Virtual Reality (2023) 27:2493–2507
https://doi.org/10.1007/s10055-023-00822-y

ORIGINAL ARTICLE

Visual working memory in immersive visualization: a change detection experiment and an image-computable model

Chiara Bassano[1] · **Manuela Chessa**[1] · **Fabio Solari**[1]

Received: 28 December 2022 / Accepted: 12 June 2023 / Published online: 27 June 2023

Abstract
Visual working memory (VWM) is a cognitive mechanism essential for interacting with the environment and accomplishing ongoing tasks, as it allows fast processing of visual inputs at the expense of the amount of information that can be stored. A better understanding of its functioning would be beneficial to research fields such as simulation and training in immersive Virtual Reality or information visualization and computer graphics. The current work focuses on the design and implementation of a paradigm for evaluating VWM in immersive visualization and of a novel image-based computational model for mimicking the human behavioral data of VWM. We evaluated the VWM at the variation of four conditions: set size, spatial layout, visual angle (VA) subtending stimuli presentation space, and observation time. We adopted a full factorial design and analysed participants' performances in the *change detection* experiment. The analysis of hit rates and false alarm rates confirms the existence of a limit of VWM capacity of around 7 ± 2 items, as found in the literature based on the use of 2D videos and images. Only VA and observation time influence performances ($p<0.0001$). Indeed, with VA enlargement, participants need more time to have a complete overview of the presented stimuli. Moreover, we show that our model has a high level of agreement with the human data, $r>0.88$ ($p<0.05$).

Keywords Change blindness · Immersive virtual reality · Depth · Visual angle · 3D spatial arrangement · Saliency

1 Introduction

During the last few years, we have seen a growing interest in using immersive Virtual Reality (VR) systems for training and assisting specialized operators in several fields of applications. A wide literature describes such systems, evaluating their contribution to the learning curve and the final performance of people working in specialized contexts (Checa and Bustillo 2020).

There are several technological solutions for implementing VR systems, including head-mounted displays (HMDs). Their use in industrial, manufacturing, or medical contexts is favored because they enable the creation of a virtual replica of an industrial facility to complete a specific, complex task, e.g., an assembling task (Guo et al. 2020), in a controlled and safe setting, as well as immersing the user in an immersive Virtual Environment (VE) by removing external distractions, facilitating active learning (Capasso et al. 2022).

The users usually act in virtual scenarios enriched by added visual information (e.g., the working instructions), displayed around them, exploiting the 3D environment surrounding them. This is different from what happens by using handheld systems, like a tablet, where information is on the display, similarly to what happens with a book. Nevertheless, the actual field of view (FOV) of most HMDs is quite limited: it is about 90 degrees for commercial VR headsets, as the HTC Vive we used in our experiment. Though, there are HDMs with larger FOV, such as the Pimax Vision 8K Plus. It is thus interesting to understand how people can use such an amount of additional information and possibly to have some guidelines to put text, objects, or instruments where the users can see them and be aware of changes and modifications of the scene.

✉ Manuela Chessa
manuela.chessa@unige.it

✉ Fabio Solari
fabio.solari@unige.it

Chiara Bassano
chiara.bassano@dibris.unige.it

[1] Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Via Dodecaneso 35, Genoa 16146, Italy

Indeed, one of the main problems researchers in VR have to solve while designing an application for training/simulation or in the field of visualization and computer graphics is where and how to present information in the immersive VE surrounding the users, so that they can notice and process it efficiently. In 2D screen applications, e.g., desktop computers or touchscreen devices, the relevant information is often notified using several visual factors (color, size, movement, duration). However, in immersive VR the space for interaction and visualization is no longer limited to a 2D screen surface, it is, instead, a 3D volume distributed around the user. Hence, additional cognitive and visual mechanisms must be exploited, such as visual working memory (VWM), attention, spatial memory, distance perception, and peripheral vision. Understanding the functioning of human perception in VR, the amount of information we can access and process efficiently consciously, the influence of the position, and the way information is presented over our ability to perceive it, can significantly improve both the quality and the quantity of information being displayed for it to be efficiently noticed and processed by the user (Seinfeld et al. 2020; Healey and Enns 2011).

In this paper, we analyze VWM in immersive VR by using an experiment that replicates the *one-shot change detection* task (Rensink 2005), in which participants are required to detect a change in the VR environment. Prior literature reports studies with 2D displays (Luck and Vogel 2013; Cohen et al. 2016). Here, we adapt the same paradigm to a simple immersive scene, considering stereoscopic visualization. Specifically, we consider two 3D layout arrangements: the vertical one, where the objects are at the same distance in front of the observer, as on a wall, and the horizontal one, where the objects' distance with respect to the observer changes, as on a table. We can control the 3D position of the objects and their angular position with respect to the observers. Although, in principle, we can put objects all around the observers, completely exploiting the 3D environment, i.e., all the 360-degree scene around them, we focus on three visual angles (VAs) subtending stimuli presentation space: 40, 80, and 120 degrees. Specifically, the vertical layout and VA 40 degrees experimental condition replicates the standard 2D experiment with the addition of stereopsis, allowing us to bridge the gap with the standard literature and defining a baseline for the interpretation of the results obtained in our experiment. Moreover, we propose a computational model based on the same virtual images shown inside the HMD, and we show that this model can replicate the results obtained by humans in detecting changes in an immersive VR scenario.

The paper has the following main contributions:

- Results of our experiment confirm that the limitations of VWM capacity (i.e., the largest number of items for which an observer can identify a change with a certain accuracy) persist in immersive visualization, that simulated variations in depth have no effect on change detection ability, and that only VA and observation time affect the performances;
- An image-computable model of the VWM can be applied to quantify how a change in a scene could be effectively detected by an observer immersed in a VR scenario, which could be exploited to design immersive visualization systems and find the better layout of visual information.

2 State of the art

2.1 Visual working memory in the literature

Experimental results indicate that the bandwidth of human perception is severely limited, and this seems to have a physiological basis(Luck and Vogel 2013). Several theories try to explain this phenomenon (Cohen et al. 2016). We have a rich experience of the world, but all this information cannot be fully captured by our capacity-limited cognitive mechanisms (Block 2011). Information is not consciously perceived until it is accessed by higher-order systems, i.e., attention, VWM, and decision-making (Kouider et al. 2010; Lau and Rosenthal 2011).

VWM is an apparatus dedicated to actively maintaining visual information to serve the needs of ongoing tasks. Over the last decades, research within cognitive psychology and visual perception has demonstrated that VWM is limited in terms of time, it decays in some seconds unless information reinforcement occurs, and of capacity, i.e., the number of information that can be stored. In the literature, studies on the assessment of VWM capacity can be divided into two main categories: reductionist approaches and real-world or natural behavior paradigms (Kristjánsson and Draschkow 2021). The former studies the mechanisms of visual cognition in a pure sense, breaking them down into fundamental operations measured with simple stimuli. The latter investigates the functional nature of visual perception and cognition within active natural behavior.

Reductionist approaches are milestones of the research in cognition and perception. However, most works found are based on the use of 2D images or videos displayed on standard desktop screens occupying ~30°–40° of the visual angle. An equivalent systematic approach exploiting immersive 360° VE does not exist yet, at least to our knowledge.

However, research on defining how VWM capacity is limited is still controversial, and the answer often depends on the task. Indeed, capacity estimates may not always generalize across tasks, and performance across tasks cannot be modeled by a common set of parameters

(Robinson et al. 2020). Many studies quantify VWM in terms of items and claim that a plausible value for the capacity could be around 4 (Cohen et al. 2016; Cowan 2001; Luck and Vogel 1997) or 7 ± 2 (Franconeri et al. 2007; Miller 1956) items, depending on the task. While other authors more generically refer to chunks (Brockmole and Henderson 2005), i.e., higher-order representations in which individual pieces of information are interassociated and stored in memory and act as a coherent, integrated group when retrieved. Authors in (Alvarez 2011) found a monotonic relation between the amount of information per item and the reciprocal number of items that can be memorized, implying a trade-off between complexity and quantity of the remembered information. Moreover, people tend to vary the VWM load accordingly to task demand to reduce the effort involved, especially during longer-duration tasks (Thornton et al. 2020).

Multiple objects tracking (Pylyshyn and Storm 1988), visual search (Wolfe 2012), cueing studies (Posner et al. 1978), visual foraging (Wolfe 2013), model arrangement reproduction (Ballard et al. 1995), recognition (Endress and Potter 2014; Brady and Störmer 2021; Zhang and Luck 2008) and change detection (Luck and Vogel 1997) are the most commonly diffused tasks used for VWM assessment (Kristjánsson and Draschkow 2021).

