Enhancing Spatial Awareness with Multi-Camera View Synthesis for Diagnostic and Manipulation Tasks

by

William S. Walmsley

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

© Copyright by William S. Walmsley, 2012
Abstract

Enhancing Spatial Awareness with Multi-Camera View Synthesis for Diagnostic and Manipulation Tasks

William S. Walmsley
Master of Applied Science
Graduate Department of Mechanical and Industrial Engineering
University of Toronto
2012

View synthesis is a technique used to generate novel views from input images captured from separate viewpoints. I have explored the use of this technique for motion parallax displays in surgical endoscopy by developing real-time interpolation software (especially for use with a three-camera endoscopic tool), and by conducting a perceptual experiment with 16 novice participants to determine the benefits of such a display in a surgical-like occluded relative-depth discrimination task. ‘Noisy-edged’ objects are proposed to aid in elimination of the relative size cue for such an investigation.

Results indicate that participants performed significantly better using a motion parallax display as compared to alternate or simultaneous presentation of multiple images, and subjective rankings of display difficulty corroborate this finding. Participant interpolation control tendencies were analyzed, showing that strategies involving lengthier movement times produced higher performance. These motion parallax display benefits also pertain to similar tasks in a variety of applications.
Acknowledgments

This thesis would not have been possible without the tremendous support of a number of people, most notably my industrious advisor, Paul—vielen Dank! I am eternally grateful to each of Paul Milgram, Dr. James Drake and Justin Hollands, for donating their precious time to my edification (and to the clarity of this thesis).

This project would never have come into existence without Dr. Peter Kim, Harmanpreet Bassan, and Thomas Looi at the CIGITI in SickKids Hospital, and I am indebted to Dr. Ted Gerstle and Dr. Walid Farhat for inspiration leading to my novel experimental paradigm.

For their practical support and encouragement, I am beholden to my colleagues at the ETC lab; for their emotional support and encouragement, I am beholden to my family, my friends, and especially those who could be said to be both.

Thanks for freeing my point of view,
Will
## Contents

**Abstract**

**Acknowledgments**

**1 Introduction**

**2 Background**

2.1 History of minimally invasive surgery

2.2 Depth perception in endoscopic surgery

2.2.1 Stereoscopy

2.2.2 Motion parallax displays

2.2.3 Multi-camera laparoscope

2.3 Image-based rendering

2.3.1 Depth extraction

2.3.2 View synthesis

**3 Experimental evaluation**

3.1 Study 1: Preliminary psychophysical calibration study

3.1.1 Calibration study design

3.1.2 Calibration study results

3.2 Study 2: Tri-camera display study

3.2.1 Display conditions

3.2.2 Hypotheses

3.2.3 Experimental setup

3.2.4 Experimental presentation

3.2.5 Participants
4 Results 36
  4.1 Analysis of subjective preferences ................. 36

5 Discussion 41
  5.1 Limitations and future work ....................... 44

6 Conclusions 47

References 49

A Algorithmic details of view interpolation 53

B Response accuracy data table 56

C Participant questionnaire 57

D In-depth analysis of subjective preferences and response behaviour 61
  D.1 Condition 0: Subjective responses ................. 61
  D.2 Condition 1: Subjective responses ................. 62
  D.3 Condition 2: Subjective responses and Viewpoint switching tendencies ....................... 62
  D.4 Condition 3: Subjective responses and Viewpoint movement tendencies ..................... 63
      Movement tendencies ............................. 64
# List of Figures

2.1 Laparoscopic surgery requires multiple small incisions, while open surgery involves a typically more traumatic single large incision. Cholecystectomy is a common laparoscopic procedure involving removal of the gallbladder 4

2.2 Illustrations of typical laparoscope form factors, with representative a) monocular tip with one lens; and b) stereoscopic tip with two lenses and one light source 6

2.3 The three-camera laparoscope prototype developed at the Toronto Hospital for Sick Children (SKH), with three hinged arms (arm length ~5 cm), which fold open post-insertion 8

2.4 The goal of image mosaicing is to fuse a series of images taken from the same viewpoint into a single image (a.k.a. mosaic, or ‘panorama’) 10

2.5 The goal of view synthesis is to render a view (purple) from a novel virtual viewpoint, based on images (red and blue) captured from separate real viewpoints 10

3.1 a) An untreated 7/8” wooden dowel, and b) final version of ‘noisy’ 1/2” dowel coated in red plasticine 16

3.2 Illustration of experimental setup for the preliminary calibration study, with one rod supported at a variable height and photographed by an overhead camera 17

3.3 Screenshots from psychophysical calibration experiment, presenting a) untreated horizontal wooden rods; b) noisy sloped plasticine-covered rods 18
3.4 Individual response scores at various rod separations, for a) untreated horizontal rods and b) sloped ‘noisy’ rods; participants 1–4 are identified by colour, and response scores account for the proportion of responses a participant made identifying that the non-control rod was closer than the control rod 20.

3.5 Psychophysical models of sensitivity to relative depths of two different varieties of rod, untreated horizontal wooden dowels and noisy angled plasticine-covered dowels; the black horizontal line at a proportion of 0.5 indicates the performance results that would be expected if the effect of the width cue were to be entirely eliminated 21.

3.6 The 4 display conditions experienced by participants in the study 23.

3.7 Full screenshots of GUI used to present each display condition (as described in Figure 3.6) 24.

3.8 The graphics tablet used by participants as an absolute position control input device to control the viewpoint in display conditions 2 and 3. Participants slid the pointer by grasping its weighted tip 25.

3.9 The experimental setup of the artificial scene from a) the side and b) a sample image obtained from one of the cameras 27.

3.10 The textured surface appearing in the background of the experimental scene 28.

3.11 Three overall (yaw) orientations of rods, used to vary the appearance of the scene during the experiment 29.

3.12 Two rods, each shown at two possible rolled orientations (90° apart), as used to vary the appearance of the scene during the experiment 30.

3.13 Occluding circle suspended by a 6cm pillar above the top rod 30.

3.14 Equilateral triangular configuration of three PS3 Eye™ webcams with near-parallel viewing directions; cameras are spaced apart by 8.8 cm 31.

3.15 By variably aligning left, right, and top images, the horopter (plane of zero disparity) was adjusted such that the background appears in the same location in each image. The black rectangle indicates the subset of the overlapping region that was presented to participants. 32.
3.16 a) A synthesized viewpoint between the viewpoints of the physical cameras; b) a close-up of the synthesized viewpoint, revealing subtle ghosting artefacts as shown with arrows

4.1 Response accuracy results for each display condition and rod separation, with a) means and standard errors; b) probability density estimate (violin) plots indicating medians and interquartile ranges

4.2 Median response time means and standard errors for each display condition and rod separation

5.1 Two images of the same scene (with rods separated by 4 cm), as captured by the left and right viewpoints of our triangular camera configuration

5.2 A transparent blue overlay on each image reveal regions that are not visible in the other image due to occlusion; circled regions reveal smaller occluded regions on rod D compared to rod A in each image, indicating that rod D is above rod A

A.1 Step 3—Image as captured by the left camera in the three-camera configuration: a) original; and b) after undistortion, rectification, and cropping

A.2 Step 4 and 5—Pairwise stereo correspondence between left and right images, and between left and top images, yields two somewhat reliable depth maps; combination and refinement results in a higher-accuracy depth map

D.1 For display condition 2, a) time allocations on each viewpoint, averaged across participants, with standard error bars shown, and b) discrete viewpoint switching tendencies

D.2 Response data for each participant in display condition 3, including normalized directional tendency angular histograms, viewpoint dwelling preference heat maps, mean time spent manipulating (s/trial), mean nonzero speed of manipulation (edge lengths/s), and response accuracy (averaged across rod separations); participants are sorted in decreasing order of accuracy achieved
List of Tables

4.1 Subjective rankings of display condition difficulties, with a ranking of 1 indicating a low difficulty, and 4 indicating a high difficulty. Participant numbers (P1 to P16) are colour coded, based on response accuracy performance rank, to provide a comparison between perceived difficulty and measured performance scores. Colour-coding is further explained in the main body of the text. 39

B.1 Individual subject response accuracies by separation and display condition; each accuracy score is computed based on 24 binary responses 56
“There is nothing insignificant in the world. It all depends on the point of view.”

—Johann Wolfgang von Goethe
1 Introduction

View synthesis is an image-based rendering technique that involves generating a view of a scene from an arbitrary viewpoint based on multiple images captured from fixed viewpoints. View synthesis has been used for applications such as free viewpoint television, in which a user is free to ‘look around’ at the transmitted video feed, and thus to perceive it from a variety of angles. We specifically are interested in the application of view synthesis to endoscopic surgery, with the present work having a specific focus on laparoscopic surgery. In light of this, the present work may be considered to pertain to ‘free viewpoint surgery’.

A novel endoscopic camera tool has been developed at The Toronto Hospital for Sick Children (SKH) that consists of three widely spaced cameras, which can provide a surgeon with multiple views into an abdominal cavity while conducting laparoscopic surgery (Bassan et al., 2011). Beyond merely providing three (simultaneous) camera views, however, we are also able, by applying established view synthesis techniques, to provide a surgeon with a manipulable viewpoint of the scene, allowing them to access a continuum of virtual viewpoints ranging within a plane defined by the physical camera locations. One primary benefit of such a manipulable viewpoint display is the sensation of motion parallax that it provides, with the consequent potential to enhance a surgeon’s spatial awareness of the observed surgical environment.

The interesting thing about such a manipulable viewpoint is that intermediate viewpoints—and the smooth transitions between them—are synthesized using only the visual information present in the three original viewpoints. This implies that human operators provided with the original three distinct viewpoints could theoretically be capable of extracting the same quantity of information as they would if they had been given access to the actual intermediate viewpoints.

The important question arising from this new capability is thus: can active manipulation of the effective display viewpoint created

---

1 Laparoscopic surgery refers to abdominal surgical interventions that involve viewing the surgical environment via an endoscopic camera tool.

2 While the author of the present work was involved in the development of the three-armed endoscope prototype, specific aspects of the hardware development remain separate from the scope of this thesis.
by combining three separate sensor viewpoints improve an operator’s spatial awareness and enhance performance on manipulation tasks that are dependent on such information?

Due to the inherent limitations of the present thesis research, it was not possible to explore a wide range of tasks related to “spatial awareness”; nor was it possible to investigate different aspects of manipulation performance. Instead we have limited ourselves to one particular characteristic of spatial awareness—by devising a generic ‘occluded relative-depth discrimination task.’ This task was designed to be consistent with some elements of intraoperative spatial estimation tasks frequently conducted by laparoscopic surgeons. Put simply, the task involved viewing—from above, with a variety of display options—two perpendicular intersecting rods (representing anatomical structures) and identifying which was on top of the other, while the intersection of the rods was covered with an occluding surface.
2 Background

The motivation behind the present work derives from the perceptual limitations faced by surgeons conducting endoscopic surgery, in particular laparoscopic surgery, which involves a relatively large operating cavity. In the context of the human factors issues that arise in such surgeries, we discuss the principles of combining multiple views into a unified representation and how they might be employed to improve spatial awareness.

