Integrating Multiple Views with Virtual Mirrors to Facilitate Scene Understanding

CARMEN E. AU and JAMES J. CLARK, McGill University

In this article, an image integration technique called Virtual Mirroring (VM) is evaluated. VM is a technique that combines multiple 2D views of a 3D scene into a single composite image by overlaying views onto virtual mirrors. Given multiple views of a scene, one view is augmented with the remaining views by placing virtual mirrors on the first view and overlaying onto them the corresponding remaining views. Unlike a standard array presentation, where 2D views are not integrated and simply placed adjacent to one another, the VM presentation preserves the relative location, orientation, and scale between views. As such, it is our contention that humans will fare better at performing certain visual tasks, such as scene identification, when viewing a 3D scene via a VM presentation than when viewing an array presentation. We performed an experiment on 12 participants, where participants were required to identify 96 scenes both with a VM and an array presentation and we compared their % correctness and response times. Moreover, we studied the effects of adding an auditory attentional load on performance. We found that regardless of load, participants were able to identify scenes using VM presentation with greater accuracy and at greater speeds.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI); Screen design; User-centered design*; J.7 [**Computer Applications**]: Computers in Other Systems—*Consumer products*

General Terms: Design, Experimentation, Human Factors, Performance

Additional Key Words and Phrases: Image integration, psychophysics, visual task, scene identification, virtual mirror, attentional load

ACM Reference Format:

Au, C. E. and Clark, J. J. 2011. Integrating multiple views with virtual mirrors to facilitate scene understanding. ACM Trans. Appl. Percept. 8, 4, Article 28 (November 2011), 14 pages.
DOI = 10.1145/2043603.2043610 http://doi.acm.org/10.1145/2043603.2043610

1. INTRODUCTION

Advances in technology have given rise to a world in which cameras are ubiquitous. With the pervasiveness of camera phones and digital cameras, it is not unusual for several cameras to be present in most crowds. In addition, security cameras, traffic cameras, and publicly accessible web cameras are omnipresent. These cameras and the excess of images (and videos) they capture has led to an increased interest in the development of applications that integrate these images into meaningful displays, which can convey greater amounts of information to the observers. As such, there have been many published works that describe techniques that seek to create such multiviewed displays [Beis

This work was funded by the EC under grant 043157, project SynTex.
Author's address: C. E. Au, 3480 Sherbrooke W., Montreal, Quebec, Canada, H3A 2A7; email: au@cim.mcgill.ca.

DOI 10.1145/2043603.2043610 http://doi.acm.org/10.1145/2043603.2043610

and Lowe 1997; Irani and Anandan 2000; Capel and Zisserman 1998; Debevec et al. 1998; Kanazawa and Kanatani 2002; Rankov et al. 2005; Brown and Lowe 2007]. In this article, we investigate the advantages, should there be any, of one such technique described as Virtual Mirroring (VM) [Au and Clark 2008]. VM is a technique for integrating images with widely differing viewpoints. The name is derived from the use of virtual mirrors to combine and integrate these views into a single image. We seek to determine whether there are any benefits to a human observer, such as better scene understanding or spatial awareness, when viewing images that have been integrated over viewing images disparately.

Previous efforts for image integration have successfully created mosaics with multiple images; however, these methods, whether direct methods [Beis and Lowe 1997; Irani and Anandan 2000], or feature-based methods [Capel and Zisserman 1998; Debevec et al. 1998; Kanazawa and Kanatani 2002; Rankov et al. 2005], require that input images have overlapping views. Brown and Lowe's image-stitching algorithm, for example, successfully uses Lowe's Scale Invariant Feature Transform (SIFT) [Lowe 2004] to create beautiful panoramic views of a scene [Brown and Lowe 2007]. Due to the fact that these approaches rely on overlapping views, the situation where images are acquired from cameras with opposing viewpoints of the same scene is not accounted for. In many real-world situations, surveillance, for example, cameras are intentionally placed to maximize coverage of a scene. As such, a technique more akin to the VM technique is required, where views are integrated by placing virtual mirrors on the scene and overlaying onto these mirrors the image content of the cameras. In Figure 1, two images, (a) and (b), are combined by overlaying (b) onto a virtual mirror and shown in (c). Thus, rather than having a situation where the views are presented to observers in multiple adjacent screens, a setup common in many multicamera surveillance systems, views can be integrated using VM, where the spatial relationships between the views are maintained. In this article we investigate whether the latter (VM) presentation has any benefits to the observer over the standard (array) presentation.

We begin by acknowledging two implementation issues that are relevant to the VM presentation, but not to the array presentation, which are the issues of calibration and of occlusion. While the VM presentation requires calibration of the camera's intrinsic (focal length) and extrinsic (position and orientation) parameters, for which there are many techniques [Zhang 2000; Kannala and Brandt 2006; Colombo et al. 2006; Hartley and Kang 2007; Zheng and Liu 2008; Tardif et al. 2009], the array presentation does not. However, the array presentation does use some positional information to determine how to place the images from left-to-right in the array. Moreover, with the VM presentation, the issue of occlusion arises. There are two types of occlusion that need to be addressed. The first type involves the mirror occluding the foreground inadvertently. As the mirror is overlaid onto an image, the algorithm must be wary of overlaying the mirror over foreground, thereby creating a view where the foreground is behind the mirror, and yet the mirror continues to reflect the foreground. The cited article on the VM technique addresses this issue by applying a background subtraction model for separating foreground and background [Au and Clark 2008]. Once the two are separated, the mirror can be correctly placed behind the foreground. The second type of occlusion occurs when the mirror occludes portions of the image that may be pertinent. We might argue that all portions of the image are pertinent, and this may be in some cases, however, it is often the case that the mirror can be placed against a wall or other less important portions of the image. The resulting trade-off in image real estate will likely be negligible to user scene understanding. Despite these two implementation issues, which may render the VM technique more computationally expensive, in this article, we seek only to investigate whether observers will be able to perform various spatial tasks, such as scene identification, with greater accuracy and speed using the VM presentation over an array presentation. The two discussed issues should not affect the experimental process. The goal is to compare the performance of observers on

Fig. 1. Example "paparazzi" images with virtual mirroring: (a) image from principal camera; (b) image from secondary camera; and (c) image from principal camera with image from secondary camera overlaid onto a virtual mirror. Image created using an image-editing program.

those spatial tasks, where scene assessments must be made relatively quickly and using only the visual cues available in the two presentation types.