In *change detection* tasks, participants are first shown an array of items or a scenario, followed by a blank screen or a mask, and are asked to remember them. Subsequently, a full set of items is presented at test (*whole-display* (Rouder et al. 2011)) with or without a novel element, and participants are asked to judge whether a change occurred in the array or not. Arrays have increasing size and VWM capacity is defined as the largest array size for which observers are able to identify a change with a certain accuracy. Stimuli provided can also be videos, in which the change occurs gradually (Simons et al. 2000), sequences of images, in which the change is contingent on an event (such as a brief flash, eye movement, or occlusion) that creates a global motion signal that masks the transient (Rensink et al. 1997), or real life settings (Simons and Levin 1998). This inability to detect expected or unexpected change between two different pictures when a brief interruption occurs between them or the change occurs so gradually that it does not automatically draw attention is called change blindness (CB). *Change detection* experiments usually fall into the category of *intentional* CB, as observers are directly asked to look for changes. The term *incidental* CB, instead, refers to those experiments where participants are instructed to perform a certain task and, in the end, they are asked to remember if they have noticed some change (Varakin et al. 2007).

2.2 Visual working memory in immersive visualization

The majority of articles found in the VR literature adopt the real-world paradigm. They usually focus on spatial representation construction and navigation of a VE, both indoor and outdoor (Read et al. 2022; Jaiswal et al. 2010; Meilinger et al. 2008; Gras et al. 2013), or on *recognition* tasks (La Corte et al. 2019), or on the *incidental* CB induction, i.e., participants are asked to accomplish a task (sort objects, walk or drive) and unexpected changes are applied to the scene (Suma et al. 2011; Karacan et al. 2010; Marwecki et al. 2019). Nonetheless, studies using the reductionist approach in VR exist, i.e., *multiple-object tracking* (Lochner and Trick 2014), *visual search* (Li et al. 2018), *cueing* (Seinfeld et al. 2020), *model arrangement reproduction* (Draschkow et al. 2021).

In the literature, we can identify two paradigms for *change detection*, the *flicker*, and the *one-shot* (Rensink 2005). In the first case, the original and modified (test) images are presented in a loop, separated by a blank screen (or retention mask), until the observer finds the change. In the second case, the sequence is shown once, and participants have to answer within a certain time, usually from 1 s to a few tens of seconds. In both paradigms, the duration of the retention mask, namely, inter stimulus interval (ISI), can vary from 20 ms to 9 s (Phillips 1974). Still, in general, experimenters prefer using ~500 ms, in order to mask the transient without impairing VWM. Stimulus duration, instead, is usually around 275–300 ms, i.e., the minimum time required to extract the gist of a scene, i.e., the observer's experience of grasping the meaning of a scene with a simple glimpse, (Cohen et al. 2016). Authors in (Steinicke et al. 2011) found that *the flicker* paradigm causes simulation sickness when used for semi-immersive and immersive VR experiments and suggested as an alternative solution the projection of two different dephased images onto the two eyes.

Change detection experiments have identified different factors influencing CB, including attention, interest, or visual stimulus saliency, i.e., object perceptual properties that catch human attention. It depends both on low-level features (color, shift, rotation, appearance/disappearance, elements distribution) and high-level characteristics (coherence/incoherence of the modified element with respect to the semantics of the scene). Recently, a limited set of visual features that are detected rapidly by low-level, fast-acting visual processes have been identified (e.g., hue and curvature). Their detection can occur in less than 200–250 ms, before the start of a saccade, which takes 200 ms. For this reason, they are referred to as preattentive processes, preceding focused attention (Healey and Enns 2011). Considering the low-level stimulus features that influence change detection ability, authors in (Gusev et al. 2014; Simons

et al. 2000) demonstrated that the appearance and disappearance of an item are more easily and fast detected than other modifications.

2.3 Models of visual working memory and saliency

There is a rich literature about descriptive models of VWM (Ma et al. 2014; Fougnie et al. 2012; Bays and Husain 2008; Luck and Vogel 1997) and models based on signal detection theory (Williams et al. 2022; Van den Berg and Ma 2018; Wilken and Ma 2004). In (Brady and Tenenbaum 2013) the authors present a probabilistic model of VWM by using summary parameters of the stimuli, and in (Van den Berg et al. 2017) a visual feature is used for leading to a Fechner model of VWM confidence. A normative proposal, where the expected performance of the task is balanced with the cost of spending neural resources for coding it, is presented in (Van den Berg and Ma 2018). In (Schneegans et al. 2020) the authors show that the discrete versus continuous nature of sampling is not critical to model fits by using a sampling interpretation of population coding of the visual parameters.

It is worth noting that such models do not process directly the same visual stimuli of the participants of an experiment.

Since visual saliency can have a role in modulating working memory storage capacity and can provide insights into working memory functions, an interest in neural modeling has emerged to understand its neural basis, e.g., (Brunel and Wang 2001; Compte et al. 2000). These models are based on low-level modeling of the neurons, e.g. biophysically realistic attractor network with spiking neurons has been proposed in (Dempere-Marco et al. 2012). Rather, we are interested in proposing a functional model that is able to capture essential aspects of visual working memory by exploiting the scene saliency as the model input.

More specifically, saliency models can be used for assessing the effectiveness of visualizations, e.g., (Matzen et al. 2017; Polatsek et al. 2018), whereas few studies use the saliency models to mimic human behavioral data by using images as input: in (Fine and Minnery 2009) the authors assess how visual salience affects performance in a Working Memory task, and in (Ma et al. 2013) the authors propose a computational model that is able to predict degrees of CB. A saliency model is used in (Pedale and Santangelo 2015) to assess the relevance of the salience in a memory task. Our model is aiming to mimic more complex human outcomes (such as hit rate and false alarm rate as a function of the conditions of the experiment) by using the same images humans observed.

In the literature, plenty of saliency detection methods have been proposed (Cong et al. 2018), also with deep learning approaches (Wang et al. 2019; Li et al. 2021; Fang et al. 2016). The existing models can be evaluated by using standard datasets (e.g., the MIT/Tuebingen Saliency Benchmark (Kümmerer et al. 2018)). Many saliency detection methods, which provide a topographic representation of the stimulus relevance for human attention (i.e., a saliency map), are coherently presented and implemented in (Wloka et al. 2018):

- AIM (Attention by Information Maximization): a bottom-up model that is based on the principle of maximizing information sampled from a scene and it is neurally plausible (Bruce and Tsotsos 2005).
- AWS (Adaptive Whitening Saliency): it is based on the adaptation of the basis of low-level features to the statistical structure of the image (Garcia-Diaz et al. 2012).
- CAS (Context Aware Saliency): it detects the image regions that represent the scene by exploiting principles observed in the psychological literature (Goferman et al. 2011).
- CVS (Covariance-based Saliency): the method uses covariance matrices of simple image features as meta-features for saliency estimation (Erdem and Erdem 2013).
- DVA (Dynamic Visual Attention): it maximizes the entropy of the sampled visual features, which represent an image patch as a linear combination of sparse coding basis functions(Hou and Zhang 2009).
- FES (Fast and Efficient Saliency): the method is based on estimating the saliency of local feature contrast in a Bayesian framework (Tavakoli et al. 2011).
- GBVS (Graph-Based Visual Saliency): a bottom-up saliency model based on graph computations, by exploiting activation map of feature vectors (Harel et al. 2007).
- IKN: (Itti-Koch-Niebur Saliency Model): the method is biologically inspired and it uses a linear combination of a set of feature maps from three complementary channels as intensity, color, and orientation (Itti et al. 1998).
- IMSIG (Image Signature): the algorithm is based on the image signature that approximates the foreground of an image within the framework of sparse signal mixing (Hou et al. 2011).
- LDS (Learning Discriminitive Subspaces): the method estimates saliency by learning a set of discriminative subspaces that select targets and suppress distractors (Fang et al. 2016).
- QSS (Quaternion-Based Spectral Saliency): the method uses spectral saliency and the color space has an important influence (Schauerte and Stiefelhagen 2012).
- RARE2012 (Multi-scale rarity-based saliency model): it is bottom-up and selects salient information based on multi-scale spatial rarity by using the colour and orientation features (Riche et al. 2013).
- SSR (Saliency Detection by Self-Resemblance): the method computes local regression kernels from the given

image and visual saliency is obtained by using a self-resemblance measure (Seo and Milanfar 2009).
– SUN (Saliency Using Natural statistics): the method describes how to combine bottom-up and top-down information within a Bayesian probabilistic framework (Zhang et al. 2008).

3 Experiment

3.1 Methods

Considering our goal is to understand the quantity of information the viewers can gather and process simultaneously in an immersive VE and the influence of the presentation modality and the information location on their ability to efficiently notice the information presented, we focused on the *change detection* task, adapted to immersive VR. We furthermore decided to adopt the reductionist approach. Moreover, we opted for the *one-shot* paradigm, ensuring we avoid simulation sickness while providing a well-established solution.

According to the prior literature, we have decided to use simple stimuli, i.e., light blue spheres on a plain light background, to avoid undesired effects of attention prioritization that would bias results, and focused on the appearance or disappearance of an item.

