Look Out! A Design Framework for Safety Training Systems A Case Study on Omnidirectional Cinemagraphs

Brendan John* Sriram Kalyanaraman† Eakta Jain‡

University of Florida

ABSTRACT

Accidents occur when a person's attention is distracted. A key aspect of safety training is directing people's attention to potential hazards. Unfortunately, creating such hazards also puts people at risk, especially during safety training in manufacturing and construction. Virtual reality provides a training mechanism by which hazardous training scenarios can be created without putting the trainee at risk. We present a general framework for safety training systems and also present results from a case study where we create and evaluate a novel safety training environment, namely, omnidirectional cinemagraphs.

Keywords: omnidirectional cinemagraphs, 360-degree panorama, virtual reality, safety training

Index Terms: Human-centered computing—Systems and tools for interaction design; Computing methodologies—Virtual reality

1 INTRODUCTION

Safety is critical for worker health and productivity in every workplace, especially in the manufacturing and construction sectors. The Occupational Health and Safety Act of 1970 mandates that employers create a safe and healthy workplace for their employees. As a result, significant resources are spent on training inspectors to identify safety hazards, as well as training employees to identify and ameliorate safety hazards on the job site. However, creating training scenarios with potential hazards naturally puts the trainees at risk. As a result, teaching aids such as pictures and videos have been employed for workplace training. Virtual reality provides a unique teaching aid for safety training that allows for simulating a hazardous environment and at the same time, creating a sense of presence and immersion that keeps the trainee engaged.

In this position paper, we discuss the key issues in designing training systems for safety critical contexts. This design framework applies to domains including manufacturing and construction, defense and first responder, mining, nuclear safety, retail, and healthcare. We also conceptualize omnidirectional cinemagraphs, a hybrid training environment that affords the ability to guide the trainee's attention while retaining presence and immersion. We present a case study where we design omnidirectional cinemagraphs for fire safety training and report findings from a pilot study.

*e-mail: brendanjohn@ufl.edu

†e-mail: sri@jou.ufl.edu

‡e-mail: ejain@cise.ufl.edu

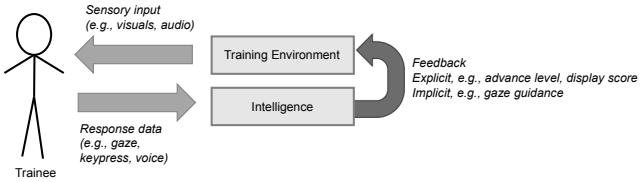

Figure 1: System overview and design variables in a safety training system.

2 DESIGN FRAMEWORK

Safety training may be of two main types: Theoretical or Experiential. Theoretical training refers to transferring knowledge about the fundamental natural world principles underlying their domain of work. These principles include the physics of how fire spreads, or how nuclear fission is controlled, or the steps to follow in case of an unusually high temperature reading. This type of training may be done via textbooks and binders containing text, pictures, graphs that the trainee needs to go through, or via lectures with a human or intelligent tutor. The trainee has declarative knowledge at this stage, i.e., they can describe a procedure or process. Experiential training on the other hand refers to hands-on experience which is aimed at the associative and automatic stages of learning procedure, i.e., being able to put their knowledge to practice without "thinking about it" under real world constraints such as limited time. This type of training may involve repeating practice exercises in a lab or teaching environment, and/or supervised hours in a working facility. The design guidelines, variables, cognitive models, and role of Artificial Intelligence in the former type of training has been extensively studied in the context of intelligent tutoring systems. In this paper, we focus on the latter, i.e., the type of training that involves hands on worker experience. While direct experience in the lab may be considered the highest fidelity training environment, there is a rich history of providing trainees with simulated environments to practice in. Simulated environments allow users unlimited hours of training at their convenience (in contrast to a lab or working facility) and thus have the advantage of reducing the time and cost of training. In addition, for safety training in particular, simulated environments make it possible to offer trainees experience with events that would rarely occur in real life, with future environments that may not yet exist, and by personalizing the training. In the paragraphs that follow, we define the variables of a simulated safety training system and create a design framework that may be adapted to specific contexts. Figure 1 illustrates this framework.

