18

Task performance using 3D displays

T. N. Bardsley and I. Sexton
Imaging and Displays Research Group
Department of Computer Science
De Montfort University
The Gateway, Leicester LE1 9BH
ENGLAND
Tel.: (-533) 577498 Fax: (-533) 541891
e-mail: tnb/sexton@dmu.ac.uk

Abstract

This paper presents the results of a recent Ph.D. research programme encompassing the design, development, and evaluation of a variety of stereoscopic display systems. Emphasis is placed upon four distinct methodologies: anaglyphic display, a commercial frame sequential display using a switchable polarisation rotator and passive spectacles, a frame sequential display using active spectacles, and an autostereoscopic (no glasses) system, developed in our laboratory which uses a flat panel display with a specially produced lenticular faceplate.

The operation and technical limitations of the various display types is discussed and a robust assessment methodology based upon visual search and spatial tracking tasks is presented. The limitations of the four display types are discussed to provide an indication of the ergonomic factors likely to contribute to a differential display performance.

The data analysis and inferences drawn from this cross-display evaluation are presented and a hypothesis is formulated which describes the causal relationship between stereoscopic display methodology and isolated aspects of observer performance. A subjective ranking of 3D display 'quality' was solicited from participants in the evaluation and this is compared with the display ordering derived from the experimental study.

Keywords

3D displays, stereoscopic, autostereoscopic, computer Graphics

J. Rix et al. (eds.), *Virtual Prototyping*

INTRODUCTION

Recent technological advances and the demands of more sophisticated methods of interaction (virtual reality is one particular example) have engendered an unprecedented interest in the techniques and applications of 3D imagery.

Although the head mounted display is probably the preferred display apparatus for true (immersive) virtual reality systems these are often costly and cumbersome. Because of this, and in view of rising concern regarding the health and safety issues surrounding such displays, alternative non immersive or desktop VR systems are often preferable. Although many commercially available examples are monoscopic implementations, the arguments supporting the use of stereoscopic displays are quite compelling.

The Imaging and Displays Research Group at De Montfort University is involved in various aspects of three dimensional and stereoscopic display research. A central theme of this research has been the development of autostereoscopic display systems; particularly display types which utilise view selecting mechanisms in the form of parallax barriers and lenticular screens. Autostereoscopic displays are quite different in their implementation to the more conventional stereoscopic systems and as a consequence the subjective impression is also quite different.

The use of stereoscopic and 3D displays has become well established in niche application areas where it is possible to clearly demonstrate an improvement in operator performance, however this is usually achieved by comparing a particular stereoscopic system to a monoscopic counterpart (Pepper *et al.*, 1997). In areas where the use of 3D displays has become established, the evidence to support the superior performance is often incontrovertible but few researchers have the opportunity to make direct comparisons of different 3D systems. Because of this, the choice of display methodology is often quite arbitrary.

There are many ways of producing a monoscopic display and many more ways of producing a stereoscopic system. This is to be expected as monoscopic displays are often components of stereoscopic systems. Each display type exhibits a set of artefacts; some of which are unique to the implementation. These typically encompass limited screen resolution (influencing depth quantization level), crosstalk, retinal rivalry, and perceptual flicker, all of which ultimately impinge upon operator performance. It is important for potential users of such displays to recognise these limitations in order to determine the applicability of a particular methodology.

Two of the displays in this investigation are frame sequential, wherein the two views comprising the stereo pair are temporally multiplexed. The system which uses active spectacles decodes the image by means of liquid crystal shutter devices arranged in lieu of the lenses in a pair of spectacles. These shutters can selectively block an image from either eye. The system which uses passive spectacles has a large liquid crystal device (in this case a Tektronix SGS410 Stereoscopic Modulator) interposed between the monitor and the viewer. This device encodes alternate frames with either a clockwise or anticlockwise polarisation. These are subsequently decoded by polarising spectacles worn by the viewer.

The anaglyph display rarely requires explanation. The left and right views are encoded using different colours on the display and are decoded by means of coloured (usually red and green) spectacles worn by the viewer.

The autostereoscopic display differs fundamentally from the others in this evaluation. The left and right views are spatially multiplexed and the viewer is not required to wear any special spectacles. Appropriate left and right views are presented to the viewers eyes by means of a lenticular sheet at the display surface. In this particular case the display device is a colour (TFT) liquid crystal panel.

It is possible to make a purely objective assessment of a 3D display; invariably such systems depend upon the channelling of appropriate views to each eye of the observer. Although the techniques used to achieve this objective differ markedly between display strategies it is a common requirement across display types. Consequently, a revealing measure may be obtained by determining how effectively the appropriate images are channelled to each of the observers eyes. Unfortunately this approach takes no account of the fundamental ergonomic differences between display types, and it was considered that more appropriate metrics should engage the concept of *fitness for purpose* whereby task related performance could be determined. A comparative assessment of the four participating display types was performed on this basis.

Two separate experimental procedures were developed to enable a quantitative assessment of display capability in supporting both static and dynamic graphical interaction. The first experiment comprised a *visual search* task which required observers to correctly identify pairs of visual stimuli which were uniquely correlated in depth. The second experiment utilised real-time animated graphics to evaluate observer performance during a continuous *spatial tracking* task.

EXPERIMENTAL OBJECTIVES

A controlled study was devised to examine the efficacy of each stereoscopic display technique (anaglyph, lenticular, LC polariser and LC shutter glasses) during a quantitative assessments of operator task performance. The hypothesis underlying the experiment being that each stereoscopic display methodology will have a specific and measurable effect on critical aspects of human performance. Objectives of the display evaluation were to determine the potential of each display technique as the basis for implementing a stereoscopic graphics workstation, and to establish a performance envelope for the recently developed lenticular prototype. A subsidiary goal was to identify a set of universally applicable performance metrics to serve as a basis for future stereoscopic display assessment.

