36**(5), 439–445 (2006)
46. Simons, D.J., Chabris, C.F.: Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception **28**, 1059–1074 (1999)
47. Simons, D.J., Rensink, R.A.: Change blindness: past, present, and future. Trends Cogn. Sci. **9**(1), 16–20 (2005)
48. Smith, E.E., Jonides, J.: Working memory: a view from neuroimaging. Cogn. Psychol. **33**(1), 5–42 (1997)
49. Smith, E.E., Jonides, J., Koeppe, R.A.: Dissociating verbal and spatial working memory using PET. Cereb. Cortex **6**(1), 11–20 (1996)
50. Smith, E.E., Jonides, J., Koeppe, R.A., Awh, E., Schumacher, E.H., Minoshima, S.: Spatial versus object working memory: PET investigations. J. Cogn. Neurosci. **7**(3), 337–356 (1995)
51. Spence, C., Driver, J.: Audiovisual links in endogenous covert spatial attention. J. Exp. Psychol. Hum. Percept. Perform. **22**(4), 1005–1030 (1996)
52. Spence, C., Driver, J. (eds.): Crossmodal Space and Crossmodal Attention. Oxford University Press, Oxford (2004)
53. Spence, C., Nicholls, M.E.R., Driver, J.: The cost of expecting events in the wrong sensory modality. Percept. Psychophys. **63**(2), 330–336 (2001)

54. Stanney, K., Samman, S., Reeves, L., Hale, K., Buff, W., Bowers, C., Goldiez, B., Nicholson, D., Lackey, S.: A paradigm shift in interactive computing: deriving multimodal design principles from behavioral and neurological foundations. Int. J. Hum.-Comput. Interact. **17**(2), 229–257 (2004)
55. Stanton, N. (ed.): Human Factors in Alarm Design. CRC Press, Boca Raton (1994)
56. Sutcliffe, A.G., Kurniawan, S., Shin, J.E.: A method and advisor tool for multimedia user interface design. Int. J. Hum.-Comput. Stud. **64**(4), 375–392 (2006)
57. Wahlster, W.: Smartkom: symmetric multimodality in an adaptive and reusable dialogue shell. In: Proceedings of the Human Computer Interaction Status Conference, vol. 3, pp. 47–62 (2003)
58. Wahlster, W., Andre, E., Bandyopadhyay, S., Graf, W., Rist, T.: WIP: the coordinated generation of multimodal presentations from a common representation. Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues, pp. 121–144 (1992)
59. Wahlster, W., André, E., Finkler, W., Profitlich, H.J., Rist, T.: Plan-based integration of natural language and graphics generation. Artif. Intell. **63**(1), 387–427 (1993)
60. Wickens, C.D.: Engineering Psychology and Human Performance, 3rd edn. Prentice Hall, New York (1999)
61. Wickens, C.D.: Multiple resources and performance prediction. Theor. Issues Ergon. Sci. **3**(2), 159–177 (2002)
62. Wickens, C.D., Dixon, S.R., Seppelt, B.: Auditory preemption versus multiple resources: who wins in interruption management. In: Proceedings of Human Factors and Ergonomics Society Annual Meeting, vol. 49, pp. 463–467. Human Factors and Ergonomics Society, 2005
63. Wolfe, J.M.: Visual attention. Seeing **2**, 335–386 (2000)
64. Zhou, M.X., Wen, Z., Aggarwal, V.: A graph-matching approach to dynamic media allocation in intelligent multimedia interfaces. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, pp. 114–121. ACM, New York (2005)