- **Trainee** This is the human user of the training system. User characteristics that impact the outcome of training may include group level characteristics such as age range, and individual characteristics such as personality or fatigue level.

- **Context** This refers to the work domain that the training applies to, and also to the type of scenarios that the training environment is intended to represent. For example, a fire safety

training may be targeted to workers in a plant, or to college freshman renting their first apartment.

- **Training Environment** This refers to the scene around the trainee including objects and people, which collectively form the stimulus for the trainee. The training environment will produce sensory input for the trainee, which includes sound, visuals and haptics. This includes objects that may be picked up, buttons that may be pressed, scene elements that react according to physics and can be knocked over.

- **Intelligence** This refers to the collection of methods and algorithms, together known as artificial intelligence or AI, that drive the stimulus presentation, process response data, and provide personalized or standardized feedback as desired. The input to these algorithms is the response data, i.e., information that is recorded about the trainee as he or she goes through the training. Response data includes explicit responses from the trainee, such as the hazards identified in the scene, and also implicit responses such as their facial expression and their gaze patterns. The output of these algorithms feeds back to the training environment. Explicit feedback includes advancing to the next level or displaying the score. Implicit feedback includes guiding the user such as using gaze guidance to direct them toward a hazard they need to identify.

The design of a simulation-based safety training system will start by identifying the trainee and context. The next step would be to identify the features of the training environment, categorized as below:

Sensory fidelity how closely the sensory cues created by the training environment match the real world.

- *Pictorial fidelity* refers to the quality of the modeling and rendering of the environment, ranging from the poly count of the meshes, to the shaders and lighting used to create scene elements such as background, objects, and people. Analogous to the mis-en-scene, or visual theme, in cinema, this feature sets the tone for the training environment. For example, a highly stylized "cartoony" look would be low in pictorial fidelity relative to a computer generated scene from a 360° texture and depth map. Pictorial fidelity will need to be carefully designed in case of augmented reality, as the design choices in this case will determine whether the scene elements feel integrated into the real world or "placed for training purposes".

- *Animation fidelity* refers to the extent which the scene elements in the training environment respect physics when either the trainee or the scene elements are in motion. Thus, an omnidirectional image is low in animation fidelity because it does not provide the correct parallax cues when the user moves closer or further. If a tool falls to the ground, the extent of its bounces is another aspect of animation fidelity. This feature also encompasses the animation of people in the training environment such as their facial expressions, body gestures, crowd behavior, etc.

- *Aural fidelity* refers to the sound generated by the training environment, including accurate spatialization, global effects such as echo, and the extent to which sound reflects interactive events such as a trainee knocking over a tool and it falling to the ground.

- *Haptic fidelity* refers to the extent to which the training environment can be "felt" by the user. This includes feeling the heat of fire, in addition to conventional haptic feedback in the form of feeling the resistance of a tool while using it.

- *Olfactory fidelity* refers to the extent to which the sense of smell is replicated in the training environment, for example, the sharp smell of rubber burning.

Interactivity, or, the ways in which the trainee is allowed to alter the training environment.

- *User degrees of freedom (DOF)* refers to how much the trainee may move in the simulated training environment. Typically, this refers to three degrees of rotational freedom and/or three degrees of translational freedom. We also include in this feature the extent to which the user is provided a body, for example, using hands to pick up an object versus using a controller with a button that picks up the object.

- *Environment responsiveness* refers to the extent to which scene elements in the training environment are responsive to user degrees of freedom. This includes whether buttons can be pressed, and how many objects can be grasped, knocked over, or stacked. This also includes whether water would splash if the training environment contained a puddle and the user stepped into it. Environment responsiveness is connected with animation fidelity, though it is distinct in that the movement of the scene elements is in response to the user. An omnidirectional static image would be low in both animation fidelity and environment responsiveness, whereas an omnidirectional video would be high in animation fidelity but low in environment responsiveness.

What distinguishes a safety training system from previously studied technology aided instruction and intelligent tutoring systems? We propose that safety training relies on three processes: ***See, Recognize, Act***. See refers to the perceptual process of noticing or attending to a potential area of interest, such as a spill, a poorly placed item, or an unusual reading on a meter. Recognize refers to the cognitive process of identifying that this is a hazard. Act refers to making the appropriate intervention. Thus, the unique aspect of safety training lies in directing the worker's attention to potential hazards. What they do about it (follow protocol, display presence of mind, etc.) is bottlenecked at detecting the safety hazard, i.e., the See stage. As a result, attention guidance is a key scaffolding in the training process.