One question that must be addressed is whether humans are able to properly understand objects and scenes when viewed through a mirror. While seeing the world through reflective surfaces may be commonplace, numerous studies have shown that humans tend to have certain misconceptions and misperceptions regarding mirror views and what they reason about the mirror reflections [Bertamini et al. 2003; Bertamini and Park 2005; Lawson and Bertamini 2006; Croucher et al. 2002]. There are

four types of errors common to human observers; the errors all pertain to what they believe should be visible in a mirror based on their vantage point. (1) First, Bertamini and colleagues showed that humans tend to overestimate what is visible in a mirror; people believe that the size of the projection of their face onto the surface of the mirror should be the same as the actual size of their face, when in reality it is closer to half the size. (2) Another common error is the belief that their reflection should reduce in size as observers step away from the mirror; in reality, the size of the reflection does not depend on distance from the mirror [Bertamini and Park 2005]. (3) The "Venus effect," aptly named after a painting of Venus apparently looking at herself in the mirror, is the belief that if a human observer sees both a target person and the reflection of that target person in a mirror, the observer believes that what he or she sees in the mirror is the same as what the target person sees. In fact, we would not see the same reflection in the mirror as Venus would. (4) Finally, humans tend to expect to see their reflection as they approach a mirror from the side before they reach the near edge of the mirror. In spite of these common misperceptions that humans have regarding mirrors and their reflections, Lawson and Bertamini were able to show that humans are still able to derive spatial information from reflective surfaces under certain circumstances [Lawson and Bertamini 2006]. More specifically, when certain spherical objects were placed in front of a planar mirror, observers were able to use the information provided by the mirror images to ascertain the size and distance of the objects being reflected with a certain degree of accuracy. What was significant about these latter findings is that they show that humans are indeed able to combine what they observe from a mirror view and what they observe in the world directly to derive deeper spatial understanding of the scene. Our setup differs from Lawson and Bertamini's in that in our case, virtual rather than actual mirrors are employed. With the current VM technique, there is a loss of realism in the virtual mirrors. This loss may affect the human observers' ability to derive useful or accurate information from the virtual mirrors. Moreover, the misperceptions humans have regarding mirror reflections may in fact hinder results or make it more difficult for the observers to perform our chosen spatial tasks. Our study will therefore confirm or invalidate our hypothesis that humans can derive more spatial information from the VM presentation.

The rationale underlying our hypothesis is that we expect that by presenting an integrated rather than non-integrated view of a scene to humans, fewer mental transformations will be required for a given spatial task. We make this assertion based on the human object and scene recognition literature. The literature on how objects are represented in our minds is divided into two main models: a view-invariant model and a view-dependent model. The view-invariant model suggests that the visual system creates a viewpoint-invariant representation of objects [Marr and Nishihara 1978; Biederman 1987; Biederman and Gerhardstein 1993; 1995; Tarr et al. 1998; Hayward and Williams 2000]. On the other hand, other work described in the literature argues for a view-dependent model, which holds that 3D objects are likely encoded by the human visual system as multiple viewpoint-specific representations that are largely two-dimensional (2D) [Tarr and Pinker 1989; Poggio and Edelman 1990; Edelman and Weinshall 1991; Ullman and Basri 1991; Vetter et al. 1994; Bülthoff et al. 1995; Wallis and Bülthoff 1999]. While strong arguments remain for both models, Wallis and Bülthoff convincingly showed that at least with objects that are unfamiliar to humans, object recognition is view-dependent [Wallis and Bülthoff 1999]. This finding is relevant to our study, as the participants are tested on unfamiliar scenes. The human scene recognition literature is derived from the object recognition literature, where subjects are presented different viewpoints of a given scene rather than an object [Castelhano et al. 2009; Christou and Bülthoff 1999; Garsoffky et al. 2002; Hock and Schmelzkopf 1980; Johnson 2002]. Based on this research, Castelhano et al. were able to argue that scene recognition was also highly viewpoint dependent [Castelhano et al. 2009]. It follows that 3D scenes are also represented in memory by several 2D views of the scene. These 2D views are either stored separately in memory or linked in some

way. Castelhano and her colleagues suggest a model for the human scene recognition process where the views are stored independently, and it is only at retrieval for decision making that the views are integrated. They do, however, leave room for the possibility that the views are stored in some linked fashion rather than independently. If it is the case that images are stored independently, then we surmise that by presenting humans with an integrated view of the scenes, as in a VM presentation, we avoid the possibility that humans are unable to relate the views at retrieval. If the images are stored in some linked fashion, there still remains the possibility that humans integrate the views incorrectly before storage, thereby storing an incorrect representation of the scene. In other words, regardless of the correct model, it is our hypothesis that with VM presentations, since the virtual mirrors are placed in the location, or very close to the location of their respective cameras, the spatial relationships between the views are preserved in the integration. Therefore, humans are less likely to make mistakes about how the views relate spatially to one another. Finally, it is our contention that by presenting an integrated view to humans, the time it takes to view a test scene and make a decision about it should be less than if they were presented with a display of images with no spatial integration.