The present research was conducted in parallel with the development at SKH of a prototype three-armed three-camera endoscope tool. Whereas our display technology development and perceptual investigation is independent of that prototype, our work is motivated by the goal of implementing view synthesis techniques with such a camera tool. Thus, discussion of design tradeoffs integral to developing such a multi-camera tool are presented, but technical details related to any actual hardware configuration are omitted. For further information regarding technical details of the present work, consult ETC Lab Technical Document #1 (2012)\(^1\).

2.1 History of minimally invasive surgery

Minimally invasive surgery (MIS) emerged as a practical surgical technique in the 1980s, serving as a replacement for many varieties of open surgery. The most common form of MIS is laparoscopy, a surgery that involves the use of an endoscopic camera to see inside the abdominal cavity. Proven benefits of laparoscopic techniques over open surgery include reduced postoperative pain, fewer complications, earlier patient discharge, fewer cosmetic marks, and improved patient satisfaction (Gadacz, 1991; Rosen & Ponsky, 2001).

As illustrated in Figure 2.1, the multiple small incisions required for MIS are typically less traumatizing than a single major open surgical incision.

Despite these advantages, surgeons conducting MIS are forced

\(^1\) Contact William Walmsley or Paul Milgram at the Ergonomics in Teleoperation and Control (ETC) Lab for access to this document: {willw,milgram}@mie.utoronto.ca
to view the surgical environment through a camera via a display monitor, thus ‘tele-operating’ with reduced visual and tactile perception. This, coupled with the ergonomic difficulties of conducting the surgery with long and awkwardly placed tools, and with camera input typically under the control of a surgical assistant, results in these procedures requiring substantial training in the ability to navigate and reconcile viewpoint mismatches when carrying out surgical procedures (Buzink et al., 2009).

More recently, progressively less invasive techniques, including single port access surgery and natural orifice translumenal endoscopic surgery (NOTES), have been under development (Romanelli & Earle, 2009; Mintz et al., 2008). In the physically constrained task of operating through a single orifice, tactile and depth perception are even more important than in traditional laparoscopy. In laparoscopy and especially NOTES, the assistant requires much skill in manipulating the endoscope and careful coordination between the surgeon and assistant is needed (Mintz et al., 2008).

We interviewed two laparoscopic surgeons at SKH; they pointed out that in common operations such as cholecystectomy, one of the common problems they face is interpreting the relative depths of vessels such as the cystic duct and cystic artery. In some cases, ultrasonographers are called in to confirm the relative depths of organs and vessels to avoid unnecessarily damaging one while removing / severing another. This is especially important when an organ or tumour encases a number of surgically pertinent vessels, thereby occluding the surgical view of them. Laparoscopic tools may also become inadvertently and unnoticeably crossed during surgery,
requiring careful manipulation to correct their orientation.

Way et al. (2003) found overwhelmingly that “errors leading to laparoscopic bile duct injuries stem principally from misperception, not errors of skill, knowledge, or judgment”. Further, as Buzink et al. (2009) point out, camera navigation is a skill unrelated to tissue manipulation; the viewing limitations that surgeons currently face require additional specific (and costly) camera training, and it is clear that enhancing surgical spatial awareness would provide clear benefits to patients and the medical community.

2.2 Depth perception in endoscopic surgery

The human visual processing system makes use of a wide range of cues to estimate depth and interpret the spatial configuration of scenes. The primary depth cues among these (where we take depth to refer to the distance of perceived objects from an observer) include (Wickens & Hollands, 2000):

- **binocular disparity**, whereby differences between horizontally displaced images received by each eye result in a single fused image that enables a direct perception of depth—a.k.a. **stereoscopic vision**
- **motion parallax**, the interpretation of relative movements of stationary objects in a scene when observed from a moving viewpoint; the **kinetic depth effect** is the analogous interpretation of the relative movements of moving objects in a scene when observed from a stationary viewpoint
- **relative size**, interpretation of object depth, based on known size, compared to other objects in the scene
- **interposition**, whereby any object, or portion of an object, that is seen to cover another portion of the visual scene is perceived to be closer to the observer than the concealed portion of the scene
- other object-centred cues including perspective, shading, and textural gradients

A great deal of literature exists on the topic of depth perception and the relative impact of different depth cues (for example: Adelson et al. (1993); Johnston et al. (1994)). For the purposes of the present thesis research, however, our primary interest, as explained more extensively below, is **motion parallax**. For the sake of completeness, we also consider in the following an important ‘rival’ of motion
parallax, binocular disparity, which is very commonly, and often exclusively, associated with depth perception—especially in relation to the current popularity of “3D displays” in a broad variety of fields.

Although in most situations the various depth cues complement each other, to facilitate reliable perception by observers of actual locations and sizes of real-life objects, this is not always the case, which has led to many studies of conflicting depth cues (for example: Drascic & Milgram (1996); Cavanagh (1987)). Since reviewing that literature is beyond the scope of this study, suffice it to say that, in a large number of studies, motion parallax cues have been shown in general to dominate binocular disparity (stereopsis) cues—(see, for instance, Ware (1995)). Because both remain nevertheless important, we consider in the following some of the key technological factors related to the realisation of displays that provide each class of perceptual depth cues.

2.2.1 Stereoscopy

WHEREAS conventional (monocular) endoscopes in laparoscopic surgery make use of a single camera view, stereoscopic endoscopes contain a pair of cameras, slightly displaced horizontally from one another, to provide stereoscopic views of the surgical scene; representative tips of each type of endoscope are shown in Figure 2.2.

The benefits of stereoscopic presentation of laparoscopic video have been extensively explored and demonstrated (Taffinder et al., 1999; Birkett et al., 1994; Munz et al., 2004; Ilgner et al., 2007). Taffinder et al. (1999) investigated both expert and novice
performance with suturing, cutting, and grasping tasks, to find that stereoscopic vision reduced the performance drop brought about by endoscopic surgery (as compared to direct vision surgery) by about 50%. Birkett et al. (1994) found a 25% speed increase in a suturing task, and found subjective improvements in a wide range of other laparoscopic operations.

Despite such findings, as well as the intuitive belief on the part of many (typically stereo display aficionados) that stereo displays “must be useful” for surgical applications, not all such investigations have led to such unequivocal results (e.g., Herron et al. (1999)). In many cases failure to find a stereoscopic advantage has been attributed to shortcomings in the technologies used for the studies. Such issues often involve discordance between two images such as vertical shift, rotation error, and size inconsistency, which are known to lead to eyestrain issues.

2.2.2 Motion parallax displays

As an alternative to stereoscopic displays, using motion parallax may be an equally valid approach to providing surgeons with highly sought-after spatial perception. Whereas on the one hand motion parallax may be achieved simply by moving a camera, the problem in laparoscopic surgery is that camera movement occurs as a rotation about the fulcrum of the insertion point. This type of camera motion is beneficial for building up a panoramic view of the interior, but does not provide the viewpoint rotation about an interior fulcrum point as would be necessary to achieve motion parallax\(^3\). One endoscope has been designed to allow physical movement of a single camera about such an interior fulcrum point, and was found to improve performance in a laparoscopic exploration task by a factor of 2 (Voorhorst et al., 1996). An additional difficulty with camera control in surgery is that it is typically the responsibility of the assistant rather than the surgeon. This issue was researched by Hubona et al. (1997), who demonstrated accuracy improvement in a variant of the mental rotation paradigm when providing users with control over object motion.

Head-mounted displays have been tested for use with overlaying virtual information over the real world (Fuchs et al., 1998; Sielhorst et al., 2006), supporting motion parallax for spatial interpretation of the virtual information. However, virtual information in this context has typically been limited to virtual reconstructions of selected anatomy, approximated based on scans obtained prior to surgery. McMillan & Bishop (1995) present an early discussion of how an

\(^3\) For more on the distinction between panoramic viewing (image mosaicing) and rotation about an interior fulcrum point (view synthesis), refer to Section 2.3.
alterable viewpoint is needed to make head-mounted viewing of a virtual world seem plausible.

Sollenberger & Milgram (1993) demonstrated the usefulness of both motion parallax and stereopsis, as well as a combination of the two, in a task involving visualizing the connectivity of complex structures such as blood vessels in the brain. While the topology of the laparoscopic surgical environment is less complex than cerebral vascular systems, the same benefits of enhanced spatial awareness should apply.

2.2.3 Multi-camera laparoscope

Figure 2.3: The three-camera laparoscope prototype developed at the Toronto Hospital for Sick Children (SKH), with three hinged arms (arm length ~5 cm), which fold open post-insertion.

Figure 2.3 depicts the three-camera laparoscope developed at the Toronto Hospital for Sick Children, with cameras mounted on hinged arms in order to provide relatively widely separated viewpoints (Bassan et al., 2011). One basic use case of this novel endoscope form factor involves a surgeon switching between views to avoid visually occluding elements such as surgical tools and layered anatomical features. Such a use should especially benefit single port access surgery (Romanelli & Earle, 2009), in which tool-occluded anatomy is more prevalent.

The arguably greater benefit of this tool, however, will be the ability to allow a surgeon to manoeuvre within the continuum of virtual viewpoints defined within the plane formed by the three cameras. By synthesizing a virtually manipulable viewpoint, it is possible to provide additional structure from motion without the need for physical camera movement, to achieve motion parallax. It is also anticipated that pairs of viewpoints will be synthesisable, to also
provide depth from stereopsis.

By allowing virtual camera motion without physical camera motion, it is anticipated that the frequency of collisions with patient tissue will be reduced. Further, for use with single port access surgery, it is anticipated that physical scope movements that often result in an inadvertent displacement of carefully arranged adjacent surgical instruments will be reduced (Romanelli & Earle, 2009).

2.3 Image-based rendering

The combination of multiple viewpoints into a continuous representation can be accomplished in a variety of ways, with a fundamental distinction being between image mosaicing and view synthesis, which are depicted in Figures 2.4 and 2.5. Image mosaicing is a better-known technique, commonly used to stitch images together into panoramas. Mosaicing conventionally relies on images being captured from the same viewpoint, with the camera simply rotated to capture various subsets of the surrounding 360° view. View synthesis, on the other hand, entails combining images from various viewpoints (viewing the same scene) to create a synthesised image from a new viewpoint, which typically lies in between the original viewpoints. For this reason, view synthesis may also be referred to as view interpolation. (Note in addition that we may also generate extrapolated viewpoints using the same approach. The problem in that case, however, is that there will more likely be information lacking from portions of the scene visually occluded from the original views.)

View interpolation has been extensively treated in the past decade, primarily within the context of providing so-called Free Viewpoint Television (FTV) (Tanimoto, 2012). The field of FTV has emerged to focus especially on some of the issues associated with the encoding and transmission of a large array of viewpoints to allow users at home to manipulate virtual camera viewpoints—see, for instance, Kimata et al. (2007). Not all of those methods are particularly applicable for our special case of few-cameras short-distance teleoperation, however. More specifically, most of the FTV technologies gloss over the methods of obtaining disparity maps for each image by relying on the assumption that multiple viewpoint videos are collected and processed ahead of time, prior to transmission. Nevertheless, the surgical aspects of the present work can still be considered as an investigation into the benefits of ‘free viewpoint surgery’.

