Virtual Reality (2021) 25:1171–1191
https://doi.org/10.1007/s10055-021-00523-4

ORIGINAL ARTICLE

# Scene Walk: a non-photorealistic viewing tool for first-person video

**Xiaomeng Wang**[1] · **Alan F. Blackwell**[1] · **Richard Jones**[2] · **Hieu T. Nguyen**[3]

Received: 6 April 2020 / Accepted: 30 March 2021 / Published online: 22 April 2021


**Abstract**
Scene Walk is a video viewing technique suited to first-person video recorded from wearable cameras. It integrates a 2D video player and visualisation of the camera trajectory into a non-photorealistic partial rendering of the 3D environment as reconstructed from image content. Applications include forensic analysis of first-person video archives, for example as recorded by emergency response teams. The Scene Walk method is designed to support the viewer's construction and application of a cognitive map of the context in which first-person video was captured. We use methods from wayfinding research to assess the effectiveness of this non-photorealistic approach in comparison to actual physical experience of the scene. We find that Scene Walk does allow viewers to create a more accurate and effective cognitive map of first-person video than is achieved using a conventional video browsing interface and that this model is comparable to actually walking through the original environment.

**Keywords** First-person video · Body-worn camera · Video viewing · 3D scene reconstruction · Camera trajectory · Cognitive Map

## 1 Introduction

As camera technology becomes more compact, with battery life and storage capacity adequate for extensive video recording, wearable and body-worn video cameras are becoming increasingly popular. By comparison to static video cameras, video recording from wearable cameras introduces new and diverse applications. These include social and recreational uses (Chen and Jones 2010; Ishiguro and Rekimoto 2012; Higuch et al. 2016), law enforcement (Jennings et al. 2014; Smykla et al. 2016) and healthcare (De et al. 2015), as well as novel utility and creative applications that are constantly emerging. However, increasing use of wearable cameras generates large archives of first-person videos, much of which is never edited, annotated or indexed, leading to difficulty in viewing and searching such videos. Manual review of video archives is extremely time-consuming and labour-intensive. Before watching a given archive video, the viewer often has no idea when and where the content of interest will appear. As a result, it may be necessary to watch the whole length of a video (possibly at higher speed, if the video does not include complex scenes or rapid movement) in order to find a single activity or object of interest. Furthermore, content of interest is often distributed across multiple videos, especially if multiple people wearing cameras were present during the same events. In such circumstances, analysts often need to view the same video multiple times in order to understand the relationships between different points of view and object perspectives.

We consider this problem from the perspective of user experience of imagery within a virtual 3D environment, suggesting that the key problems are (a) information about camera pose at the time of capture being lost through the

The original online version of this article was revised:The Acknowledgements section has been included.

✉ Xiaomeng Wang
xw337@cam.ac.uk

Alan F. Blackwell
afb21@cam.ac.uk

Richard Jones
richard.jones16@boeing.com

Hieu T. Nguyen
hieu.t.nguyen9@boeing.com

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, United Kingdom

[2] Boeing Defence UK, Bristol BS16 1EJ, United Kingdom

[3] Boeing Research and Technology, Huntsville, AL, USA

camera projection that renders light from the 3D scene only as a series of 2D image frames and (b) successive image frames from wearable cameras having complex variations as a result of unpredictable "shaky" body movements rather than intentional pointing of the camera. We suggest that, since the combination of spatial information loss and low video quality makes first-person video especially hard to understand (del Molino et al. 2017), novel viewing tools can be used to compensate for these problems and support analytic tasks.

We apply a human factors perspective, based on the neuroscience of human spatial reasoning. We propose a novel approach in which navigation through an archived video that was originally recorded in a 3D environment can be understood by analogy to previous research in the field of "wayfinding". According to this analogy, if a previously unseen video has been recorded in a space that the viewer is not familiar with, then the viewer must implicitly construct a "cognitive map" of that space in order to understand the video. We suggest that the cognitive tasks involved in exploring or returning to a specific location in a video will draw on many of the same cognitive resources that are involved in exploring or returning to a physical place that one has visited before, and that this experience can be analysed and evaluated using methods drawn from wayfinding research. As noted by Dalton et al. (2019), "the study of wayfinding is central to research on human spatial cognition". We provide more detailed explanations of the terms "wayfinding" and "cognitive map", by reference to neuroscience and human factors literature, in section "Cognitive Maps in Virtual Environments".

Based on this cognitive perspective of video navigation as wayfinding within a 3D virtual environment, we propose a novel alternative approach to the problem of navigating first-person video, demonstrated in an application called Scene Walk. Our approach is to help users review first-person videos by reconstructing a partial spatial model of the context in which the video was recorded, and presenting this through a non-photorealistic virtual scene visualisation. Our hypothesis is that visualisation of a 3D scene model can be combined with a custom video player to help users generate a cognitive map of the contextual and spatial information that would otherwise be lost (in standard approaches such as video summarisation and fast-forwarding) as a result of the 3D to 2D projection inherent in camera recording, exacerbated by frame selection or sampling.

In summary, first-person videos are especially demanding of the viewer's cognitive map (because of changing camera pose and potential presence of multiple cameras), while also making it more difficult to acquire a cognitive map in the first place (because of camera shake and lack of editorial guidance). We expect that 3D scene-based approaches such as Scene Walk will have particular relevance to future proliferation of wearable cameras, and the growing archives of first-person video that will result. Later, in this paper we illustrate one potential application scenario—emergency response where video is recorded when walking into an unseen building. However, wearable cameras and egocentric videos are also used in a wide variety of other settings where users need to review video in relation to their movements during recording, including life-logging, memory support in dementia, or sports and recreation.

The technical approach of Scene Walk is to use recorded data to reconstruct stable elements of the 3D scene in which recordings were made by wearable cameras, and combine that stable 3D scene with a visualisation of the actual trajectory that was followed by the camera-wearer. In addition, a 2D video player window is inserted into the 3D scene, showing the sequence of 2D image frames that are associated with the trajectory, from the perspective of the camera position and pose at the time each frame was captured. Users can either view the stable 3D scene as a whole and follow the trajectory of the camera-wearer, or watch the 2D video projection from the reconstructed virtual perspective of the camera-wearer as they walked through the scene. The user can switch between these two modes, integrating their cognitive map of an egocentric first-person view with their location and context within a stable 3D environment.

In summary, we make the following contributions:

- A novel design approach in which a moving 2D playback screen is inserted into a non-photorealistic rendering of the 3D context, thus giving users a direct and intuitive understanding of the original camera projection in a way that is integrated with contextual virtual environment cues to form a cognitive map.
- A novel video viewer, Scene Walk, that uses this design approach in combination with a visualised trajectory of the camera within the virtual scene, allowing the user to either gain an overview of the camera path in a stable virtual context, or investigate the scene as if they were the camera-wearer.
- An experimental method derived from wayfinding research, that can be used to evaluate a viewer's cognitive map of video recorded with a moving camera. We present evidence from a user study, demonstrating that the Scene Walk system does allow viewers to create a more accurate and effective cognitive map of first-person video than is achieved using a conventional video browsing interface.

## 2 Related work

Previous researchers have explored strategies that could help analysts, editors and other users to more effectively and efficiently review or search for relevant content in video from wearable cameras (Betancourt et al. 2015; Bolanos et al. 2017). One

substantial focus of research is to automatically identify, extract and compile the most relevant clips from the archive through video summarisation (Lee et al. 2012; Lin et al. 2015; Ho et al. 2018). Video summarisation systems allow the user to view the video quickly, but with the danger that information may be lost if the summarisation relevance metric does not correctly anticipate the interest of the actual user. Video summarisation can also discard many familiar continuity cues that would help viewers to form a cognitive map of the 3D scene from changing camera pose in cinematography, such as "pan" (horizontal rotation of the camera pose to help the viewer understand the spatial relationship between two viewpoints) and "dolly" (linear movement of the camera through the scene).

An alternative strategy, as already mentioned, is to speed up the video replay ("fast-forwarding") by sampling a subset of image frames (Poleg et al. 2015; Silva et al. 2018). Fast-forwarding techniques may be easier to interpret because they retain the visible camera movements that are familiar film conventions. Nevertheless, treating all image frames as the same importance, without considering the relevance to the viewer, means that viewers must attend to all image content when the camera is moving. Furthermore, the design and implementation of frame sampling algorithms is not straightforward—they can easily exacerbate problems of camera shake, making close attention to a rapidly changing scene even more cognitively demanding than viewing a well designed summary (indeed, the role of a professional film editor is precisely to assist interpretation through selection of relevant content and transition scenes for continuity—although this manual process is in itself time-consuming, as well as introducing significant elements of editorial interpretation).

The primary goal of research in video summarisation is to automatically select key frames or video segments that can adequately represent the content of the original video when seen by a human viewer. This depends on constructing a relevance model for frames (and segments), in which the relevance reflects correspondence between the content of each frame, and the (anticipated) user's understanding of the scene content. For application to first-person video, researchers have investigated various elements within the image frame to evaluate the potential of those elements to improve user experience. For example, based on the social context of first-person video recording, the people and objects that the camera-wearer interacts with can be used as a cue to the relative importance of the frames (Lee et al. 2012). The scene context can also be estimated, for example using a classification approach, inferring a level of importance for that category of content in relation to expected user interest (Lin et al. 2015). Observed eye-gaze information can be used more directly to identify points of interest within a frame and thus to predict relative importance of frames in relation to the user's understanding of that content (Xu et al. 2015). Where such measures of the user's attention to the image are available (or can be reliably inferred from other evidence), deep learning methods can be used to train the necessary classifiers. For example, Yao et al. (2016) propose video summarisation using a highlight prediction score that is generated by a pairwise deep ranking method. Ho et al. (2018) proposed a deep learning framework which learned the distinctive spatiotemporal context information across multiple videos. However, in all of these methods, the main disadvantage of video summarisation is that video content meaningful to the user might be incorrectly omitted, either because the relevance model fails to generalise to new content, or because a new user's attention may be focused on specific types of content that have not been captured in prior training. In creative or research tasks, where novel interpretation is a key requirement, such failures are almost guaranteed.

The other technical approach that has most often been applied to browsing first-person videos is the fast-forwarding method, with specific enhancements to preserve scene interpretation by the user. Kopf et al. (2014) described the conversion of first-person videos to "hyperlapse" videos, which reduce frequent camera shake by constructing a smoother camera path for the output video, cropping and stitching frames to be consistent with this virtual camera. Poleg et al. (2015) proposed a frame sampling technique which selects frames having a similar direction of movement, in order to achieve a smooth fast forward video. Other fast forwarding technologies incorporate semantic information into the frame sampling algorithm, in order to ensure that the playback is consistent with the viewer's understanding of the scene. Higuchi et al. (2017) developed an interface providing adaptive playback speeds based on egocentric cues for video browsing. Silva et al. (2018) combined a weighted frame sampling strategy with a transition smoothing method to generate a smoother and more consistent video. All these methods allow users to watch the resulting first-person videos more quickly and comfortably. However, the frame selection unavoidably leads to information loss. As with video summarisation, an algorithm that correctly recognises and anticipates the user's attention and interest should result in fast-forwarding that is easier to watch so long as it is consistent with that model. But if the user has individual interests that are not anticipated, fast-forwarding can easily make video content more difficult to watch and assimilate.