2. EXPERIMENT

In this experiment we seek to evaluate our hypothesis that humans can recognize scenes with greater accuracy, meaning with higher % correct proportions, and at greater speeds. Moreover, we seek to determine whether an attentional load disturbs or improves performance. We compared participants' ability to correctly identify scenes by observing multiple views of the scene. The goal of this experiment is to determine whether observers have better scene understanding when viewing the scene using VM presentation or an array presentation. Our hypothesis is that by presenting the views in an integrated fashion that maintains the spatial relationships between the views, viewers will be required to make fewer mental transformations between the views, and will therefore have a simpler time of understanding the scenes.

2.1 Method

2.1.1 *Participants.* The participants for the experiments were all McGill University students between the ages of 20 and 30. The participants were recruited using a classified advertisement in the McGill University website. In total, there were 12 participants, of whom 9 were male and 3 were female. A small monetary incentive (CDN$10) was provided to participants who completed the study, regardless of their performance. The participants were naive to the purpose of the experiment, and had no *a priori* knowledge about the setup. All participants had normal or corrected to normal vision. The Research Ethics Board of McGill University approved the experimental protocol, and participants gave their informed consent before participating.

2.1.2 *Displays and Apparatus.* The images were acquired by using a camera stage fabricated specifically for the experiment as shown in Figure 2. The stage comprises of three camera mounts placed at the vertices of an equilateral triangle and oriented in such a way that the intersection of their optical axes meet at the center of that triangle. To acquire the images, we used the camera on the Nokia 5500 Sport phone with a 2-megapixel camera, although the type of camera used is not relevant to the experiment. To complete the stage, a black triangular screen was also fabricated. The screen served to black out the surrounding environment in order to ensure that only objects placed on the scene were imaged.

For the testing phase, a Dell Inspiron 640m 2.0GHz laptop was used with a 14.1" screen set at a 1440×900 pixel resolution. Subjects were presented with questions, which we will describe in greater

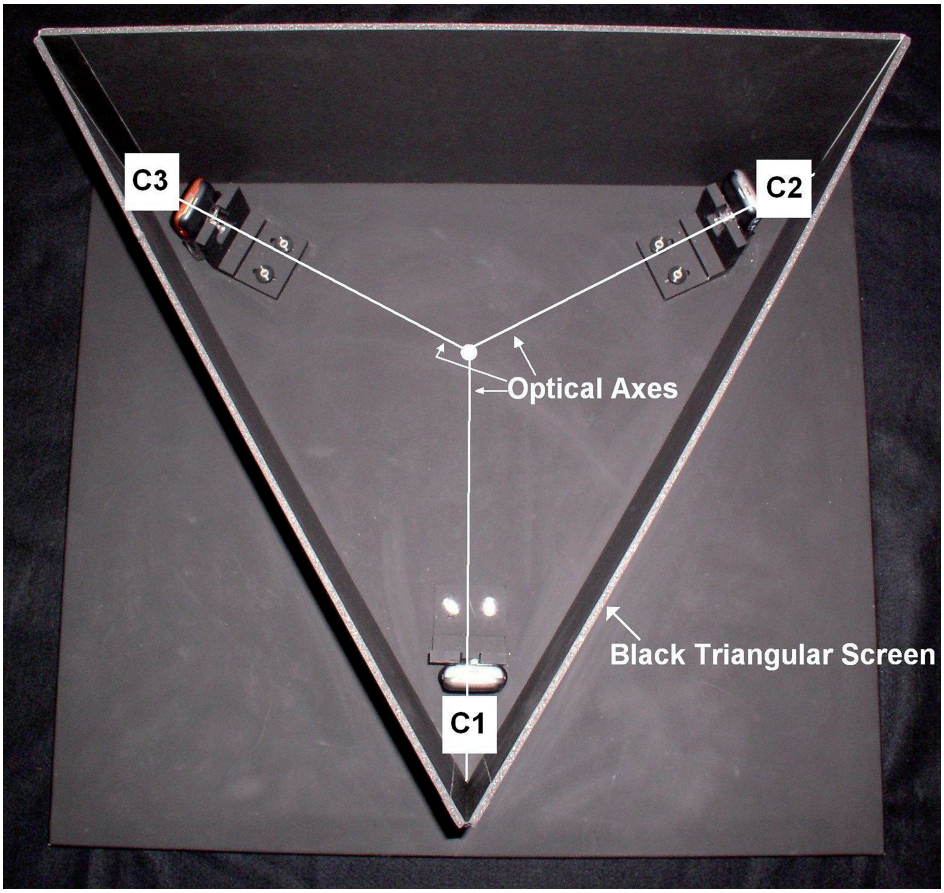

Fig. 2. Diagram of experimental stage where C1 is Camera 1, C2 is Camera 2, and C3 is Camera 3.

detail in Section 2.1.4. Answers were selected by either pressing the left or right arrow keys of the laptop keyboard.

2.1.3 *Methods.* Using the camera stage, 96 scenes were set up and imaged. The scenes were divided into two main scene types: relational and reconstructional. As the names suggest, the two scene types were conceived to test for specific aspects of the visual task. The relational scenes were devised in such a way that each of the three cameras would only capture one specific portion of the scene and there was no overlap between the three views. The purpose of this type of scene was to test whether subjects could better understand how the three nonoverlapping views relate when viewing a VM presentation over an array presentation of a given scene. The motivation for this type of scene is the surveillance example described in Section 1. A casino security guard may see from one camera's point of view that there is a person, Person A, making signals, however, the person to whom Person A is signalling may only be visible from another camera's viewpoint. It is thus necessary for the guard to relate, with a fair amount of speed, the two separate views in order to ascertain if any event that warrants response is taking place. The reconstructional scenes were devised to test subject ability to reconstruct a 3D scene from multiple views and to recognize it. The motivation behind this type of scene is to determine whether a subject can better reconstruct a scene in their minds using the VM presentation. Both of