Previous psycho-optical investigations have revealed certain aspects of human performance which are sensitive to the provision of binocular stereopsis. Human factors experiments, originally intended to explore the superiority of stereoscopic presentation over conventional monoscopic display formats, have shown marked improvements in spatial positioning accuracy [Pepper *et al.*, 1977; Kim *et al.*, 1987; Reinhart, 1991; Takemura *et al.*, 1989; Beaton, 1990), enhanced speed and accuracy during depth correlation judgements (Reinhart *et al.*, 1990; Zenyuh *et al.*, 1988; Miller and Beaton, 1991; Yeh and Silverstein 1990) and a significant reduction in task completion times (Pepper *et al.*, 1977; Cole *et al.*, 1990). In the light of these observations, it seems reasonable to assume that the relative magnitude of such performance increases will be contingent upon the efficiency with which individual display methodologies implement their stereoscopic effect.

Two separate task scenarios were devised to investigate differences in stereoscopic display performance during the presentation of both static and dynamic visual stimuli. The requirement for interaction with dynamic stereoscopic imagery is universally recognised

(Butts and M^{c}Allister, 1988), and an investigation of display performance in this context is easily justified. A static visual search task was included to investigate each display's disposition for one form of visual stimulus over the other.

A *target tracking* task was conducted to monitor the *spatial positioning error* associated with the continuous pursuit of a dynamic visual stimulus as it manoeuvres within the confines of a 3D display viewing volume. A composite measure of tracking performance was obtained by computing the *mean Euclidean distance error* during tracking operations. Six additional error metrics were calculated based upon the *mean* and *absolute* tracking errors associated with each of the independent X, Y and Z display axes. Measures of absolute deviation form the target trajectory were used to provide a useful indication of the error magnitude attributable to each display axis, while the polarity of mean tracking scores was intended to reveal any systematic bias in positional deviation.

A second *target acquisition* task was designed to examine the potential of each display for facilitating three-dimensional *visual search* and *depth correlation* judgements. The response measures associated with this particular task were *decision response time* and *accuracy* during depth correlation judgements. The degree of accuracy was determined in two ways. Firstly by computing the percentage of correct depth correlations for each display and secondly, by recording the degree of error when objects were incorrectly matched in depth.

Both the tasks and the visual stimuli used to depict them, were designed to draw upon the perceptual and motor skills necessary for comprehending and interacting with three-dimensional graphical environments. As the purpose of the study was to provide a measure of stereoscopic effect, all monocular cues with the exception of motion parallax, size constancy, linear perspective and object interposition were eliminated. Adopting a subset of the available monocular cues was intended to isolate the influence of binocular stereopsis, while maintaining important binocular/monocular cue consistency.

Both visual search and interactive cursor positioning operations can be characterised by their high levels of task demand (Reinhart, 1991; Zenyuh *et al.*, 1988). This is fortunate as both the workload and the degree of stress experienced during task performance are considered to be influential factors which determine the benefit and efficacy of retinal disparity cues (Miller and Beaton, 1991). The high cognitive load associated with both the target acquisition and tracking tasks, help to ensure the adopted human performance metrics are well sensitised to small differences in display quality, and therefore constitute good indicators of stereoscopic display performance.

A subjective assessment of display quality and performance was obtained by interviewing participants at the end of the experimental procedure. Subjects were asked to rank the four displays in order of preference based on their confidence level while performing the experimental tasks. In the event that subjects could not differentiate the displays using this criteria, they were instructed to consider additional factors such as viewing comfort, visual fatigue and any disturbing visual characteristics the displays may have.

METHOD

Subject population

Thirty two potential participants were screened for satisfactory stereopsis using a 16 depth level random-dot stereogram (RDS) test stimulus. The absence of extrastereoscopic cues in the RDS image make it an unfakeable test for stereopsis (Reinhart, 1991) and this permits early rejection of subjects with stereoanomalous vision. All stereodeficient candidates were eliminated from the study. Twenty nine subjects (23 male and 6 female) passed this initial screening procedure by virtue of being able to discriminate the smallest values of positive and negative parallax embodied within the RDS test image. At twenty nine, the number of subjects participating in the evaluation study was reassuringly large in comparison to the numbers commonly employed in similar investigative procedures (Kim *et al.*, 1987; Williams and Parrish, 1990), consequently both the experimental power and the prospects for generating statistically significant results were greatly improved.

All test subjects taking part in the evaluation study were unpaid volunteers drawn from the staff and students at De Montfort University and from enthusiastic members of the public. Subjects were aged between 17 and 48 years old; a mean age of 25½ characterising the group. Previous exposure to stereoscopic imaging techniques varied widely between subjects. Some individuals had experienced the full gamut of cinematic, holographic and virtual reality based entertainment systems, while others had little or no prior exposure to stereoscopic media. None of the participants reported recent or extensive exposure to any of the stereographic displays involved in the evaluation study.

Environmental conditions and apparatus

The experimental trials took place over a three week period during which time every effort was made to ensure consistent environmental conditions for all participating subjects. Ambient lighting conditions were carefully controlled to optimise the performance of all the displays involved in the experiment. Particular attention was given to the location and brightness of individual light sources to avoid introducing first and second order reflections from CRT faceplates and screen mounted optical components.

The red and green colour components used to represent anaglyphic stereo pairs were optimised to yield screen colours which closely matched the chromatic filter characteristics of the anaglyph spectacles. Maximising transmission for corresponding colours and reducing the leakage of unwanted colour components helped to improve filter extinction ratios and reduced the detrimental effects of image crosstalk.

A choice of either full-frame or clip-on glasses were supplied for use with the Tektronix polarizing display. This allowed subjects with corrected vision to retain their prescription lenses and participate more effectively in the experimental study.

Care was taken to avoid unfairly compromising any of the displays involved in the investigation, however, none of the steps taken to ensure optimal display viewing conditions were considered to be unusual or contrived. After all, careful regulation of the viewing environment and the provision of well implemented stereoscopic display techniques are prerequisites for a robust evaluation strategy (Merritt, 1983).