Previous research has considered the problem of how first-person video recording can be related to the 3D context in which the recording is made. For example, Sugita et al. (2018) describe an approach in which a previously identified workspace is scanned in advance, in order to construct a 3D model of the environment. Content from a moving camera is then indexed against this model, so that the model orientation can be changed to correspond to the current camera

view. Use of an explicit scanning phase means that Sugita et al. (2018) are able to construct a complete visual model of the known space, unlike our own research in which we use a non-photorealistic approach to render the necessarily incomplete model that results when the environment has not previously been scanned. Kono et al. (2017) use image features from a BWC to infer the current position of the camera-wearer on a 2D map of the environment. This approach uses videogame conventions to relate local activity to a larger area in real time, but does not currently support retrospective analysis of archived video, or user understanding of a 3D scene structure.

Other research has explored the use of 3D models to browse collections of still photographs (Snavely et al. 2006; Ballan et al. 2010; Arev et al. 2014; Nuernberger et al. 2018). In photograph browsing applications, the main goal is to help users understand the angle from which a photograph has been captured. As with Scene Walk, the 3D context is rendered in a non-photorealistic manner, in order to emphasise the spatial geometry by contrast to the actual images. However, these photograph browsing interfaces do not allow the user to directly navigate the 3D model in the non-photorealistic virtual style employed in Scene Walk.

## 3 Cognitive maps in virtual environments

Cognitive map is the term used to describe the complex set of mental representations that are used by humans to reason about their surroundings—to integrate observations and experiences, construct and interpret relationships between places, and plan actions. As a functional definition, the term cognitive map does not correspond to an isolated brain mechanism, but to the joint operation of these diverse capabilities related to spatial reasoning and environmental modelling (Kitchin 1994). The multidisciplinary areas of research that contribute to the study of cognitive maps include, among many others, studies of reasoning about space, for example, as in the work of Tversky (1993) and Kaplan (1973), computational models of spatial reasoning (O'Neill 1991), studies of urban environments as in the work of Lynch (Lynch 1960), studies of cartographic representations as in the work of MacEachren (1992) and studies of developmental deficits and educational strategies as in the work of Golledge et al. (1985).

Support for cognitive maps is a central consideration in the design, evaluation and optimisation of user interfaces that represent spatial and environmental data, including geographical information systems (GIS), computer-aided design (CAD) systems for architectural and urban design, navigation systems ranging from satellite route planning to VR gaming applications, as well as applications for data review and analysis in 3D environments of the kind that we present in the current paper. Because cognitive maps integrate knowledge across multiple levels of detail, experimental tasks explore task performance at different granularities, ranging from assessment of local accuracy in recall of a map (e.g. McNamara 1986) to integrative studies of large-scale understanding of the environment (e.g. Herman and Siegel 1978). A large body of empirical research confirms that cognitive maps do not consist of a single representation, but combine different levels of sensory and reasoning capabilities, influenced by a range of social, educational and developmental factors.

The diverse neuropsychological resources that are studied as aspects of cognitive mapping are critical to many areas of human performance, including reasoning about spatial facts, solving problems in relation to spatial constraints and planning navigation paths. This last topic has been a longstanding concern for applied psychology research under the rubric of wayfinding—the human competence that involves finding a path from one's current location to a desired destination, by reasoning about space (Golledge 1999). Support for wayfinding is a critical affordance of the built environment and is an important priority in architectural practice and information design for public spaces (Arthur and Passini 1992; Gibson 2009).

Early research into human performance in virtual environments was concerned with understanding navigation and route-finding errors in the virtual environment, and with understanding how users acquire cognitive maps of a virtual environment in order to improve wayfinding. One important application of such research was to understand whether a virtual environment could be used to train people for later wayfinding in the real building that had been modelled (e.g. Witmer et al. 1996), a functional area of performance that requires transfer of knowledge from the virtual to the real environment by constructing and retaining a cognitive map. In order to compare cognitive maps in relation to virtual and real environments, past studies have therefore carried out controlled experiments in which participants navigate through either a real building or a virtual simulation of that building, and then carry out reasoning or memory tasks in order to compare the accuracy and completeness of the cognitive maps in the two cases, for example, as studied by Ruddle and Payne (Ruddle et al. 1997) in a "desk-top" (non-immersive) virtual environment.

Controlled studies carried out with these earlier generations of technology confirmed that learning of cognitive maps in non-immersive virtual environments was impaired by technical factors such as poor visual fidelity and reduced peripheral vision (aspects that soon improved, and continue to be improved in studies of more recent generations of immersive VR, e.g. Ruddle et al. 1999). Learning of cognitive maps in the physical world is also dependent on other

sensory inputs that have not yet been easily simulated in VR, such as vestibular, haptic, locomotor and proprioceptive feedback (Lackner and DiZio 2005), and these types of sensory input are shown to improve performance in immersive VR when approximated in a walking interface while using six-DoF tracker with a head-mounted display (Ruddle and Lessels 2009).

A range of experimental tasks have been developed in order to compare the accuracy and completeness of cognitive maps acquired in real and virtual environments, some adapted from earlier research that compared learning from paper maps to actual navigation (e.g. Thorndyke and Hayes-Roth 1982). We have applied related experimental measures for the research reported in this paper, although our research focuses on a more specific use case rather than generic learning. In typical learning studies, participants repeatedly navigate the real and virtual environments, often for several hours, in order to recall the full structure of the environment (Wilson et al. 1997; Ruddle et al. 1997). As noted by those researchers, there are design opportunities to replace the need for comprehensive learning with improved guidance, interaction and navigation aids (for example, as explored by Burigat and Chittaro (2007) for outdoor environments), and our own Scene Walk interaction method offers an example of such a novel approach. Other recent research has taken a different emphasis, using immersive virtual reality to simulate and evaluate improved design for wayfinding. However, it is still unclear whether immersive VR will be effective for that purpose (Kuliga et al. 2020), or even that common navigational tools such as street view visualisations offer measurable benefits over conventional maps (Qiu et al. 2020).

In summary, research into cognitive maps has been carried out using a wide variety of technologies and representational techniques, some more realistic than current commercially available virtual reality products, and many less realistic. At the most realistic extreme, studies involve actual real buildings, studying the formation and application of cognitive maps as people move around in the real physical world. The least realistic representational extremes have included studies of cognitive maps in relation to digital sketch maps and traditional 2D paper maps or building plans. Many of the earlier studies that we build on in the current project have involved "desktop virtual environments" (as they are defined in the literature Ruddle et al. 1997) which are non-immersive 3D perspective renderings, that are navigable, and may be more or less photorealistic, depending on the maturity of real-time graphics hardware at the time of that particular study, and on the degree of detail and fidelity in the scene model and rendering algorithms.

Our project extends this previous literature with two new variants within this range of virtual and representational realism. The first is to study acquisition of cognitive maps when the scene is represented via first-person video recordings of an actual environment. 2D video recordings are photorealistic, non-immersive and non-navigable (other than playing backward or forward in the video timeline). The closest to this approach in previous research is the study by (Witmer et al. 1996) which involved presenting the route through a building with a series of photographs accompanied by a verbal description of the route being followed.

Our second new contribution within the range of realism options is the Scene Walk method, which is non-photorealistic and non-immersive, but is fully navigable. Our evaluation of non-photorealistic interactive scene rendering can be compared to the work of Ruddle and Lessels (2009), which compared a detailed VR model to an "impoverished" model with far less visual detail, and concluded: "Participants" performance was largely unaffected by the amount of detail in the visual scene, indicating that full body-based information is necessary for efficient navigation, but a rich visual scene is not".

In the evaluation study reported below, which uses the same experimental techniques as many of the studies in the previous literature reviewed here, we compare these two new representational alternatives (first-person video and Scene Walk) to the benchmark of navigation in the actual physical world. The physical world is photorealistic, immersive and navigable, in addition to offering other aspects of realistic sensory experience that are not yet available in commercial VR products including haptic, locomotor feedback. Comparison of the physical world to these two alternative methods for virtual scene representation allows us to study the extent to which photorealism and navigability are important factors in the formation of cognitive maps, as design parameters that can be varied in future virtual reality applications.

## 4 Scene Walk

The goal of the *Scene Walk* project is to enhance a user's ability to form a cognitive map of the 3D environment within which first-person video has been recorded. We use non-photorealistic rendering to help viewers understand how the images they are seeing relate to the 3D spatial context in which the video was recorded, and to the route that was followed by the person wearing the video camera. Such capabilities are likely to be useful in tools for indexing or editing recreational first-person video, or for situations where analysts must review archives of video that has been captured from wearable cameras, for example, during forensic analysis of recordings made by body-worn cameras (BWCs) on the uniforms of police or emergency service personnel during crime-scene investigation or emergency response.

In forensics or emergency response situations, as in many kinds of sport and recreational activity, wearers of BWC devices are not primarily focused on making video

recordings—they are engaged in other tasks, and capture of video is incidental to those tasks. As a result, the captured video has little systematic survey or narrative structure. As the camera-wearer moves and turns their body, some points in the environment are likely to be recorded multiple times, while other points may never be captured at all. Information about the environment is thus necessarily incomplete, and watching such video can be confusing for the viewer. Other quality challenges inherent in footage recorded from wearable cameras, such as camera shake, are also more severe in situations where the person wearing the camera is engaged in other tasks, and not primarily attending to the quality of the video recording.

As an example of a scenario where first-person video of this kind might be recorded, emergency response personnel must often enter buildings that they have not previously visited, and for which they do not have access to maps or plans of internal layout. Prior information is limited to the location of the building itself, and its external dimensions. The path taken by emergency personnel through the building is therefore exploratory, with some sections of their path motivated by route-finding, and some directed towards specific objects of interest (for example, an appliance that was the origin of a fire, casualties, safety hazards, weapons or other illegal materials). Points of interest are likely to be far more strongly represented in the recorded video, because personnel walk back and forth in areas that are most relevant to their task. Video recorded in those localities will therefore include considerable redundancy and will capture pictures of the same objects from many different angles. Relatively featureless parts of the environment (such as passages with no doors) are likely to be only partially recorded in this kind of scenario, because personnel move more quickly through contexts that have no features of interest and do not require exploratory path-finding.

In scenarios of this type, the person who later views the recorded video is likely to be a different person from the one who was wearing the camera during recording. Typical examples might be a reporter, a forensic analyst, or members of a legal team. The viewer is likely to watch extended recordings that were made while camera-wearers were walking back and forth in the same area, often including multiple separate videos recorded during the same period (by several members of an emergency response team, each wearing their own BWC). Finding all the information potentially relevant to particular locations, or particular objects of interest, will involve reviewing the full length of every individual video, probably repeatedly, in order to recognise and annotate points at which the camera-wearer returns to the same locations. As with wayfinding tasks, successful recognition and labelling depends on the viewer forming a (probably incomplete) cognitive map of the internal layout of the building, in order to decide which 2D images correspond to which 3D locations. The *Scene Walk* system aims to provide users with the necessary tools to form a cognitive map of the internal layout of a building, and relate this to the 2D image frames that were captured from wearable cameras, in order to perform analytic tasks with the resulting video archives.