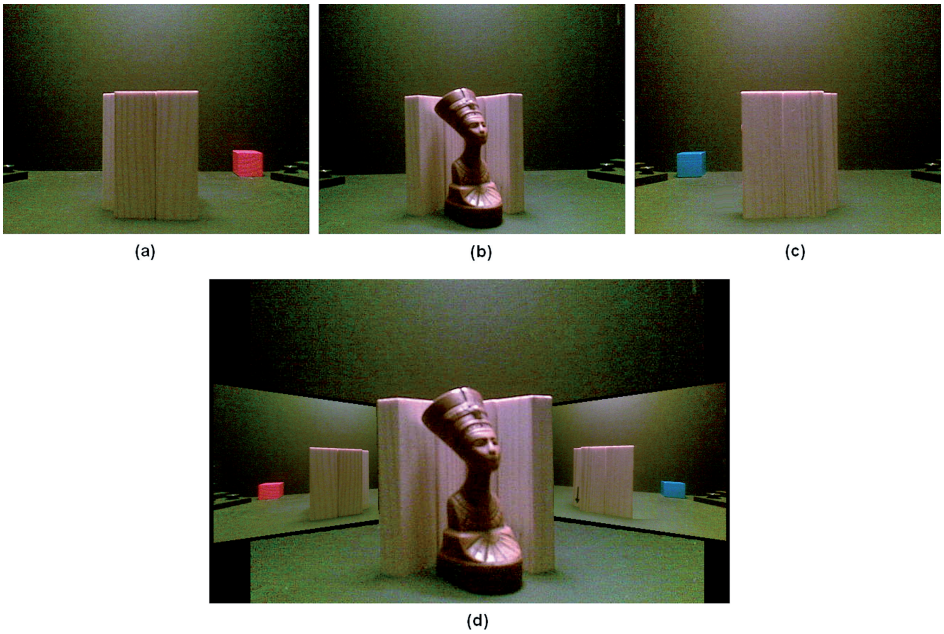

Fig. 3. Example of directional type scene. (a) View from Camera 1 of a red block; (b) View from Camera 2 of the bust; (c) View from Camera 3 of a blue block, make up the Array view of the scene; and (d) is the VM view of the same scene.

these scene types are static. In this article, we seek to determine whether observers can make relatively quick scene assessments given the available cues. In our motivating surveillance example, the guard has already determined that Person A is making signals, from a single view of that person. He or she must then determine to whom Person A is signalling. To make this determination, the guard can only rely on static cues, such as the direction Person A is facing. In this experiment, we are studying the ability of observers to relate the views with such static cues. Moreover, for the reconstructional scene types, we are attempting to determine the observer's ability to reconstruct a scene based on the static cues available in the scene, such as object structure.

Relational Scenes. Figure 3 shows an example of a relational scene. The scenes comprise an Egyptian bust and two different colored wooden cubes; in total there were three cubes: red (R), blue (B), or green(G). For each scene, two cubes were chosen from among the three, resulting in three possible linear combinations: R-B, G-B, and R-G. In each of the directional scenes, the bust would be facing one of the two cubes, either directly or through a wall. For example, in Figure 3, the bust is facing one of the cubes. It is not immediately obvious which cube the bust is facing (red (Figure 3(a)) or blue (Figure 3(c))). The setup is such that each camera can only see either the bust or one of the cubes. In our explanatory example, one camera, Camera 2, has a view of the bust, and it is clear from this view that the bust is facing its left (or the right of the camera); however, it is not clear from this view which block is on the left of the bust. It is from the views of the other two cameras, Camera 1 and Camera 3, that we can see the blocks, red and blue, respectively. The task would then be to identify which block the bust is facing either from a VM or an array presentation. In fact, the bust is facing the red block, as can be seen in the diagram in Figure 3. The purpose of this scene type is to determine whether the subject would fare better at identifying the block the bust is facing using the array view or the VM

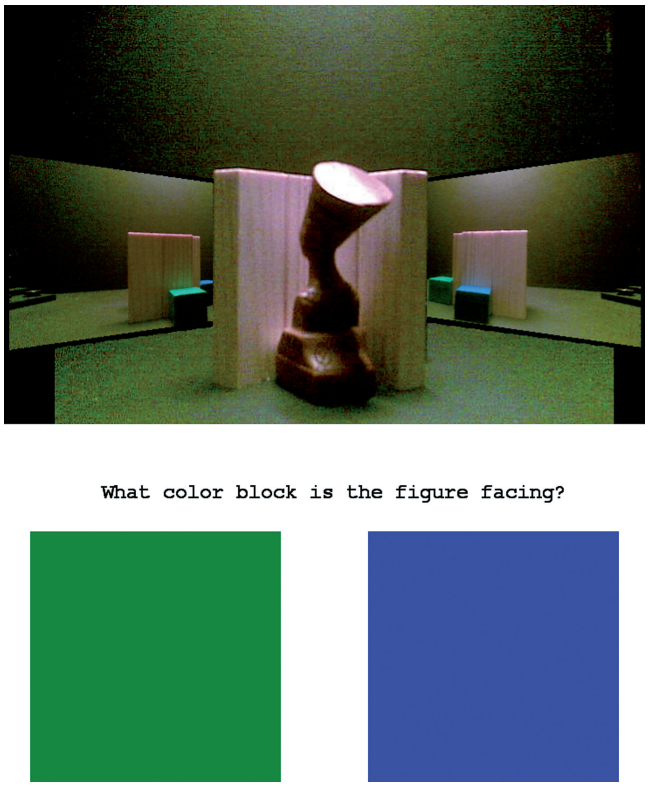

Fig. 4. Sample test question. Participants choose between the two images at the bottom of the screen.

view. The motivation behind this type of scene is the casino security guard example described earlier. There are 48 different relational scenes. Figure 4 shows a sample question for this scene type.