Attention directability refers to whether the training environment affords the ability to implement attention guidance cues.

- *Visual guidance* refers to any visual stimuli that intends, and is able to, influence the viewer's eyes to orient in a particular direction. Visual cues can either be *overt* in that they are perceived by the viewer, such as a large red arrow or character pointing, or *subtle* in that they are not perceived by the viewer's visual system. A common subtle cue for guidance is to present a flicker in the viewer's periphery such that the flicker triggers an eye movement, but the cue is removed before the eye reaches the target. Visual cues can either be *diegetic*, which means they are a natural part of the environment (e.g., a character pointing or a flickering light,) or *non-diegetic* (e.g., a bright green arrow or marker imposed on the scene). Visual cues can be *static*, e.g., using color, or *dynamic*, e.g., using motion to guide attention.

- *Auditory* refers to any sound that influences the viewer's eye to orient in a particular direction. This includes spoken dialogue from a narrator or even a character in the scene, and spatial audio within the 3D environment.

The final step in the design of a safety training system is defining the parameters of the intelligence in the system. This design variable may be parametrized by the following features:

Viewing Media	Design Features								
	Sensory fidelity					Interactivity		Attention directability	
	Pictorial	Animation	Aural	Haptic	Olfactory	User DOF	Responsiveness	Visual	Auditory
Omnidirectional Image	High	None	None	None	None	Rotational	None	Yes	Yes
Omnidirectional Video	High	High	High	None	None	Rotational	None	Yes	Yes
Omnidirectional Cinemagraph	High	High	High	None	None	Rotational	Low	Yes	Yes
Omnidirectional Video Texture	High	High	High	None	None	Rotational	Low	Yes	Yes
3D Scene (Synthetic)	Med.	High	High	None	None	Rotational Translational Interactive	High	Yes	Yes
3D Light Field Captured Scene	High	High	None	None	None	Rotational Translational	None	Yes	Yes

Table 1: We categorize various VR training media based on the proposed design features. Omnidirectional cinemagraphs, shaded in gray, were selected for the case study reported in Section 5.

Input data refers to the features of the data collected to assess, scaffold, and motivate the trainee.

Time scale refers to how often the data is logged. Discrete data include responses to questions at predetermined points in the training module. Continuous data include responses that are logged at a much higher temporal resolution than human reaction time, for example, the pupil diameter of the trainee is recorded at 120Hz.

Type of data refers to whether the data are nominal, ordinal, or numeric (ratio).

Stealth of collection refers to whether the data are collected through explicit prompts, for example as the number of hazards identified in a scene, or implicitly, such as the average number of fixations needed to locate a hazard in a scene.

Type of feedback refers to the affordances provided to the AI to influence the trainee during a training session. Explicit feedback includes displaying a score, advancing to the next level, or providing verbal positive feedback, which may impact the level of immersion and presence the trainee feels in the simulated training environment. Implicit feedback includes subtle cues such as in-obtrusive gaze guidance to direct the trainee to a hazard they need to identify.

Level of personalization refers to the degree to which the input data influence the algorithm that controls the training environment. For example, a simple state machine that cycles through a preset series of training setups is low in personalization. On the other hand, an algorithm that combines gaze and facial expression to infer areas of weakness for the trainee, and then presents guided feedback, displays a high level of personalization.

Though the design framework discusses attention directability as a feature, the framework itself is agnostic to the *effectiveness* of attention guidance mechanisms. In the next sections we discuss previously proposed attention guidance mechanisms and make the case for cinemagraphs for our choice of training environment.