The Spaceball Technologies Spaceball™ was used to provide subject input responses during spatial tracking operations. The Spaceball was chosen in preference to a glove-based positioning device, as the users arm is fully supported during use, and this reduces the influence of arm fatigue on tracking performance. The Spaceball comprises a rubberised sphere about the size of a tennis ball mounted on a stable platform. The Spaceball is sensitive to the fingertip pressures and torsional forces applied to it, resolving these simultaneously into X, Y and Z translations and rotational components. Although the Spaceball is capable of simultaneously encoding six degrees of freedom (S6DOF), rotational effects were disabled during the tracking task. The Spaceball was programmed to provide output responses that were proportional to the magnitude of applied translational forces. This meant that cursor velocity could be freely controlled during pursuit of the target stimulus.

Target tracking

Task structure

The three-axis tracking task was designed to measure subject performance during the pursuit of a target symbol which moved in a random and continuous fashion throughout the confines of a three-dimensional viewing volume. The hypothesis under investigation being that choice of stereo display methodology will significantly influence the accuracy with which subjects perform spatial tracking operations.

Subjects were required to manipulate a three-dimensional cursor using the Spaceball positioning device which afforded simultaneous control over all three cursor axes. Participants were instructed to follow the target symbol as closely as possible in order to minimise their X, Y and Z-axis tracking error.

A continuous target motion was chosen in preference to a static stimulus to increase the cognitive load associated with the spatial positioning task. Other investigators have employed static target stimuli which results in an oversimplification of the cursor alignment task by allowing a method of limits approach to cursor alignment (Reinhart, 1991; Takemura *et al.*, 1989; Drascic, 1991). Although use of a static target stimulus would have enabled the assessment of cursor positioning times, the associated task demand was considered too low to constitute a good measure of interactive stereoscopic display capability.

Visual stimulus

Visual representation of the manual pursuit tracking task comprised a wire-frame perspective cube inscribed within the boundaries of the stereoscopic display screen, a flat-shaded target symbol and a wire-frame *full-space jack* (FSJ) cursor. Each of these components was assigned a unique colour and rendered in order of decreasing depth on a black background to replicate the extrastereoscopic cue of object interposition. A pictorial representation of the target tracking task appears in figure1.

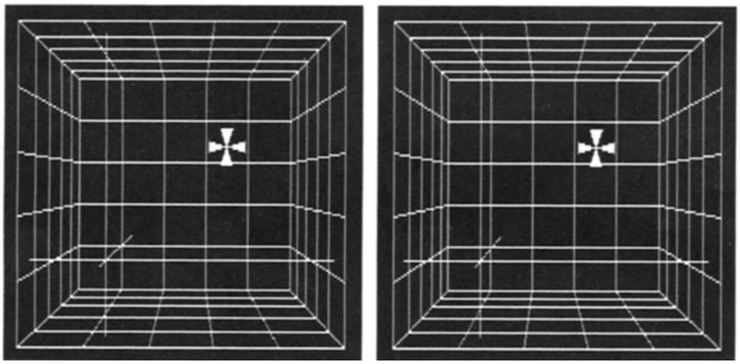


Figure 1 Stereogram depicting the target tracking stimulus (arranged for transverse viewing)

Both the simplistic perspective groundplane (Kim *et al.*, 1987; Yeh and Silverstein 1990; Grotch, 1983), and the more complex viewing volume reference cube (Butts and M^{c}Allister, 1988; Barham and M^{c}Allister, 1991; Beaton *et al.*, 1987), have been utilised extensively in the depiction of stereoscopic data and three-dimensional task scenarios. Such enhancements are routinely used to increase the perceived realism and facilitate relative spatial judgements in both 2½D and true 3D images. A view volume reference cube was provided in the target tracking task to delineate the extent of the three-dimensional workspace and indicate the intended limits of target and cursor motion.

Four criteria were used to determine an appropriate target shape for the pursuit tracking task. (1) The 'focus' of the target must remain visible regardless of its position and orientation within the viewing volume. (2) The precise focus of the target must be unambiguous and should be, at most, 1 pixel wide about all 3 display dimensions. (3) The target must undergo an appropriate change in angular subtense as it travels back and forth within the confines of the viewing volume. (4) Interposition cues must accurately reflect the relative depth of target and cursor stimuli. The planar Maltese Cross depicted in figure 1 was thought to be a reasonable compromise which satisfied all four design constraints. The planar nature of the stimulus gave it a zero depth extent thereby ensuring an unequivocal depth location. The focus located at the intersection of the cross axes was precisely one pixel wide and the target stimulus was made opaque to afford good object interposition characteristics.

The target trajectory for the three-axis manual tracking task was randomised to inhibit predictability and avoid track repetition. Target motion was composed of three independent single-axis trajectories, each comprised of a series of sinusoids having random frequency. Target position was updated during successive animation frames at a rate of 10 Hz. At no time was the target symbol permitted to exceed the bounds of the viewing volume.

A full-space jack (FSJ) cursor was chosen to indicate subject response during the target tracking operations. This cursor format was chosen as it embodies motion parallax, linear perspective and size constancy cues effectively, and can be rendered quickly. The FSJ is a modified two-dimensional cross hair cursor in which the depth axis is represented by a third vector lying perpendicular to the display surface. Each axis of the FSJ cursor is extended to meet the boundaries of the viewing volume to indicate absolute position within the 3D workspace. The cursor *hot-point*, from which all positioning errors are measured, lies at the intersection of these three vector components. Cursor position was updated between animation frames using force data supplied by the Spaceball; cursor velocity being directly proportional to the applied force on each Spaceball axis.

Target acquisition

Task structure

A target acquisition study was designed to reveal the influence of the anaglyphic, lenticular, polarizing and optical shutter based display techniques on subject performance during visual search and depth correlation operations. The underlying hypothesis being that choice of stereo display methodology will significantly effect both the speed of visual search and the accuracy of depth correlation judgements.

Subjects were presented with a sequence of static images, each containing numeric symbols scattered randomly throughout the confines of the display viewing volume. Subjects were required to identify verbally, which of these numeric symbols occupied the same depth plane as a single reference symbol lying elsewhere in the viewing volume. Only one numeric symbol was co-located in depth with the reference stimulus; the remainder serving as distracters, both to increase task demand during visual search operations and further confound estimates of cue/target depth correlation.