Reconstructional Scenes. An example of the second type of scene, the reconstructional scene, is shown in Figure 5. In these scenes, wooden blocks were placed in different configurations. Each camera captured its own unique view of the block configurations. The subject was required to identify the scene in question when presented with two similar looking scenes—of which one is the corresponding scene. In Figure 6, we show an example question for this scene type. As shown, the subjects are in fact required to choose from two bird's eye views of similar-looking scenes. This view was chosen in order to properly show the scene configurations in a single image, without using view integration to avoid biasing. In total, 48 different reconstructional scenes were set.

Attentional Load. An attentional load was added to the experiment—for which we used an audio track. To make the recording, integers between 0 and 100 were read out in random order and at varying rates. The rates could vary between approximately 0.3 to 4 seconds between numbers. The track was 10 minutes and 53 seconds long and could be run in a loop. The participants were asked to repeat the numbers spoken in the audio track. They were supervised throughout the session that they carried out the tasks as instructed.

The purpose of the attentional load was to determine whether attention is important or required for carrying out the visual task in either mode. If there is no deficit noted when attentional loading, then it is likely attention is not required for the given task. It has long been theorized that there is a sort of

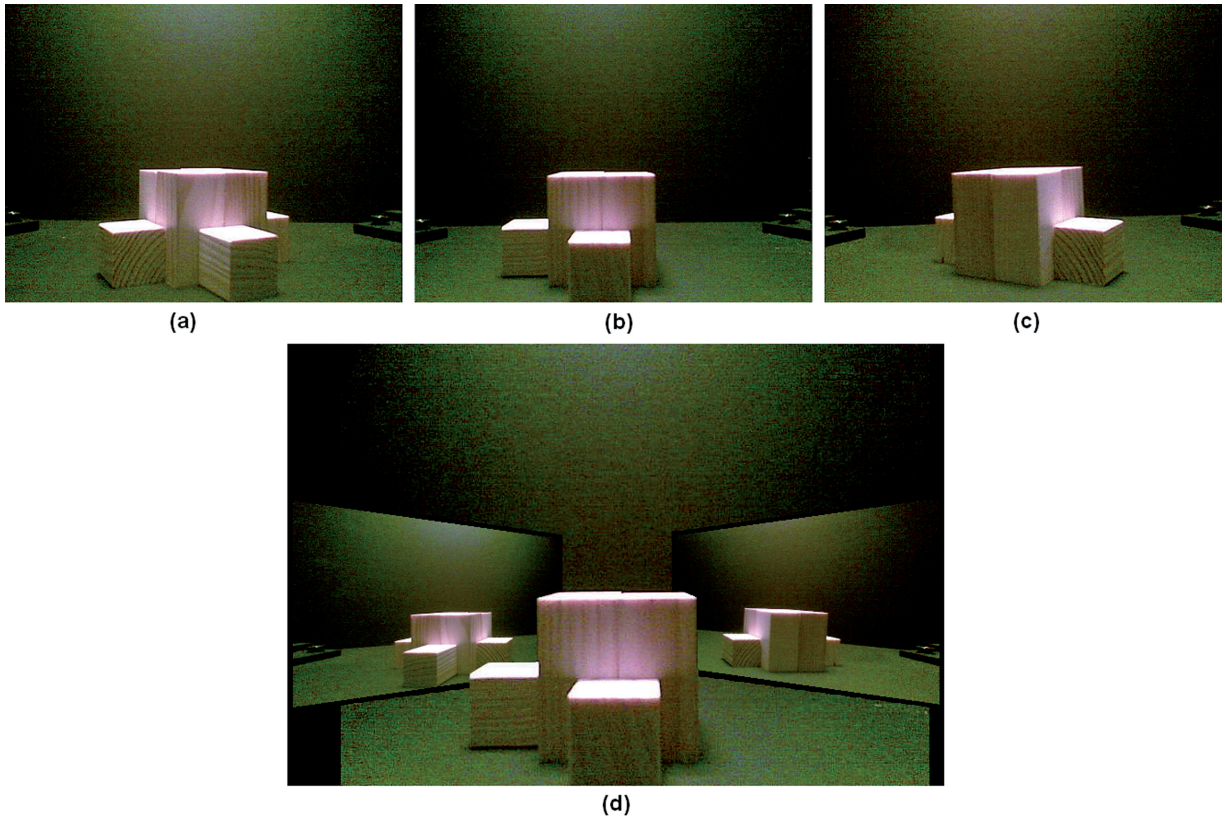

Fig. 5. Example of block type scene. (a), (b), (c) make up the Array view of the scene and (d) is the VM view of the same scene.

"bottleneck" when humans attempt to process a large amount of sensory information [Broadbent 1958]. Studies have shown that adding an attentional load can affect the ability for sensory processing [Alais and Blake 1999; Rees et al. 1997]. The decision to use an audio load was motivated by the surveillance example, where, at any given time, the guard may have different audio cues calling his or her attention elsewhere.

2.1.4 *Procedure.* The VM and array presentations were assessed in two separate ways: using 48 relational and 48 reconstructional scenes. Therefore, there were 96 VM presentation questions and 96 array presentation questions for a total of 192 questions in the test set. Each of the participants were asked all 192 questions in random order. Additionally, the participants were also required to answer the same test set with the attentional load described in the previous section. Half the subjects (6 subjects) performed the study by answering the test set first without attentional load, followed by an optional 5-minute break, and then answering the test set with the load. The other half of the subjects answered the test set first with the attentional load, then without the load, with the optional 5-minute break in between. The subjects were not permitted to leave the room. It should be noted that none of the participants chose to rest more than 1 minute.