3 Background on Attention Guidance

Overt methods to direct and guide a viewer's attention include adding markers such as circles or arrows to the scene that explicitly tell the viewer where to look [33]. These markers are non-diegetic as they are superimposed on the scene, however they are effective in guiding attention at the cost of presence and immersion within the environment [22]. Beyond markers, Smith and McNamara have implemented color-based guidance effects that depend on the head orientation of the viewer [27]. The virtual scene is untouched when the viewer is oriented towards the designated area of interest, and if the viewer looks away from this point a yellow-green tint is applied to the scene with increasing intensity, guiding the viewer back towards the intended region. An alternative technique that does not rely on eye or head tracking is to use the scene content itself for guidance, i.e., diegetic cues. Previous work has used cues such as a flickering light source, the main character pointing in particular direction, or spatial audio [22,23]. Diegetic cues must be implemented when generating stimuli, and do not use gaze or head tracking to verify that the user has followed the cue.

Researchers have also proposed subtle gaze guidance, a suite of techniques to subtly flicker the luminance and contrast of a target location to nudge the human visual system into shifting attention to the target location [8,28,30]. These techniques have been shown to be effective at guiding gaze to a specific location very quickly, but they only work well in cases where eye tracking is fast (120-250Hz) and accurate enough that the guiding flicker has ended by the time the user's eye reaches the target. This approach is prone to errors, especially for 60 Hz eye trackers such as ones built into AR/VR headsets.

Past work has shown that motion is reliable cue for guiding attention [6,14,18]. The way in which a motion cue attracts attention depends on other motion(s) within the stimulus [20]. For example, detecting a moving target surrounded by static distractors is far easier than detecting a static target among moving distractors [5]. Furthermore, when both targets and distractors are moving it is easier to detect a fast target among slow distractors than a slow target among fast distractors [10]. For this reason we consider a training environment where we have control over the motion within the scene to implement responsive motion cues, which are cinemagraphs and video textures [17]. Both forms of media utilize pre-recorded video to 'loop' the content endlessly. However, the primary difference between them is that cinemagraphs provide the ability to designate static and dynamic regions, whereas video textures animate the entire video frame. Because it is easier to attract attention to moving regions when the surrounding area is static or moving slowly, cinemagraphs provide more effective motion cues.

4 Omnidirectional Cinemagraphs

Cinemagraphs are created by selecting a region of a video and identifying a start and end frame that would visually loop forever when played continuously. Cinemagraphs may contain multiple looping regions, while the rest of the video remains static. As can be seen from Table 1, omnidirectional cinemagraphs are unique in that they provide high pictorial fidelity while also providing responsiveness. In our case study we focus on the following propositions:

- Cinemagraphs are as or more effective than marker-based gaze guidance.

- Cinemagraphs lead to greater presence/immersion than marker-based gaze guidance.

- Cinemagraphs lead to "desired" cognitive processing of the scene, where "desired" could mean "more safety hazards are identified" or some other learning outcome compared to omnidirectional video or images.

Generation of cinemagraphs from an input omnidirectional video presents a unique set of technical challenges. Established graphics research has targeted traditional 2D video content, providing solutions that allow a user to select a looping region by scribbling over the input video [2], optimizing loops that appear continuous to the human eye [16], and leveraging semantic segmentation to identify which regions should be animated or left static [19]. Omnidirectional content is typically stored in the equirectangular format, presenting distortions and discontinuities that prevent traditional approaches to be applied off the shelf. Liu et al. present a method for adapting video textures to omnidirectional content by independently finding optimal loops for overlapping vertical segments of the input video, and optimizing the global result using a graph cut [17]. Video textures vary from cinemagraphs, in that cinemagraphs contain static regions and objects within the scene.

Humans can manually create convincing cinemagraphs with multiple regions within 30 minutes to an hour, but automated algorithms [3, 15, 31] will perform better at picking the exact starting and ending frames for each region. Instead, the optimal solution would depend on algorithms that generate multiple candidate cinemagraphs and allow the user to select or modify the most applicable candidate.

Section 5 describes the cinemagraph generation, study protocol, and results. Section 6 discusses these results as a proof of concept, as well as the benefits and potential pitfalls of VR cinemagraphs for safety training. Lastly, Section 7 summarizes our results relative to the proposed framework and motivates future work on safety training.