Visual stimulus

A number of factors are known to influence the speed and accuracy of relative stereoscopic depth judgements. The number of depth planes (Reinhart, 1991), the spatial proximity between cue and target stimuli (Yeh and Silverstein, 1989; Reeves and Tijus, 1990) and the number of distracter symbols present in the viewing context (Miller and Beaton, 1991) all influence the potency of binocular disparity cues.

Numeric symbols were utilised throughout the target acquisition study to enable participants to quickly and unambiguously identify their choice of co-located stimuli. Planar symbols with a parallel screen orientation were used to avoid generating stimuli that spanned several depth planes simultaneously, thereby confusing the depth matching process. A four point star symbol was chosen, quite arbitrarily, to represent the reference or *cue* stimulus. The precise viewing context of the target acquisition task is illustrated in figure 2.

All stimuli occupied the same visual angle regardless of their apparent depth within the viewing volume, and the spatial location of each symbol was carefully controlled to eliminate any character overlap. The elimination of retinal image size and object interposition cues was a necessary precaution to increase the cognitive load associated with visual search and depth correlation operations while to emphasising the impact of binocular stereopsis.

The depth axis was partitioned into 21 discrete depth planes encompassing the fusible extent of the display viewing volume. Each depth plane was assigned a unique value of screen parallax. Crossed and uncrossed disparity levels for each of the 21 depth planes ranged from -10 to +10 display pixels; the zero parallax condition coinciding with the plane of the display screen. The location and proximity of co-located symbols were randomised for successive stimulus conditions to prevent systematic cue/target separations. All non-correlated distracter symbols were randomly distributed throughout the viewing volume.

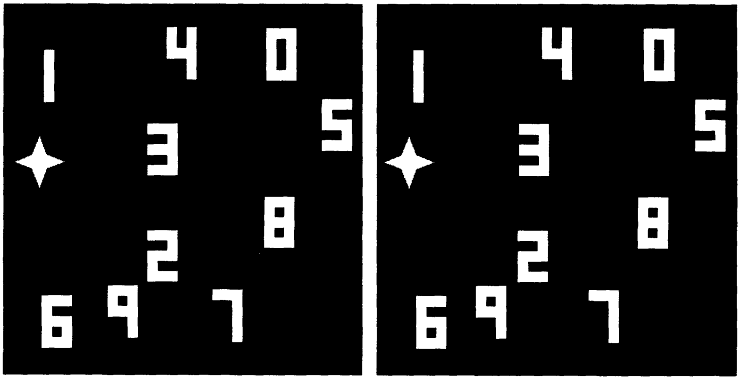


Figure 2 Stereogram of the target acquisition stimulus (arranged for transverse viewing)

Experimental procedure

All 29 test subjects undertook an initial training period to aid familiarisation with the Spaceball input device. This precaution was necessary to saturate the learning effects inherent in mastering an unfamiliar input device, and helped to ensure a consistent dexterity level for all subjects participating in the target tracking evaluation. During this training procedure, 2 subjects failed to meet the minimum performance criteria for steady-state tracking error and were subsequently eliminated from the target tracking study.

The remaining 27 subjects performed the spatial tracking and target acquisition tasks consecutively on each of the four stereoscopic displays. The order of exposure to each display was randomised between subjects to distribute the impact of learning and carry-over effects associated with task repetition.

Subjects performed three tracking runs of 60 seconds duration. Partitioning the task in this way helped to reduce the influence of fatigue on tracking performance. An initial warm-up period of between 10 and 15 seconds was provided before an audible tone announced the onset of data collection. Following the 60 second recording period, a second audible tone indicated the completion of each tracking run.

All 29 subjects who passed the initial RDS screening procedure completed 20 replications of the target acquisition task on each of the four stereoscopic display systems. Unique stimulus configurations were constructed for each task replication, and the display screen was blanked during the calculation and rendering of new target and distracter locations. Subjects indicated their choice of target symbol by calling out a number between zero and nine. This verbal response was immediately entered on a numeric keypad by the evaluation supervisor. This elaborate response procedure was required to maintain the attention localisation of each subject and to eliminate the influence of keypad search intervals on response time measures. During subject briefing, participants were instructed to make accuracy their prime concern while remaining mindful of the fact that their decision response times were also being monitored.

Dependent measures

All spatial tracking measures were computed from recorded differences in target and cursor trajectories. Uni-directionally, *absolute tracking errors* were calculated for X, Y and Z to establish the error magnitude associated with each display axis. *Mean errors* were also produced to highlight any systematic bias left-right, above-below or fore-aft of the target position.

Positional deviations in the X-Y plane were measured in world coordinates, while errors in depth were quantified in terms of the number of discrete depth planes separating the target and cursor location; each depth plane corresponding to a specific value of screen parallax. Tracking errors in the depth dimension were measured in discrete depth planes to compensate for apparent differences in display resolution between the X-Y plane and the depth axis. Small changes in an object's world-coordinate position are quickly reflected in X-Y screen location, however, such changes do not become apparent on the depth axis until their magnitude is sufficient to generate a variation in screen parallax. Consequently, measuring Z-axis tracking precision at a level which exceeds the display's capacity for representation makes little sense.

A composite measure of spatial tracking performance was obtained by averaging the *mean Euclidean distance* errors separating the target and cursor throughout the tracking interval. The average of all such instantaneous errors serving as a composite measure of tracking accuracy. Object-to-object distances were calculated thus,

$$Error = \sqrt{(X_t - X_c)^2 + (Y_t - Y_c)^2 + (Z_t - Z_c)^2} \quad , \qquad (1)$$

where t and c denote major axis coordinate positions of the target and cursor respectively.

Human performance during the target acquisition study was characterised by the length of time taken to search the viewing volume for a cue/target match, and by the accuracy attained in nominating co-located target symbols. Four experimental measures were used in quantifying subject performance.

Decision response time was measured from the instant each stimulus condition was presented to the moment subjects announced their response. Response times were measured in real-time clock increments giving an effective resolution of 1/18.2 sec.