Before the test, the participants were given an instruction sheet. The sheet did not provide any information about the setup of the stage or how the images were acquired. The instructions first described

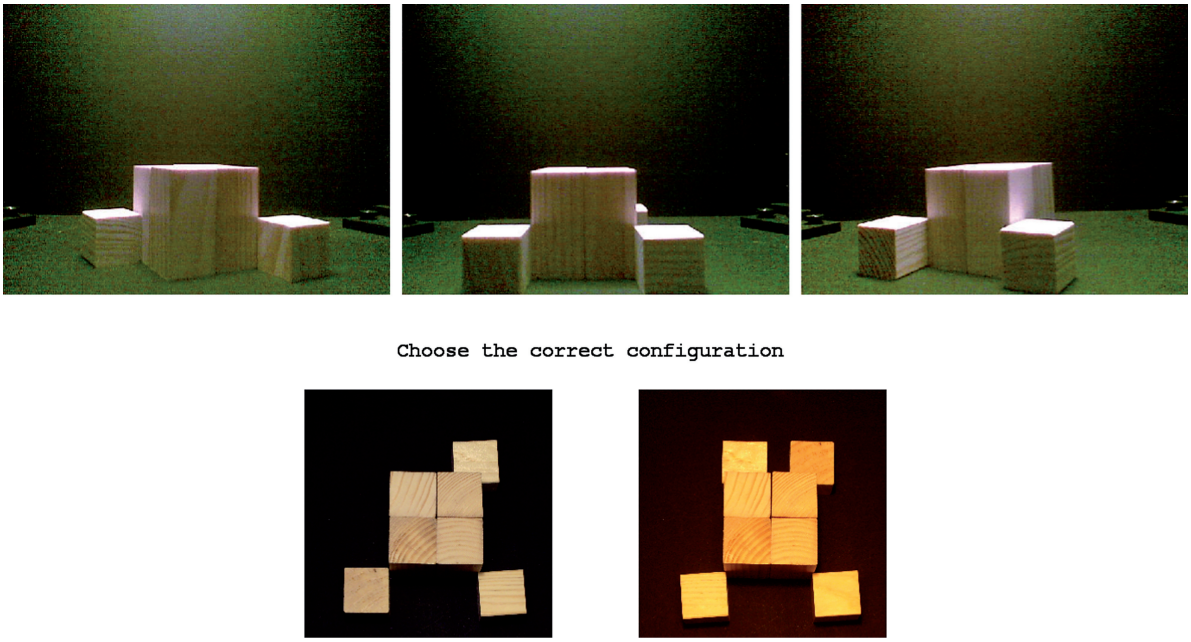

Fig. 6. Sample test question. Participants choose between the two images at the bottom of the screen.

how a question screen would appear. As shown in Figures 4 and 6, the top portion of the screen contained the test scene and the bottom two images are the answer choices to the question. The instruction sheet then described how to make a selection: pressing the left or right arrow key to select the left or right image, respectively. Participants were also informed they should make their selection as quickly as possible, with a time limit of 10 seconds per question, and that in total there would be 2 sets of 192 trials (384 trials). The attentional load was described and they were informed whether they would be starting with or without the audio attentional load. Following the instruction sheet, for the purpose of full clarity, the participants were given a trial run, where four example questions (two array presentations and two VM presentations) were asked. The four trial questions were not the same as those in the main test set, but were presented exactly as the experiment questions would be presented. The participants were informed that images presented by the VM presentation contained mirrors, but they were not informed about the construction or placement of the mirrors.

The binary result (correct or incorrect) as well as the response time were recorded.

2.2 Results

We present the results of the experiment in Table I. When we combine the scene types, we see an 8.3% improvement in accuracy of the VM presentation over the array presentation. The relational scenes produced over a 10% improvement, while the reconstructional produced a 6% improvement. Response times of relational scenes saw a mean decrease of 1.2 seconds and those of reconstructional scenes saw a mean decrease of 2.7 seconds. These results are the mean results of having combined the data of with and without attentional load and in which order the subjects were given the test set. In Table II we present the results divided by presentation type (VM or Array), no load (NL) or loading (AL), and order, whether NL then AL (NL $>$ AL) or AL then NL (AL $>$ NL). Again, the accuracy

Table I.
Total percent correct and response times of the experiment for both scene types (relational (Rel) and reconstructional(Rec)), broken down into scene types and combined.

	Percent Correct			Response Time (s)		
	Rel+Rec	Rel	Rec	Rel+Rec	Rel	Rec
VM	84.16	82.20	86.11	4.02	3.60	4.40
Array	75.87	71.53	80.21	4.68	4.79	4.67

Table II.
Percent correct and response time of the experiment broken down into the no load (NL) and attentional load (AL) conditions. And the order of loading, whether NL comes first then AL (NL > AL) or AL comes first then NL (AL > NL).

			Percent Correct			Response Time (s)		
		Order	Rel+Rec	Rel	Rec	Rel+Rec	Rel	Rec
VM	NL	NL > AL	89.06	89.24	88.89	3.46	3.0	3.85
		AL > NL	82.64	80.90	84.38	4.85	4.66	5.02
	AL	NL > AL	84.90	82.29	87.50	4.59	3.94	5.18
		AL > NL	80.03	76.39	83.68	3.19	2.74	3.57
Array	NL	NL > AL	82.47	80.90	84.03	4.12	4.21	4.21
		AL > NL	74.31	68.40	80.21	5.58	5.87	5.35
	AL	NL > AL	71.35	64.24	78.47	5.20	5.18	5.27
		AL > NL	75.35	72.57	78.12	3.84	3.91	3.84

of subjects with the VM presentation surpassed that with array presentation regardless of loading or ordering.

These patterns were confirmed by the repeated measures analysis. We divided the group into the two scene types: relational and reconstructional. We begin by presenting the results of the relational scenes. Scene identification accuracy was measured in a $2 \times 2 \times 2$ (VM/array $\times$ NL/AL $\times$ order) mixed ANOVA. The analysis revealed main effects of presentation type, $F(1,10) = 7.899$, $p = 0.018$. There were interactions between load and order, $F(1,10) = 9.442$, $p = 0.012$ and second order interactions between load, order, and presentation type, $F(1,10) = 14.259$, $p = 0.004$. While loading did not significantly deteriorate accuracy, we did find that ordering did affect the results.