An arbitrarily large dense array of cameras, as is typically used for FTV applications, would be useful in a surgery context to simplify
Figure 2.4: The goal of image mosaicing is to fuse a series of images taken from the same viewpoint into a single image (a.k.a. mosaic, or ‘panorama’)

Figure 2.5: The goal of view synthesis is to render a view (purple) from a novel virtual viewpoint, based on images (red and blue) captured from separate real viewpoints
interpolation between adjacent viewpoints. This is the motivation behind micro lens arrays, which achieve a similar effect using a single large image sensor and a dense array of lenses, such that the resulting image can be subsequently computationally decomposed to interpret the image as seen from this array of slightly different viewpoints (Gortler et al., 1996). We are more interested in a wider viewpoint baseline, which may be unwieldy to achieve by using a large array of image sensor(s) and lenses. However, investigation of the practical feasibility of such a wide baseline microlens array certainly warrants further thought and investigation.

In relation to our work with 3 cameras, it is worth considering that the simplest case of view interpolation could have involved only 2 cameras, but interpolation would be restricted to a single dimension. Three cameras is thus the minimum number required to enable planar interpolation, where any direction of viewpoint translation within the planar interpolation region is available when manipulated by a human operator. For the case of four cameras, a difficulty arises from some of the simplifying assumptions made by efficient view interpolation algorithms, which require that the optical centres of the cameras be coplanar (lie on a mutual plane). These assumptions may still be relied upon, provided that the camera tool can be manufactured to precisely achieve this geometry.

View synthesis, to provide physically valid virtual views, requires information about the spatial configuration of every feature in the scene. A typical way of achieving this is using stereo vision techniques. While we will omit technical details from this discussion (refer to ETC Technical Document #1 (2012)), it is important to understand the fundamental working of view synthesis using stereo vision, which can be broken into two parts: depth extraction and actual view synthesis.

2.3.1 Depth extraction

Depth extraction relies on stereo correspondence, whereby the locations of pixels or other image features in one image are compared to features in another image captured by an adjacent camera. Similar to human perception, the disparity (i.e., distance) between an image feature in adjacent images provides a measure of the depth of that feature in the scene. In this way, a 3D reconstruction of the scene (or, at least, of the portions of the scene that are visible to both cameras) can be created.

Three cameras have been used in a variety of configurations to augment this depth image, with typical arrangements being a parallel
horizontal configuration (Okutomi & Kanade, 1993) or right-angled L-shaped configuration (Stewart & Dyer, 1988). These are useful primarily for obtaining better depth estimates in one of the images (i.e., the reference image; the one in the middle). However, when more accurate depth estimates are required for each image, and especially when viewpoints are spaced further apart, as is desired for our wide-baseline endoscope, stereo correspondence may be performed pairwise between sets of captured images.

Stereo correspondence can tend to fail in situations such as in surgery, where the visual scene is typically highly specular (shiny). Coping with such issues is an active field of research, with the best techniques making use of robust feature descriptors instead of using conventional stereo matching techniques (Stoyanov et al., 2010; Tola et al., 2010). Other means of coping with these issues include integrating additional sensor technologies, for instance laser range-finding, or by using structured light techniques to project patterns onto the surgical scene (McKinlay et al., 2004). However, as Scharstein (1999) describes, the problem of view synthesis is less constrained than the problem of full 3D reconstruction, as a novel view can still be synthesized and appear physically valid, even if depth estimates are inaccurate or imprecise.

2.3.2 View synthesis

The task of actually synthesizing a novel view is far simpler computationally than the task of depth estimation, and makes use of original captured images with associated estimated depth maps to reproject pixels as necessary to their correct locations in a novel view. A common issue in view synthesis is that depth and/or image information may be missing for portions of the scene that were not visible within the original views, but are expected to be visible from the new viewpoint. Depending on the application, some synthesis techniques choose to leave these regions of unknown image content blank, while others may make use of hole-filling techniques that blend in surrounding content to fill in these unknown regions. For a recent review of various view synthesis techniques, refer to Rogmans et al. (2009).

Consistent with the title of this dissertation—“Enhancing Spatial Awareness with Multi-Camera View Synthesis for Diagnostic and Manipulation Tasks”—I would like to reiterate the two principal components of the research performed towards that end:

1. I developed a real-time video interpolation software prototype, based on view synthesis with a prototype three-camera system;
2. I conducted an experiment to evaluate the potential utility of this prototype, using a pseudo-surgical perceptual task.

In writing the present dissertation, I have elected to focus in the remainder of the document, for reasons of both ease of reading and confidentiality\(^4\), on only the experimental evaluation. Technical details of the interpolation software can be found in *ETC Technical Document #1 (2012)*. For limited technical discussion describing the interpolation software actually used in the experimental evaluation, refer to Appendix A. The results of the evaluation experiment are presented in the following section.

\(^4\) My concern with confidentiality is due to the fact that, at the time of writing this thesis, an invention disclosure has been submitted to the University of Toronto.
3 Experimental evaluation

Due to constraints of the version of the view synthesis software that was deemed sufficiently robust for experimentation with human subjects, the decision was made to relinquish the need to carry out real-time manipulations, which would have necessitated high fidelity updates to continuously changing images in real-time, in favour of generating high fidelity updates to stored images in real-time. The main consequence of this decision was that any experimental paradigm selected, or contrived, would have to involve observations only—in other words, an experiment involving spatial perception. Furthermore, to investigate the potential for improvement in spatial awareness resulting from use of a manipulable viewpoint display in the context of endoscopic surgery, three possible experimental environments were considered: a real surgical environment, a simulated real environment, or a simulated virtual environment. Because our emphasis in the present research is on using real images, to maintain the inherent imperfections that arise when using view synthesis from stereo vision, we therefore eliminated the third option of completely simulated images.

In order to gather data from this evaluation that would be likely to provide generalisable insights, it was deemed necessary to maintain experimental control over the scene. This led to elimination of the first option of using a real surgical environment. As such, our clear choice for the present investigation was the second option: an artificial, but not computer-generated, scene. Nevertheless, in light of the constraints outlined above, further work to investigate the benefits of such a display in a real surgical environment clearly is recommended for the future.

Following consultation with two experienced laparoscopic surgeons, we proposed a generic relative-depth discrimination task based on the surgical need to interpret the topology of occluded tube structures. The task consisted of looking down from an above viewpoint at two perpendicularly arranged rods, and identifying which rod appears to be placed on top of the other—i.e., closer to the observer. To
render the task non-trivial, a circular object was suspended above the intersection of the two rods. With the rods placed very close together, the task was designed to be very difficult for a viewer who is presented with a single view of the scene. In keeping with our understanding of this particular surgical problem, the task was expected to be less difficult when the viewer was provided with at least one additional alternate view of the scene, taken from a different viewpoint. By varying the means of displaying a number of viewpoints, this task could be used to investigate the effects of a variety of viewing options. We chose to focus on only the motion parallax depth cue, although, similar to Sollenberger & Milgram (1993), this experimental paradigm could easily be used to investigate the relative advantages of multiple depth cues (e.g., stereopsis), including their interaction.

3.1 Study 1: Preliminary psychophysical calibration study

Initial experimentation raised the concern that, when using simple wooden dowels of identical width as rods, a viewer would be capable of too easily identifying the relative separation of the rods based solely on the observed widths (i.e., diameters) of the rods—that is, that the rod that appears wider must be the closer of the two. Because the focus of our study was to examine the potential advantages of any viewpoint control capability that could be provided by our view synthesis software (including motion parallax), we needed to minimize the potential strength of the relative width cue. This is especially important in the context of surgery, since for actual tubular vessels (that is, in lieu of our artificial rod configuration), the width cue is not reliable. Consequently, we conducted a psychophysical investigation to understand the sensitivity to the relative widths of identical rods at varying distances from the observer. This first study can thus be regarded as a calibration exercise, to determine the operating parameters of the latter experiment.

In anticipation of high sensitivity of our subjects to the dowel width cue, which would have made the experiment untenable for reasons outlined above, a solution was devised to reduce this sensitivity. As illustrated in Figure 3.1, our solution involved coating a dowel in plasticine, thereby perturbing its surface with noise. A preliminary ‘noisy dowel’ was created, with pilot testing revealing little apparent change in sensitivity due to the high spatial frequency and low amplitude of the surface variation. A revised noisy dowel was then sculpted, with lower spatial frequency and
greater amplitude of surface variation, as shown in Figure 3.1b.

Figure 3.1: a) An untreated 7/8” wooden dowel, and b) final version of ‘noisy’ 1/2” dowel coated in red plasticine

In the following we present the results of a comparative psychophysical study to determine the sensitivity of subjects to dowel widths within our experimental paradigm, for both varieties of rod—untreated dowels and ‘noisy dowels’. The objectives of this study were twofold:

1. To confirm the anticipated problem of not using noisy dowels—that is, the anticipated need, if non-noisy dowels were to be used, to place untreated dowels unrealistically close to each other (with separations potentially even less than one dowel width) in order to remove the use of width cues in the main experiment (study 2);

2. To derive a psychophysical curve for the noisy dowels, which could then be used for ‘calibrating’ the actual evaluation experiment using noisy dowels.

As a final tactic for reducing sensitivity to rod width, we also placed our noisy dowels with (non-horizontal) slopes, a.k.a. pitches\(^1\), as shown in Figure 3.2. This tactic effectively added a constant width variation (due to linear perspective, as each dowel sloped away from the camera) to the noisy width variation along the length of the rod, which diminished the ability to compare a perceived minimum width of a point along one rod to a perceived minimum width at a different point on the other rod.

\(^1\) A rod has three axes around which it can be rotated; ‘pitch’ is the slope of the rod away from the camera, while ‘roll’ refers to an orientation as defined about the long axis of the rod; ‘yaw’ refers to the overall orientation of the rod or configuration of rods as perceived from above.

3.1.1 Calibration study design

A single rod of each of the two rod types was used: one 7/8” diameter wooden dowel and one 1/2” diameter dowel coated in 1/4–1/2” red plasticine, as described above. With a single rod suspended
between two calibrated retort stands, a camera placed directly above was used to capture images of each rod, as depicted in Figure 3.2.² Prior to the experiment, images were captured of each rod placed at 7 different distances from the camera (shown by the double arrows in the figure), and at varying orientations (specifically, varying roll, with consistent pitch and yaw) to prevent participants from realizing that the same rod was being used each time.

We recruited 4 male student participants, aged 21, 22, 23, and 23, who volunteered for the study. Each participant was presented with (pre-recorded) images on a computer screen showing one type of rod at a time, and each participant experienced both types of rod. Participants sat at a viewing distance of approximately 50 cm from the screen with minimal head movement. Though they did not witness the actual experimental setup, we explained to participants that they were looking down upon two rods suspended above a black table top, and explained the nature of each of the rods (i.e., the materials they were constructed of and whether or not they were sloped). Using the Method of Constant Stimuli (refer to Gescheider (1985)), participants were instructed to identify which rod appeared closer to them in each trial, by pressing ‘q’ or ‘p’ on a computer keyboard to indicate a selection of the correspondingly labelled rod (see Figure 3.3). Image presentation was vertically flipped in each successive trial to prevent carry-over effects from adjacent trials.