Finally, we aim to model the processing involved in the VWM by exploiting the same images observed by the experiment participants. Thus, we use the saliency maps provided by the fourteen methods (i.e, different ways of objectively estimating the saliency in scene images), presented in Sect. 2.3, as the input to our model. From these fourteen saliency maps as input, we model the behavioral data by developing the processing steps for obtaining a modeled hit rate and false alarm rate similar to the ones in our experiment.

3.2 Task

The adaptation of the *one-shot* paradigm for VR scenarios introduces many aspects that were not involved in previous experiments with 2D stimuli visualization: such as stereopsis, motion parallax (if the user can move the head), possible occlusions, and shadows.

We consider two different layouts, see Fig. 1. In the vertical case (later indicated by V), we replicate the standard 2D experiments but replace surfaces with volumes and 2D stimuli with spheres, introducing stereopsis (even if displayed objects are at the same distance and have the same dimension, so there are no disparity differences). The layout and VA 40° experimental condition replicate the standard 2D experiment, allowing us to bridge the gap with the standard literature and define a baseline for interpreting the

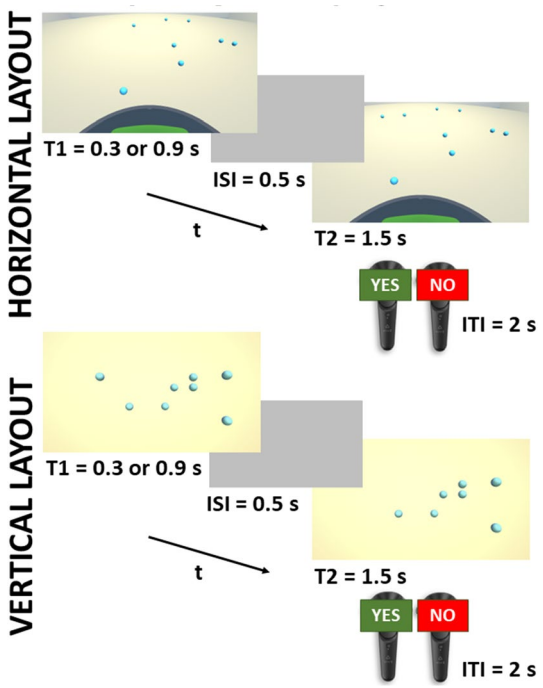

Fig. 1 Schematic of a single trial of the experiment. The *one-shot change detection* paradigm is implemented for the horizontal (top) and vertical (bottom) layout, with the interaction modality used in the experiment

results obtained in our experiment. In the horizontal case (later indicated by H), we exploit the 3D nature of VR, by adding simulated variations in object locations in depth: the visual size of objects decreases with distance (perspective), and occlusions may appear, though we designed the system not to have physically overlapping objects.

As shown in Fig. 1, to replicate the *one-shot change detection* paradigm, we display a memory array for an observation time T1 (300 ms or 900 ms), then, a gray canvas covers the entire headset FOV for an ISI of 500 ms. Finally, a test array is shown for a time T2 of 1.5 s. During the test, the original set of spheres is shown with or without any modification, e.g., a sphere could be added or removed, or no change is applied (control condition). In the *change detection* task, participants have to answer the question *"Has something changed in the scene?"* by pressing the trigger button of one of the controllers, "Yes" or "No", within 1.5 s. Participants can decide in which hand they want to handle them. The answer timer duration was calculated using the mean response time obtained in a pilot study. Once the participant has answered or the answer timer has expired, the test array disappears, and a new scene is automatically presented after an inter trial interval (ITI) of 2 s.

The VE, developed in Unity, is simple to prevent any interest prioritization based on stimulus features or the semantics of the scene. It comprises a bare room with no

windows or furniture except for a green office chair and a semi-circular desk or a semi-cylindrical wall, on which items, i.e., light blue spheres, are generated. The green chair facilitates the localization of the starting position and increases the sense of presence in the VE.

As illustrated in Table 1, VWM capacity has been measured at the variation of four different factors: the set size (4, 6, 8, 10, 12 objects), the distribution of the objects, both in terms of spatial layout (vertical or horizontal) and VA (40, 80 and 120 degrees) and the observation time (300 and 900 ms).

Since VR technologies allow us to exploit a 3D scene, i.e., a 360° space surrounding the viewer, stimuli presentation is no longer limited to a space subtending a VA of 30°–40°, as in the experiments with standard PC screens found in the literature. Indeed, we decided to gradually enlarge the VA, starting from 40°. In VA 40° case, stimuli are presented within the near peripheral vision area, where visual acuity is high, and humans are more sensitive to colors and shapes. VA 80°, instead, refers to the mid-peripheral vision, which is more sensitive to motion signals, and corresponds to exploiting entirely HTC Vive's FOV, which is namely 110°, but, due to lenses distortion, is limited to 90°. In VA 120° trials, participants must turn their heads to get a complete overview of the presented stimuli. Thus, the representation of the scene results from integrating information across different views.

3.3 Scene layout and stimuli arrangement

Items are distributed following a 6×n grid, subtending a 30° × VA (40°, 80° or 120°) visual angle and composed of 5° × 5° bins. The light blue spheres are placed at the center of the bins (Fig. 2a and b left). They have a radius of 2.5 cm to occupy around 4° of the VA at a distance of 70 cm and fit the 5° × 5° bins without overlapping.

The n bins columns are symmetrically arranged with respect to the participant center of view, from −20° to +20°, or from −40° to +40°, or from −60° to +60°, depending on the VA. The vertical 30° VA, instead, has been defined differently for the two layouts. In the vertical layout case, the head position is assumed to be 70 cm from the semi-cylindrical wall (Obj_Dist). Hence, the height of the area (V_Size) subtending a VA of 30° has been calculated using Eq. (1),

$$\text{V\_Size} = 2\ \text{Obj\_Dist} \tan\left(\frac{\text{VA}}{2}\right) \tag{1}$$

This area has been divided into 6 rows, symmetrically arranged with respect to the participant's center of view (Fig. 2a right). In the horizontal case, instead, the surface where items are spawned is not orthogonal to the participant's eyes, thus the area subtending a vertical VA of 30° is

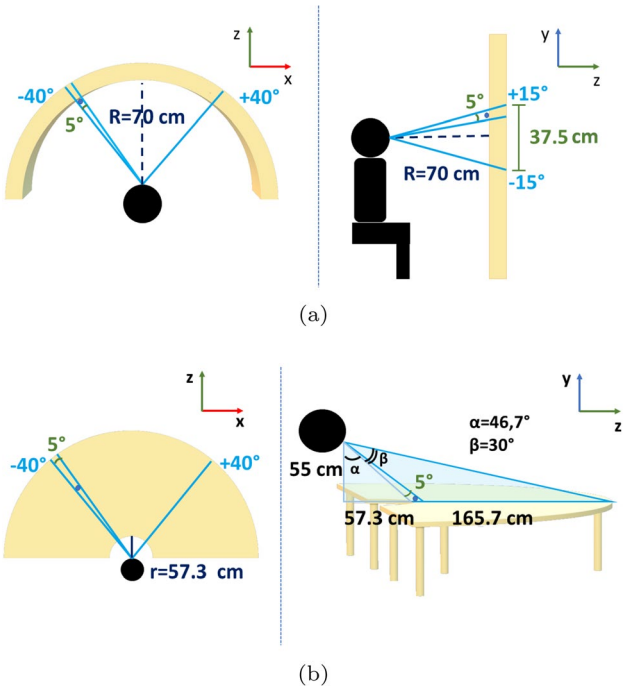

Fig. 2 Spatial distribution of items in the (**a**) vertical and (**b**) horizontal layout cases from the top (left) and lateral (right) points of view (example layout for VA 80°). In both layouts, the spheres are arranged in 5° × 5° bins computed as described in the text

Table 1 Independent and dependent variables of the change detection experiment

Independent variables
Set size (4, 6, 8, 10, 12 items)
Spatial layout (V and H)
VA (40, 80, 120 degrees)
Observation time (S and L)
Dependent variables
HR and FAR

defined by combining Eq. (1) and simple geometry. Considering the head had a fixed position of 55 cm above the table and the perspective of the scene, we first needed to calculate the minimum distance at which spheres occupy around 4° of the VA without overlapping, (Fig. 2b right). Then, we could define the area subtending a vertical VA of 30° and the position of the 6 rows. In both cases, the participant's head position is approximated with the headset position recorded at the beginning of each experimental block, and the table/wall location is adjusted accordingly.

Finally, observation times are chosen considering the previous literature: a 300 ms observation time (later indicated by S) is enough for participants to generate the gist of the scene (Cohen et al. 2016) in the VA 40° and 80° trials. It avoids neither preattentive processing, requiring 200–250

ms (Healey and Enns 2011), nor subitizing, i.e., the direct perceptual apprehension of the numerosity of a group, which is in the order of 30 ms for each extra element (Svenson and Sjöberg 1983), but prevents participants from counting spheres. As in the VA 120° case, participants have to turn their heads, a higher observation time is necessary. Considering that the VA has been tripled, we have decided to increase the observation time proportionally and set it to 900 ms (later indicated by L).