The number of *search errors* expressed as the 'percentage of correct depth correlations', provided a useful indication of the 'hit-rate' attained by subjects when using each of the stereo displays. In each of the 20 task replications, there is a 1 in 10 probability that participants will select the correct target symbol by chance alone. Consequently hit rates in excess of 10% could be considered evidence of a positive display influence during depth correlation.

A measure of *absolute depth plane error* was generated to reflect the magnitude of search errors arising from incorrect cue/target associations. *Mean depth plane errors* were calculated to reveal any directional bias fore or aft of the cue stimulus during depth correlation judgements.

Experimental design

Response-time and correlation-error data obtained in the target acquisition experiment were collapsed across trial replications to provide averaged performance metrics for each subject using each of the four displays.

Data from the target tracking and target acquisition experiments were submitted to separate, doubly multivariate, analysis of variance (MANOVA) procedures for repeated measures. Separate analyses were used to avoid complications arising from differences in the sample sizes involved (27 and 29 subjects respectively). Four levels of the single within-subjects variable *display type* (anaglyphic, lenticular, polarizing and shutter-glasses) were used in a mixed effects design. Subjective ranking scores obtained during post experimental interview were assessed using a simple one-way analysis of variance procedure.

Examination of the descriptive statistics and normal plots obtained for each dependent variable, revealed all data to be normally distributed with good homogeneity of variance characteristics (Devore, 1990; Norusis, 1986). As all data appeared to satisfy the basic assumptions required for the proper application of analysis of variance procedures (Keppel, 1973), Greenhouse-Geisser probabilities were not used to adjust univariate test results. Post-hoc Student-Newman-Keuls multiple comparisons tests, conducted at the 0.05 significance level were used to identify critical differences between performance means on each display.

RESULTS

Only the most salient results to emerge from the target tracking and acquisition studies will be presented in the sections which follow. The interested reader is referred instead to the full thesis for a more detailed discussion (Bardsley, 1994).

Target tracking performance

All multivariate test criteria (Pillai's Trace F[21,222]=5.6, $p<0.001$, Wilks' Lambda F[21,207]=7.7, $p<0.001$ and Hotelling's Trace F[21,212]=10.3, $p<0.001$) suggest that the single within-subjects factor *display type* has a significant effect on all combined measures of tracking performance. Univariate F-tests conducted with (3,78) degrees of freedom were used to determine the effect significance in relation to individual measures of tracking performance.

Euclidean distance tracking error

The composite measure of tracking accuracy, Euclidean distance error, was significantly influenced by variations in display type (F[3,78]=46.93, $p<0.001$). A clear indication of this effect can be seen in figure 3 which portrays the mean Euclidean distance error scores associated with each display type.

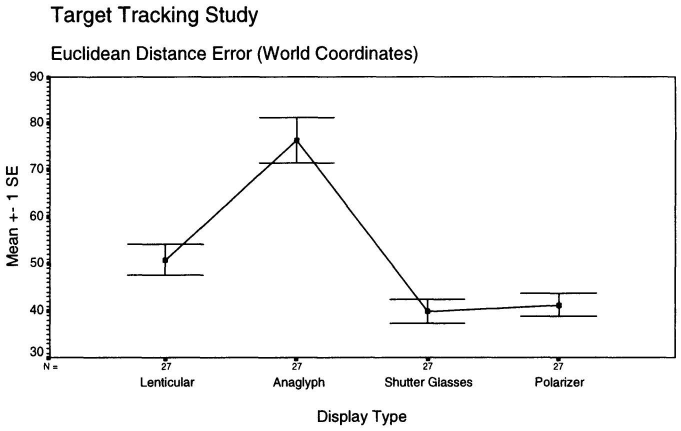


Figure 3 Euclidean distance tracking error

These apparent differences in display performance were investigated using a Student-Newman-Keuls multiple range test conducted at the 0.05 significance level. Results confirmed the world coordinate Euclidean distance error to be significantly greater for the anaglyphic display (76.3) than for the lenticular (50.8), shutter glasses (39.9) or polarizing (41.3) variants. No significant differences were found between either the lenticular, shutter glasses or polarizing displays, suggesting an equitable performance in respect of this composite measure of spatial tracking accuracy.

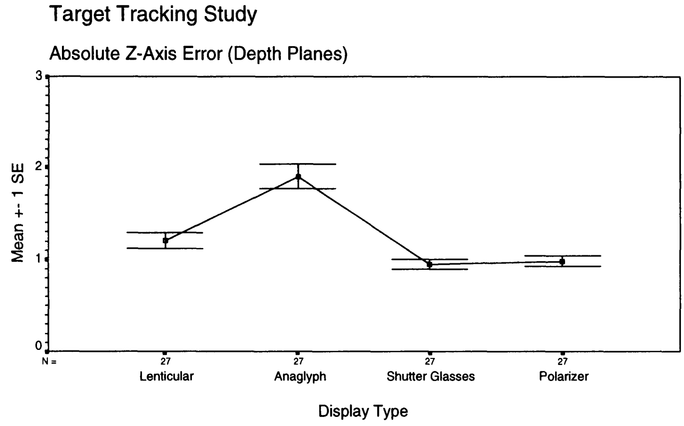


Figure 4 Average magnitude of depth axis tracking errors

Z-axis tracking error

In accordance with expectation, the largest impact on Euclidean distance error was made by the cursor/target displacements arising along the depth axis. It is generally acknowledged that performing accurate cursor positioning in depth is inherently more difficult than conventional manipulations in the X-Y plane (Beaton *et al.*, 1987). This observation seems to hold true regardless of the stereoscopic technique or the spatial positioning device employed, however stereoscopic display fidelity is expected to influence the magnitude of depth placement errors.

Figure 4 illustrates the absolute average tracking error measured in discrete depth planes along the Z-axis for each display. The most striking feature is the profile similarity of figures 3 and 4, establishing Z-axis deviation as the major component of Euclidean distance error. Not surprisingly, the type of stereoscopic display was found to have a significant influence on depth axis tracking error (F[3,78]=46.55, $p<0.001$). The only critical difference in performance means lay between the anaglyph display (1.9) and the lenticular (1.2), shutter glasses (0.96) and polarizing (0.99) variants. This result confirms the anaglyphic display as the worst performer during continuous spatial positioning operations, leaving the lenticular, shutter glasses and polarizing displays equal on merit.