### 4.1 Implementation

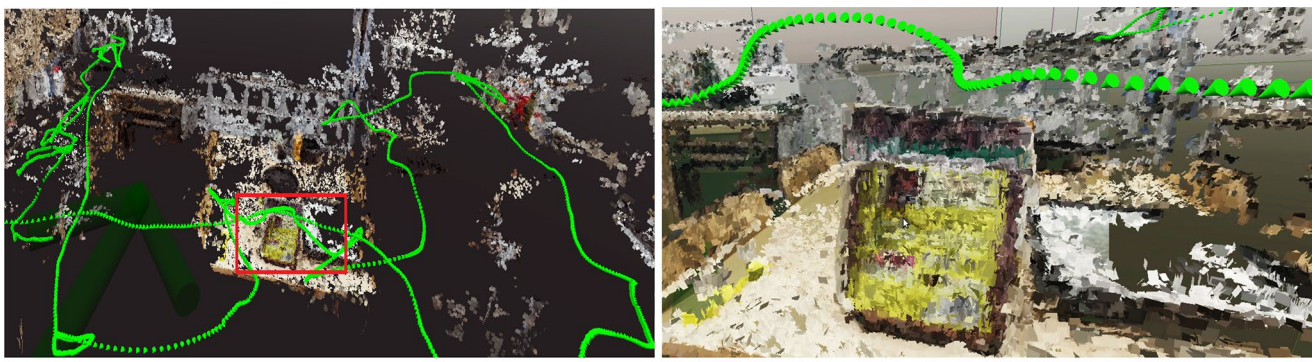
In our motivating scenario of emergency response personnel exploring an unknown building while wearing BWCs, a typical exploration path in one room of this building might be as shown in Fig. 1, with each camera pose rendered as a green cone. The tip of the cone is the direction that the camera was facing. As discussed, multiple images will often be captured that capture the same regions within the 3D frame of reference, but have been recorded by the BWC from different view angles, as the camera-wearer turns around, retraces steps or walks back and forth. These multiple viewpoints, while occurring incidentally, are a resource that can be used to reconstruct the 3D scene. In order to construct a 3D model, we therefore use a combination of camera sensor data and feature-based post-processing of the image frames to identify and match corresponding image patches that have been captured from different viewpoints.

The wearable camera market is currently very dynamic, with rapid development of device specification and feature sets. We make the assumption that future devices in this class will include at least a minimal level of low-cost and low-power sensors, including motion sensors, such as gyroscope and accelerometer, as in BWC models such as Panasonic Arbitrator-BWC and Philips DVT3120 VideoTracer BWC. We note that geographic location sensors such as compass and GPS, even if sufficiently accurate to be used for image localisation at the scale of our motivating scenario, would not be reliable or effective in indoor settings. We also note that other motion sensors (e.g. barometer) may provide additional motion information of BWC, such as vertical movement. However, following the design of Scene Walk relying on a minimal level of sensors, we only assume the necessary motion sensors, gyroscope and accelerometer are available in order to provide motion information with six full degrees of freedom (DoF). For our study, we used a programmable Android platform (Huawei P20) to implement an image capture protocol including these sensor annotations. The camera sensor is with field of view of 66°. The frame rate is 30 frames/second, and its image resolution is $1280 \times 720$. The motion sensors are microelectromechanical systems (MEMS) gyroscope and accelerometer, which are widely available, especially on mobile devices, for monitoring the device movement. We developed a simple data collection application using the ARCore library, which captures the image frame sequence, timestamp at which each frame was captured, and the six-DoF camera pose at that time.

In post-processing, we applied the patch-based multiple view stereo (PMVS) method (Furukawa and Ponce 2010) to reconstruct the 3D scene using the recorded image frames and the camera poses. Given a sequence of image frames, the features of each image are detected using Harris and difference of Gaussians operators. For each feature in an image, matching features that satisfy an epipolar consistency constraint are found in other images. Reconstructed scene patches are then initialised by triangulating from the matching feature pairs. Only reconstructed scene patches which are visible in at least three image frames are regarded as a successful reconstruction. This initial set of reconstructed scene patches is usually sparse, with the patch centres and normal vectors optimised by minimising a photometric discrepancy function. The whole reconstruction process of the scene patches includes three steps: matching, expansion and filtering. In the expansion step, the PMVS method generates new scene patches utilising the parameters of their neighbouring existing patches. The neighbouring relationship is determined by the position of the image projection of the patches. After the expansion, a filtering step is applied to delete outliers from the reconstructed patches by considering consistency of visibility (for example, removing patches estimated to be outside the spatial bounds of the building, or outside the viewing angle of the camera). Further iterations of expansion and filtering can be performed to add more reconstructed patches, which are rendered within the 3D scene as a cloud of coloured image patch fragments, as shown in Fig. 1. The reconstructed scene more focuses on the static structure and elements of an environment captured by BWCs. Note that presence of specular highlights and obstacles (e.g. people walking through the scene) will be discarded in patch optimisation of PMVS method and will not appear in the reconstructed scene model. Note also that since the PMVS method is feature-based, it may not work well when reconstructing featureless scenes such as blank walls (although such scenes may include very little contextual information of any kind).

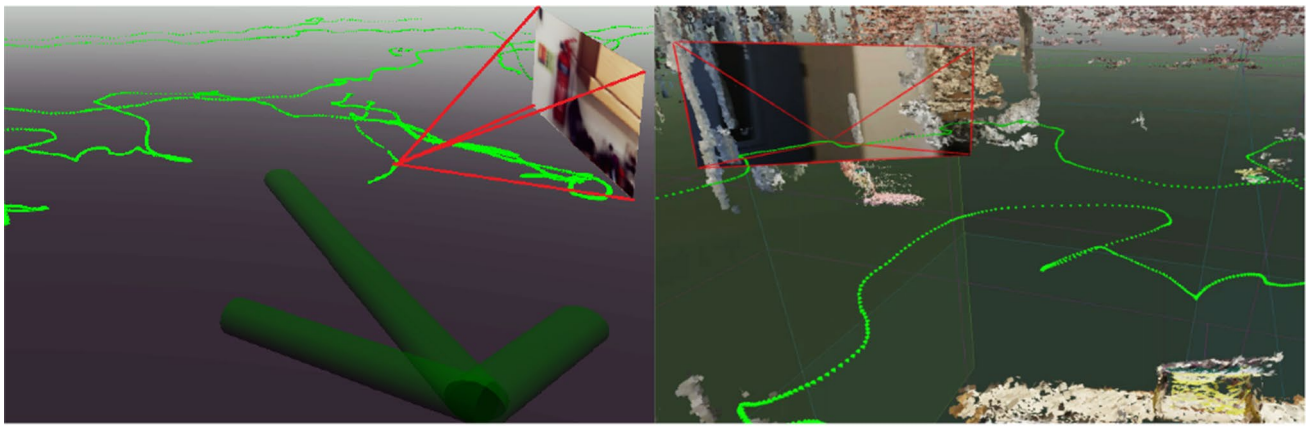
The 3D rendering and view control facilities of the Scene Walk prototype are provided by the Godot open source game engine. The elements of the Godot scene are the cloud of coloured image patch fragments, and the trajectory of camera pose vectors, as shown in Fig. 1. We note that this rendering is not intended to be a photorealistic scene reconstruction (for example, as might conventionally be constructed using 3D scanning, structure from motion and texture mapping), but a non-photorealistic partial reconstruction that is designed to provide sufficient information about the 3D context so that viewers can form a cognitive map supporting their interpretation of the video.

Each image frame from the recorded sequence is indexed by timestamp to correspond with one camera pose location. At any time, a single image frame is rendered within the 3D scene as a 2D video player, projected onto a virtual screen from the perspective of the corresponding camera pose, as shown in Fig. 2. The position, shape and orientation of the virtual screen are determined by the camera projection, while the distance from the camera position to the virtual screen (and thus size of the projected image) can be adjusted by the user with keyboard controls. When the video plays, the projection point of the virtual screen image advances to each successive location along the motion trajectory, giving the impression of a video screen that is moving forward along the trajectory while the video plays. This moving screen can naturally be interpreted by the user as the first-person view that was being seen by the camera-wearer walking along



**(a)** **(b)**

**Fig. 1** The non-photorealistic 3D scene reconstruction, with trajectory of camera poses within the scene. The 3D scene is rendered as a cloud of coloured image patch fragments, and the sequence of six-DoF camera poses as a trail of green cones, each pointing in the direction the camera was facing at that location. View (**a**) corresponds to a portion of the scene in which the camera-wearer was walking around a single room (a kitchen/common room in a student residence), looking towards objects at the sides of the room, and on a table in the centre of the room. View (**b**) is a close-up of the area bounded by the red rectangle in view (**a**), showing the level of detail maintained by the model—here, an object that is resting on the table



(a) (b)

**Fig. 2** The rendered camera trajectory (a sequence of green cones) and the 2D video player screen as projected from one of the camera positions. The current projection point and player screen can move along the trajectory as the video plays. View (**a**) shows only the camera trajectory and projected screen, without the rendered 3D scene, and corresponds to a video recorded while walking along several corridors of an office building, turning to look into rooms and at distinctive objects such as the fire extinguisher that is currently visible on the projection screen. View (**b**) includes the image patches of the rendered 3D scene, and corresponds to a portion of the video recorded as the camera-wearer had walked around a single room, and is about to walk out through the doorway that is currently visible on the projection screen. In this figure, the red lines (not part of the actual system display) have been added to explain the projection geometry from the current green cone to the current screen position. Note that the distance between the projection point and the projected screen can be adjusted by the user via keyboard controls, enlarging the video projection at the expense of obscuring scene context, to optimise the balance of projection and context according to the required level of detail versus context in different parts of the video. The large dark green arrow visible beneath the floor in view (**a**) provides a global north reference orientation within the coordinate system of Scene Walk. The typical video scenes chosen for this figure emphasise the relative lack of informative detail that is present in such recordings, and the difficulty that is faced by viewers in trying to establish a cognitive map from images of relatively featureless walls interspersed by images of recognisable or distinctive objects that the camera-wearer may have stopped to look at (Color figure online)

that route when the recording was being made. The user can either press a "play" key on the keyboard to let this virtual screen move automatically along the trajectory in time with video playback, or else manually step forward and backward to review the content of specific image frames. The Scene Walk with all rendered elements (the cloud of image patch fragments, the 2D video player and the camera trajectory) is shown in Fig. 3a. A stitched photograph of the real scene (the kitchen and dining area of a common room) is shown in Fig. 3b to give readers a general idea of what was shown in the Scene Walk.

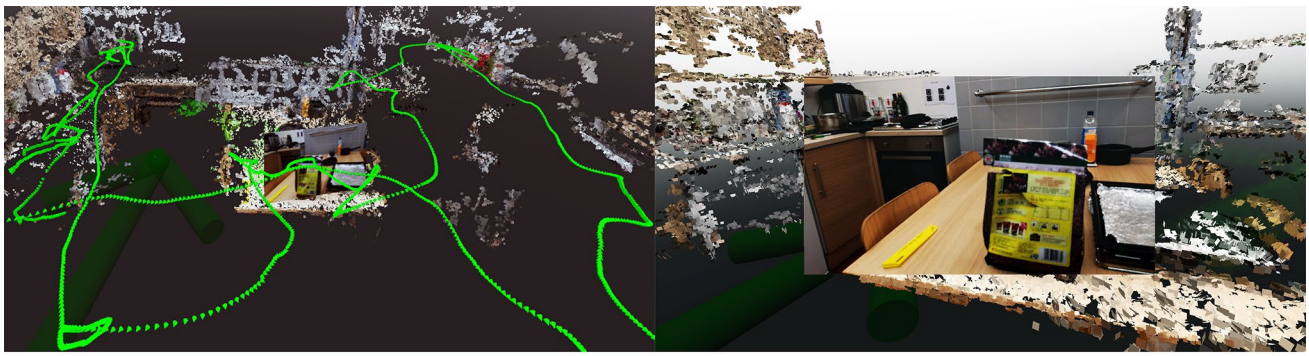
Scene Walk offers two modes in which the moving video playback screen can be followed. In one, the user's viewpoint stays static relative to the 3D building coordinates and reconstructed point cloud. From this "God-view" perspective, the virtual projection screen moves around in the building as they watch, as shown in Fig. 4a. The God-view mode allows users to understand the moving viewing position in relation to the overall scene and trajectory. Alternatively, the user can choose to follow the point of view of the person wearing the camera. In this "first-person view" mode, the virtual projection screen stays static at the centre of the display, and the building coordinates and point cloud move around to show the 3D location and context around the screen, as shown in Fig. 4b. The first-person view mode allows more detailed inspection of the image frames captured by the camera and also allows the viewer to understand which aspects of the overall scene were visible to the person wearing the camera.