3.4 Measures

During task execution, we collect the given answers, the spheres' distribution, and the head positions and rotations. Participants who demonstrated not having understood instructions properly, i.e., they could not answer on time in the majority of trials, were excluded from further analysis. We discarded 10.2% of trials in the *change detection* test.

Hit rate (HR), see Eq. (2), and false alarm rate (FAR), see Eq. (3), derived from the signal detection theory, are then calculated from the given answers per each participant and experimental condition. HR and FAR are common metrics for the evaluation of performance in *one-shot change detection* tasks: HR considers the number of times people correctly report a change (TP, i.e., true positive) divided by all trials with change (TP+FN), where FN denotes false negative; while FAR is computed as the number of times participants perceive a change in the control condition (FP) divided by all the trials without any change (FP+TN). In the *change detection* case, answers are given in a binary form (Yes/No).

$$\mathrm{HR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{3}$$

From previous literature, we expect high HRs and near-to-null FARs when the number of items to be remembered does not exceed VWM capacity. Once this capacity is overcome, instead, people should start answering randomly, thus HRs should decrease and FARs should increase (Keshvari et al. 2013).

In order to determine the effect of set size, spatial layout, VA and observation time on HR and FAR (see Table 1), we performed an N-ways ANOVA.

After having calculated the mean HRs per experimental condition averaged across participants, we have defined a threshold of 0.75 from (Franconeri et al. 2007; Cowan et al. 2005), slightly adapted to keep more data, and excluded from further analysis trials referred to the number of items for which the mean HR falls below the established value.

On this subset of data, we have computed the HR associated with each possible position of the 6×n grid, in order to understand the influence of the absolute position of the modified item on the probability of correctly detecting the change. Subsequently, we have considered separately the mean HR distribution at the variation of the horizontal and vertical angular distance from the participant VA center.

3.5 Image-computable model

The human visual system has the ability to attend to only salient locations in an observed scene (Itti and Koch 2001). This ability allows humans to only allocate perceptual and cognitive resources on task-relevant visual input. We developed an image-based computational model that accounts for the human performances of our VWM experiment. This model takes the same images observed by the participants of our experiment as input and recreates the observed data of human VWM: specifically, HR and FAR of our experiment. The proposed model aims to capture only the essential aspects of the neural and cognitive processes involved in VWM. Moreover, we discuss which aspects of human VWM are not described by the model and might be interesting starting points for further investigation.

Saliency maps might be an important factor in modeling memory tasks, as reported in the literature (Foulsham and Underwood 2007, 2008; Underwood and Foulsham 2006; Stirk and Underwood 2007), thus we consider saliency methods as the front-end of our model. From this input, by exploiting the information embedded in the saliency map, we model the processing steps that allow obtaining HR and FAR similar to the human ones. It is worth noting that the model does not use saliency maps to predict where people will look in an image but to predict human performances in terms of HR and FAR.

Since we use existing methods of the literature to estimate saliency maps, as the input to our model, we must pay attention to ensuring consistency and procedural correctness for the results obtained by the different methods. With this aim, we use the software package SMILER provided by (Wloka et al. 2018). In this package, fourteen methods (from the classical ones, e.g., (Itti et al. 1998), to the learning ones, e.g. (Fang et al. 2016), see Sect. 2.3) are implemented in MATLAB and we use them in our simulations. It is worth noting that we do not consider saliency methods for dynamic scenes, since in *one-shot change detection* paradigm the transition between the two views is hampered.

Figure 3 shows the saliency maps of the HMD views (as observed by a participant of our experiment) for the VA 80°, horizontal and vertical layouts, by considering the methods LDS and SSR, as an example of the fourteen methods we use as input to our model (see Sect. 2.3). The spheres are

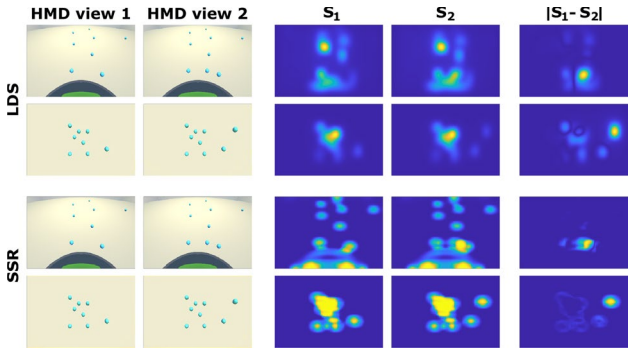

Fig. 3 The HMD views and the related saliency maps for the LDS and SSR methods. The rightmost column shows the absolute value of the differences between the saliency maps $S_1(x, y)$ and $S_2(x, y)$ of the HMD views before (view 1) and after (view 2) the change

detected, but the table structure produces a response too. The saliency map values are in the range $[0 - 1]$.

For the *change detection* task, we have observed that HRs decrease and FARs increase as a function of the set size, i.e., when the number of items increases, thus we devise how this behavior can be based on the information embedded in the saliency maps.

The saliency map $S(x, y)$, where x and y are the image coordinates, encodes the number of spheres (i.e., items) as salient objects (e.g., see Fig. 3). So, the active areas of the saliency map can embed information about the objects in the scene. When the objects are close, the related salient areas merge into one salient region that is proportional to the number of objects. Since it is based on the images observed in the HMD, we consider the same view of the observer, including the perspective cue. Thus, the whole map activity could be a measure of the number of items, at least as observed by the participants in the experiment. Since there is a difference of one item in the *change detection* test, we can consider the difference $D(x, y)$ between the saliency map $S_1(x, y)$ of the view of the scene before the gray canvas covers the entire participant's view and the saliency map $S_2(x, y)$ of the view after the canvas. The absolute value of such a difference can be a measure of residual saliency and could drive the change detection.

To show how the considered saliency methods are able to detect (i) the salient elements (i.e., spheres) of the experiment scenes and (ii) the differences between the two presented scenes, we consider two methods (i.e., LDS and SSR) in Fig. 3. The first two columns show the HMD views before and after the occluding canvas for the VA 80°, both horizontal and vertical layouts. The absolute value of the differences between the saliency map $S_1(x, y)$ and $S_2(x, y)$ is shown in the last column of Fig. 3: there is variability among the saliency maps of the different methods that will reflect in their capacity to account for human data.

It is worth noting that the saliency methods process the same stimuli viewed by the subjects during the experiments.

To sum up, the absolute value of the difference between the saliency maps (before and after the change) might drive the change detection, since it can be considered a measure of residual saliency. The whole map activity might take into account the number of items in the scene, thus it can be used to mimic the behavior of the HR when the number of items increases. However, to take into account the inherent uncertainty in human judgments, we have to consider also the noise present in the human neural processes.

The modeled HR (HR_M) can be described by

$$HR_M = c_1 \frac{\max(|S_1(x,y) - S_2(x,y)|)^{\alpha}}{\left(\sum_x \sum_y S_1(x,y)\right)^{\beta}} + n_1, \tag{4}$$

where c_1 is a normalization term to limit the modeled HR in the range of human HR; $\max(\cdot)$ finds the maximum difference over all pixel coordinates; $|\cdot|$ denotes the absolute value; α and β are static non-linearity; n_1 is the noise from a normal distribution, with a mean of zero and its standard deviation is a fraction of the average HR_M.

Consequently, the modeled FAR (FAR_M) can be described by

$$FAR_M = c_2 \ \max(|S_1(x,y) - S_2(x,y)|)^{\gamma} \cdot \left(\sum_x \sum_y S_1(x,y)\right)^{\delta} + n_2, \tag{5}$$

where c_2 is a normalization term to limit the modeled FAR in the range of human FAR; γ and δ are static non-linearity; n_2 is the noise from a normal distribution, with a mean of zero and its standard deviation is a fraction of the average FAR_M.

3.6 Apparatus

The experimental setup is composed of the HTC Vive and an Alienware Aurora R5 with a 4 GHz Intel Core i7-6700K processor, 16 GB DDR4 RAM (2,133MHz), and an Nvidia GeForce GTX 1080 graphic card.

In change detection experiments, the use of a chin-rest is standard practice as it ensures a constant eyes-stimulus distance. Conversely, in some of the trials, e.g., those with VA 120° and long observation time, participants are required to rotate their heads, thus we decided not to use it. In this manner, the same setup is maintained in all experimental conditions. However, at the beginning of each experimental block, participants are asked to place their head in the starting correct location, indicated by a white sphere in the VE, leaning their back against the back of the chair they are sitting on. They are encouraged to find a comfortable position and maintain it for the entire duration of the block since

they are only allowed to rotate their head. Additionally, an experimenter supervised the experimental session to ensure that the instructions were correctly followed.