Image presentations included 7 different rod depth pairings, with

² The apparatus used was a simplified version of the apparatus used in our subsequent study, which is shown in Figure 3.9.

Figure 3.2: Illustration of experimental setup for the preliminary calibration study, with one rod supported at a variable height and photographed by an overhead camera
Figure 3.3: Screenshots from psychophysical calibration experiment, presenting a) untreated horizontal wooden rods; b) noisy sloped plasticine-covered rods.
a control rod always appearing at a middle depth, and the other rod appearing an equal number of times at each of 7 depths. With depths \(\{1,2,3,4 \text{ (control)}, 5,6,7\}\), rod depth pairing in any given trial could be either \(\{1,4\}, \{2,4\}, \{3,4\}, \{4,4\}, \{5,4\}, \{6,4\}, \text{ or } \{7,4\}\). Thus, rods were sometimes presented at equal depth, and higher depths were presented equally as often as lower depths. Presentations were flipped at random, such that the middle depth rod appeared equally often on the left as on the right. Each participant was presented with each depth pairing 24 times, for a total of 168 trials with each variety of rod. Presentation order was counterbalanced between participants.

3.1.2 Calibration study results

According to the method of constant stimuli (Gescheider, 1985), we aggregated performance results for each participant over each of the 7 separations. Figures 3.4a and 3.4b present the percentage scores obtained by each of the 4 participants for each of the stimulus levels and for each of two varieties of rod. When the rod separation was positive, the control rod was farther from the camera than the non-control rod, so a response score of 1 indicates perfect response accuracy; conversely, when the rod separation was negative, the control rod was closer to the camera than the non-control rod, and so a response score of 0 indicates perfect accuracy. Those percentage scores were converted to z-scores; on the basis of a cumulative normal distribution model, a straight line was fit to each participant’s z-score data for each rod variety. The quality of those fits ranged from \(R^2 = .815\) to \(R^2 = .969\), with mean \(R^2 = .89\). Each of the linear models for each participant’s data were then transformed back to the probability domain and were plotted as ogive curves, as shown in the figures.

For each type of rod (horizontal untreated, and sloped ‘noisy’), response proportions were aggregated across participants at each separation and a linear fit was produced. These fits yielded \(R^2 = .968\) and \(R^2 = .974\) for untreated and noisy rods, respectively. They were then transformed back to the probability domain to yield overall depth sensitivity models as ogive curves shown in Figure 3.5.

Results indicate higher sensitivity to the width cue with untreated horizontal wooden rods (steep lighter curve in Figure 3.5) relative to the noisy sloped rods (shallow darker curve in Figure 3.5). Interpreting these graphs: with untreated horizontal rods at a separation of 2 cm, for example, our model results predict that participants would be able to correctly identify the rod that is closer to them on average 83% of the time; with noisy rods, this estimate
20. Enhancing spatial awareness with multi-camera view synthesis

Figure 3.4: Individual response scores at various rod separations, for a) untreated horizontal rods and b) sloped 'noisy' rods; participants 1-4 are identified by colour, and response scores account for the proportion of responses a participant made identifying that the non-control rod was closer than the control rod.
Figure 3.5: Psychophysical models of sensitivity to relative depths of two different varieties of rod, untreated horizontal wooden dowels and noisy angled plasticine-covered dowels; the black horizontal line at a proportion of 0.5 indicates the performance results that would be expected if the effect of the width cue were to be entirely eliminated.
drops to 64% of the time. By slightly increasing the untreated rod separation to 4 cm, a participant should be able to identify the closer rod 97% of the time based on the width cue alone, whereas with noisy rods, this would be only 76%.

The conclusion of this sensitivity study was that, to minimize the extent to which a participant may be able to use the width cue in identifying the relative depths of rods, it was decided that:

1. ‘Noisy’ and sloped plasticine-covered rods would be used, rather than untreated horizontal rods.

2. Rod separations greater than ~4 cm (corresponding to greater than 75% accuracy) would be avoided, to allow sufficient latitude for measurement of further performance gains afforded by alternate depth cues.

In conclusion, based on results from our preliminary psychophysical calibration study, we resolved to conduct our three-camera relative-depth discrimination task with noisy sloping rods at a separation of 4 cm. In addition, to minimize any ceiling effects, we introduced a second separation condition of 2 cm, which was deemed to be extraordinarily difficult, regardless of the viewing condition.

3.2 Study 2: Tri-camera display study

While the calibration study dealt with single stationary viewpoints, many other factors needed to be considered for the tri-camera display study. Before we delve further into the design of the artificial scene, we first discuss the variety of ways in which an operator may interact with multiple viewpoints. The pressing experimental conundrum was, when granted access to a given set of viewing angles of a scene, an operator is provided with a fixed amount of information regardless of the display configuration or nature of the control they have over the views. While view interpolation affords an operator with differently encoded information about the correspondence between image features (presented as optical flow between features), that same correspondence information is technically available to the operator also when visually inspecting features across multiple images without any synthesized motion between them. As such, the primary question addressed was whether active manipulation of the display configuration and control mechanisms can improve operator performance in our relative depth discrimination task solely by rendering the information in a different manner.

3 Due to perspective scaling, a scene with 2 cm-diameter rods separated by 2 cm, at 50 cm away from the camera produces the same display geometry as a scene with 0.2 cm-diameter rods separated by 0.2 cm at 16 cm away from the camera. Though not a factor in this calculation, it is worth noting that the wide angle lens used for this experiment had a diagonal field of view of 92°. In other words, an effort was made in designing the scale of the present experiment to ensure that the magnification, field of view, and dimensions of the visual images observed by participants corresponded reasonably closely with the corresponding dimensions of objects that a laparoscopic surgeon might encounter in an actual operative display.
3.2.1 Display conditions

For the present case of three cameras providing three distinct viewpoints, the choices that immediately sprang to mind as being feasible in a practical surgical setting were:

a) Presenting all three viewpoints at the same time, on separate screens;

b) Presenting one viewpoint at a time, on one screen, and providing the surgeon the means to switch between viewpoints at will;

c) Picture-in-picture: Presenting one primary viewpoint at a time, but with one or more additional viewpoints presented at a reduced size in the corner(s) of the display.

An additional display choice, and the one that we hypothesized would provide the most useful depth perception, was the manipulable viewpoint display. Of particular interest with this display, when constructed by interpolating between viewpoints using image-based rendering techniques, is that it makes use of and presents the operator with an identical set of visual information as the viewing conditions a), b), and c) listed above. However, as with many visualization technologies, the key benefit is not just to present disparate information sources, but also to render them in a form suitable for the human operator. In other words, the key question is: can performance be increased by synthesizing visual momentum between three disparate views (to allow a sensation of visual momentum between views), as compared to allowing human operators to synthesize information from the disparate views on their own?

![Diagram of display conditions](image)

We chose to investigate a total of 4 display conditions, as illustrated in Figure 3.6. Condition 0, as a control condition, presents a monocular view of the scene, captured from a real (non-synthesized) viewpoint in the centre of the three peripheral
viewpoints. For purposes of the present experiment, this centre viewpoint was actually created by shifting the experimental scene (relative to the cameras) by a few centimetres, not the viewpoint, so as not to interfere with the precise camera calibration. Condition 0 corresponds to the status quo of how a surgeon would perceive the surgical scene with a standard monocular endoscope, and provides less visual information than in conditions 1, 2, or 3.

Condition 1 presents a trinocular view of the scene, with all three viewpoints presented simultaneously, with images arranged on the screen analogously to the configuration of the physical cameras. Condition 2 presents one view at a time, with the operator in control of selecting which viewpoint to view at any given time. Condition 3 is a motion parallax display: a single view is presented, but the operator is free to continuously adjust the virtual viewpoint to anywhere in between the three captured viewpoints. (Screen shot examples of these four display conditions are presented in Figure 3.7.)

The shape of the explanatory icons in Figure 3.6 for conditions 2 and 3 relate to the shape of the input device used by participants—a graphics tablet modified to constrain the user to a triangular region of input, as shown in Figure 3.8. Control was provided to participants (instead of more controllably presenting predefined viewpoint movements) because we postulated that control is an important factor in understanding perceived between-viewpoint motion. (I was also not aware of any accepted optimal strategy for such a method of viewpoint manipulation.) Tablet input positions were sampled at 30 Hz for subsequent analysis.

---

4 Conditions 1 and 2 correspond to display options a) and b) above. For the sake of simplicity, we chose not to investigate the picture-in-picture option (c), as it would be too similar to condition 2. In any case, the entire set of possible sub-image placements and sizings would warrant a separate investigation on its own.
The modified graphics tablet was chosen to provide approximately isotonic absolute position control over the viewpoint, matching the human instinct for natural viewpoint manipulation by head movement as closely as possible. No indication of input device movements was provided on the screen, since we believed that this would have been distracting, as any cursor motion would have been opposite to the direction of movement of objects in the viewed scene.

3.2.2 Hypotheses

With the task of classifying relative depths of rods in an observed scene, our expectation was that participants would achieve greater accuracy with greater rod separation. Further, we expected participants to achieve successively greater accuracy in each subsequent display condition; in particular:

• In condition 0 there are essentially no depth cues available to the operator, and chance (50%) performance is expected, assuming no use of the relative width cue. (Results from our calibration study (see above) predicted that operators could, on average, achieve up to 64% accuracy with a 2 cm rod separation, or 76% accuracy with a 4 cm separation, if they were to base their responses on the relative width cue.)

• Condition 1 provided additional scene information as captured

Figure 3.8: The graphics tablet used by participants as an absolute position control input device to control the viewpoint in display conditions 2 and 3. Participants slid the pointer by grasping its weighted tip

\[\text{\textcopyright 2023 by the author. All rights reserved.} \]
by disparate viewpoints, such that very careful examination of the relative displacement of scene objects could have provided some information about relative depths.

- Condition 2 was expected to result in higher accuracy than condition 1, simply because visual distance between image features in any two viewpoints was reduced by presenting all images in the same location on the display, using temporal separation. Perhaps more importantly, by rapidly switching between viewpoints, an operator could conceivably generate some semblance of motion parallax.

- In Condition 3, the fluid transition between viewpoints was expected to generate motion parallax that resembled actually ‘looking around’ the scene, leading to the hypothesis that accuracy in this condition would be the highest.

With regard to time taken to examine the images and respond, we expected lower response times for the 4 cm separation conditions relative to the more difficult 2 cm conditions. However, with each increase in the amount of information provided and ability to control the viewpoint—that is, in progressing from Condition 0 through to Condition 3—it was hypothesised also that participants may wish to spend more time exploring the views available to them.