All user controls, as shown in Table 1 (including stepping back and forth along the video timeline, and switching between view modes), use standard keyboard and mouse game controls, as implemented in the Godot engine. Users are not restricted to simply following the path of the projected video, but can use game controls to view the 3D model from any angle.

A useful further affordance of the Scene Walk interaction paradigm is that semantic labels can be applied to objects and scene elements when the video is paused in first-person view, by clicking on the plane of the virtual projection screen. In order to label an object, the user draws a bounding polygon in this plane, defining the area within the video frame that corresponds to the object of interest, as shown in Fig. 5. Projection from the current camera viewpoint allows this appearance model to be associated with the part of the 3D scene model that is projected onto this polygon, and to derive features of its appearance that can be used to train a classifier, or for retrieval of similar content in related videos. Note that evaluation of this labelling facility as an analysis

**(a)** **(b)**

**Fig. 3** **a** The complete rendered elements of the Scene Walk. The 2D video player was brought slightly backward to show its matching with the rendered cloud of image patch fragments. **b** This stitched image shows the real scene (the kitchen and dining area in a common room) corresponding to that shown in **a**



**(a)** **(b)**

**Fig. 4** Two views of the same kitchen/common-room scene that is shown in Fig. 1. View (**a**) shows the God-view mode, in which the geometry of the camera path can be seen relative to the point cloud rendering of the walls and benches at the sides of the room and the table in the centre. The projection screen moves around within this 3D model, and can be interpreted as an image of what the camera-wearer was looking at (within the field-of-view constraint of the camera) at a given point in time. View (**b**) shows the same point in time, but as seen in first-person view mode. Here, the projection screen is always centred in the display, so that the viewer replays the first-person visual experience (within the field-of-view constraint of the camera) as seen by the camera-wearer when moving around while the video was recorded. In first-person view mode, the point cloud rendering at the sides of the video screen (extending the field of view of the camera) moves around to give the impression that the viewer is moving through a 3D environment while a video screen "floats" directly in front of them. As in Fig. 2, the global coordinate reference is indicated by the large dark green arrow beneath the floor in View **a** (Color figure online)

**Table 1** The user controls in Scene Walk

| Description | Keyboard control |
|---|---|
| Global navigation using WSAD and mouse | *W-forward *S-back *A-left *D-right *Mouse motion—turning left/right/up/down |
| Play/pause the video with "play/pause" button | P key |
| Enter or exit the "first-person view" by pressing "Switch View" button | Space key |
| Jump 0.1 seconds along the trajectory by pressing "forward"/"backward" button | Period/comma key |
| Jump 10 seconds along the trajectory by pressing "Shift"+"forward"/"backward" button | Shift + period/comma key |

and labelling tool is not a focus of the current paper and will be the subject of future work.

## 5 Wayfinding evaluation study

As explained in "Introduction", we characterise the cognitive benefits of the Scene Walk interaction paradigm in terms of helping the user to construct a cognitive map. On that basis, as explained in "Introduction", we propose that the cognitive task of finding locations within a first-person moving video that has been recorded in a 3D spatial context can be investigated by analogy to wayfinding within that 3D context.

Many previous experimental investigations of cognitive maps have involved wayfinding tasks in which the spatial context is presented in several different modalities, in order to compare the relative advantages of those modalities for constructing cognitive maps. In previous research as reviewed earlier in this paper, alternative modalities have included paper maps (Thorndyke and Hayes-Roth 1982), sketch maps (Tversky 1993), photographs with verbal description of a route (Witmer et al. 1996), desktop "virtual environments" (non-immersive 3D scene renderings) (Ruddle et al. 1997), immersive VR headsets (Ruddle et al. 1999), Google Street View renderings (Qiu et al. 2020), immersive VR with a walking interface (Ruddle and Lessels 2009) and physical experience in the real world (Kuliga et al. 2020). We introduce first-person video recordings as a further modality from which viewers must also acquire an implicit cognitive map, if they are going to be able to understand the 3D space within which the recording was made, and be able to find locations within that space by "wayfinding" to a point in the video.

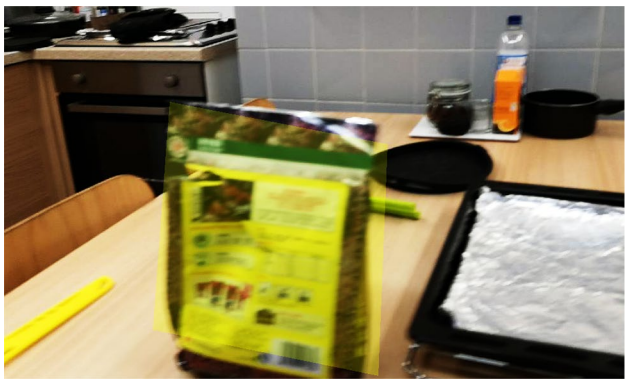


**Fig. 5** An object of interest can be labelled by the user drawing a polygon onto the projection screen. This frame from the video shows a close-up of the same kitchen/common-room scene already shown in Figs. 1 and 4, at a point in the video where the camera-wearer was standing in front of a kitchen table (kitchen chairs, a sauce bottle, and a cooker can be seen in the background). Note that the wider context of the room, previously seen in the point-cloud rendering, is not so easy to recognise in this close-up of the video alone. The yellow highlight shows where the user has drawn a polygon around one of the objects on the table, to label this part of the image as representing an object of interest

In our evaluation, we designed an experiment that compares cognitive maps acquired in three different modalities. The control condition provides the most immersive sensory experience, using the modality of walking through a real-world building. Our expectation is that walking in the real world provides maximal sensory cues that could be achieved in immersive virtual reality, including full haptic, locomotor and proprioceptive experience of the environment. The second condition is a moving first-person video recording, which is a photorealistic projection of the same building, acquired by making an actual photographic video recording while a person walked through the building, but with free movement of the camera-wearer as would occur in the emergency response scenario described above. Our expectation is that this modality would be more challenging than most previous cognitive map and wayfinding experiments. The third condition is the Scene Walk prototype, used to navigate the same first-person video as in the second condition, but rendered non-photorealistically as a scene model that has been extracted purely from geometric information that was implicit in the video. Our expectation is that the 3D scene context and navigation tools will assist the cognitive processes involved in forming a cognitive map, and in navigating to a specific location in the video.

In comparing these conditions, we expect cognitive map formation and wayfinding in the Scene Walk condition to be superior to the video condition, and inferior to the fully immersive real-world condition. Based on previous findings by (Ruddle and Lessels 2009), that an "impoverished" VR model was just as effective as a high-resolution one, we anticipate that the non-photorealistic rendering of Scene Walk may be as effective as more complete and detailed VR models that have been studied in the past. Our design goal is to understand to what extent this non-photorealistic scene rendering is able to provide sufficient support for cognitive map formation, such that user performance might approach the fully immersive case of real-world experience. The overall structure of the experiment is similar to other studies that have used three presentation conditions such as (for example) that of (Richardson et al. 1999) or (Witmer et al. 1996). However, where these more conventional studies of wayfinding have typically compared a virtual environment to a 2D map and to real-world experience, we compare real-world experience to our Scene Walk viewer rather than a conventional VR environment and to a video recording rather than a map.

### 5.1 Experimental design

Our goal in this project was to create a video viewer that would assist users to formulate a cognitive map of the 3D environment in which first-person video had been recorded. Our two primary hypotheses were, firstly, that the Scene Walk prototype would help users to develop a cognitive map that was comparable to the experience of actually walking through a physical scene and, secondly, that the Scene Walk prototype would help users develop a cognitive map that is superior to watching the equivalent video in a conventional video player.

Drawing on methods from wayfinding research (e.g. Richardson et al. 1999; Witmer et al. 1996), we designed a study to compare three experimental conditions: physically walking through a building; watching a first-person video that was recorded by someone walking along the same route; and using the Scene Walk prototype to "walk" along the route. Video recordings were made of three different routes, and a within-subjects design involved each participant following one of the routes in a physical walk, one with Scene Walk, and one by watching the video. Assignment of routes to each condition, and presentation order of the conditions, was balanced across participants. After following each route, the participant completed simple tasks so that we could assess the accuracy of the resulting cognitive map, including recall of the route, and reporting on the position and direction of a distinctive object.

### 5.2 Preparation

Three different routes were designed, each of equivalent difficulty and length (as validated in a pilot study), in a university campus building that was unfamiliar to experiment participants. Each route included multiple turns, since changes in direction are the key feature in route-learning tasks, and included two flights of stairs, since navigating different levels in a building requires formation of a 3D model. We carried out several pilot studies to decide suitable length and level of difficulty for the routes. Each route was approximately 200 meters along indoor corridors, starting at the end of one corridor, including 8 or 9 turns, and two flights of stairs, before ending at an office door. The directions of the individual turns in the sequence were different, but equivalent in complexity, for all routes. Recordings were made by the experimenter, using the capture software previously described, running on a Huawei P20 phone worn in the middle of the chest. The 3D scene models of three routes were reconstructed from recorded image sequences and camera poses using PMVS method. The number of 3D patches in each scene model is 800,000 on average. Each patch is $7 \times 7$ pixels. Each route covers a volume of about $1.6 \times 2.4 \times 200$ $m^3$, as it is mostly office corridors along the route.

We recruited 12 participants from the students and staff at the University of Cambridge, including five females and seven males, aged from 18 to 40 years old ($M = 28$, SD $= 4.65$). All participants were physically able. All three routes used in the experiment were in an office building of the Department of Physics. None of the participants was familiar with the routes or the building. All participants reported informally that they found the route learning task to be challenging.

The experiment procedure was approved by the ethics committee of the University of Cambridge Department of Computer Science and Technology. Participants were compensated for their participation with a gift voucher.

### 5.3 Procedure

Each participant completed sessions in three conditions, navigating one route in a *Walking* session, one in a *Video* session, and one in a *Scene Walk* session. In the Walking session, the participant was asked to walk through the building, following the experimenter. The experimenter walked at a comfortable normal speed, averaging about 1.4 m/s. In the Video session, the participant was asked to watch the video played once through, using a default video player on a standard laptop model (Dell P75F003). The screen size of the laptop is 15.6", and its screen resolution is $1920 \times 1080$. The frame rate of the video is 30 frames/second, and its image resolution is $1280 \times 720$. The video was recorded by the same experimenter while walking with the same speed as that in the Walking session. Participants were allowed to play and pause the video player at any time, but not to rewind and play back.

The Scene Walk session used the same Dell P75F003 laptop, with the Scene Walk prototype running under the Godot game engine, and the complete 3D reconstructed scene, the camera trajectory and the 2D video player rendered. At the start of the Scene Walk session, the experimenter briefly explained the operation of Scene Walk, and the participant was able to refer to a reference list of Godot navigation controls (attached in "Appendix 1") while practicing on a short trial video (with only two turns). All participants learned to use the Scene Walk controls within 2–5-min practice. When the participant was ready, he/she was asked to navigate the route from the beginning to the end using Scene Walk. They were able to switch freely between first-person mode and God-view mode whenever they wished. They completed the route only once and were not allowed to navigate backward, or to repeat sections.