3.7 Procedure

We used a full factorial design by considering the independent variables shown in Table 1: we obtained 1080 trials per experiment. In order to avoid simulator sickness and the negative effect of fatigue and prolonged task workload on results, we grouped the trials into 12 blocks of 90 trials and divided the experiment into 3 sessions composed of 4 blocks, with a minimum inter-session break of half-day. Each block could last a minimum of ~6 min and a maximum of ~13 min, according to the observation time (S or L) and the time to answer (1.5 or 5 s). In each block, the spatial layout, the VA and the observation time were fixed, while the set size and the kind of change varied. We randomized the experimental runs within subjects and the block presentation order between subjects. Moreover, each participant saw different spheres configuration, as the arrays of items were randomly generated.

An optional demo scene precedes the actual trial to familiarize participants with the interface and the task.

3.8 Participants

Eighteen participants accomplished the *change detection* test (age range 20–36, 25.1 ± 3.8 years), by completing 1080 trials each.

They were all students, PhDs, and researchers at the University of Genoa and had to sign an informed consent. They all reported having normal or corrected-to-normal vision and no deficit in stereo vision.

4 Results

4.1 Experiment: change detection test

First, we evaluated the influence of set size, layout, VA, and observation time on participants' ability to detect changes. In all experimental conditions, the HR tends to decrease as the number of items increases (see Fig. 4) and reaches the threshold value around 6–8 items, confirming the VWM capacity limit found in the literature (7 ± 2 items) (Franconeri et al. 2007; Miller 1956).

Concurrently, the standard deviation increases, indicating a higher uncertainty of answers. N-ways ANOVA statistical analysis highlights a significant effect of the set size ($F(4, 965)=146.62$, $p < 0.0001$), with all groups marginal means being significantly different from each other ($p < 0.02$). FAR, instead, slightly increases with the number of items ($F(4, 965) = 3.68$, $p < 0.02$). However, large standard deviations highlight a high data variability and make differences not statistically significant. Only marginal means referred to 4 and 12 items differ significantly ($p<0.02$). Layouts alone seem to have no influence, but we found a joint interaction of layout and VA over HRs ($F(2, 967) = 3.99$, $p < 0.02$): in the vertical layout case, the marginal mean referred to VA 120° is different from VA 40° ($p < 0.0001$) and 80° ($p < 0.02$) means. Better performance is associated to the longer observation time (HR: $F(1, 968) = 33.54$, $p < 0.0001$; FAR: $F(1, 968) = 17.17$, $p < 0.0001$) and smaller VAs (HR: $F(2, 967) = 11.29$), $p < 0.0001$; FAR:$F(2, 967) = 56.46$), $p < 0.0001$). Results in the VA 40° and 80° trials are comparable and differ from those obtained with VA 120° ($p < 0.02$), where HRs decrease faster and FARs are higher. Moreover, in the FAR case, a joint influence of VA and observation time has been highlighted ($F(2, 967) = 4.18$, $p < 0.02$).

On the subset of data exceeding 0.75 accuracy, we have computed the mean HR distribution as a function of the

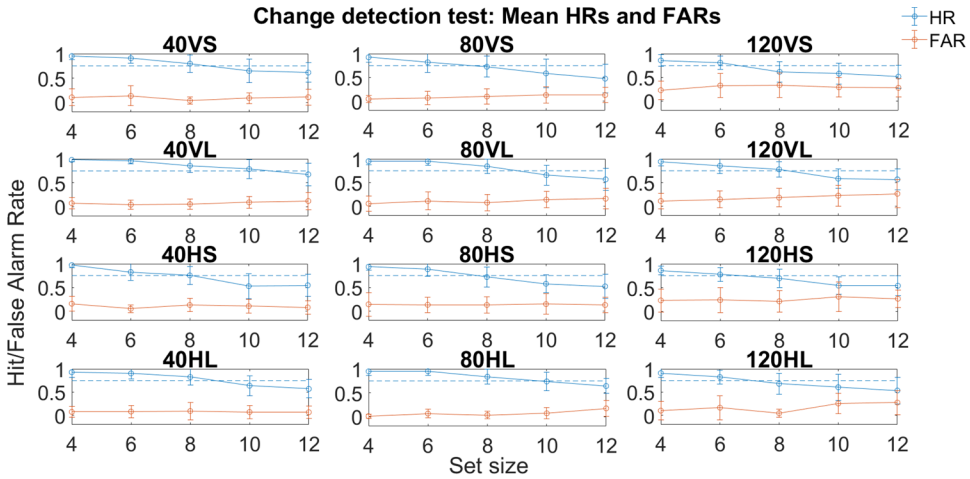

Fig. 4 Mean and standard deviation of HRs and FARs at the variation of the set size in the different experimental conditions, considering the VA (40, 80 or 120), the layout (V or H), and the observation time (S or L) - e.g., 40VS is for visual angle (VA) 40°, vertical (V) layout and short (S) observation time. The dashed line represents the 0.75 threshold

distance from the VA centre. Considering the horizontal VA, as each position has an equal probability to be selected to spawn the modified item, to guarantee the same density of samples, we have grouped data referred to adjacent positions, considering bins of different sizes, 5° for VA 40°, 10° for VA 80°, 15°for VA 120°. For the vertical VA, we evaluate the distance from the center of view in the vertical layout case, and the distance from the user position in the horizontal layout case.

Results in Fig. 5 (HR distribution, dark lines, and gray areas) show that performance is better and with a lower variability when a longer observation time is provided and decreases with VA enlargement. No particular effect of the absolute spatial position of the modified object has been found, except for the VA 120° with short observation time, where accuracy decreases in the periphery of the horizontal VA (Fig. 5, 3rd and 6th column, top). Furthermore, neither the layout nor the distance from the user influences the results, implying that stereopsis and depth perception do not affect participants' ability to detect changes.

We analysed head rotations as a measure of participants' tendency to look around. As headset FOV is ~90°, head rotations provide us an important information: if participants were actually able to see the change, in other words if the change was inside their view. We calculated the normalized histogram of head rotations around the vertical axis during the observation of the memory array, shown in Fig. 5 (yellow bars). As expected, with VA 40° and 80°, rotations are clustered around a central value because all stimuli are presented inside the visual field. Also in the VA 120° trial with short observation time, Fig. 5 (3rd and 6th column, top) head rotations are limited; whereas, having more time, participants explore the entire scene and performance improves, especially in the periphery of the horizontal VA (Fig. 5, 6th column, bottom).

4.2 Model: correlation with human data

The proposed model is tested with the same stimuli and procedures as the human observers (like the model was an individual human participant). In particular, we model the long observation time of our experiments.

We use the default parameters for all the methods. A fitting procedure between the experiment results and the model outputs has been carried out to have both data in the same range of values. However, without an optimization procedure, since we do not adapt our model to each condition, we use the same parameters (see Table 2) for all the conditions.

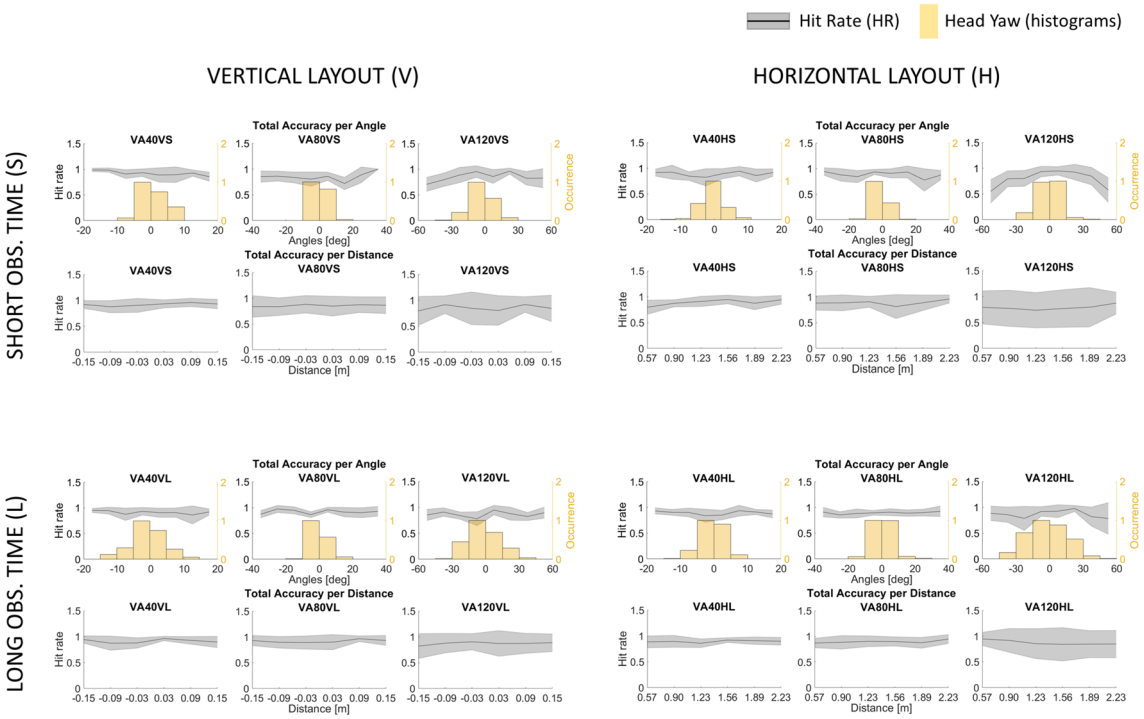

Fig. 5 Change detection test. The 4 subplots represent the combination of V layout (left), H Layout (right), S observation time (top), and L observation time (bottom). In each subplot, the three VA (40°, 80°, 120°) are considered (left-middle-right). For each VA, the mean Hit Rate (HR black solid line) and its standard deviation (gray area) are plotted with respect to the horizontal angle (top, in degrees) and the distance (bottom, in meters). Distance is along the Y axis for the V layout (since the spheres are all at the same depth) and along the Z axis for the H layout (since the spheres are all on the table at the same height), see Fig. 2. Yellow bars represent the normalized histogram of yaw head rotations during the presentation of the memory array. (colour figure online)