A peculiar trend was detected during analysis of the mean z-axis error scores (measured in depth planes) for each subject participating in the tracking study. All data displayed a strong negative bias, indicating that subjects consistently positioned the stereoscopic cursor well in front of the target stimulus. Although the foreground positional bias was evident for all stereoscopic displays, the magnitude of the effect was display dependent (F[3,78]=56.91, $p<0.001$). Student-Newman-Keuls tests revealed the effect to be greatest for the anaglyph technique (-1.6), while for the first time, the lenticular display (-0.75) registered an effect size worse than either the shutter glasses (-0.34) or polarizing (-0.41) techniques.

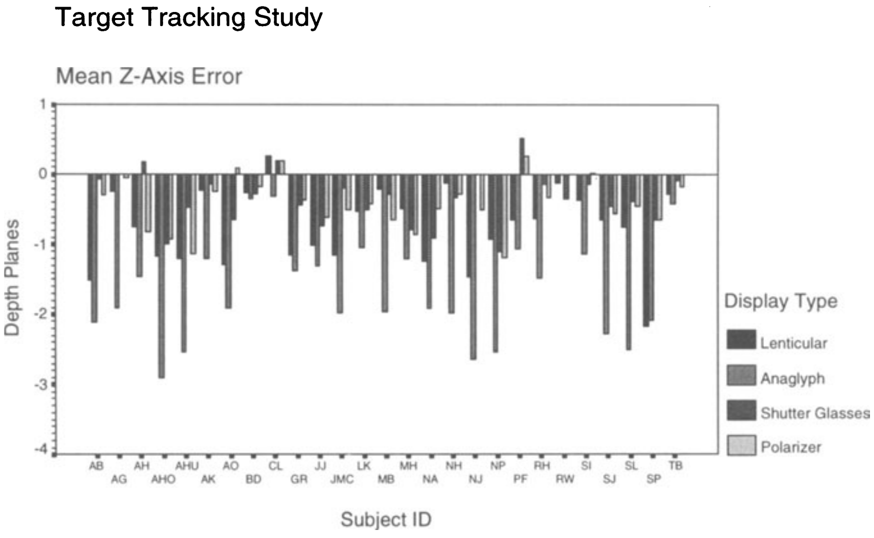


Figure 5 Mean Z-axis tracking errors

Figure 5 shows the mean z-axis deviation for each subject when using each of the four display types. The figure provides a clear indication of the consistent tendency to underestimate target depth.

Target acquisition performance

All multivariate test criteria (Pillai's Trace F[12,249]=3.7, p<0.001, Wilks' Lambda F[12,214]=3.8, p<0.001 and Hotelling's Trace F[12,239]=3.9, p<0.001) suggest that the single within-subjects factor *display type* has a significant effect on all combined measures of target acquisition performance. Univariate F-tests conducted with (3,84) degrees of freedom were used to determine the effect significance in relation to individual measures of target acquisition performance.

Decision response times

Surprisingly, the observed variation in decision responses times for each of the stereoscopic displays barely achieved statistical significance (F[3,84]=4.05, p=0.01), and Student-Newman-Keuls tests revealed that mean response times for the lenticular (7.64 seconds), anaglyph (7.34 seconds), shutter glasses (6.49 seconds) and polarizing (7.44 seconds) displays were not significantly different at the 0.05 level.

Percentage of correct depth correlations

The number of search errors committed by subjects when trying to identify co-located cue and target symbols was significantly influenced by stereoscopic display methodology used (F[3,84]=9.29, p<0.001). Figure 6 shows the average correlation performance or 'hit rate' attained on each display, expressed as a percentage of the total number of depth judgements performed.

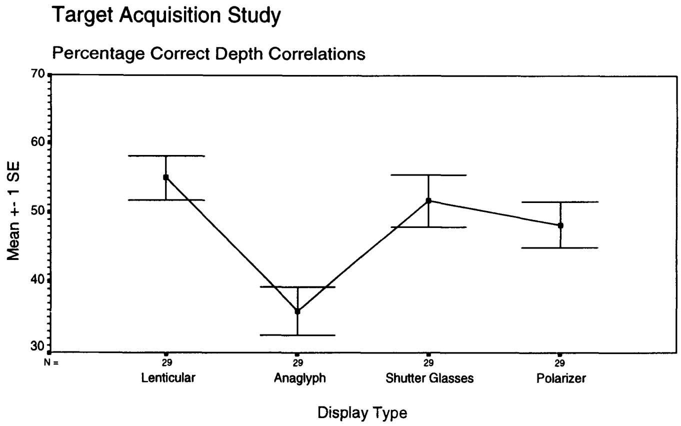


Figure 6 Percentage of correct depth correlations

The apparent superiority of the lenticular prototype was investigated using Student-Newman-Keuls multiple comparisons conducted at the 0.05 significance level. Test results failed to confirm any significant performance advantage for the lenticular display, indicating instead that the lenticular (55%), shutter glasses (52%) and polarizing (48%) variants performed equally well in respect of depth correlation accuracy. The anaglyph display was (predictably by now) the worst performer, with a mean percentage score of 35.

The hit rates attained by all subjects using each display were well above the anticipated 10% threshold. That is to say, each of the four display techniques substantially increased the number of correct depth correlation judgements beyond levels normally associated with random stimulus selection (2 in 20).

Depth correlation error

The absolute magnitude of errors arising from incorrect cue/target depth correlations was found to be dependent on display type (F[3,84]=11.11, p<0.001). The degree of absolute correlation error associated with each display type is shown in figure 7. On average, depth judgement errors are little more than 1 depth plane in magnitude for all of the displays participating in the study. Depth judgement errors reported in previous investigations are typically one unit pixel disparity in size (Drascic, 1991; Yeh and Silverstein, 1989). As depth planes and pixel disparities are synonymous in the current display context, the observed correlation errors are within expected limits.