Before the experiment started, the participants read a detailed explanation of the procedure, and signed a consent declaration. Participants then completed three sessions, with presentation order of the three conditions balanced across participants. Each session followed the previous one

immediately, with no break interval. At a predetermined point in each route, the experimenter would tell the participant to note a distinctive object, by pointing to it and saying that "This is the distinctive object, the *XXX* [saying the name of the object]". The participant did not know the distinctive object in advance, and the distinctive object in each route is different. There was no other interaction between the experimenter and the participant during the route navigation. After following each route in a session, the participant completed the three wayfinding assessment tasks for that session. Mean completion time for the whole experiment was 54 min, within which each of the three sessions lasted for about the same duration.

### 5.4 Tasks

In each session, after following a route either physically or virtually, we tested the cognitive map of the participant using a variety of recall and judgment tasks. The first task was to verbally describe the route, as if telling another person to follow the same route. The sequence of instructions given was recorded for further analysis. For the second task, participants were taken into a room of known dimensions—this room was used for measurement only and had no overlap with any route in either session. They stood at a fixed location in the room and were asked to imagine that they were facing forwards at the start point of the route. They were then asked to point with a laser pointer in the direction that they imagined for the distinctive object, while the experimenter took a photograph of the laser spot for later measurement. The third task was to estimate the length of a direct line (in meters) from the start of the route (the point where the participants imagined standing) to the location of the distinctive object. The ground-truth distance and the ground-truth direction of the distinctive object were determined using the construction plans of the building where the route recordings were made.

There was an additional task for the Video and Scene Walk conditions, which was to find an image frame in the video that contained the distinctive object. In the Video condition, the participant located the image frame by clicking on the progress bar in the video player. In the Scene Walk condition, the participant located the image frame by jumping along the trajectory using navigation keys. For this task, elapsed time was measured from when the participant started the search to when they found an image frame containing the object.

Finally, after the Scene Walk condition, we asked the participant how often they had played video games with the navigation controls as used in Godot and in Scene Walk. All participants claimed never to have played a game of this type, although some had played first-person games with a God-view mode. None of these had played games regularly in recent years.

## 6 Results

### 6.1 Route recall accuracy

Accuracy of route recall was measured by comparing the sequence recalled by the participant to the actual route. For comparability between routes, the accuracy statistic is the ratio of the number of turns correctly described, to the actual number in the route. Although the participant reported the turns in the route sequentially, we counted the largest number of continuously correctly described turns. For example, if only the fifth turn was missed in the route recall of a ten-turn route, and other turns were reported correctly, then the accuracy of route recall would be 0.9 rather than 0.4. Figure 6 shows the distributions of route recall accuracy in three conditions (Walking, Video and Scene Walk). We have attached the raw data in Appendix 1. Table 2 reports mean and standard deviation for each of the three methods (Walking, Video and Scene Walk) and also the differences between conditions for each participant, corresponding to our two hypotheses comparing Scene Walk to actual walking and to conventional video browsing. Figure 7 shows the distributions of the hypothesised differences in accuracy between the paired samples in these conditions: "Walking−Scene Walk" and "Video−Scene Walk". A Shapiro–Wilk test

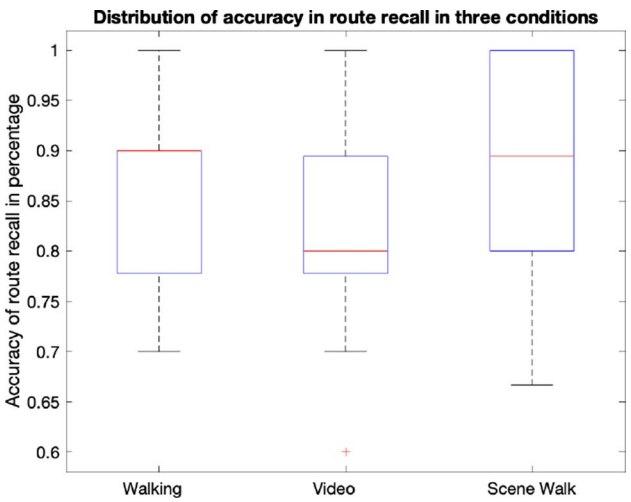



**Fig. 6** Distribution of route recall accuracy in three conditions: Walking, Video and Scene Walk. In these standard Matlab box plots, the box extends from the first to third quartile, the red line shows the median, whiskers show the data range excluding outliers (approximately 99th percentile), and a red plus shows outliers that are more than 1.5 box lengths away from the bottom or top of the box (Color figure online)

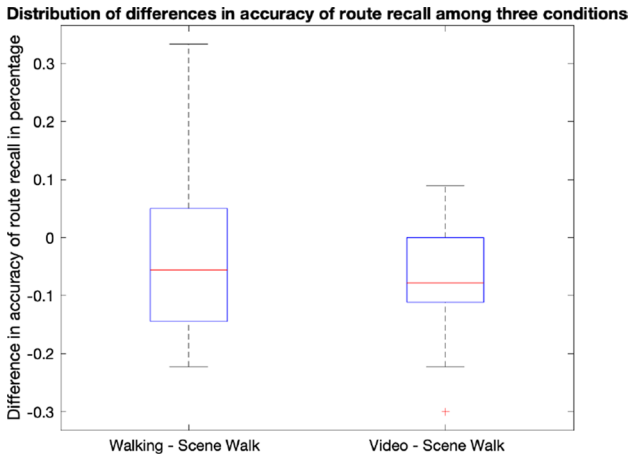



**Fig. 7** Distribution of differences in route recall accuracy, comparing Scene Walk to the two comparison conditions: Walking and Video

**Table 2** Descriptive statistics for route recall accuracy

| Condition | Mean | SD | N |
|---|---|---|---|
| Walking | 0.86 | 0.09 | 12 |
| Video | 0.82 | 0.11 | 12 |
| Scene Walk | 0.90 | 0.11 | 12 |
| Walking−Scene Walk | −0.03 | 0.16 | 12 |
| Video−Scene Walk | −0.08 | 0.11 | 12 |

**Table 3** Paired-sample *t* tests for route recall accuracy

| Hypothesis | *t*-value | *p*-value |
|---|---|---|
| Walking versus Scene Walk | −0.75 | 0.24 |
| Video versus Scene Walk | −2.44 | 0.02 |

confirms that these differences are normally distributed. Table 3 shows the results of the paired-sample *t*-tests for the two hypotheses.

The mean accuracies for the Walking and Scene Walk conditions are very similar, and the paired-samples *t*-test does not find evidence for a significant difference ($M = 0.86$ and $M = 0.90$, $p = 0.24$). Recall accuracy in the Video condition is significantly poorer by comparison to Scene Walk ($M = 0.82$, $p = 0.02$).

## 6.2 Distance estimation

We asked participants to estimate the distance from the beginning of the route to the location of the distinctive object. Figure 8 shows the distributions of distance estimation in three conditions (Walking, Video and Scene Walk).

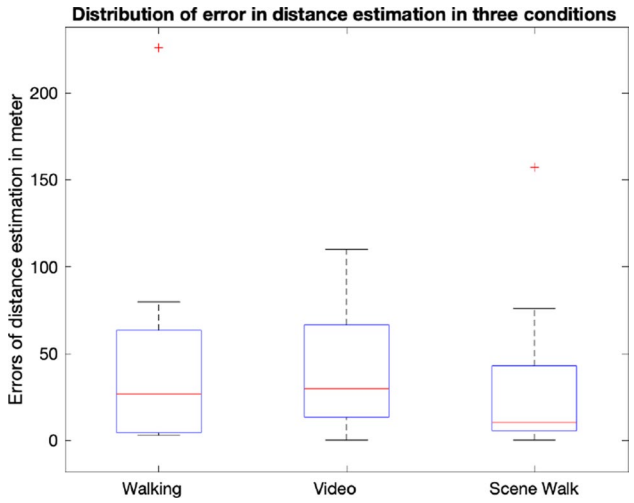



**Fig. 8** Distribution of error of distance estimation in three conditions: Walking, Video and Scene Walk. In standard Matlab box plots, a red plus shows outliers that are more than 1.5 box lengths away from the bottom or top of the box (Color figure online)

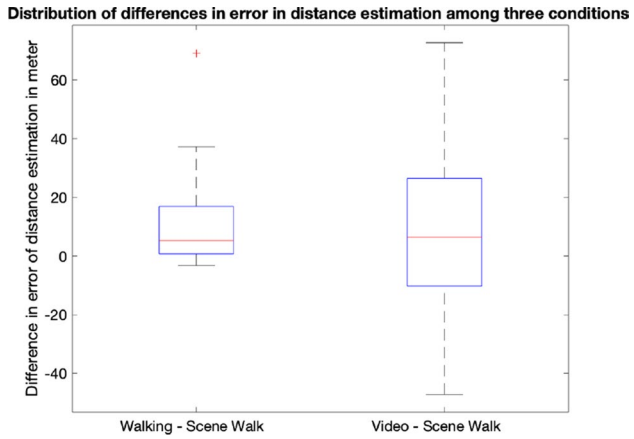



**Fig. 9** Distribution of differences in error of distance estimation, comparing Scene Walk to the two comparison conditions: Walking and Video

We calculate an error statistic as the difference (in metres) between their estimate and the actual ground-truth distance. Figure 9 shows the distribution of the differences in the error statistic between the paired samples for our two hypotheses: "Walking−Scene Walk" and "Video−Scene Walk". A Shapiro–Wilk test confirms that the differences of Video−Scene Walk are normally distributed. However, one extreme value in the differences of Walking−Scene Walk means that this statistic cannot be assumed to follow a normal distribution. We therefore used the nonparametric Wilcoxon signed ranks test to test the hypothesis for "Walking−Scene Walk". Descriptive statistics are shown in Table 4. Results of the Wilcoxon signed ranks test and paired-sample *t*-tests are shown in Tables 5 and 6, respectively.

**Table 4** Descriptive statistics for distance estimation error (metres)

| Condition | Mean | SD | N |
|---|---|---|---|
| Walking | 45.12 | 62.98 | 12 |
| Video | 40.72 | 35.23 | 12 |
| Scene Walk | 31.84 | 45.95 | 12 |
| Walking−Scene Walk | 13.28 | 20.90 | 12 |
| Video−Scene Walk | 8.88 | 31.79 | 12 |

**Table 5** Wilcoxon signed ranks tests for distance estimation error

| Hypothesis | *z*-value | *p*-value |
|---|---|---|
| Walking versus Scene Walk | −2.51 | 0.01 |
| Video versus Scene Walk | −0.82 | 0.21 |

**Table 6** Paired-sample *t* tests for distance estimation error

| Hypothesis | *t*-value | *p*-value |
|---|---|---|
| Walking versus Scene Walk | 2.20 | 0.02 |
| Video versus Scene Walk | 0.97 | 0.18 |

In distance estimation, the mean error for the Scene Walk condition ($M = 31.84$) is lower than that for the Video condition ($M = 40.72$). However, the standard deviation of error for Scene Walk ($SD = 45.95$) is considerably higher than for Video, which also has a large standard deviation ($SD = 35.23$). As a result of this high variance, the paired-sample *t*-test does not allow us to report a significant difference in means ($p = 0.18$). While we expected performance using Scene Walk to be comparable to that when walking physically through the building, performance on this task in the Walking condition has even larger mean error ($M = 45.12$) and greater variability ($SD = 62.98$) than either of the other conditions. Both Wilcoxon signed ranks test and paired-sample *t*-test find that this difference in means is significant ($p = 0.01$ and $p = 0.02$, respectively).