Response times of the relational scene results were also measured in the same $2 \times 2 \times 2$ mixed ANOVA. The analysis again revealed main effects of presentation type, $F(1,10) = 21.332$, $p = 0.001$. There were no other interactions of significance.

For the reconstructional scenes, the $2 \times 2 \times 2$ mixed ANOVA was measured for accuracy. The analysis revealed the main effects of presentation type, $F(1,10) = 18.828$, $p = 0.001$. There were no other interactions recorded.

Response times of these scenes revealed only main effects for presentation types, $F(1,10) = 6.342$, $p = 0.03$.

Overall, the proportion correct was greater with a VM presentation, irrespective of scene type. Moreover, subjects had a shorter response time with the VM presentation. We note that these results do not reflect a speed/accuracy tradeoff, as participants answered with both greater % correctness and more quickly with the VM presentation.

3. GENERAL DISCUSSION

The analysis of our experiment supports our claim that humans are more accurate and faster at identifying scenes with the use of a VM presentation over an array presentation. It is our belief that these results are due to the fewer mental transformations required to link the views when subjects view a scene with a VM presentation. With this presentation type, the spatial relationships between views are inherent in the resulting composite image containing the views; the virtual mirrors are positioned in such a way as to preserve the position and orientation of the corresponding cameras that captured the images. Alternatively, with the array presentation, neither the relative scale, rotation nor location between the views are preserved. Earlier in this article, we cited Castelhano et al.'s model for human scene recognition—where the different 2D views of a 3D scene are linked at some point in the decision process of scene recognition. Our data is consistent with this model: while with the VM presentation the views are inherently linked—with array presentation, the views must be linked mentally. It is our belief that the poorer performance observed with the array presentation is due to the subjects being required to mentally link the 2D views. In doing so, extra time is required in the decision process, thus explaining the slower response times observed. The lower accuracy can also be attributed to the extra time required, as the task was time-sensitive, and this factor could have added pressure to subjects, thereby degrading their performance. While this theory is certainly plausible, we note that the response times were well below the allowable 10 seconds. Hence, we seek an alternative explanation. We already stated that we believe identifying scenes with array presentations require some mental transformations. It is possible that the VM presentation also requires a certain amount of mental transformations; however, these transformations are known transformations, since humans are accustomed to observing the world through a mirror. Therefore, while both presentation types may require mental transformations, those required for VM presentations are more familiar, and therefore suggest that humans can perform the needed transformations more accurately and quickly.

For both the VM and array presentation, when the test order was no load (NL), then with load (AL), the % correct was lower with the load than without. More specifically, in this order, accuracy dropped by 4% with VM presentation and over 11% with the array presentation. The greater drop in accuracy noted in the array presentation over the VM presentation is consistent with the notion that greater attention is required with the array presentation. When subjects were tested with the load first, accuracy saw little change between AL and NL in the array presentation (1% decline) and a slight improvement in the VM presentation (2% improvement). From our data we cannot conclude that adding an attentional load had any significant impact on the accuracy. The changes in performance appear to be more affected by ordering than by loading. While we expected the accuracy to be lower for AL regardless of ordering, our data does not reflect this. Our hypothesis is that when subjects began with the attentional load, the task was more difficult and perhaps facilitated learning, thus explaining why for the AL then NL order accuracy was not as affected between the first and second run of the test than with the NL then AL ordering. Additionally, we examined the performance of each participant in bins of 30 seconds, and no learning curve was observed. As for the response times, regardless of ordering or loading, the speeds were slowed significantly on the second test. This slowing in response times can be attributed to subject fatigue. While no major conclusions can be drawn from the data about the attentional load, it can be stated that regardless of presentation type, ordering, or loading, subjects are able to identify scenes with greater accuracy and speed when using a VM presentation over an array presentation.

Finally, after their sessions, participants were asked some qualitative questions about how they felt the VM presentation fared over the array presentation. Specifically, they were asked "Which presentation type did you find easier." Out of the 12 participants, 11 responded favorably for the VM presentation, while only 1 participant felt he performed better with the array presentation. Upon

investigation of his result, he had, in fact, performed better with the VM presentation. Of the 11 who responded favorably, 5 participants felt that the VM presentation required getting used to—however, once after viewing no more than 10 scenes in the VM presentation, they became familiar with viewing scenes using the virtual mirrors. The qualitative results are also consistent with our hypothesis that by presenting multiple 2D views of a 3D scene in an integrated fashion, which maintains the spatial relationships between views, viewers will fare better at certain spatial tasks.

REFERENCES

ALAIS, D. AND BLAKE, R. 1999. Neural strength of visual attention gauged by motion adaptation. *Nature Neurosci. 2*, 1015–1018.

AU, C. E. AND CLARK, J. J. 2008. Multiple view integration and display using virtual mirror. In *Proceedings of the 5th Canadian Conference on Computer and Robot Vision (CRV).* 286–293.

BEIS, J. AND LOWE, D. 1997. Shape indexing using approximate nearest-neighbor search in high dimensional spaces. In *Proceedings of the Conference on Computer Vision Pattern Recognition.* 1000–1006.

BERTAMINI, M., LATTO, R., AND SPOONER, A. 2003. The Venus effect: People's understanding of mirror reflections in paintings. *Perception 32*, 5, 593–599.

BERTAMINI, M. AND PARK, T. E. 2005. On what people know about images on mirrors. *Cognition 98*, 1, 85–104.

BIEDERMAN, I. 1987. Recognition-by-components: A theory of human image understanding. *Psych. Rev. 94*, 115–147.