5 Case Study

Cinemagraphs are an artistic visual somewhere between a picture and a movie: only one or more chosen objects move while the rest of the picture is static. The dynamic regions can loop simultaneously and independently from each other. These regions could for example be played sequentially to tell a story. While cinemagraphs are very popular for web advertisements and artistic photography, our novel contribution lies in using them for safety hazard detection. We created a cinemagraph of a *Kitchen* scene where an empty kitchen was staged with potential hazards: a steaming pot of water left on an open stove, a dishcloth placed over one of the stove burners, a paper towel placed underneath a lit candle, a rotating fan blowing towards the candle and paper towel rack, and a sink with the water left running. Figure 2 contains the static equirectangular frame used for our study with Areas of Interest (AOIs) labelled.

5.1 Research Questions

The aim of our case study is to answer the following research questions:

- RQ_1: Can 360 degree cinemagraphs lead to greater presence/immersion than static images?

- RQ_2: Do 360 degree cinemagraphs create more positive attitudes than static images?

- RQ_3: Do 360 degree cinemagraphs draw gaze to task relevant objects longer than static images?

5.2 Stimuli Generation Method

In order to create omnidirectional cinemagraphs, we started by recording omnidirectional video using an off-the-shelf VIRB 360 camera. We uploaded the footage to the VIRB editing application at native resolution and quality, and then converted them into an MP4 file at 4K resolution for further editing. We input our omnidirectional video in equirectangular format into the Microsoft Cliplets tool [11]. The Cliplets application allows users to modify 10 seconds of footage and select areas within the scene to set as static or dynamic. We then exported two versions of the scene: the cinemagraph we designed, and a ten second video of a static frame from the same scene. The ten seconds obtained from Cliplets were then looped once to create a 20 second video.

5.3 User Study Protocol

Ten participants in a between-subjects design were recruited under an IRB approved protocol. The cinemagraph (Condition Dynamic) and the static frame video (Condition Static) were displayed using Unity on an Oculus DK2 head mounted display (HMD). An SMI eye tracker integrated with the Oculus DK2 was used to record eye movements at 60 Hz.

After informed consent, the participant sat in a swivel chair and the eye tracker was calibrated to within 2° visual angle. Each participant was tasked with identifying fire hazards in the scene (total of five) and was given 20 seconds to complete this task. After this, the participant was asked to fill out survey items relating to presence, immersion, cognitive absorption, and attitude/preference. Finally, demographics were collected and the participant was debriefed.

The survey items were aggregated by summing to yield one numerical value representing each of the following concepts:

1. Spatial Presence [32]: A measure of spatially feeling like you are present in the virtual environment.

2. Transportability [4, 7]: A measure of being transferred from the current location to a different location or perspective.

3. Telepresence [13]: A more general measure of feeling as if you have been teleported to a different location.

4. Cognitive Absorption [1]: A measure of how engaged the user is within the virtual experience.

5. Attitude [12]: A measure of how positively you interpreted the experience.

Responses ranged from 1-7, with 1 being strongly disagree, and 7 being strongly agree. Questions that imply a negative effect are inverted, so that summing the scores represents a greater agreement with the concept. Eye tracking data collected during scene viewing was parsed into fixations using I-DT [26]. The fixation data was analyzed using Areas of interest (AOIs) around each potential hazard, as marked in Figure 2. Net Dwell Time, Time to First Fixation, and Number of Fixations was computed for each AOI, along with the Average Fixation Duration. The Time To First Fixation metric indicates how long it took for the participant to find that AOI. Smaller values are better for the purpose of safety training. Increased Number of Fixations and Net Dwell Time would also indicate that the AOI held their attention.

5.4 Results

Hazard Identification Participants seemed to have identified more hazards when viewing the cinemagraph compared to the static scene. Participants who saw the cinemagraph identified three, four, three, five, and two hazards respectively, out of the total of five. Participants who saw the static scene identified values of two, five, and two. The other two participants did not fill out that portion of the response form.

Figure 2: Static *Kitchen* scene with labelled hazards.

Condition	SPSL		Transportability		Telepresence		Cognitive Absorption		Attitudes	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Static	37	12	38	12	23	7	53	14	55	13
Dynamic	42	12	44	9	30	11	76	11	71	11

Table 2: Condition Dynamic (cinemagraph) was found to be higher on all measures of immersion and presence compared to Condition Static (a static picture). Participants had a generally more positive attitude to Condition Dynamic with respect to the appeal and likeability of the virtual experience.