## 6.3 Angle estimation

The third task required the user to estimate the angle at which the distinctive object was located, relative to the position and orientation where they were standing at the start of the route. Figure 10 shows the distributions of angle estimation in three conditions (Walking, Video and Scene Walk). We calculate an error statistic as the difference (in degrees) between vectors representing their estimate (as calculated from the position of the laser spot in a photograph) and the actual direction. Figure 11 shows the distribution of the differences in the error statistic between the paired samples for our two hypotheses: "Walking−Scene Walk" and "Video−Scene Walk". A Shapiro–Wilk test confirms that these differences are normally distributed. Descriptive statistics and results of paired-sample *t*-tests are shown in Tables 7 and 8, respectively.

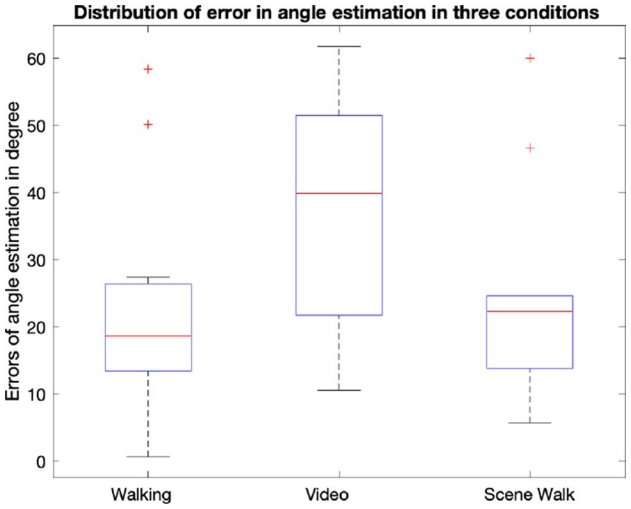



**Fig. 10** Distribution of error of angle estimation in three conditions: Walking, Video and Scene Walk. In standard Matlab box plots, a red plus shows outliers that are more than 1.5 box lengths away from the bottom or top of the box (Color figure online)

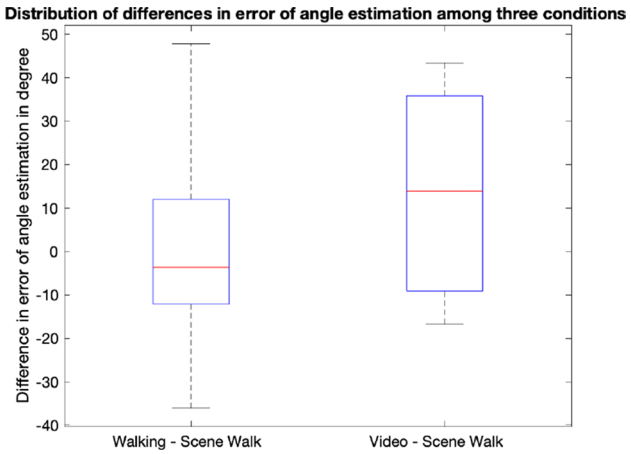



**Fig. 11** Distribution of differences in error of angle estimation, comparing Scene Walk to the two comparison conditions: Walking and Video

The mean of the errors in the Scene Walk and Walking conditions is similar, ($M = 23.58°$ and $M = 23.02°$), with the difference non-significant ($p = 0.47$). Mean error in the Video condition is significantly larger than in the Scene Walk condition ($M = 36.89°$, $p = 0.03$).

**Table 7** Descriptive statistics for angle estimation error (degrees)

| Condition | Mean | SD | N |
|---|---|---|---|
| Walking | 23.02° | 16.45° | 12 |
| Video | 36.89° | 18.06° | 12 |
| Scene Walk | 23.58° | 15.69° | 12 |
| Walking−Scene Walk | −0.56° | 24.21° | 12 |
| Video−Scene Walk | 13.31° | 22.11° | 12 |

**Table 8** Paired-sample *t* tests for angle estimation error

| Hypothesis | *t*-value | *p*-value |
|---|---|---|
| Walking versus Scene Walk | −0.08 | 0.47 |
| Video versus Scene Walk | 2.09 | 0.03 |

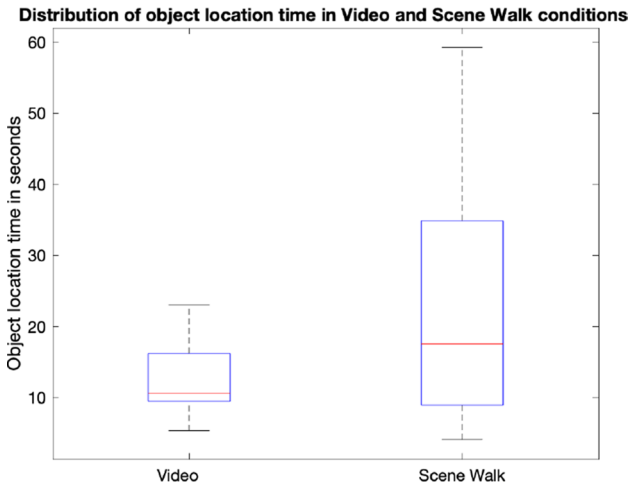



**Fig. 12** Box plots showing the distribution of time taken to locate a video frame containing a distinctive object in two conditions: Video and Scene Walk

## 6.4 Time to locate target frame

For the Video and Scene Walk techniques, we measured the time taken to locate a frame containing the distinctive object. Figure 12 shows the distributions of object location time in two conditions (Video and Scene Walk). The distribution of difference between times in the Video and Scene Walk condition is presented in Fig. 13. A Shapiro–Wilk test finds that the differences between the paired samples are not from a normal distribution, so a nonparametric Wilcoxon signed ranks test is used for hypothesis testing. Descriptive statistics and results of the Wilcoxon signed ranks test are shown in Tables 9 and 10, respectively. The mean time taken to locate a target frame was 12.49 s for the Video condition and 24.07 s for the Scene Walk condition. The Wilcoxon signed ranks test finds that this difference is marginally significant

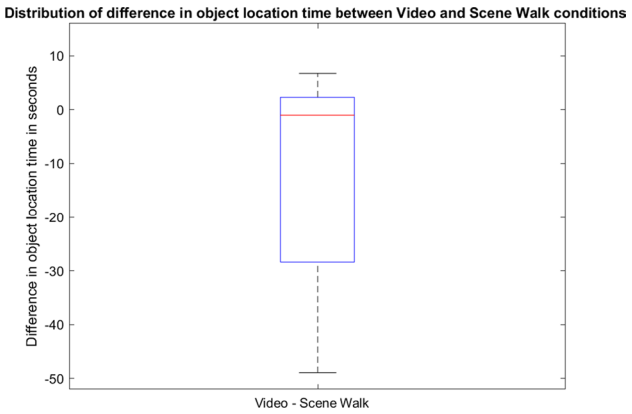



**Fig. 13** Box plots showing the distribution of differences in time taken to locate a video frame containing a distinctive object between two conditions: Video and Scene Walk

**Table 9** Descriptive statistics for object location time (seconds)

| Condition | Mean | *SD* | *N* |
|---|---|---|---|
| Video | 12.49 | 5.24 | 12 |
| Scene Walk | 24.07 | 19.59 | 12 |
| Video−Scene Walk | −11.58 | 19.39 | 12 |

**Table 10** Wilcoxon signed ranks test for object location time

| Hypothesis | *z*-value | *p*-value |
|---|---|---|
| Video versus Scene Walk | −1.26 | 0.10 |

($p = 0.10$), meaning that while no statistically reliable conclusions can be drawn, the observed difference cannot be discounted as resulting from random variation.

## 6.5 Discussion of quantitative results

The results reported in the previous section show that route recall accuracy in the Scene Walk condition is significantly better than in the Video condition. This confirms the finding of previous studies where navigating in a virtual environment supported improved route recall by comparison to photographs of the route (Witmer et al. 1996). Our finding that the non-photorealistic rendering style of Scene Walk has no disadvantage by comparison to fully photorealistic video recording confirms the finding by (Ruddle and Lessels 2009) that virtual environments with "impoverished" visual detail are equally effective for the acquisition of cognitive maps.

In addition, the error in estimating the direction of the distinctive object is significantly smaller in the Scene Walk condition by comparison to the Video condition. These

results support our hypothesis that Scene Walk allows users to create a more accurate cognitive map of the recorded scene, by comparison to watching first-person video. In contrast, comparing Scene Walk to the Walking condition where participants actually walk through the physical building, we observed that the mean values are similar and that tests for difference of means are non-significant. This does not allow us to discount any effect, but lends support to our hypothesis that the cognitive map formed by using Scene Walk is comparable to that formed when walking physically through a building. This suggests that value of Scene Walk in supporting this task is at least as effective by comparison to real-world navigation as has been demonstrated for VR displays in studies such as (Kuliga et al. 2020).

In the task estimating the distance to the distinctive object, comparison of errors between the Scene Walk condition and the Video condition does not find a significant difference. We noted that the mean error in the Scene Walk condition is lower than in Video condition; however, the standard deviation in the Scene Walk condition is higher than in the Video condition. This result suggests that participants do not have a consistent basis for estimation of distance. We note that in the physical Walking condition, the error in distance estimation is even larger than the other two conditions, suggesting that estimating the distance to an object located elsewhere in a multi-storey building is a difficult task even in the physical world. This is confirmed by participants' informal reports, since they remarked when completing this task that they found it very difficult to make estimates of distance in three dimensions, between two points that are on different levels of a building. This finding can be compared to the observation by Ruddle et al. (1997), when comparing participants' ability to estimate distances within a virtual environment (VE). As reported by Ruddle et al. "Despite being given a sense of scale, our navigation participants showed wide variability in their ability to estimate absolute distances, and this did not correlate with other measures of their spatial knowledge (route-finding ability, distance correlations and direction estimates). In addition, they did not demonstrate a specific tendency to either over- or underestimate the distances. Although we do not know how well they were able to estimate distances in the real world, their mean performance was substantially worse than that of the (Thorndyke and Hayes-Roth 1982) navigation participants. It may be that absolute distance estimation is inherently difficult in a VE."

Comparison of the time taken to locate an object using Scene Walk and a standard Video browser shows that the task completion times using Scene Walk are longer on average, although this difference is only marginally significant. This result does not support our hypothesis, in which we had expected to observe superior performance in the Scene Walk condition. However, we note that performance in this task was highly variable, with participants reporting a variety of strategies. Some completed the task very quickly in the Video condition, but reported that this was accidental. For example, after remembering that the object was in the middle of the route, one participant simply clicked in the middle of the video progress bar and found the object there. While using the Scene Walk prototype to locate the frame, participants reported that they did find the spatial position of the object instantly by viewing the scene reconstructed in Scene Walk, but were not sufficiently familiar with the navigation controls to quickly move to that place. For practical application of the Scene Walk approach, implementing a reverse index from the 3D model to the camera views contributing features at that location could in future make it possible to navigate to an object by clicking in the scene, which would make this task trivially easy.