3.2.3 Experimental setup

The complete experimental setup used for this experiment, as outlined at the beginning of this chapter, is depicted in Figure 3.9a, comprising a triangular configuration of webcams looking down at an occluding circle suspended above two sloping ‘noisy’ plasticine-coated wooden dowels, which in turn were suspended above a flat textured surface. Figure 3.9b shows the experimental scene from the point of view of an above camera.
Figure 3.9: The experimental setup of the artificial scene from a) the side and b) a sample image obtained from one of the cameras.
Textured backdrop  Because of limitations with the stereo matching algorithm used to interpret images of the scene for the purposes of view interpolation, it was deemed unsuitable to provide a uniform background (such as a simple monochrome backdrop) or a background with repeating patterns (such as a checkerboard) for the visual scene. Additionally, neither uniform nor repeating patterns are typically found in surgical contexts. We therefore elected to use a textured surface as a background. Our objective was to create a pseudorandom texture with minimal regions of non-uniform intensity for the sake of adequate performance of the stereo matching algorithm. Because it was not feasible to generate this pattern objectively, we compromised by subjectively applying layers of spray paint, alternating between black and white, with progressively decreasing sizes of features in successive layers. The resulting pattern is depicted in Figure 3.10. The same background surface was used for all images presented during the experiment.

Rod orientation  Two wooden rods were coated in plasticine as previously described for the design of the calibration experiment, with diameter varying between $\frac{1}{4}^\prime$–$\frac{3}{4}^\prime$. Rods were suspended about 30 cm above the textured surface at a 10° slope (pitch), with the top rod either 2 or 4 cm above the lower rod at the point of intersection. The rods were oriented so as to appear at right angles to each other from the viewpoint of the overhead cameras (as shown in Figure 3.11), to avoid interfering with participants’ perception of the continuity of each rod, as would be predicted by the Poggendorff effect.6

6 The Poggendorff effect is a powerful illusion whereby the ends of a slanted line that is partially covered by an occluding surface will be perceived as disconnected from each other (Poggendorff, 1863).

Figure 3.10: The textured surface appearing in the background of the experimental scene
The configuration of the rods was varied between trials in three ways, to prevent participants from improving over the course of the experiment by learning to recognize characteristics of individual rods or subtle difference cues that emerge as carryover effects between successive trials. For example, it would be inappropriate for a participant to perceive a near-identical scene in two successive trials, where the only apparent change is the relative depths of the rods; in such a case, the rod that appears to suddenly decrease in width upon switching to the second trial is more likely below the other rod.

The first variation entailed switching among three overall yaw rotations: while maintaining perpendicularity, the rods were together rotated around their yaw axis, as shown in Figure 3.11, for overall angles of 15°, 45°, or 75°. These angles were chosen so that the rods would be as far as possible away from having any rod be parallel to a viewpoint-switching direction: having one of the rods parallel to a viewpoint-switching direction would potentially have deceived participants into perceiving that only the other rod was moving when the viewpoint was changed (in Conditions 2 and 3), thus potentially biasing the response towards concluding that the other, apparently moving rod was on top, or closer. This bias was further avoided by also ensuring that the rods were sufficiently textured (with the plasticine), so as to provide salient feature points that appear to move even when overall a particular rod did not move significantly in the direction orthogonal to its length.

![Figure 3.11](image)

Figure 3.11: Three overall (yaw) orientations of rods, used to vary the appearance of the scene during the experiment

The second variation entailed manipulating the pitch direction of each of the rods—each sloped either 10° or -10° away from the horizontal—that is, the longitudinal axis of the camera system (see Figure 3.9a)—for a net total of 4 possible sloping configurations.

The third variation involved rotating each rod about its longitudinal roll axis, to present a different surface appearance. With two possible roll angles for each rod, 4 possible presentation combinations were possible, as shown in Figure 3.12.
**Occluding circle** The occluding circle used to cover the intersection region of the two rods was constructed from a mixture of white and red plasticine, with a diameter of about 8 cm. As depicted in Figure 3.13, the circle was designed to be somewhat distinct from the visual appearance of the rods, while remaining similar enough to reasonably be imagined as a volumetric structure that the rods could be passing through—that is, as if the circle was the top surface of an organ or tumour.

In each case, the circle was placed upon a short pillar, 6 cm above the top rod. There were two reasons for not placing the circle directly on the top rod: to avoid having the circle cast any distinctive shadows upon either rod, and to avoid having the circle appear at the identical depth as one of the rods. Care was taken also to consistently place the circle directly above the intersection of the rods.
rods at a consistent orientation relative to the camera, with the plane of the circle parallel to the textured background surface. The circle suspended above the two rods is shown in Figure 3.13.

**Cameras** Three PS3 Eye™ cameras were used for the experiment,7 configured as an equilateral triangle, with a distance of 8.8 cm between the optical centres of each pair of cameras as shown in Figure 3.14. The cameras were equipped with 2.8 mm lenses, to provide diagonal fields of view (FOV) of 92° each,8 equivalent to a typical laparoscope and providing a broad depth of field.

The camera plane was set 50 cm above the intersection of the rods, a distance that was chosen to ensure consistent camera focus within the range of depths the rods appeared at. Overall, the geometry provided images in which the rods appeared as though they were slightly separated narrow tubes being viewed with high magnification, as might be the case in a surgical task. (See, for example, Figure 3.16.) Placing the rods at this distance allowed for fine control over the rod spacing, as well as control over the rod surface variation while avoiding fine details such as shadowing that might bias responses by inadvertently providing additional cues. More importantly, the geometry of the present experimental paradigm was designed not to be limited to the surgical analogy, but also to provide an easy to control set of stimuli that provide a difficult yet accomplishable binary response task.

7 In other words, the endoscope prototype shown in Figure 2.3 was not used for this evaluation experiment in order both to accommodate construction of the target stimuli at an appropriate scale and to facilitate calibration of the cameras.

8 Note that after cropping out unwanted portions of each image for the purposes of the experiment, as illustrated in Figure 3.15, images presented to participants provided a field of view of approximately 60°, much narrower than that originally captured by the cameras.

Figure 3.14: Equilateral triangular configuration of three PS3 Eye™ webcams with near-parallel viewing directions; cameras are spaced apart by 8.8 cm
Image processing  All images were captured in advance of the experiment. As shown in Figure 3.15, the three untreated camera views overlap. Because it is only in the overlapping regions that depth information can be reliably extracted, users were presented with a subset of images chosen from the overlap in the image contents of the three cameras. Images were also slightly offset to provide an appropriate horopter set at the depth of the background surface.\(^9\)

\[^9\text{i.e., the background textured surface in the three images was aligned such that when switching between images, the background does not appear to change. Note that, e.g., an alternative to setting the horopter at the background could have been for the occluding circle to appear in the same location in each image, in which case the images would be spread further apart.}\]

For the case of display condition 3, where the viewpoint can be manipulated to any interpolated view between the three existing views, a real-time forward warping GPU-accelerated view synthesis algorithm was used, as described in Appendix A. Image rectification and stereo correspondence were performed offline in advance of the study using a high-accuracy algorithm as described by Rhemann et al. (2011). The resulting synthesized images were highly accurate, though not without some slight ghosting artefacts, especially at depth discontinuities. For further information about the software used to process and present images in this experiment, consult Appendix A. Refer to Figure 3.16 for a sample synthesized view, interpolated between the left and top images.

3.2.4 Experimental presentation

In total, 48 experimental scenes were configured and captured. Each image set consisted of 4 images: left, right, top, and centre, to be
Figure 3.16: a) A synthesized viewpoint between the viewpoints of the physical cameras; b) a close-up of the synthesized viewpoint, revealing subtle ghosting artefacts as shown with arrows.
presented as necessary depending on the display condition. The primary factors under investigation were Display Condition and Rod Separation.

Each of the 48 experimental scenes was shown once with each display condition, for a total of 192 trials performed by each participant. Trials were grouped into 4 blocks of 48, with an equal number of trials from each display condition presented within each block. Within a block, each display condition appeared as groups of 12 consecutive trials. The order of display condition groups within each block was counterbalanced with a latin square. Further, the ordering of blocks within the experiment was counterbalanced between participants, to account for effects of both learning and fatigue. Additional scene variation (varying pitch, yaw, and roll) was semi-randomly balanced amongst trials, with no two identical rod orientations appearing in a row, to avoid carryover between trials.

Participants sat at a standard computer workstation with a 24” LCD monitor, and were free to adjust their sitting positions as best suited them. They were requested to “identify which rod appear[ed] closer [to them]” in each trial, and they responded by pressing the ‘A’ or ‘D’ buttons on a keyboard to select the correspondingly labelled rod on the screen. Figure 3.7 depicts the appearance of the experimental interface, as seen by participants.

The experiment was preceded by a quick introduction and training session, comprising viewing the experimental scene in person (see Figure 3.9a), and subsequently following on-screen instructions to understand the task with its 4 display conditions. Participants completed 12 training trials10 prior to commencing the 192-trial experiment, and filled in a demographics / subjective feedback questionnaire upon completion (see Appendix C). For the actual experiment, participants did not receive any advice about what cues to look for, though they were given sufficient information to understand that the width cue was to be treated as unreliable, as implied in the phrasing: “Each trial you will see different rods, of variable width”. Participants did not receive feedback about their response accuracy performance during the training or experiment.

3.2.5 Participants

We recruited 16 male11 participants from the University of Toronto campus via email and posters, with all participants either undergraduate or graduate students. Participant ages ranged from 19 to 41 (Mean = 25.7 years, SD = 6.0). All participants were right-handed, and 8/16 participants wore corrective lenses. With each of

---

10 The 12 training trials consisted of 3 trials from each display condition, but with only the larger (4 cm) rod separations to better emphasize the available depth cues while learning.

11 Only male participants were chosen, to control for gender differences.
the 192 trials tending to last about 10–15 seconds (no time limit was imposed), the entire participant experience, including training, lasted approximately one hour. Participants were compensated $15 for their participation.
4 Results

A within-subjects repeated measures ANOVA was used to investigate the effect of display condition on response accuracy, in terms of percent correct. There was a main effect of display, $F(3, 45) = 68.25, p < .001$, and rod separation, $F(1, 15) = 24.25, p < .001$. There was no interaction found between the two factors (although there is a slight indication that each successive display condition afforded a greater increase in performance between rod separations). Sample means with standard error bars are presented in Figure 4.1a; while medians, quartiles, and sample distributions are presented as a violin plot in Figure 4.1b.

Within-subjects contrasts among pairs of successive display conditions revealed a large significant effect only between conditions 2 and 3, with $F(1, 15) = 87.97, p < .00001$.

For tables of participant response accuracy results, consult Appendix B.

Because participants were provided with unlimited time to complete each trial, the median\textsuperscript{1} response times of each participant were also treated as a performance measure. A within-subjects repeated measures ANOVA revealed a main effect of display, $F(3, 45) = 31.27, p < .001$, and of separation, $F(1, 15) = 5.75, p < .05$, and an interaction between display and separation, $F(3, 45) = 3.74, p < .05$. Within-subjects contrasts revealed no significant effects between pairs of successive display conditions, except a difference between conditions 0 and 1, $F(1, 15) = 52.18, p < .00001$. Sample means are presented with standard error bars in Figure 4.2.

4.1 Analysis of subjective preferences

(Consult Appendix D for an in-depth discussion of the analyses discussed in the previous section, as well as further insight to be gained from both subjective feedback and viewpoint manipulation tendencies, especially as they relate to measured task performance.)