Table 2 The specific values of the model parameters employed in our simulations. The standard deviation is std

Parameter	Value	Description
α	1.5	HR_M non-linearity, Eq. (4)
β	1	HR_M non-linearity, Eq. (4)
n_1 std	0.04	Average HR_M fraction, Eq. (4)
γ	1.5	FAR_M non-linearity, Eq. (5)
δ	1	FAR_M non-linearity, Eq. (5)
n_2 std	0.08	Average FAR_M fraction, Eq. (5)

A visual inspection of the modeled HRs and FARs allows us to qualitatively assess the similarity with the corresponding human data. For a quantitative evaluation of the similarity between the model and human performances, we use the Pearson correlation score (r) and the related p value. This metric is widely used in saliency prediction and in assessing agreement with human data (Ma et al. 2013; Maiello et al. 2020; Sitzmann et al. 2018).

Table 3 shows the correlations between modeled HRs (see Eq. (4)) and human HRs. Our model by using the SSR methods has a high level of agreement with the human data ($r>0.88$ and $p<0.05$). On average also the methods AIM, AWS, DVA, IKN and SUN show a good level of agreement. It seems, thus, that such methods embed information about the stimuli that can be exploited by our model to mimic human data.

Figure 6 shows the modeled HRs (blue) and human HRs (red) for two methods in order to allow a visual inspection of the correlation data of Table 3.

On the contrary with respect to the HRs, the correlations between modeled FARs (see Eq. (5)) and human FARs do not show a good agreement. The reason might be that the human FARs do not increase as a function of set size (there are some fluctuations), thus the correlation is not able to get a similarity. However, for some methods, specifically for SSR, the 95% bootstrapped confidence intervals of the mean are overlapped, thus the proposed model shows behaviors similar to human ones. Moreover, the model is able to capture the increase of human FARs as a function of the VA. There is a good agreement for the SSR (r=0.89, p=0.02) method.

5 Discussion

In general, our experiment in immersive VR shows that HRs decrease and FARs increase with the number of items, confirming the previous literature and the hypothesis that accuracy is high when the number of elements to be remembered do not exceed VWM capacity, and decreases once the capacity is overcome, as participants start guessing the answer. In particular, participants can correctly detect changes with 0.75 accuracy when they are asked to memorize a maximum of 6–8 items.

A combined effect of time and VA positively influences performances. In fact, in trials with VA 40° and 80° and 300 ms observation time, participants can only rely on preattentive processes to detect changes, whereas having 900 ms to observe stimuli, they can build a gist of the scene, which usually requires around 275–300 ms (Cohen et al. 2016), but also try to memorize some structures or spatial distributions of elements. The similarity of performances in trials with VA 40° and 80° also suggests that a change in VR is equally detectable when it is applied to the near peripheral or mid-peripheral vision area. However, as we did not use an eye tracker, we can not ensure where the user was looking at. In VA 120° case, instead, the short observation time does not allow participants to turn their heads, as confirmed

Table 3 Correlations r (p value) between modeled HRs and human HRs. The significant agreements with the human data are in bold

Method	VA40H	VA40V	VA80H	VA80V	VA120H	VA120V
AIM	0.77 (0.13)	0.87 (0.05)	0.95 (0.01)	0.90 (0.04)	0.98 (0.00)	0.95 (0.01)
AWS	0.77 (0.13)	0.86 (0.06)	0.96 (0.01)	0.82 (0.09)	0.93 (0.02)	0.94 (0.02)
CVS	0.59 (0.29)	0.84 (0.08)	0.93 (0.02)	0.87 (0.06)	0.72 (0.17)	0.82 (0.09)
DVA	0.76 (0.13)	0.87 (0.05)	0.93 (0.02)	0.86 (0.06)	0.95 (0.01)	0.97 (0.01)
GBVS	0.67 (0.21)	0.76 (0.14)	0.93 (0.02)	0.83 (0.08)	0.94 (0.02)	0.89 (0.04)
IKN	0.66 (0.22)	0.74 (0.15)	0.98 (0.00)	0.88 (0.05)	0.87 (0.06)	0.88 (0.05)
IMSIG	0.67 (0.22)	0.80 (0.11)	0.85 (0.07)	0.82 (0.09)	0.80 (0.11)	0.95 (0.01)
LDS	0.64 (0.25)	0.88 (0.05)	0.83 (0.08)	0.77 (0.12)	0.84 (0.08)	0.90 (0.04)
QSS	0.61 (0.28)	0.80 (0.11)	0.91 (0.03)	0.87 (0.06)	0.80 (0.11)	0.99 (0.00)
SSR	**0.93 (0.02)**	**0.88 (0.05)**	**0.96 (0.01)**	**0.92 (0.02)**	**0.90 (0.04)**	**0.89 (0.05)**
CAS	0.59 (0.29)	0.87 (0.05)	0.92 (0.03)	0.88 (0.05)	0.41 (0.49)	0.94 (0.02)
FES	− 0.73 (0.16)	0.69 (0.20)	− 0.25 (0.69)	0.22 (0.72)	− 0.70 (0.19)	0.28 (0.65)
RARE2012	0.68 (0.21)	0.84 (0.07)	0.91 (0.03)	0.89 (0.04)	0.94 (0.02)	0.91 (0.03)
SUN	0.93 (0.02)	0.96 (0.01)	0.87 (0.05)	0.93 (0.02)	0.83 (0.08)	0.94 (0.02)

The statistically significant ($p<0.05$ for all VAs) agreements with the human data are in bold

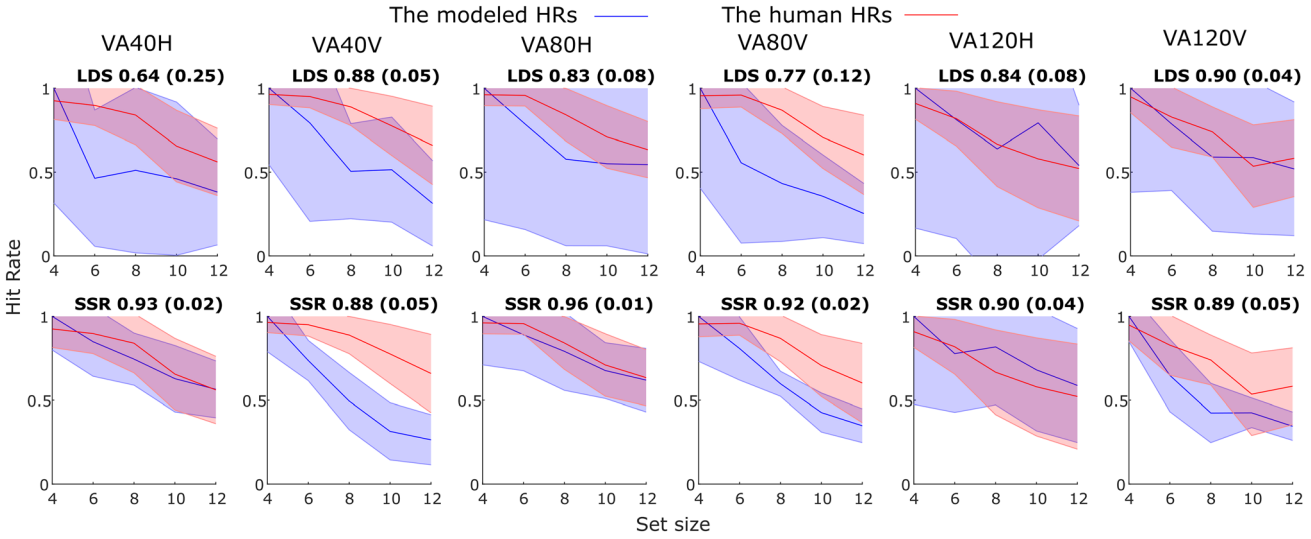

Fig. 6 The modeled HRs (blue) and human HRs (red) for the methods LDS and SSR. The mean is denoted by the line and the standard deviation by the shaded area. The horizontal axis denotes the set size. Correlations r (p value) are reported in the title of each subfigure. (colour figure online)

by Fig. 5, thus they can see two third of the scene and infer an answer based on what they were able to see. While in the long observation time condition, they can rapidly turn their head and have an overview of the entire scene. Performances with VA 40° and 80° and short observation time are comparable to those with VA 120° and long observation time, meaning that the integration of multiple views or the additional workload due to head rotation required in the second case does not compromise participants' ability to memorize items. Thus, the increase of time proportional to VA enlargement can be a good strategy to be also adopted in future experiments. Instead, the absence of an influence of distance and layout on results suggests that additional cues, such as perspective and depth perception in the horizontal case, do not improve participant ability to detect changes.