### 6.6 Observation of user strategies

We observed the participants to see how they used the two view modes of Scene Walk (God-view and first-person view) during the route navigation session. Eight of the 12 participants used the God-view more than the first-person view during the session. Participants reported that God-view helped them observe the complete scene and route. With the moving video player, they can know their position relative to the whole route easily, demonstrating that Scene Walk is able to retain the advantage observed by Ruddle et al. (1997) for overall view of building layout in a floor plan. In addition, they reported that the visualised trajectory of camera poses was helpful in route recall, confirming route encoding strategies observed by Taylor and Tversky (1992). The first-person view provided more details when the detailed information was required, for example, when the viewer was approaching turns. As a result, a typical strategy was to use the God-view and switch to first-person view when they found this necessary. Four of the 12 participants used first-person view more than the God-view. As reported by these participants, they felt the video viewer was a safe choice because it seemed more familiar in their daily life. Nevertheless, these participants reported that God-view was helpful when locating the distinctive object, and they switched to this view when trying to understand the spatial location of the distinctive object. They also reported that at each turn, it was helpful to use God-view to understand their present spatial position and that in first-person view, it became difficult to remember their direction and position in the spatial environment after a few turns. This confirms that Scene Walk retains the advantages of cognitive map acquisition when aided by a map or floor plan (Thorndyke and Hayes-Roth 1982; Ruddle et al. 1997).

We noted the number of times that participants switched between views during each session. The Scene Walk system always started in God-view by default, after which we counted each cycle that the user switched to first-person view and back. The mean number of view switch cycles across all participants is 4.56 (SD = 2.59). We conclude from this frequency of switching that both God-view (rich in 3D information) and first-person view (rich in 2D information) are of value to the user when completing wayfinding tasks.

## 7 Discussion of Scene Walk

Overall, we found statistical evidence that the Scene Walk interface was beneficial for participants, in performing tasks that depend on a cognitive map of the scene where first-person videos had been recorded. We summarise three important benefits that are provided by Scene Walk.

### 7.1 Reconstructed 3D scene and spatial structure

The non-photorealistic rendering of a reconstructed 3D scene means that the viewer has immediate access to the whole of the 3D model and can also use viewing controls to view it freely from different angles. An important benefit by comparison to conventional video players is that the complete spatial structure of the 3D scene is available to the viewer. The performance observed in the user study demonstrates clear benefits. The available spatial structure: (a) helped users to form an accurate cognitive map of the route they had followed; and (b) assisted them in understanding the spatial position of a distinctive object that had been observed on the route. These benefits resulted in superior performance for route learning and angle estimation tasks when using Scene Walk, confirming advantages that have been observed for VR environments over photographs with verbal route descriptions (Witmer et al. 1996), while retaining the advantages observed for maps and floor plans (Ruddle et al. 1997).

### 7.2 Complete 2D image details with free view control

The Scene Walk system keeps all recorded image frames, meaning that no recorded information is lost in this interaction method. However, the navigation approach in Scene Walk means that the user does not have to review all image frames in order to view first-person video content. By using the 3D scene and spatial structure to navigate the video, the user can decide when they need specific image details and what image details they are looking for. This confirms that it is productive to apply a "wayfinding" analogy to the problem of navigating within a first-person video recording to find images captured from a specific location. Free switching of view mode allows users to access image details at any time, by switching to first-person view mode.

### 7.3 Camera trajectory registration to 3D structure

The camera trajectory in Scene Walk is rendered into the 3D scene as a visualisation of the locations and poses from which the video was recorded. The visualised camera trajectory offers the viewer an overview of the path followed by the camera-wearer, providing orientation cues that assist with navigation and formation of cognitive maps as demonstrated by (Burigat and Chittaro 2007). The 2D video player, moving forward along the camera trajectory, allows the viewer to directly understand the spatial position of the camera-wearer relative to their point of view within the scene and thus to construct a cognitive map that corresponds to the experience of walking through a physical scene.

### 7.4 Possibility of extending to outdoor videos

In this usage scenario and experimental evaluation, we have applied Scene Walk to video collected in the interior of a building. However, the interaction method can straightforwardly be applied in outdoor environments, either where existing models are available [for example, using CityGML models (Kolbe et al. 2005; Gröger and Plümer 2012)], or where partial surface models are constructed using the methods described in this paper. The 3D reconstruction technique that we have described is not limited to indoor scenes, and an interesting question for future research will be whether similar patch-based non-photorealistic rendering would be equally valuable for users to construct cognitive maps from first-person video, for VR-assisted wayfinding in outdoor settings. Previous research by Burigat and Chittaro (2007) demonstrates the value of desktop virtual environments in an outdoor navigation application and also confirms that orientation cues such as our God-view reference direction and camera pose markers support navigation performance in that context.

## 8 Limitations and future work

In order to evaluate the accuracy of users' cognitive maps, we used several different performance measures, only some of which demonstrated advantages for the Scene Walk approach. While route recall and angle estimation were both assisted by Scene Walk, the third measure (distance estimation to a distinctive location) was not assisted, so this may be a limitation of the system as described (although we note that previous research has also found this task to be

challenging). Given that users specifically report how difficult it is to estimate direct distance between points within a building, we note that it would be relatively straightforward to add a scale reference into the Scene Walk viewer in the future. If future applications of first-person video did include a requirement for distance estimation, the Scene Walk interaction model would provide a good basis for enhancement with scale reference functionality, offering the same advantages for distance estimation that were found for maps over virtual environments, in previous research by Richardson et al. (1999) and Ruddle et al. (1997). We note that it would be very difficult to add such capabilities to conventional video viewers, in which spatial structure is not visualised.

In this work, we have made conservative projections regarding the sensor capabilities of mass-market wearable cameras in the near future, based on the devices that were available in consumer and professional markets at the time the research scope was defined. We developed a custom app (implemented on a commodity Android device), using only the onboard pose sensors that we predict will be deployed in models for the target market. We note that this combination of sensors is already available on some professional models of body-worn cameras, but not yet ubiquitous in low-cost consumer models (e.g. entry-level models of the popular GoPro "Hero" range). At the conclusion of this phase of research, it does appear that recent high-end consumer products are starting to include the sensing capabilities we predicted. However, if the market for wearable cameras develops such that ground-truth pose sensors are not included in consumer models, this would require additional methods for data analysis.

This work has explored the potential advantages of treating video navigation, from a human cognition perspective, as a special case of wayfinding. This perspective motivated the creation of the Scene Walk technique, as a new kind of video viewing interface where the moving video is inserted within a static non-photorealistic 3D scene. We have discussed a variety of video viewing use cases in which Scene Walk may be valuable. However, this work has not considered applications other than video viewing. It is possible that the cognitively motivated approach used in Scene Walk could also offer benefits in other applications, but we do not suggest here that Scene Walk is a general-purpose approach suited to all applications.

In this work, we have specifically investigated the value to users of a non-photorealistic scene model, constructed using only image patches that triangulate camera views from uncalibrated cameras. It is of course possible to create more complete and photorealistic VR models, through methodical scanning, use of calibrated cameras, and supplemental range-finding hardware. Many products are already available, some at the upper end of the consumer price range, that support such VR model construction. It therefore seems likely that in future, photorealistic 3D models of familiar environments will become more widely available. We expect that the interaction methods we have described are likely to remain effective when better quality data are available, and this expectation is confirmed by previous research showing that cognitive map advantages are still observed when the level of detail is reduced (Ruddle and Lessels 2009), although this expectation would benefit from further assessment using data from actual consumer devices.

Potential future applications of the Scene Walk technique include a variety of situations in which first-person video is recorded from wearable cameras, and must be retrieved or analysed from archives of video content. In future work, more sophisticated machine vision methods would allow the 3D scene model to be reconstructed more efficiently, and with higher accuracy, potentially exploiting new sensing capabilities in future wearable camera products. In the limit, a high-accuracy reconstructed 3D scene model might be able to provide complete contextual information without the assistance of 2D image frames—although this still does not solve the problem of time-varying data in cases where the structure of the scene is changing in places not within the view of any camera at the time of change, or that some parts of the scene may never have been observed. We have demonstrated the value of non-photorealistic rendering, even with partial data, by integrating a 2D rendering of the captured photographic frames into a partially reconstructed scene. Observation of users during our studies demonstrates that users are able to use the Scene Walk technique to switch effectively between photographic frames and non-photorealistic visualisation of spatial context, and this approach can in future be extended to a variety of non-photorealistic partial model content.

Based on current state of the art, it also seems possible that partial scene models could be constructed in real time, allowing the Scene Walk method to be used for real-time mission control or situation assessment from streamed video data. Our experimental findings suggest that integrating 2D video recordings into an interactive 3D model in this way could offer substantial advantages for professional analysts, editors and other users of real-time first-person video content, especially where it is impractical to carry out full 3D scans of the environment in advance.

## 9 Conclusion

We have introduced Scene Walk, a video viewing technique designed for first-person video recorded from wearable cameras. The Scene Walk prototype integrates a 2D video player and visualisation of the camera trajectory into a non-photorealistic partial rendering of the 3D environment as reconstructed from image content. We applied methods from wayfinding research to assess the effectiveness of this approach in comparison to physical experience of the scene. Results demonstrate that Scene Walk prototype supports construction of a cognitive map that enables performance comparable to walking through a physical scene. Results also demonstrate that viewing first-person video with the Scene Walk prototype allows users to construct a cognitive map enabling significantly improved task performance by comparison to a conventional video viewer. These results support our hypothesis that a partial virtual scene model reconstructed from image geometry can be an effective video interaction tool, even where fully realistic scene reconstruction is not feasible.

## Appendix

### A Raw experimental data

See Tables 11, 12, 13 and 14.

**Table 11** The raw data of route recall accuracy in three conditions: Walking, Video and Scene Walk

| Participants | Walking | Video | Scene Walk |
|---|---|---|---|
| Participant 1 | 0.78 | 0.90 | 1.00 |
| Participant 2 | 1.00 | 1.00 | 1.00 |
| Participant 3 | 0.78 | 0.70 | 1.00 |
| Participant 4 | 0.78 | 0.80 | 0.80 |
| Participant 5 | 0.90 | 0.78 | 1.00 |
| Participant 6 | 0.90 | 0.89 | 0.80 |
| Participant 7 | 0.90 | 1.00 | 1.00 |
| Participant 8 | 0.90 | 0.78 | 0.90 |
| Participant 9 | 1.00 | 0.60 | 0.67 |
| Participant 10 | 0.70 | 0.80 | 0.89 |
| Participant 11 | 0.80 | 0.80 | 0.89 |
| Participant 12 | 0.90 | 0.80 | 0.80 |

**Table 12** The raw data of error of distance estimation in three conditions: Walking, Video and Scene Walk

| Participants | Walking | Video | Scene Walk |
|---|---|---|---|
| Participant 1 | 4.82 | 13.88 | 7.08 |
| Participant 2 | 29.82 | 16.12 | 22.92 |
| Participant 3 | 59.82 | 57.08 | 56.12 |
| Participant 4 | 79.82 | 57.08 | 76.12 |
| Participant 5 | 25.47 | 43.59 | 5.77 |
| Participant 6 | 27.92 | 0.18 | 13.88 |
| Participant 7 | 3.88 | 79.82 | 7.08 |
| Participant 8 | 226.12 | 109.82 | 157.08 |
| Participant 9 | 9.88 | 15.92 | 0.82 |
| Participant 10 | 3.88 | 12.92 | 5.18 |
| Participant 11 | 67.08 | 76.12 | 29.82 |
| Participant 12 | 2.92 | 6.12 | 0.18 |

**Table 13** The raw data of error of angle estimation in three conditions: Walking, Video and Scene Walk

| Participants | Walking | Video | Scene Walk |
|---|---|---|---|
| Participant 1 | 11.63 | 59.70 | 46.74 |
| Participant 2 | 0.68 | 46.09 | 6.75 |
| Participant 3 | 21.11 | 61.82 | 18.40 |
| Participant 4 | 27.43 | 56.94 | 23.13 |
| Participant 5 | 25.47 | 43.59 | 5.77 |
| Participant 6 | 16.26 | 18.15 | 23.45 |
| Participant 7 | 15.95 | 36.60 | 17.16 |
| Participant 8 | 23.94 | 43.27 | 60.02 |
| Participant 9 | 9.90 | 11.73 | 24.68 |
| Participant 10 | 50.20 | 28.76 | 21.63 |
| Participant 11 | 58.37 | 25.40 | 10.54 |
| Participant 12 | 15.30 | 10.60 | 24.69 |

**Table 14** The raw data of object location time in two conditions: Video and Scene Walk

| Participants | Video | Scene Walk |
|---|---|---|
| Participant 1 | 9.67 | 7.96 |
| Participant 2 | 9.22 | 6.5 |
| Participant 3 | 23.07 | 23.1 |
| Participant 4 | 15.26 | 11.95 |
| Participant 5 | 10.83 | 4.08 |
| Participant 6 | 5.36 | 29.69 |
| Participant 7 | 7.5 | 40 |
| Participant 8 | 19.68 | 59.23 |
| Participant 9 | 10.32 | 59.25 |
| Participant 10 | 11.84 | 9.96 |
| Participant 11 | 17.14 | 25.12 |
| Participant 12 | 9.96 | 11.94 |

## B Manual for Scene Walk

- Global navigation using WSAD and mouse
  - * W—forward
  - * S—back
  - * A—left
  - * D—right
  - * Mouse motion—turning left/right/up/down
- 3D scene toggled with "3D scene" (1 button)
- 2D video toggled with "2D video" (2 button)
- Camera trajectory toggled with "trajectory" (3 button)
- Play/pause the video with "play/pause" button (P button)
- Enter or exit the "first-person view" by pressing "Switch View" button (space button)
- Jump 0.1 seconds along the trajectory by pressing "backward" or "forward" button, jump 10 seconds along the trajectory by pressing "Shift "+"backward" or "Shift "+"forward" button.