\textsuperscript{1} With response time data, we aggregated median values for each participant, to account for outlying cases where participants were anomalously distracted.
Figure 4.1: Response accuracy results for each display condition and rod separation, with a) means and standard errors; b) probability density estimate (violin) plots indicating medians and interquartile ranges.
Figure 4.2: Median response time means and standard errors for each display condition and rod separation.
Subjective feedback about each display condition was obtained from participants via a questionnaire, in which they were asked to describe their impressions of each condition and the strategies they employed, as well as rank the conditions by difficulty of use (refer to Appendix C for the questionnaire). In addition to self-reported subjective strategies, the usage tendencies exhibited by participants when given control over the camera viewpoint in conditions 2 and 3 were also analyzed.

Subjective difficulty rankings reported by participants are presented in Table 4.1, with locations in the table indicating difficulty ranking, and colouring of participant numbers indicating the relative response accuracy (rank) achieved by that participant in that condition, compared to the response accuracy achieved by that participant in the other display conditions. Colouring (with varying intensities of red) is provided to allow comparison between perceived difficulty and measured performance in each condition; light red indicates lowest accuracy rank, while black indicates highest accuracy rank. Thus, for any given display condition, the emergent feature of the proportion of participants that appear in higher rows denotes the perceived difficulty of that condition; the emergent feature of the overall darkness of each column denotes the response accuracy achieved in that condition.

<table>
<thead>
<tr>
<th>Rank Score</th>
<th>Display Condition 0</th>
<th>Display Condition 1</th>
<th>Display Condition 2</th>
<th>Display Condition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P9</td>
<td>P1 P16</td>
<td>P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>P2 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15</td>
<td>P1 P9 P16 P14 P15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>P6 P15</td>
<td>P1 P10 P11 P12 P13 P14 P15</td>
<td>P2 P4 P5 P8 P10 P11 P12</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16</td>
<td>P9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Subjective rankings of display condition difficulties, with a ranking of 1 indicating a low difficulty, and 4 indicating a high difficulty. Participant numbers (P1 to P16) are colour coded, based on response accuracy performance rank, to provide a comparison between perceived difficulty and measured performance scores. Colour-coding is further explained in the main body of the text.

Participant rankings of display difficulties revealed a low perceived difficulty for condition 3, and high perceived difficulty for condition 0. Overall, participants preferred condition 2 to condition 1, but with less agreement (11 out of 16 preferred condition 2 to condition 1). This was evaluated with a Friedman’s ANOVA, revealing an overall significant difference in ranked difficulties between conditions, with \( \chi^2 = 31.35, p < 0.000001 \). Wilcoxon tests between successive display conditions revealed that condition 3 was rated as less difficult than condition 2, with \( T=11, p=.002 \), and that
condition 1 was subjectively ranked as less difficult than condition 0, with $T=111.5$, $p=.02$.

Participants provided written feedback about their impressions of each display condition. They generally found condition 0 to be most difficult due to the lack of available depth cues, conditions 1 and 2 to be less difficult because of the increased availability of depth cues, and condition 3 to be least difficult because the “sense of movement” it provided helped; only in condition 3 did the majority of participants report using the relative displacement of the rods and occluding circle as a depth cue. Participants who preferred condition 1 over condition 2 reported the benefit of having all information available at the same time, while participants who preferred condition 2 were averse to the effort involved in looking between images that were physically separated on the display monitor.

Response behaviours extracted from recorded cursor movement data (see details in Appendix D) in conditions 2 and 3 revealed a tendency to switch between the left and right views, and less often switch to the top view. This preference may be an artefact of the nature of human binocular perception, but it would be interesting to investigate a non-triangular camera configuration (i.e., a 4-camera rectilinear configuration) to determine if participants would be more likely to make use of the vertical dimension when not confined to a triangular control region.

With condition 3, participants generally reported preferring faster viewpoint movement speeds (rather than slower), though no correlation was found between movement speed and response accuracy. A strong significant correlation was found between response accuracy and time spent manipulating the viewpoint, Pearson’s $r = .516$, $p < .05$; indicating that participants who spent more time observing the motion between viewpoints were able to achieve higher accuracy, while spending more time observing unmoving intermediate viewpoints was not shown to have an effect on accuracy. Finally, circular motion tendencies were more prevalent among participants with higher response accuracies, while limited directional tendencies were more prevalent among participants with lower response accuracies.

1 This phrasing was used by one participant, though other participants described the same phenomenon in different words.

3 Examples of limited directional tendencies (which are further described in Appendix D) include motion only in a horizontal direction, and motion only along the triangle edges in a clockwise manner.
5 Discussion

The clear conclusion from both the objective and subjective data obtained from our simulated estimation task is that the motion parallax provided by viewpoint synthesis (condition 3) was most effective in enhancing depth perception performance, and thus holds good promise for facilitating understanding of remotely observed scenes.

Referring to our calibration experiment, which examined participants' sensitivity to rod separations using only rod diameter as a cue, we determined that for noisy rods at separations of 2 cm and 4 cm we should expect depth estimation performances of approximately 64% and 76% respectively (refer to Figure 3.5). For our control condition 0 in Study 2, however, the corresponding accuracy results were 46% and 52% (i.e., chance performance) for the 2 cm and 4 cm separations, with no performances higher than 58% and 71% respectively. This result suggests that our preliminary psychophysical calibration effort in Study 1 was indeed effective in guiding our decision in Study 2 to use ‘noisy dowels’ and rod separations of 2 cm and 4 cm; however, a greater factor in encouraging participants to avoid use of the relative size cue was likely the instruction that the various rods were “of variable width”. In other words, we were apparently successful in deterring participants from relying on only the width cue to come to a conclusion about the relative depth of rods in the experimental scene.

While response accuracy performance data did not show differences between display conditions 0, 1, and 2, there is nevertheless some indication in Figure 4.1a that a difference among these conditions might occur with larger rod separations. As such, if the relative advantages of these conditions are to be conclusively determined, a wider rod separation should be used.

The long tails of the estimated density functions for the 4 cm results in Figure 4.1b indicate the occasional high performance
score for conditions 1 and 2, relative to condition 0. In other words, a few participants managed to achieve high scores (88–96%) in display conditions 1 and 2. Subjective feedback revealed that those participants explicitly searched for relative displacements between rods and the occluding circle. This stands in contrast to most participants who reported relying on their instincts and tended to perform at chance or worse. Performance below chance indicates that participants tended to provide wrong responses more often than correct responses, a result attributed to available, albeit misinterpreted, depth cues. While further training could teach participants to avoid misinterpreting cues, and convert below-chance scores to corresponding above-chance scores, no participant performed with accuracy in conditions 0, 1, or 2 low enough to indicate that training could rectify his performance to exceed his performance with condition 3.

To understand how those participants were able to perform at such high levels, we must scrutinize the visible differences between any given pair of left and right camera views taken of the same scene, independent of any perceived motion. As shown in Figure 5.1, due to the closer distance to the viewer of the occluding circle compared to the rods below it, the extent of the displacement of the occluding circle between the left and right images is more than the displacement of the rods beneath it. Similarly, because one of the rods is closer to the viewer than the other rod, it will shift more than the other (further) rod (although still less than the occluding circle above). The result is that, as the circle above moves, a smaller portion of the closer rod is revealed and a larger portion of the farther rod is revealed. This is illustrated in Figure 5.2, which depicts the regions of the rods that are only visible in one image and not the other. In this case, it is apparent that rod A is on top of rod D because less of it is revealed by the displacement of the occluding circle. Interestingly, this overlay is equivalent to the shadow that would result from placing a light source at the location of the other camera.

An alternative (or additional) cue involves observation of the rod displacement relative to the background texture. This disparity may be perceived either consciously (in conditions 1, 2, and 3) or unconsciously (in conditions 2 and, especially, 3) by observing the relative movements of the rods and circle.

The ordering of the response times for the various conditions supports our hypothesis that participants would take less time to respond in display condition 0 because less information was available. This was in spite of the fact that they reported that condition 0 was a more difficult task; this suggests that they
Figure 5.1: Two images of the same scene (with rods separated by 4 cm), as captured by the left and right viewpoints of our triangular camera configuration.

Figure 5.2: A transparent blue overlay on each image reveal regions that are not visible in the other image due to occlusion; circled regions reveal smaller occluded regions on rod D compared to rod A in each image, indicating that rod D is above rod A.
preferred to rely on their gut feeling rather than spend excessive time on those difficult trials, for which they felt that excessive scrutiny of the images would be unrewarding. For conditions 1, 2 and 3, on the other hand, increases in the response times seemed to be dependent on the nature of the control and the sensation of having fully explored the information available.

**In summary**, despite the fact that some participants did manage to successfully make use of only relative rod displacement information evident on the screen, the primary message to be derived from the results of Study 2 is that for the majority of participants display conditions 1 and 2 did not afford an understanding of the visual scene that was sufficient to allow them to achieve performance scores comparable to those of condition 3, in which motion parallax was present. (In particular, the majority of people performed slightly worse than chance on conditions 1 and 2, indicating that they were so confounded by what they were seeing that they performed worse than if they had simply guessed randomly!)

### 5.1 Limitations and future work

If future investigations of remote viewing wish to make use of a similar relative-depth discrimination task, and especially if the use of ‘noisy-edged’ objects is desired, it would be worthwhile to expand our calibration experiment, a psychophysical investigation of sensitivity to the relative width cue, by including varying degrees of surface noise. In our calibration experiment, we also did not distinguish between the relative width cue reduction benefits of a) rod ‘noisiness’ and b) rod slope.

In our tri-camera experiment, the triangular nature of the camera configuration may have influenced participants’ movement tendencies. The tendency to move along the edges of the triangle (with 60° diagonal lines) is an arbitrary directional choice that may have biased participants to see one rod being displaced less than another rod, because one rod is closer to orthogonal to the moving direction and the other is closer to parallel to it. We accounted for this in two ways:

- We gave participants control over movement within the entire triangle and in any desired direction;

- The overall orientation (yaw) of the rods was selected appropriately, such that rods appeared as far away as possible from being orthogonal or parallel to the edges of the triangle.  

Note that even those participants who performed well above chance in conditions 1 and 2 still performed even better in condition 3.

Orienting any given rod 30° away from being parallel or orthogonal to an edge of the triangle would cause the other rod to become either parallel or orthogonal to another edge of the triangle. Consequently, the appropriate angles are 15° between erroneous angles.
Movement tendencies of at least one participant (P3—refer to Appendix D), and subjective feedback from another (both of whom achieved higher accuracy scores than most in condition 3), revealed a preference to move the viewpoint in directions parallel to one of the rods in the scene. However, the majority of participants appeared to limit themselves primarily to movement parallel to an edge of the triangular configuration, and some even more restrictively limited themselves to horizontal movements. This may have impaired the performances of some. Further investigations with, for example, a rectilinear configuration (with four cameras) may have the potential to prevent participants from making arbitrary choices of viewpoint movement direction.

Given a system that can synthesize novel viewpoints as reliably as that used in the present investigation, it should be possible to generate instead of a single viewpoint, a pair of viewpoints, that can be used for stereoscopic presentation to each eye of the human operator. A comparison of the relative benefits and the interaction between stereoscopic viewing and viewpoint movement for such a task is recommended for future investigations. The effects of these two depth cues are likely additive, as found (in different context) by Sollenberger & Milgram (1993).