RQ_1 **Presence and Immersion** Cinemagraphs (Condition Dynamic) received generally higher scores, as shown in Table 2. Participants reported a slightly higher level of Spatial Presence, Transportability, and Telepresence for omnidirectional cinemagraphs compared to a static omnidirectional image.

RQ_2 **Engagement** Participants generally felt more positively about the experience and reported a higher level of Cognitive Absorption (how engaged they felt in the scene) and Attitudes (how much they liked the scene).

RQ_3 **Gaze guidance** Viewers started viewing at the center of the panorama, which is an open bathroom door next to the fan (see Figure 2). Both the candle with the paper towel and the faucet with running water are positioned on the user's left from this starting point, while the other areas of interest (AOIs) are positioned to the right. For participants viewing the cinemagraphs (Condition Dynamic), the Time to First Fixation on the candle and running water faucet is shorter. Shorter times imply that the participants' gaze reached those hazards quicker. Time to First Fixation increased for the other hazards likely as a result of this. Based on the trends in the Net Dwell Time and Number of Fixations, we believe that the motion created by the cinemagraph changed how participants attended to the scene, though this data is rather preliminary to determine conclusively what these changes are, and how desirable they are for training purposes. Table 3 provides a summary of gaze related metrics. We also checked the average fixation duration and found no difference between the two conditions.

6 Discussion

With only five participants in each condition, the values in Table 2 represent preliminary trends rather than conclusive findings. They suggest that omnidirectional cinemagraphs could be an effective medium for workplace safety training as they provide a sweet spot between a static omnidirectional image and omnidirectional video. The ability to direct attention while maintaining a high level of immersion suggests applications in a variety of safety and training domains. Cinemagraphs are able to guide attention and allow the content to accommodate the pace of the user, something that is crucial for both training and entertainment [17].

For example, consider a scenario where students in construction management are being taught to identify hazards common at a job site. Omnidirectional images and videos can be used to transport students to a photo-realistic representation of the job site, however such images will typically lack the dynamics of the real world while the video would contain many distractors that pull attention away from the intended hazards. Also, some hazards can only be identified when motion is present. For example, an excavator that is backing up is a hazard, but in a static image, it would appear as if it was parked. The benefit of the cinemagraph technique is to isolate specific hazards while freezing the rest of the scene. This allows the user to focus on the intended regions without being distracted by extraneous visual clutter. Though our case study did not do this, a future version might even present each hazard region sequentially as part of the training process. Eventually the number of hazards and distractors being animated can be increased to simulate the full experience of a

Figure 3: One frame from the dynamic *Kitchen* scene with labelled hazards.

AOI	Net Dwell Time (s)				Number of Fixations				Time to First Fixation (s)			
	S		D		S		D		S		D	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Hazard 1: Faucet	1.38	0.59	1.12	0.34	3.75	1.26	3.50	1.22	6.27	6.77	3.19	2.97
Hazard 2: Candle	0.66	0.95	0.96	0.67	1.00	0.82	1.83	1.17	12.5	6.18	4.51	2.66
Hazard 3: Fan	0.28	0.35	0.53	0.38	1.50	1.29	1.67	1.21	2.16	2.26	8.09	8.54
Hazard 4: Pot	1.61	1.45	1.12	0.43	5.25	4.35	3.50	1.38	3.69	1.77	7.37	5.39
Hazard 5: Burner	0.72	1.05	0.26	0.40	1.75	1.71	0.83	1.17	7.93	5.75	8.38	5.85

Table 3: Time to First Fixation was shorter on the first two hazards for Condition Dynamic (D) compared to Condition Static (S), though it was longer for the others. The faucet was not in view at the start of the cinemagraph, and the candle was near the periphery. The fan was directly in front of the viewer.

busy construction site. Other potential application domains include education (e.g., training pre-service teachers to manage classrooms, training special education providers), and medical care (e.g., training nurses for triage or other emergency situations). Robust evaluation across safety and training application domains is required for any training environment as the design framework we have introduced helps organize the space of safety training, but does not comment on the impact of the design choices on the learning outcomes in any given context.

The outcomes of safety training systems, while traditionally assessed through responses such as hazard identification rates, warrant further research as well. For example, attention-based metrics such as Time to First Fixation on hazards, and time duration between the first fixation on a hazard and a key press are metrics that quantify how effectively the trainee proceeds from See, to Recognize, to Act.