**Acknowledgement** This research was funded by Boeing Corporation. The development team who created the Scene Walk prototype were Jacob Coxon, Agnieszka Koc, Adam Kucz, Zijun Yan and Jae Yeun Yoon. We are also grateful to the participants in the experiment described.



## References

Arev I, Park H, Sheikh Y, Hodgins J, Shamir A (2014) Automatic editing of footage from multiple social cameras. ACM Trans Graph 33:1–11. https://doi.org/10.1145/2601097.2601198

Arthur P, Passini R (1992) Wayfinding: people, signs, and architecture

Ballan L, Brostow G, Puwein J, Pollefeys M (2010) Unstructured video-based rendering: interactive exploration of casually captured videos. ACM Trans Graph 29:1. https://doi.org/10.1145/1833351.1778824

Betancourt A, Morerio P, Regazzoni CS, Rauterberg M (2015) The evolution of first person vision methods: a survey. IEEE Trans Circuits Syst Video Technol 25(5):744–760. https://doi.org/10.1109/TCSVT.2015.2409731

Bolanos M, Dimiccoli M, Radeva P (2017) Toward storytelling from visual lifelogging: an overview. IEEE Trans Hum Mach Syst 47(1):77–90. https://doi.org/10.1109/THMS.2016.2616296

Burigat S, Chittaro L (2007) Navigation in 3d virtual environments: effects of user experience and location-pointing navigation aids. Int J Hum Comput Stud 65(11):945–958

Chen Y, Jones GJF (2010) Augmenting human memory using personal lifelogs. In: Proceedings of the 1st Augmented Human International Conference, AH '10, pp 24:1–24:9. https://doi.org/10.1145/1785455.1785479

Dalton R, Hölscher C, Montello D (2019) Wayfinding as a social activity. Front Psychol 10(142). https://doi.org/10.3389/fpsyg.2019.00142

De D, Bharti P, Das SK, Chellappan S (2015) Multimodal wearable sensing for fine-grained activity recognition in healthcare. IEEE Internet Comput 19(5):26–35. https://doi.org/10.1109/MIC.2015.72

Furukawa Y, Ponce J (2010) Accurate, dense, and robust multiview stereopsis. IEEE Trans Pattern Anal Mach Intell 32(8):1362–1376. https://doi.org/10.1109/TPAMI.2009.161

Gibson D (2009) The wayfinding handbook: information design for public places. Princeton Architectural Press, Princeton

Golledge RG (1999) Human wayfinding and cognitive maps. In: Golledge RG (ed) Wayfinding behavior: cognitive mapping and other spatial processes. Johns Hopkins University Press, Baltimore, pp 5–45

Golledge RG, Smith TR, Pellegrino JW, Doherty S, Marshall SP (1985) A conceptual model and empirical analysis of children's acquisition of spatial knowledge. J Environ Psychol 5(2):125–152

Gröger G, Plümer L (2012) CityGML: interoperable semantic 3d city models. ISPRS J Photogramm Remote Sens 71:12–33. https://doi.org/10.1016/j.isprsjprs.2012.04.004

Herman JF, Siegel AW (1978) The development of cognitive mapping of the large-scale environment. J Exp Child Psychol 26(3):389–406

Higuch K, Yonetani R, Sato Y (2016) Can eye help you?: Effects of visualizing eye fixations on remote collaboration scenarios for physical tasks. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp 5180–5190. https://doi.org/10.1145/2858036.2858438

Higuchi K, Yonetani R, Sato Y (2017) Egoscanning: Quickly scanning first-person videos with egocentric elastic timelines. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp 6536–6546. https://doi.org/10.1145/3025453.3025821

Ho HI, Chiu WC, Wang YCF (2018) Summarizing first-person videos from third persons' points of views. In: European conference on computer vision, pp 72–89

Ishiguro Y, Rekimoto J (2012) Gazecloud: A thumbnail extraction method using gaze log data for video life-log. In: 2012 16th International symposium on wearable computers, pp 72–75. https://doi.org/10.1109/ISWC.2012.32

Jennings WG, Fridell LA, Lynch MD (2014) Cops and cameras: officer perceptions of the use of body-worn cameras in law enforcement. J Crim Justice 42(6):549–556. https://doi.org/10.1016/j.jcrimjus.2014.09.008

Kaplan S (1973) Cognitive maps in perception and thought. Cognitive mapping and spatial behavior, Image and environment, pp 63–78

Kitchin RM (1994) Cognitive maps: what are they and why study them? J Environ Psychol 14(1):1–19

Kolbe T, Gröger G, Plümer L (2005) CityGML: interoperable access to 3d city models. Geo-inf Disaster Manag. https://doi.org/10.1007/3-540-27468-5_63

Kono M, Miyaki T, Rekimoto J (2017) Jackin airsoft: Localization and view sharing for strategic sports. In: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, Association for Computing Machinery, New York, NY, USA, VRST '17. https://doi.org/10.1145/3139131.3139161

Kopf J, Cohen MF, Szeliski R (2014) First-person hyper-lapse videos. ACM Trans Graph 33(4):1–10. https://doi.org/10.1145/2601097.2601195

Kuliga S, Mavros P, Brösamle M, Hölscher C (2020) Comparing human wayfinding behavior between a real, existing building, a virtual replica, and two architectural redesigns. In: German Conference on Spatial Cognition, Springer, pp 160–179

Lackner JR, DiZio P (2005) Vestibular, proprioceptive, and haptic contributions to spatial orientation. Ann Rev Psychol 56:115–147

Lee YJ, Ghosh J, Grauman K (2012) Discovering important people and objects for egocentric video summarization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 1346–1353. https://doi.org/10.1109/CVPR.2012.6247820

Lin Y, Morariu VI, Hsu W (2015) Summarizing while recording: Context-based highlight detection for egocentric videos. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp 443–451. https://doi.org/10.1109/ICCVW.2015.65

Lynch K (1960) The image of the city, vol 11. MIT press, Cambridge

MacEachren AM (1992) Application of environmental learning theory to spatial knowledge acquisition from maps. Ann Assoc Am Geograph 82(2):245–274

McNamara TP (1986) Mental representations of spatial relations. Cognit Psychol 18(1):87–121

del Molino AG, Tan C, Lim J, Tan A (2017) Summarization of egocentric videos: a comprehensive survey. IEEE Trans Hum Mach Syst 47(1):65–76. https://doi.org/10.1109/THMS.2016.2623480

Nuernberger B, Höllerer T, Turk M (2018) Hybrid orbiting-to-photos in 3d reconstructed visual reality. pp 1–10, https://doi.org/10.1145/3281505.3281528

O'Neill M (1991) A biologically based model of spatial cognition and wayfinding. J Environ Psychol 11:299–320. https://doi.org/10.1016/S0272-4944(05)80104-5

Poleg Y, Halperin T, Arora C, Peleg S (2015) Egosampling: Fast-forward and stereo for egocentric videos. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4768–4776. https://doi.org/10.1109/CVPR.2015.7299109

Qiu X, Wen L, Wu C, Yang Z, Wang Q, Li H, Wang D (2020) Impact of learning methods on spatial knowledge acquisition. Front Psychol 11:1322

Richardson AE, Montello DR, Hegarty M (1999) Spatial knowledge acquisition from maps and from navigation in real and virtual environments. Memory Cognit 27(4):741–750

Ruddle RA, Lessels S (2009) The benefits of using a walking interface to navigate virtual environments. ACM Trans Comput Hum Interact 16(1):1–18

Ruddle RA, Payne SJ, Jones DM (1997) Navigating buildings in "desktop'' virtual environments: experimental investigations using extended navigational experience. J Exp Psychol Appl 3(2):143

Ruddle RA, Payne SJ, Jones DM (1999) Navigating large-scale virtual environments: what differences occur between helmet-mounted and desk-top displays? Presence Teleoperators Virtual Environ 8(2):157–168

Silva M, Ramos W, Ferreira J, Chamone F, Campos M, Nascimento ER (2018) A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2383–2392. https://doi.org/10.1109/CVPR.2018.00253

Smykla JO, Crow MS, Crichlow VJ, Snyder JA (2016) Police body-worn cameras: perceptions of law enforcement leadership. Am J Criminal Justice 41(3):424–443. https://doi.org/10.1007/s12103-015-9316-4

Snavely N, Seitz S, Szeliski R (2006) Photo tourism: exploring photo collections. 3d acm trans graph 25(3):835–846. ACM Trans Graph 25:835–846. https://doi.org/10.1145/1141911.1141964

Sugita Y, Higuchi K, Yonetani R, Kamikubo R, Sato Y (2018) Browsing group first-person videos with 3d visualization. In: Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces, Association for Computing Machinery, New York, NY, USA, ISS '18, p 55–60. https://doi.org/10.1145/3279778.3279783

Taylor HA, Tversky B (1992) Spatial mental models derived from survey and route descriptions. J Memory Language 31(2):261–292

Thorndyke PW, Hayes-Roth B (1982) Differences in spatial knowledge acquired from maps and navigation. Cognit Psychol 14(4):560–589

Tversky B (1993) Cognitive maps, cognitive collages, and spatial mental models. In: European conference on spatial information theory, Springer, pp 14–24

Wilson PN, Foreman N, Tlauka M (1997) Transfer of spatial information from a virtual to a real environment. Hum Factors 39(4):526–531

Witmer BG, Bailey JH, Knerr BW, Parsons KC (1996) Virtual spaces and real world places: transfer of route knowledge. Int J Hum Comput Stud 45(4):413–428

Xu J, Mukherjee L, Li Y, Warner J, Rehg JM, Singh V (2015) Gaze-enabled egocentric video summarization via constrained submodular maximization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2235–2244. https://doi.org/10.1109/CVPR.2015.7298836

Yao T, Mei T, Rui Y (2016) Highlight detection with pairwise deep ranking for first-person video summarization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 982–990. https://doi.org/10.1109/CVPR.2016.112