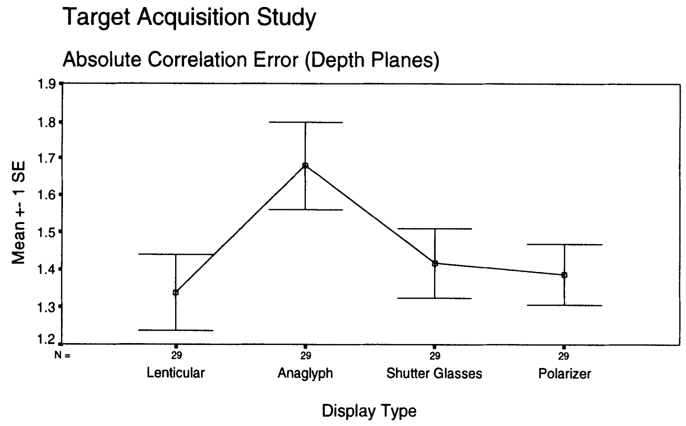


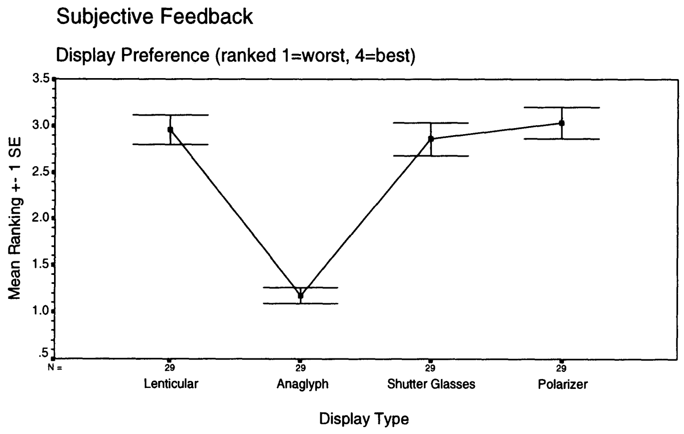
Figure 7 Average depth correlation error

The comparatively poor performance of the anaglyphic display (1.68) was confirmed during post hoc analysis. However, means comparisons also concluded that error magnitudes for the lenticular (1.34), shutter glasses (1.42) and polarizing (1.39) displays were not critically different. No significant bias to either the front or rear of the cue stimulus was detected, suggesting correlation performance to be independent of the relative depth placement of cue and target stimuli.

Subjective ranking scores

Following the conclusion of the target tracking and acquisition exercises, participants completed a subjective ranking scale to determine their display preference. Subjects were instructed to base their decision primarily upon the confidence level experienced while performing the tracking and acquisition tasks. If subjects were unable to differentiate between specific displays using this criteria, then additional considerations such as viewing comfort, fatigue, image quality and other extraneous factors were to be taken into account. A integer between 1 and 4 was uniquely assigned to each display format in order of increasing merit.

Display type was found to have a significant effect on subjective ranking scores (F[3,84]=25.84, p<0.001), the means of the scores awarded for each display appear in figure 8. Post hoc analysis failed to detect any significant differences in mean display rankings between the lenticular (2.97), shutter glasses (2.86) and polarizing (3.03) systems, however the anaglyph display was clearly isolated in terms of subject preference with a mean ranking of 1.17.



Subjective Feedback

Display Preference (ranked 1=worst, 4=best)

Figure 8 Average display ranking scores

DISCUSSION

The most salient result to emerge from the preceding analysis is the comparative performance of the lenticular, shutter glasses and polarizing display systems. Selected human performance measures indicate these three displays are inseparable in terms of their ability to support spatial tracking and visual search operations; a verdict supported by subjective ratings of display performance. The comparatively poor results obtained with the anaglyph display system suggests the evaluation strategy and response measures used were an effective tool for discriminating stereoscopic display performance.

It is tempting, at first, to attribute poor spatial tracking performance on the anaglyph display to difficulties in replicating the monocular interposition cues so vital to depth perception.

However, Barham (1991) suggests that stereopsis is used to achieve initial cursor placement only, after which interposition cues become dominant, allowing finer depth adjustments. The sheer magnitude of the Euclidean distance and Z-axis tracking errors associated with the anaglyph technique indicate that more complex perceptual difficulties exist when using this display. After all, the target acquisition task used a visual stimulus that was totally devoid of object interposition cues and this produced an equally damning verdict on anaglyph display performance in respect of depth correlation accuracy.

The tendency for subjects to underestimate object depth during the target tracking study gives cause for concern. Drascic (1991) has noted a similar phenomenon in his experiments to compare the positioning accuracy afforded by a stereovideo representation of a real world pointer and a computer generated stereographic pointer. A small but significant foreground bias was detected during the alignment of both the real and virtual pointer with a static target stimulus. Drascic offers no concrete explanation for the observed phenomenon, claiming the lack of an experimental precedent and the failure of researchers to report mean positioning errors is hindering the search for a solution. A cursory inspection of the subject profiles obtained during post experimental interview would indicate that an individual's propensity for underestimating object depth is inversely proportional to their stereo experience level. Consequently, a longer stereo acclimatisation period may well see a reduction in foreground positional bias during spatial tracking operations.

The lack of significant response time effects associated with the target acquisition task was disappointing, as it suggests that stereoscopic display discrimination can not take place on the basis of timed assessments of task performance. However, it appears that this observation is not without precedent. During an investigation of monocular and binocular cue saliency levels, Reinhart (1990) discovered that the inclusion of binocular disparity cues did not produce significantly faster response times for subjects performing simple relative depth judgements. Zenyuh (1988) also concluded that a stereoscopic display format did not significantly improve response times during visual search and object counting operations. Unfortunately neither investigator offers a conclusive explanation for their observation. There evidence to suggest that a degree of cognitive capture occurs during exposure to stereoscopic display techniques (Miller and Beaton, 1991), and this is likely to confound measurements of human performance. Response times can be expected to decrease as subjects gain proficiency during speeded depth judgements while progressing form one display to the next. Such variations can mask the impact of differential display performance on decision response times. Again, extensive stereo pre-training should be considered to exhaust such learning effects and heighten the influence of disparity cues.