As described in Section 3.2.3, and with Figure 3.15, images may be re-aligned to create any desired horopter, or plane of zero disparity. (In other words, images may aligned such that all pixels at any single chosen depth appear in an identical location in all images.) In our experiment, images were aligned such that the horopter coincided with the plane of the textured background surface. This caused the background texture to be constant for all images captured, in spite of the different viewpoints. Ordinarily, without post-processing image alignment, the horopter is located at infinity for the case of parallel camera viewing directions; however, the consequence of a horopter at infinity is that only objects in the distant background will appear unchanged as viewpoint images vary. For the purpose of our relative-depth discrimination task, setting the horopter to a further distance would have resulted in exaggerated displacements of objects in the foreground. On the other hand, setting the horopter to a closer distance, such as to coincide with a depth near the intersection of the rods, would have resulted in a smaller displacement of the foreground objects between images. Whether the choice of horopter has an effect on a human operator’s ability to perform a relative-depth discrimination task such as that examined here remains to be seen, and is recommended for further study.5

It is interesting to note that the majority of participants tended to tilt their heads throughout the study, apparently to compensate for

5 In a surgical setting, it may be useful to set the horopter at a standard distance, either around the far wall of the abdominal cavity or towards the region where manipulation tasks are likely to take place. It may also conceivably be useful to provide the surgeon with control over the horopter.
the variable overall rotations of the rods on the screen. This suggests that rotation may be a desirable control axis for the interaction viewpoint; an additional mode of control that may warrant further investigation.

Further to the above point, the graphics tablet used as a viewpoint control input device in this study would be an unlikely candidate for use in an actual surgical setting. Instead, a more natural means of controlling a virtual viewpoint might involve a direct mapping of head position to viewpoint location. Such a head-tracking input mechanism would not only be practically feasible, but could provide hands-free operation, which is an especially important constraint in the sterilized environment of an operating room. Such an input mechanism may also be especially beneficial for cases in which a surgeon’s hands are fully occupied, such as suturing.

Finally, our investigation of manipulable viewpoints using multiple camera inputs has been limited to the specific case of a relative-depth discrimination task. While we expect that analogous benefits will arise for a variety of tasks requiring spatial awareness of a scene, including tasks requiring manual inputs, further investigations are necessary to demonstrate this, and especially to measure the extent of such benefits for each such task.

---

6 Care would have to be taken in designing such a setup, however, to ensure that the display system used is compatible with the head-slaved viewpoint control. In particular, using a fixed display screen (with rotation direction compatible with the surgeon’s ‘looking around’ inputs) might become tedious, whereas using a head-mounted display might be considered cumbersome.

7 Our expectations for extrapolating to motor control tasks are fueled by the research of Taffinder et al. (1999) and Birkett et al. (1994) on suturing tasks performed with enhanced depth perception via stereoscopic displays. Although Birkett et al. (1994) found speed improvements for suturing tasks, they found only subjective preference for enhanced depth displays in a large variety of less complex surgical tasks. To that end, the work of Derossis et al. (1998) may provide useful criteria for classifying laparoscopic skills and objectively evaluating performance on a variety of surgical tasks.
6 Conclusions

The objective of the present work was to evaluate the benefits of a motion parallax display achieved using view synthesis with multiple cameras, particularly for use in the context of minimally invasive surgery. To evaluate such a display, I proposed a new experimental paradigm comprising an occluded relative-depth discrimination task, whereby participants are expected to determine which of two intersecting rods appears above the other. This task was inspired by the occurrence of occluded tube structures in laparoscopic surgery. To isolate the motion parallax depth cue in this task, an occluding surface was used to preclude participant use of the interposition cue, and ‘noisy-edged’ rods were used to prevent participants from using the relative size cue.

I used this paradigm to demonstrate a large performance improvement among a sample of 16 participants when using a manipulable viewpoint display with intermediate viewpoints synthesized between three original viewpoints (condition 3), as compared to a switchable viewpoint display with no intermediate viewpoints (condition 2), a simultaneous display of three unchanging viewpoints (condition 1), and a single unchanging viewpoint (condition 0). Subjective participant feedback confirmed that the manipulable viewpoint display provided the lowest perceived difficulty.

Conditions 1 and 2 were not found to afford significant performance improvements in such a relative-depth discrimination task as compared to condition 0. As such, conditions 1 and 2 are determined to be an ineffective display mode for endoscopic surgery for the purposes of diagnosing relative depths of occluded objects. However, the presentation of a plurality of viewpoints (as occurs in conditions 1 and 2) may be useful for other purposes, as when they are providing new information by pointing at different regions of the surgical field, or in cases where a tool may occlude one or more viewpoints.

We recommend the motion parallax display investigated herein (condition 3) for use with at least some aspects of laparoscopic
surgery; further investigation is required to determine the benefits of a motion parallax display for other important surgical tasks, particularly motor control tasks such as suturing which involve manipulation (as opposed to merely diagnosis or exploration). Other physical benefits of a multiple-camera system in use with a virtually manipulable viewpoint include a reduction in likelihood of camera movement causing injury to adjacent tissue or interfering with carefully arranged surgical tools. However, our findings remain valid regardless of whether virtual viewpoints or real viewpoints obtained by a mechanically-controlled camera are used.

Finally, results from this generic task are not expected to be dependent on the type of surgery used, and are likely independent of surgery altogether; conclusions drawn here may be applicable to a broad range of other tele-operative tasks, for example in fields of aerospace or search-and-rescue robotics.
References


Poggendorff, J. C. (1863). Biographich-literarisches Handwörterbuch zur Geschichte der exacten Wissenschaften (herausg. von BW Feddersen [and others]).


A Algorithmic details of view interpolation

This appendix outlines the various processes used to achieve the view interpolation effect in the primary investigation.

1. **Calibration:** The three cameras were calibrated with two separate procedures, as described by Bouguet (2010) and Svoboda et al. (2005). First, the radial distortion of each lens was precisely determined by capturing multiple images of a calibration pattern in a variety of orientations for each camera and using Bougeut’s calibration software. Next, the geometric relationship between the cameras\(^1\) was determined by capturing synchronized camera videos of a moving point light source and using Svoboda’s multi-camera self-calibration software.\(^2\)

2. **Image pre-processing:** For additional reliability of computer vision algorithms, each image was preprocessed by capturing multiple (RAW Bayer format) instances within the span of a few seconds and averaging pixel intensities between instances to reduce Gaussian pixel noise. Subsequently, colour-balancing was performed to reduce the variation in hue and intensity between images.

3. **Image Undistortion, Rectification, Cropping:** Using camera calibration information, each set of images of the experimental scene was undistorted, rectified to row-align image pairs for the purposes of stereo correspondence (refer to step 4), and cropped to exclude unnecessary image portions, as illustrated in Figure A.1.

4. **Stereo Correspondence:** Images were processed offline using a (slow because non-optimized, though theoretically fast) version of a box-filter-based stereo correspondence algorithm by Rhemann et al. (2011). Rather than performing trinocular rectification and conducting a simultaneous three-view correspondence search, images were rectified into three different baselines (i.e., images were transformed to align the horizontal image axis with each
given edge of the triangular configuration) for the purposes of three pairwise stereo correspondence searches between each subset of the three captured images. As illustrated in Figure A.2, both disparity maps and reliability maps were computed for each image in two different baselines.

5. **Disparity map refinement**: Disparity maps for each image, as determined by stereo matching to each of the two other images, were combined into a single accuracy map based on detected reliability of the computed disparity of each pixel; this procedure allowed for proper depth estimation of pixels that were only visible in two of the three images. Depth maps were refined by a combination of median filtering and a heuristic refinement step unique to the specific experimental scene used, involving altering erroneously computed depths of groups of pixels based on expected hues of objects at various depths. The result of disparity refinement is illustrated in Figure A.2.

6. **Forward Warping**: Images, disparity maps, and reliability maps were stored after being pre-computed offline, and view interpolation was implemented in real-time on the GPU. The first stage of view interpolation involved forward warping, as described by Rogmans et al. (2009); however, our algorithm was instead modified to combine a set of three images, and weight them by novel view proximity to each original view and by pixel reliability.

7. **Hole-filling**: A hole-filling step was used to fill in intensities of pixels lacking information with the intensity of the lower disparity surrounding region\(^3\), also implemented in real-time on the GPU. Alternatively, regions lacking information could be rendered as a distinct bright solid colour to indicate to the viewer a lack of information. In our case, these regions were small enough to justify utilizing hole-filling. Figure 3.16 in the main body depicts an interpolated image, with slight ghosting artefacts visible.

\(^3\) In accordance with the lower depth recommendation for hole-filling (Scharstein, 1999).
Figure A.1: Step 3—Image as captured by the left camera in the three-camera configuration: a) original; and b) after undistortion, rectification, and cropping.

Figure A.2: Step 4 and 5—Pairwise stereo correspondence between left and right images, and between left and top images, yields two somewhat reliable depth maps; combination and refinement results in a higher-accuracy depth map.
## B Response accuracy data table

Statistics of individual subject response accuracies are presented here by separation and display condition. Each accuracy score is computed as a proportion based on 24 binary responses.

<table>
<thead>
<tr>
<th>Display</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separation</td>
<td>2 cm</td>
<td>4 cm</td>
<td>2 cm</td>
<td>4 cm</td>
</tr>
<tr>
<td>Participant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.375</td>
<td>0.541</td>
<td>0.458</td>
<td>0.375</td>
</tr>
<tr>
<td>2</td>
<td>0.583</td>
<td>0.375</td>
<td>0.542</td>
<td>0.500</td>
</tr>
<tr>
<td>3</td>
<td>0.542</td>
<td>0.333</td>
<td>0.500</td>
<td>0.583</td>
</tr>
<tr>
<td>4</td>
<td>0.500</td>
<td>0.666</td>
<td>0.750</td>
<td>0.833</td>
</tr>
<tr>
<td>5</td>
<td>0.500</td>
<td>0.500</td>
<td>0.417</td>
<td>0.500</td>
</tr>
<tr>
<td>6</td>
<td>0.500</td>
<td>0.458</td>
<td>0.375</td>
<td>0.417</td>
</tr>
<tr>
<td>7</td>
<td>0.458</td>
<td>0.458</td>
<td>0.458</td>
<td>0.875</td>
</tr>
<tr>
<td>8</td>
<td>0.333</td>
<td>0.375</td>
<td>0.375</td>
<td>0.333</td>
</tr>
<tr>
<td>9</td>
<td>0.417</td>
<td>0.667</td>
<td>0.375</td>
<td>0.458</td>
</tr>
<tr>
<td>10</td>
<td>0.333</td>
<td>0.542</td>
<td>0.458</td>
<td>0.500</td>
</tr>
<tr>
<td>11</td>
<td>0.458</td>
<td>0.375</td>
<td>0.500</td>
<td>0.708</td>
</tr>
<tr>
<td>12</td>
<td>0.458</td>
<td>0.542</td>
<td>0.417</td>
<td>0.500</td>
</tr>
<tr>
<td>13</td>
<td>0.458</td>
<td>0.583</td>
<td>0.417</td>
<td>0.458</td>
</tr>
<tr>
<td>14</td>
<td>0.541</td>
<td>0.708</td>
<td>0.667</td>
<td>0.875</td>
</tr>
<tr>
<td>15</td>
<td>0.500</td>
<td>0.667</td>
<td>0.375</td>
<td>0.500</td>
</tr>
<tr>
<td>16</td>
<td>0.458</td>
<td>0.458</td>
<td>0.417</td>
<td>0.500</td>
</tr>
</tbody>
</table>
C Participant questionnaire

Post-Study Questionnaire

Please fill out this questionnaire. If you have any questions, please feel free to ask the investigator.