BIEDERMAN, I. AND GERHARDSTEIN, P. 1993. Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *J. Exper. Psych. Human Percept. Perform. 23*, 1162–1182.

BIEDERMAN, I. AND GERHARDSTEIN, P. 1995. Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff. *J. Exper. Psych. Human Percept. Perform. 21*, 1506–1514.

BROADBENT, D. 1958. *Perception and Communication*. Pergamon, London.

BROWN, M. AND LOWE, D. G. 2007. Automatic panoramic image stitching using invariant feature. *Int. J. Comput. Vision 74*, 1, 59–73.

BÜLTHOFF, H. H., EDELMAN, S. Y., AND TARR, M. J. 1995. How are three-dimensional objects represented in the brain? *Cereb. Cortex 5*, 3, 247–260.

CAPEL, D. AND ZISSERMAN, A. 1998. Automated mosaicing with super resolution zoom. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* IEEE, Los Alamitos, CA.

CASTELHANO, M. S., POLLATSEK, A., AND RAYNER, K. 2009. Integration of multiple views of scenes. *Attention, Percep. Psychophysics 71*, 490–502.

CHRISTOU, C. AND BÜLTHOFF, H. 1999. View dependence in scene recognition after active learning. *Memory and Cognition 27*, 996–1007.

COLOMBO, C., COMANDUCCI, D., AND BIMBO, A. D. 2006. Camera calibration with two arbitrary coaxial circles. In *Proceedings of the 9th European Conference on Computer Vision (ECCV'06).* Springer, Berlin.

CROUCHER, C. J., BERTAMINI, M., AND HECHT, H. 2002. Naive optics: Understanding the geometry of mirror reflections. *J. Experimental Psychology: Human Percept. Perform. 28*, 546–562.

DEBEVEC, P., YU, Y., AND BOSHOKOV, G. 1998. Efficient view-dependent image-based rendering with projective texture mapping.Tech. rep. CSD-98-1003, University of California, Berkeley.

EDELMAN, S. AND WEINSHALL, D. 1991. A self-organizing multiple-view representation of 3D objects. *Biol. Cybern. 64*, 209–219.

GARSOFFKY, B., SCHWAN, S., AND HESSE, F. 2002. Viewpoint dependency in the recognition of dynamic scenes. *J. Exper. Psych. Learn. Memory, Cognition 28*, 1035–1050.

HARTLEY, R. AND KANG, S. B. 2007. Parameter-free radial distortion correction with center of distortion estimation. *IEEE Trans. Patt. Anal. Mach. Intell. 29*, 8 , 1309–1321.

HAYWARD, W. AND WILLIAMS, P. 2000. Viewpoint dependence and object discriminability. *Psych. Sci. 11*, 7–12.

HOCK, H. AND SCHMELZKOPF, K. 1980. The abstraction of schematic representations from photographs of real-world scenes. *Memory and Cognition 8*, 543–554.

IRANI, M. AND ANANDAN, P. 2000. About direct methods. In *Proceedings of the International Workshop on Vision Algorithms*. Springer, Berlin, 267–277.

KANAZAWA, Y. AND KANATANI, K. 2002. Image mosaicing by stratified matching. In *Proceedings of the Workshop on Statistical Methods in Video Processing*.

KANNALA, J. AND BRANDT, S. 2006. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. Patt. Anal. Mach. Intell. 28*, 8 (Aug.), 1335–1340.

LAWSON, R. AND BERTAMINI, M. 2006. Errors in judging information about reflections in mirrors. *Perception 35*, 1265–1288.

LOWE, D. 2004. Distinctive image features from scale-invariant keypoint. *Int. J. Comput. Vision 20*, 91–110.

MARR, D. AND NISHIHARA, H. 1978. Visual information=processing: Artificial intelligence and sensorium of sight. *Technol. Rev. 81*, 28–49.

NAKATANI, C., POLLATSEK, A., AND. JOHNSON, S. H. 2002. Viewpoint dependent recognition of scene. *Q. J. Exper. Psych. Sect. A 55*, 115–139.

POGGIO, T. AND EDELMAN, S. 1990. A network that learns to recognize three-dimensional objects. *Nature Neurosci. 343*, 263–266.

RANKOV, V., LOCKE, R. J., EDENS, R. J., BARBER, P. R., AND VOJNOVIC, B. 2005. An algorithm for image stitching and blending. In *Proc. SPIE*, vol. 5701, 190–199.

REES, G., FRITH, C., AND LAVIE, N. 1997. Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science 278*, 1616–1618.

TARDIF, J.-P., STURM, P., TRUDEAU, M., AND ROY, S. 2009. Calibration of cameras with radially symmetric distortion. *IEEE Trans. Patt. Anal. Mach. Intell. 31*, 9, 1552–1566.

TARR, M. AND PINKER, S. 1989. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psych. 21*, 233–282.

TARR, M., WILLIAMS, P., HAYWARD, W., AND GAUTHIER, I. 1998. Three-dimensional object recognition is viewpoint dependent. *Nature Neurosci. 1*, 275–277.

ULLMAN, S. AND BASRI, R. 1991. Recognition by linear combinations of models. *IEEE Trans. Patt. Anal. Mach. Intell. 13*, 193–254.

VETTER, T., POGGIO, T., AND BÜLTHOFF, H. 1994. The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology 4*, 1, 18–23.

WALLIS, G. AND BÜLTHOFF, H. 1999. Learning to recognize objects. *Trends Cognitive Sci. 3*, 1, 22–31.

ZHANG, Z. 2000. A flexible new technique for camera calibration. *IEEE Trans. Patt. Anal. Mach. Intell. 22*, 11, 1330–1334.

ZHENG, Y. AND LIU, Y. 2008. The projective equation of a circle and its application in camera calibration. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08).* 1–4.

Received May 2010; revised November 2010; accepted December 2010