Furthermore, human behavioral data were modeled through the proposed image-computable model. Equation (4) describes the modeled HR, it takes into account the difference of the scene saliencies between the image before and after the change, modulated by the inverse of the whole saliency of the scene. The rationale is that the HR decreases as a function of the number of items, and the saliency map of the whole scene embeds information about the number of salient objects (i.e., the items), as observed by the experiment participants. Moreover, we need also to consider the inherent uncertainty of the neural processes, thus, we add a normal noise. This model catches the essential aspects of human behavior since it can replicate human HR with a high level of agreement (measured by the Pearson correlation). It is worth noting that only a few saliency methods (e.g., the SSR method) can provide our model with the information necessary to replicate human data. We can observe that a high level of agreement with human data is obtained using a low internal noise level.

The modeled FAR is described by Eq. (5), a consequence of Eq. (4): we hypothesize that the effect of the whole saliency of the scene should be the inverse on FAR since it increases as a function of the set size. Here the agreement with human data is low. The reason might be that the FAR is quite flat, and the correlation is not able to provide a reliable measure. By looking at the bootstrapped confidence, intervals of the mean are overlapped, thus, the model mimics the human pattern. Moreover, the model also correlates well with the increase of human FARs as a function of the VA.

Overall, our results show that the proposed model is able to replicate human data with a good level of agreement.

6 Conclusion

The goal of the current work is to investigate people ability to gather and process the information presented in an immersive VE. In particular, we focus on the assessment of VWM capacity. The literature abounds with articles concerning the assessment of VWM. They usually use the *change detection* paradigm, but stimuli are displayed on 2D screen surfaces, or when immersive VR technologies are used, they employ real world paradigms and focus on the ecological assessment of VWM.

Thus, we adopted a reductionist approach and adapted the standard *one-shot change detection* paradigm to be used in an immersive VR context. We devised an

experiment, i.e., the *change detection* test, and analysed the influence of set size, objects distribution, both in terms of spatial layout and VA, and observation time on the human ability to detect changes. The results show that there is a limitation of VWM capacity in immersive visualization, as previously shown for 2D stimuli, and that depth cue does not affect change detection ability.

Furthermore, we have modeled the human behavioral data of VWM through the proposed image-computable model. We provided the model with the same input images of the experiments with human participants. The model aims to replicate the same human pattern of HR and FAR. To accomplish it, we used the saliency maps of the observed scenes as the input of the model, then the proposed model estimates the quantities of interest. To compute the saliency maps we use an available framework that provides fourteen methods since we are interested in modeling the VWM and not the saliency maps of the scene only. The proposed model can replicate the human performances, as HR and FAR, with a good agreement.

The work has limitations and possible future improvements. First, we considered a limited range of VAs, the maximum considered angle is 120 degrees. Further experiments to test the entire 360-degree world surrounding the user are necessary. Another limitation is that we did not consider some cues, e.g., lower visual acuity in the periphery and out-of-focus depth ranges that may alter users' perceptual acuity, thus VWM. Moreover, we did not check eye movements with an eye tracker. A further development will be to measure eye position and depth fixation to consider such effects.

Finally, we performed the experiment with a simple scene, and only blue spheres in two spatial configurations were considered. A more complex scene and using different objects could enhance the influence of stereoscopy and the 3D layout, with respect to standard 2D stimuli. Indeed, the contribution of perspective cues, motion-parallax, visual object occlusions, and (disparity-based) stereopsis would become more evident. The considered stimuli allowed us to confirm the existence of a limit of VWM capacity of around 7 ± 2 items, as found in the literature based on the use of 2D videos and images, and to devise an image-computable model capable of replicating the human results. The same experiment with more complex objects, shapes, textures, and colors, and with more complicated spatial arrangements, could further validate the devised model. The final aim is to better design the spatial arrangement of information in immersive visualization.

Acknowledgements The authors would like to thank all the colleagues and students involved in the experimental sessions.

Author contribution The work has been conducted when C.B. was with the Department of Informatics, Bioengineering, Robotics and Systems Engineering at the University of Genoa.

Funding Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement. This work has been partially supported by the Interreg Alcotra projects PRO-SOL We-Pro (n. 4298) and CLIP E-Santé (n. 4793).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose

References

Alvarez GA (2011) Representing multiple objects as an ensemble enhances visual cognition. Trends Cogn Sci 15(3):122–131

Ballard DH, Hayhoe MM, Pelz JB (1995) Memory representations in natural tasks. J Cogn Neurosci 7(1):66–80

Bays PM, Husain M (2008) Dynamic shifts of limited working memory resources in human vision. Science 321(5890):851–854

Berg Van den R, Ma WJ (2018) A resource-rational theory of set size effects in human visual working memory. eLife 7(e34):963

Berg Van den R, Yoo AH, Ma WJ (2017) Fechner's law in metacognition: a quantitative model of visual working memory confidence. Psychol Rev 124(2):197

Block N (2011) Perceptual consciousness overflows cognitive access. Trends Cogn Sci 15(12):567–575

Brady TF, Störmer VS (2021) The role of meaning in visual working memory: real-world objects, but not simple features, benefit from deeper processing. J Exp Psychol Learn Memory Cogn 21:2644

Brady TF, Tenenbaum JB (2013) A probabilistic model of visual working memory: incorporating higher order regularities into working memory capacity estimates. Psychol Rev 120(1):85

Brockmole JR, Henderson JM (2005) Object appearance, disappearance, and attention prioritization in real-world scenes. Psychon Bull Rev 12(6):1061–1067

Bruce N, Tsotsos J (2005) Saliency based on information maximization. In: Advances in neural information processing systems, pp 155–162

Brunel N, Wang XJ (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comput Neurosci 11(1):63–85

Capasso I, Bassano C, Bracco F, et al. (2022) A VR multiplayer application for fire fighting training simulations. In: International conference on extended reality. Springer, Berlin. pp 130–138

Checa D, Bustillo A (2020) A review of immersive virtual reality serious games to enhance learning and training. Multim Tools Appl 79(9):5501–5527

Cohen MA, Dennett DC, Kanwisher N (2016) What is the bandwidth of perceptual experience? Trends Cogn Sci 20(5):324–335

Compte A, Brunel N, Goldman-Rakic PS et al. (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb Cortex 10(9):910–923

Cong R, Lei J, Fu H et al. (2018) Review of visual saliency detection with comprehensive information. IEEE Trans Circ Syst Video Technol 29(10):2941–2959

Cowan N (2001) Metatheory of storage capacity limits. Behav Brain Sci 24(1):154–176

Cowan N, Elliott EM, Saults JS et al. (2005) On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. Cogn Psychol 51(1):42–100

Dempere-Marco L, Melcher DP, Deco G (2012) Effective visual working memory capacity: an emergent effect from the neural dynamics in an attractor network. PLoS ONE 7(8):1–20

Draschkow D, Kallmayer M, Nobre AC (2021) When natural behavior engages working memory. Curr Biol 31(4):869–874

Endress AD, Potter MC (2014) Large capacity temporary visual memory. J Exp Psychol General 143(2):548

Erdem E, Erdem A (2013) Visual saliency estimation by nonlinearly integrating features using region covariances. J Vis 13(4):11–11

Fang S, Li J, Tian Y et al. (2016) Learning discriminative subspaces on random contrasts for image saliency analysis. IEEE Trans Neural Netw Learn Syst 28(5):1095–1108

Fine MS, Minnery BS (2009) Visual salience affects performance in a working memory task. J Neurosci 29(25):8016–8021

Fougnie D, Suchow JW, Alvarez GA (2012) Variability in the quality of visual working memory. Nat Commun 3(1):1–8

Foulsham T, Underwood G (2007) How does the purpose of inspection influence the potency of visual salience in scene perception? Perception 36(8):1123–1138

Foulsham T, Underwood G (2008) What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. J Vis 8(2):6–6

Franconeri SL, Alvarez GA, Enns JT (2007) How many locations can be selected at once? J Exp Psychol Hum Percept Perform 33(5):1003

Garcia-Diaz A, Fdez-Vidal XR, Pardo XM et al. (2012) Saliency from hierarchical adaptation through decorrelation and variance normalization. Image Vis Comput 30(1):51–64

Goferman S, Zelnik-Manor L, Tal A (2011) Context-aware saliency detection. IEEE Trans Pattern Anal Mach Intell 34(10):1915–1926

Gras D, Gyselinck V, Perrussel M et al. (2013) The role of working memory components and visuospatial abilities in route learning within a virtual environment. J Cogn Psychol 25(1):38–50