While we have considered safety training with respect to hazard identification, much of the design framework also applies to criminology. Criminology is the study of why crime occurs and environmental criminology specifically studies environments that increase or decrease the chances of crimes such as assault and shoplifting. Hayes and colleagues have formulated the See and Recognize concepts to theorize about the factors that influence the decisions of shoplifters [9,29]. For example, if a shoplifter will 'See' an anti-theft device and 'Recognize' that it increases the risk that they will be caught, they might avoid shoplifting.

Cinemagraphs pose both practical and technical challenges, chief among them being that video content itself must be "loopable" to create seamless transitions between the starting and ending frames. For example, a bird flying in a straight line from one point to another will always have a discontinuity between the starting and ending position. One potential solution would be to mirror the animation such that the bird moves back and forth, risking a loss of presence and realism when the viewer sees a bird flying backwards. Ultimately the content creator can alleviate these issues at capture time by either staging events, or capturing enough video to ensure the intended region can be looped. Cinematography techniques for guiding viewer attention and interaction within omnidirectional content are still being explored [21, 23–25], however the development of cinemagraphs and video textures [17] enables authors to escape the linearity of existing approaches.

Open technical challenges in omnidirectional cinemagraph generation include automated methods to improve visual quality and aesthetics, authoring tools that provide intuitive control to the artists for selecting regions or navigating around the distortions created by the equirectangular format, and software systems to render cinemagraphs with the intended spatial and temporal properties.

7 Conclusion

Cinemagraphs are a type of media that fall between an image and video that provide looping behavior that can be used to implement training in VR environments. They provide control over motion as a guidance technique and contain photo-realistic content. Our case study explored a proof of concept application in fire safety awareness within a kitchen environment. Results suggested that cinemagraphs impacted users gaze behavior with respect to hazards and improved overall attitude of the experience, however benefits over traditional omnidirectional videos have yet to be explored. Cinemagraphs allow authors to create non-linear VR experiences that retain immersion and tap into the full potential of VR simulation for safety training. In the future we plan to further explore new application domains, such as medical and construction training, while also developing metrics and methods to evaluate the effectiveness of cinemagraphs and other gaze guidance techniques in this context.

Acknowledgments

The authors wish to thank Dr. Read Hayes for discussions that conceived the ***See, Recognize, Act*** processes, and CJ Taylor for his contributions in conducting the user study. Authors acknowledge funding from the National Science Foundation (Award #IIS-1566481), and the National Science Foundation Graduate Research Fellowship (Awards DGE-1315138 and DGE-1842473).

References

[1] R. Agarwal and E. Karahanna. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, pp. 665–694, 2000.

[2] J. Bai, A. Agarwala, M. Agrawala, and R. Ramamoorthi. Selectively de-animating video. *ACM Trans. Graph.*, 31(4):66–1, 2012.

[3] J. Bai, A. Agarwala, M. Agrawala, and R. Ramamoorthi. Automatic cinemagraph portraits. In *Proceedings of the Eurographics Symposium on Rendering*, pp. 17–25. Eurographics Association, 2013.

[4] S. Dal Cin, M. P. Zanna, and G. T. Fong. Narrative persuasion and overcoming resistance. *Resistance and persuasion*, 2:175–191, 2004.

[5] M. Dick, S. Ullman, and D. Sagi. Parallel and serial processes in motion detection. *Science*, 237(4813):400–402, 1987.

[6] M. Girelli and S. J. Luck. Are the same attentional mechanisms used to detect visual search targets defined by color, orientation, and motion? *Journal of Cognitive Neuroscience*, 9(2):238–253, 1997.

[7] M. C. Green and T. C. Brock. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701, 2000.

[8] S. Grogorick, M. Stengel, E. Eisemann, and M. Magnor. Subtle gaze guidance for immersive environments. In *Proceedings of the ACM Symposium on Applied Perception*, p. 4. ACM, 2017.

[9] R. Hayes, S. Strome, T. Johns, M. Scicchitano, and D. Downs. Testing the effectiveness of anti-theft wraps across product types in retail environments: a randomized controlled trial. *Journal of Experimental Criminology*, 15(4):703–718, 2019.