In retrospect, the inclusion of a fifth, monoscopic display format, would have provided a useful experimental control throughout the display evaluation study. Performance metrics obtained on a monoscopic display format would establish a base-line performance reference to assist in the verification of any apparent stereoscopic display advantages.

CONCLUSIONS

The validity of comparing dissimilar display implementations was initially of some concern. All of the displays in the evaluation share common objectives - the portrayal of a spatial image - it might be argued that the manner in which these objectives are achieved differs sufficiently to preclude a direct comparison.

Simple differences in physical display attributes such as screen resolution for example, can have a profound influence upon display efficacy. A consequence of finite two dimensional display resolution is that the depth dimension is quantized; in the absence of suitable anti-aliasing techniques this quantization yields discrete depth planes corresponding to integral pixel disparities. Because screen pixel pitch varies between display implementations, the range and separation of the available depth planes will too. It may be argued that relative depth judgements are easier to perform when depth resolution is coarsely quantized, a perverse consequence of this being that a lower depth resolution will enhance depth correlation judgements. Clearly an objective display comparison must incorporate a comprehensive range of response measures and task scenarios.

Preliminary results obtained from this investigation, indicate that stereoscopic display techniques can indeed be differentiated on the basis of a quantitative assessment of human task performance. Unfortunately, the inability to discriminate between lenticular, shutter glasses and polarising display systems suggests that the simple tasks adopted during the evaluation were insufficient to resolve subtle differences in performance. Extending the nature and scope of the experimental tasks is expected to yield an even more decisive measure of display quality.

Finally, the authors would like to express their appreciation to the 29 observers who took part in this investigation.

BIBLIOGRAPHY

R. L. Pepper, R. E. Cole, and D. C. Smith, "Operator performance using conventional or stereo displays," *SPIE Three dimensional imaging*, vol. 120, pp. 92-99, 1977.

W. S. Kim, S. R. Ellis, and M. E. Tyler, "Quantitative evaluation of perspective and stereoscopic displays in three axis manual tracking tasks," *IEEE Transactions on systems man and cybernetics*, vol. Smc-17, pp. 61-72, 1987.

W. F. Reinhart, "Depth cueing for visual search and cursor positioning," *SPIE Stereoscopic displays and applications II*, vol. 1457, pp. 221-232, 1991.

H. Takemura, A. Tomono, and Y. Kobayashi, "A study of human-computer interaction via stereoscopic display," *Proceedings of III Conference on Human-Computer Interaction*, vol. 1, pp. 496-503, 1989.

R. J. Beaton, "Displaying information in depth," *SID Digest*, vol. 1990, pp. 355-358, 1990.

W. Reinhart F, R. Beaton J, and H. Snyder L, "Comparison of depth cues for relative depth judgements," *SPIE Stereoscopic displays and applications* , vol. 1256, pp. 12-21, 1990.

J. P. Zenyuh, J. M. Reising, S. Walchi, and D. Biers, "A comparison of a stereographic 3-D display versus a 2-D display," *Proceedings of the human factors society-32nd annual meeting*, vol. 1, pp. 53-57, 1988.

R. Miller H and R. Beaton J, "Some effects on depth-perception and course-prediction judgements in 2-D and 3-D displays," *SPIE Stereoscopic displays and applications II*, vol. 1457, pp. 248-258, 1991.

Yei-Y. Yeh and L. D. Silverstein, "Visual performance with monoscopic and stereoscopic presentations of identical three-dimensional visual tasks," *SID Digest*, vol. 1990, pp. 359-362, 1990.

R. E. Cole, J. O. Merritt, S. Fore, and P. Lester, "Remote manipulator tasks impossible without stereo TV," *SPIE Stereoscopic displays and applications* , vol. 1256, pp. 255-265, 1990.

R. W. Butts and D. F. McAllister, "Implementation of true 3D cursors in computer graphics," *SPIE Three dimensional imaging and remote sensing imaging*, vol. 902, 1988.

S. P. Williams and R. Parrish V, "New computational control techniques and increased understanding for stereo 3D displays," *SPIE Stereoscopic displays and applications* , vol. 1256, pp. 73-82, 1990.

J. O. Merritt, "Common problems in the evaluation of 3D displays," *SID Digest*, pp. 192-193, 1983.

D. Drascic, "Pointing accuracy of a virtual stereographic pointer in a real stereoscopic video world," *Proceedings ITEC '91 International symposium on 3D imagery*, pp. 623-626, 1991.

S. L. Grotch, "Three-dimensional and stereoscopic graphics for scientific data display and analysis," *IEEE CG&A*, vol. November, pp. 31-43, 1983.

P. T. Barham and D. F. McAllister, "A comparison of stereoscopic cursors for the interactive manipulation of B-Splines," *SPIE Stereoscopic displays and applications II*, vol. 1457, pp. 18-26, 1991.

R. J. Beaton, R. J. DeHoff, N. Weiman, and P. W. Hildebrandt, "An evaluation of input devices for 3D computer display workstations," *SPIE True 3d imaging techniques and display technologies*, vol. 761, pp. 94-101, 1987.

Yei-Y. Yeh and L. D. Silverstein, "Depth discrimination in stereoscopic displays," *SID Digest*, pp. 372-375, 1989.

A. Reeves and C. A. Tijus, "The pop-out effect in simple 3D visual matching task," *Proceedings of the 3rd cognitive symposium*, pp. 559-564, 1990.

J. L. Devore, *Probability and statistics for engineering and the sciences*, 3rd ed. California: Duxbury Press, 1990.

M. J. Norusis, *Advanced statistics SPSS/PC+ for the IBM PC/XT/AT.* Chicago: SPSS Inc., 1986.

G. Keppel, *Design and analysis: A researchers handbook*, 1st ed. New Jersey: Prentice-Hall Inc., 1973.

T. N. Bardsley, "Design and evaluation of an autostereoscopic computer graphics display," Ph.D. Thesis, De Montfort University Leicester, 1994.