Guo Z, Zhou D, Zhou Q et al. (2020) Applications of virtual reality in maintenance during the industrial product lifecycle: a systematic review. J Manuf Syst 56:525–538

Gusev AN, Mikhailova OA, Utochkin IS (2014) Stimulus determinants of the phenomenon of change blindness. Psychol Russ 7(1):122

Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: Advances in neural information processing systems, pp 545–552

Healey C, Enns J (2011) Attention and visual memory in visualization and computer graphics. IEEE Trans Vis Comput. Graph 18(7):1170–1188

Hou X, Zhang L (2009) Dynamic visual attention: Searching for coding length increments. In: Advances in neural information processing systems, pp 681–688

Hou X, Harel J, Koch C (2011) Image signature: highlighting sparse salient regions. IEEE Trans Pattern Anal Mach Intell 34(1):194–201

Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194–203

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259

Jaiswal N, Ray W, Slobounov S (2010) Encoding of visual-spatial information in working memory requires more cerebral efforts than retrieval: evidence from an EEG and virtual reality study. Brain Res 1347:80–89

Karacan HU, Cagiltay K, Tekman HG (2010) Change detection in desktop virtual environments: an eye-tracking study. Comput Hum Behav 26(6):1305–1313

Keshvari S, Van den Berg R, Ma WJ (2013) No evidence for an item limit in change detection. PLoS Comput Biol 9(2):e1002927

Kouider S, De Gardelle V, Sackur J et al. (2010) How rich is consciousness? the partial awareness hypothesis. Trends Cogn Sci 14(7):301–307

Kristjánsson Á, Draschkow D (2021) Keeping it real: looking beyond capacity limits in visual cognition. Atten Percept Psychophys 83(4):1375–1390

Kümmerer M, Bylinskii Z, Judd T, et al. (2018) MIT/Tübingen Saliency Benchmark. https://saliency.tuebingen.ai/

La Corte V, Sperduti M, Abichou K et al. (2019) Episodic memory assessment and remediation in normal and pathological aging using virtual reality: a mini review. Front Psychol 10:173

Lau H, Rosenthal D (2011) Empirical support for higher-order theories of conscious awareness. Trends Cogn Sci 15(8):365–373

Li CL, Aivar MP, Tong MH et al. (2018) Memory shapes visual search strategies in large-scale environments. Sci Rep 8(1):1–11

Li X, Shan Y, Chen W et al. (2021) Predicting user visual attention in virtual reality with a deep learning model. Virtual Real 25(4):1123–1136

Lochner MJ, Trick LM (2014) Multiple-object tracking while driving: the multiple-vehicle tracking task. Atten Percep Psychophys 76(8):2326–2345

Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390(6657):279

Luck SJ, Vogel EK (2013) Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends Cogn Sci 17(8):391–400

Ma LQ, Xu K, Wong TT et al. (2013) Change blindness images. IEEE Trans Vis Comput Graph 19(11):1808–1819

Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nat Neurosci 17(3):347–356

Maiello G, Chessa M, Bex PJ et al. (2020) Near-optimal combination of disparity across a log-polar scaled visual field. PLoS Comput Biol 16(4):e1007699

Marwecki S, Wilson AD, Ofek E, et al. (2019) Mise-unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In: Proceedings of the 32nd annual ACM symposium on user interface software and technology, pp 777–789

Matzen LE, Haass MJ, Divis KM et al. (2017) Data visualization saliency model: a tool for evaluating abstract data visualizations. IEEE Trans Vis Comput Graph 24(1):563–573

Meilinger T, Knauff M, Bülthoff HH (2008) Working memory in wayfinding—a dual task experiment in a virtual city. Cogn Sci 32(4):755–770

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol Rev 63(2):81

Pedale T, Santangelo V (2015) Perceptual salience affects the contents of working memory during free-recollection of objects from natural scenes. Front Hum Neurosci 9:60

Phillips W (1974) On the distinction between sensory storage and short-term visual memory. Percep Psychophys 16(2):283–290
Polatsek P, Waldner M, Viola I et al. (2018) Exploring visual attention and saliency modeling for task-based visual analysis. Comput Graph 72:26–38
Posner MI, Nissen MJ, Ogden WC (1978) Attended and unattended processing modes: the role of set for spatial location. Modes Perceiving Process Inf 137(158):2
Pylyshyn ZW, Storm RW (1988) Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spat Vis 3(3):179–197
Read T, Sanchez CA, De Amicis R (2022) The influence of attentional engagement and spatial characteristics on time perception in virtual reality. Virtual Real 58:1–8
Rensink RA (2005) Change blindness. Neurobiology of attention. Elsevier, Amsterdam, pp 76–81
Rensink RA, O'Regan JK, Clark JJ (1997) To see or not to see: the need for attention to perceive changes in scenes. Psychol Sci 8(5):368–373
Riche N, Mancas M, Duvinage M et al. (2013) Rare 2012: a multi-scale rarity-based saliency detection with its comparative statistical analysis. Signal Process Image Commun 28(6):642–658
Robinson MM, Benjamin AS, Irwin DE (2020) Is there a k in capacity? assessing the structure of visual short-term memory. Cogn Psychol 121(101):305
Rouder JN, Morey RD, Morey CC et al. (2011) How to measure working memory capacity in the change detection paradigm. Psychon Bull Rev 18(2):324–330
Schauerte B, Stiefelhagen R (2012) Quaternion-based spectral saliency detection for eye fixation prediction. European Conference on Computer Vision. Springer, Berlin, pp 116–129
Schneegans S, Taylor R, Bays PM (2020) Stochastic sampling provides a unifying account of visual working memory limits. Proc Natl Acad Sci 117(34):20959–20968
Seinfeld S, Feuchtner T, Pinzek J et al (2020) Impact of information placement and user representations in VR on performance and embodiment. IEEE Trans Vis Comput Graph 65:1–13
Seo HJ, Milanfar P (2009) Static and space-time visual saliency detection by self-resemblance. J Vis 9(12):15–15
Simons DJ, Levin DT (1998) Failure to detect changes to people during a real-world interaction. Psychon Bull Rev 5(4):644–649
Simons DJ, Franconeri SL, Reimer RL (2000) Change blindness in the absence of a visual disruption. Perception 29(10):1143–1154
Sitzmann V, Serrano A, Pavel A et al. (2018) Saliency in VR: how do people explore virtual environments? IEEE Trans Vis Comput Graph 24(4):1633–1642
Steinicke F, Bruder G, Hinrichs K et al. (2011) Change blindness phenomena for virtual reality display systems. IEEE Trans Vis Comput Graph 17(9):1223–1233
Stirk JA, Underwood G (2007) Low-level visual saliency does not predict change detection in natural scenes. J Vis 7(10):3–3
Suma EA, Clark S, Krum D, et al. (2011) Leveraging change blindness for redirection in virtual environments. In: 2011 IEEE Virtual Reality Conference. IEEE, pp 159–166
Svenson O, Sjöberg K (1983) Speeds of subitizing and counting processes in different age groups. J Genet Psychol 142(2):203–211
Tavakoli HR, Rahtu E, Heikkilä J (2011) Fast and efficient saliency detection using sparse sampling and kernel density estimation. Scandinavian conference on image analysis. Springer, Berlin, pp 666–675
Thornton IM, Nguyen TT, Kristjánsson Á (2020) Foraging tempo: human run patterns in multiple-target search are constrained by the rate of successive responses. Quart J Exp Psychol 58:1747021820961640
Underwood G, Foulsham T (2006) Visual saliency and semantic incongruency influence eye movements when inspecting pictures. Quart J Exp Psychol 59(11):1931–1949
Varakin DA, Levin DT, Collins KM (2007) Comparison and representation failures both cause real-world change blindness. Perception 36(5):737–749
Wang W, Shen J, Xie J et al. (2019) Revisiting video saliency prediction in the deep learning era. IEEE Trans Pattern Anal Mach Intell 43(1):220–237
Wilken P, Ma WJ (2004) A detection theory account of change detection. J vis 4(12):11–11
Williams JR, Robinson MM, Schurgin MW et al. (2022) You cannot"count''how many items people remember in visual working memory: the importance of signal detection-based measures for understanding change detection performance. J Exp Psychol Hum Percep Perform 48(12):1390
Wloka C, Kunić T, Kotseruba I, et al. (2018) Smiler: Saliency model implementation library for experimental research. arXiv preprint arXiv:1812.08848
Wolfe JM (2012) Saved by a log: how do humans perform hybrid visual and memory search? Psychol Sci 23(7):698–703
Wolfe JM (2013) When is it time to move to the next raspberry bush? foraging rules in human visual search. J Vis 13(3):10–10
Zhang L, Tong MH, Marks TK et al. (2008) Sun: a bayesian framework for saliency using natural statistics. J Vis 8(7):32–32
Zhang W, Luck SJ (2008) Discrete fixed-resolution representations in visual working memory. Nature 453(7192):233–235