[10] R. B. Ivry and A. Cohen. Asymmetry in visual search for targets defined by differences in movement speed. *Journal of Experimental Psychology: Human Perception and Performance*, 18(4):1045, 1992.

[11] N. Joshi, S. Mehta, S. Drucker, E. Stollnitz, H. Hoppe, M. Uyttendaele, and M. Cohen. Cliplets: juxtaposing still and dynamic imagery. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pp. 251–260. ACM, 2012.

[12] S. Kalyanaraman and S. S. Sundar. The psychological appeal of personalized content in web portals: does customization affect attitudes and behavior? *Journal of Communication*, 56(1):110–132, 2006.

[13] T. Kim and F. Biocca. Telepresence via television: Two dimensions of telepresence may have different connections to memory and persuasion. *Journal of computer-mediated communication*, 3(2):JCMC325, 1997.

[14] H. Li, G. Chen, G. Li, and Y. Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7274–7283, 2019.

[15] J. Liao, M. Finch, and H. Hoppe. Fast computation of seamless video loops. *ACM Transactions on Graphics (TOG)*, 34(6):197, 2015.

[16] Z. Liao, N. Joshi, and H. Hoppe. Automated video looping with progressive dynamism. *ACM Transactions on Graphics (TOG)*, 32(4):77, 2013.

[17] S. J. Liu, M. Agrawala, S. DiVerdi, and A. Hertzmann. View-dependent video textures for 360° video. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19. ACM, 2019.

[18] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation*, 3(1):5–24, 2011.

[19] T.-H. Oh, K. Joo, N. Joshi, B. Wang, I. S. Kweon, and S. B. Kang. Personalized cinemagraphs using semantic understanding and collaborative learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5170–5179. IEEE, 2017.

[20] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision research*, 39(19):3157–3163, 1999.

[21] S. Rothe, H. Brunner, D. Buschek, and H. Hußmann. Spaceline: A way of interaction in cinematic virtual reality. In *Proceedings of the Symposium on Spatial User Interaction*, pp. 179–179, 2018.

[22] S. Rothe, D. Buschek, and H. Hußmann. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction*, 3(1):19, 2019.

[23] S. Rothe, H. Hußmann, and M. Allary. Diegetic cues for guiding the viewer in cinematic virtual reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2017.

[24] S. Rothe, B. Kegeles, M. Allary, and H. Hußmann. The impact of camera height in cinematic virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2018.

[25] S. Rothe, V. Sarakiotis, and H. Hussmann. Where to place the camera. In *25th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2019.

[26] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pp. 71–78. ACM, 2000.

[27] M. Smith and A. McNamara. Gaze direction in a virtual environment via a dynamic full-image color effect. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1–2. IEEE, 2018.

[28] S. Sridharan, J. Pieszala, and R. Bailey. Depth-based subtle gaze guidance in virtual reality environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pp. 132–132. ACM, 2015.

[29] S. K. Strome, R. Hayes, and K. Grottini. Situational determinants of enhanced public view monitor (ePVM) noticeability in retail environments: a randomized controlled trial. *Security Journal*, 31(3):749–763, 2018.

[30] Q. Sun, A. Patney, L.-Y. Wei, O. Shapira, J. Lu, P. Asente, S. Zhu, M. McGuire, D. Luebke, and A. Kaufman. Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[31] J. Tompkin, F. Pece, K. Subr, and J. Kautz. Towards moment imagery: Automatic cinemagraphs. In *Visual Media Production (CVMP), 2011 Conference for*, pp. 87–93. IEEE, 2011.

[32] P. Vorderer, W. Wirth, F. R. Gouveia, F. Biocca, T. Saari, L. Jäncke, S. Böcking, H. Schramm, A. Gysbers, T. Hartmann, et al. Mec spatial presence questionnaire. *Report to the Europoean Community, Project Presence: MET (IST-2001037661)*, 2004.

[33] A. Yoshimura, A. Khokhar, and C. W. Borst. Visual cues to restore student attention based on eye gaze drift, and application to an offshore training system. In *Symposium on Spatial User Interaction*, pp. 1–2, 2019.