Sex
[ ] Male  [ ] Female
Age

Occupation
[ ] Employee  Field: [ ]
[ ] Undergrad  Field: [ ]
[ ] Graduate  Field: [ ]
[ ] Other  Please specify: [ ]

Describe your handedness
[ ] left-handed  [ ] right-handed  [ ] ambidextrous

Do you have any visual disabilities?
[ ] yes  [ ] no
If yes, please describe the disability. (This includes whether you normally use corrective lenses.)

Please rank the display conditions in order of least (1) to greatest difficulty (4) of use:
Stationary viewpoint(s):

Changing viewpoints:

- Single Image [ ]
- Switchable Viewpoint [ ]
- Three Images [ ]
- Movable Viewpoint [ ]
Please describe the following about your experiences with each display condition:

**Single Image:**

Describe the strategy (or strategies) you used to figure out which rod was on top: (i.e. what did you look for?)

Describe any positive and/or negative feedback about your experience with this condition:

**Three Images:**

Describe the strategy (or strategies) you used to figure out which rod was on top:

Describe any positive and/or negative feedback about your experience with this condition:
Switchable Viewpoint:
Describe the strategy (or strategies) you used to figure out which rod was on top: (i.e. what did you look for, and how did you choose to move between viewpoints?)

Describe any positive and/or negative feedback about your experience with this condition:

Movable Viewpoint:
Describe the strategy (or strategies) you used to figure out which rod was on top:

Describe any positive and/or negative feedback about your experience with this condition:
Please describe your overall experience with this study, and/or use this space to describe more about any of the previous questions.

Thank you!
D In-depth analysis of subjective preferences and response behaviour

Subjective feedback about each display condition was obtained from participants via a questionnaire in which they were asked to rank the conditions by difficulty (refer to Appendix C). In addition to self-reported subjective strategies, the usage tendencies exhibited by participants when given control over the camera viewpoint in conditions 2 and 3 were also examined. Subjective difficulty ranking are presented in the main body of the text, in Section 4.1.

D.1 Condition 0: Subjective responses

Subjects almost unanimously reported the highest difficulty with this display condition, with only one participant (P9) ranking it as the least difficult, and two participants (P6, P16) ranking it as less difficult than condition 1. These participants each reported via written feedback that they found this condition less difficult due to its simplicity, since they felt that the overwhelming amount of information in the other conditions was, as P16 described it, ‘too much to process’. These three participants did in fact achieve equal or higher accuracy with condition 0 (though still near chance performance with 48%, 54%, and 46%) as compared to condition 1, but P9 did not surpass performance compared to conditions 2 and 3.

Four participants reported attempting to rely at least partially on the relative widths of the rods. Only one of those participants achieved higher than chance performance, however. Ten participants reported using their ‘gut feeling’, ‘instinct’, or ‘first impression’, which in some cases may have involved subconsciously responding to some perceived relative width cue. Other reported attempted strategies involved comparing the relative sharpness (focus) of the
rods, or darkness of apparent cast shadows.

**D.2 Condition 1: Subjective responses**

For display condition 1, four participants (P4, P7, P12, P14) reported observing the relative displacement of the rods between images. Among those participants, P4, P7, and P14 achieved the highest response accuracies for this condition. Two participants attempted to mimic conditions 2 and 3 by rapidly flicking their eyes between images on the screen, while two others attempted to 'look at all three images at the same time'. Other participants reported using similar sharpness or shadowing as with condition 0. Of these, three participants attempted to 'find the best view', and were bothered by occasionally perceiving contradictory cues in different images captured of the same scene.

Participants who tended to prefer condition 1 over condition 2 reported the benefit of having all information available at the same time, while participants who conversely preferred condition 2 were averse to the effort involved in looking between images that were physically separated on the display monitor.

**D.3 Condition 2: Subjective responses and Viewpoint switching tendencies**

Top performing subjects in this condition, P4 and P14, both consciously watched for relative rod displacements, with P14 preferring a circular motion to switch between the three images in rapid succession. Though P4 performed better in condition 2 than in all other conditions, he preferred condition 1 over condition 2 due to 'needing to remember' before and after switching between images in condition 2.

All subjects who reported speed preferences preferred moving faster; P16 attempted to rapidly switch back and forth only between the left and right images, in an attempt to approximate the sensation of '3D'. This did not prove beneficial, as he achieved just chance performance; however, his performance may also be attributed to him spending just 2.2 seconds on each trial.

Overall viewpoint selection tendencies as recorded during the study are presented in Figure D.1. Participants overall tended to switch between left and right viewpoints twice as often as they switched between each other pair of viewpoints. These overall measured strategy tendencies are important to reveal the residual
effects of a non-rectilinear camera configuration; with a rectilinear configuration (i.e., with 4 cameras), users may have been more encouraged to make use of the vertical switching dimension. However, the left/right preference may also be an artefact of the nature of human binocular perception. The benefit of the vertical dimension was intended to be the provision of a greater variety of available directions to perceive motion in, but this may only have been useful in display condition 3 where the variety of directions are not constrained to be parallel to the edges of the triangle.

Figure D.1: For display condition 2, a) time allocations on each viewpoint, averaged across participants, with standard error bars shown, and b) discrete viewpoint switching tendencies

D.4 Condition 3: Subjective responses and Viewpoint movement tendencies

As with the previous condition, all subjects who reported speed preferences preferred moving faster; participants also generally preferred changing the viewpoint in a continuous manner, with no reported preferences to the contrary. One participant preferred circling around the centre area looking for small differences, but felt that this was only useful when not moving slowly; another participant preferred moving rapidly along the edges of the triangle to maximize the difference between perceived images. 12 participants identified that they were using the relative motion of the circle and rods to inform their decision. P13, who reported simply relying on instinct for condition 3, achieved 92% accuracy with 2 cm rod separation, and 100% accuracy with 4 cm rod separation.

The lowest performance was achieved with condition 3 by P1, with 58% at 2 cm and 75% at 4 cm, which still is above chance and exceeding his performances in other conditions. He reported attempting to use control over the viewpoint to ‘find the best view’, but did not prefer this condition because he found it hindered the
task of finding such a view. P16 also did not prefer this condition, because he (versely to general participant consensus) found the gradual changes ‘harder to see’ than the rapid changes of condition 2. Finally, P9 did not prefer this condition because he was bothered by the artefacting that occurred in the synthesized viewpoints, and recommended using a real moving camera to achieve the same effect. His preferences aside, P9 still performed significantly better with condition 3 than with any other condition.

Two participants, P12 and P14, reported some slight difficulties with sliding the pen around the input tablet due to a build-up of plasticine, which was rectified midway through the experiment in each case. This may have had some negative effect on their performances in condition 3, and to a lesser extent also in condition 2.

Movement tendencies

To further explore the subjective preference for faster viewpoint movement, we investigated whether any correlation existed between the mean speed of viewpoint motion and overall response accuracies for display condition 3. No significant correlation was found, however. To explore the varying amounts of time participants spent controlling the viewpoint, we tested the correlation between overall response accuracies and median response times. Once again, no significant correlation was found. However, a correlation was found between time and response accuracy by isolating time spent actually manipulating the viewpoint: Pearson’s $r = .516, p < .05$. This indicates that participants who spent more time observing the motion between viewpoints were able to achieve higher accuracy, while spending more time observing unmoving intermediate viewpoints was not shown to have an effect on accuracy.

The reduction in motion tendencies displayed by low-scoring participants is manifested not only by a reduced time spent moving, but also by the variety of directions they chose to move in. Figure D.2 shows a complementary set of response data for each participant in condition 3, including directional tendencies, viewpoint dwelling preferences, time spent manipulating, mean speed of manipulation, and response accuracy.

- Participant data are ordered in descending order of overall response accuracy averaged across both rod separations.
- Directional tendency diagrams are constructed as angular histograms with 24 angular bins; the length of each arrow

\[^{1}\text{Mean speed of viewpoint motion reflects the average nonzero speed only; instances of zero speed were ignored in the calculation to avoid biasing speed measures with dwelling tendencies.}\]
indicates the total distance travelled in that direction throughout the study (using display condition 3 only). Arrow lengths are scaled within each diagram proportional to the length of the longest arrow in that diagram; as such, only the relative sizing of arrows within each diagram is pertinent.²

- Viewpoint dwelling tendencies are visualized as heat maps, where bluer is less and redder is more. These graphs generally also provide strong indications of where movement occurred within the control triangle.
- For each participant, mean time per trial spent manipulating the viewpoint is shown as a green bar graph, scaled in units of seconds (ranging from 0 to 10 seconds).
- The mean nonzero speed of manipulation is plotted as an orange bar graph, with units of triangle edge lengths per second (ranging from 0 to 3 edge lengths/second).
- Finally, response accuracy averaged over rod separations is presented as a purple bar graph, presented only in the range of 50–100%.

In Figure D.2, the overall correlation between time and accuracy can be seen as a general trend downwards in time taken amongst this accuracy-ordered participant list; in other words, the green bars generally become shorter as the purple bars become longer. The four lowest-accuracy-achieving participants either spend very little time moving, or else move in a limited range of directions:

- P6 (bottom row, middle) exhibited a lot of motion along each edge of the triangle but spent a mean of only 1.14 s moving per trial;
- P1 (bottom row, far right) intriguingly also moved primarily in directions parallel to the triangle edges, but only in a clockwise direction;
- P9 (bottom row, far right) alternated between horizontal movement in the middle of the triangle and edge-based movement around the perimeter;
- P16 (bottom row, third-right) clearly relied only on horizontal motion.

Circular motion tendencies³ appear to correspond to higher performance, with P12, P15, and P11 exhibiting this most clearly in the top row, and P3 (also in the top row, middle) demonstrating more elliptical tendencies.

Top-performing participants P12 and P14 spent much time moving at high speeds, and made full use of the triangle to perceive the scene.

² Note also that each arrow provides no indication of where within the triangle a movement occurred, only in what direction it occurred. The complementary heat maps of viewpoint preference provide some insight into where some movements tended to occur.

³ Circular motion tendencies are exhibited by directional tendencies fully distributed amongst all 24 angular bins. Note, however, that these may equally well correspond to random movement, but observations of behaviour during the study indicate otherwise.
equally from a variety of viewpoints. However, high performance does not necessitate large amounts of time devotion, as P13 (top row, left) achieved 96% accuracy with only 2.3 seconds spent moving per trial; P13 also reported ‘not trying to think about it too much’.
Figure D.2: Response data for each participant in display condition 3, including normalized directional tendency angular histograms, viewpoint dwelling preference heat maps, mean time spent manipulating (s/trial), mean nonzero speed of manipulation (edge lengths/s), and response accuracy (averaged across rod separations); participants are sorted in decreasing order of accuracy achieved.