Quantification of Contrast Sensitivity and Color Perception using Head-worn Augmented Reality Displays

Mark A. Livingston* Jane H. Barrow Ciara M. Sibley

3D Virtual and Mixed Environments
Naval Research Laboratory

ABSTRACT

Augmented reality (AR) displays often reduce the visual capabilities of the user. This reduction can be measured both objectively and through user studies. We acquired objective measurements with a color meter and conducted two user studies for each of two key measurements. First was the combined effect of resolution and display contrast, which equate to the visual acuity and apparent brightness. The combined effect may be captured by the contrast sensitivity function and measured through analogs of optometric exams. We expanded the number of commercial devices tested in previous studies, including higher resolution and video-overlay AR displays. We found patterns of reduced contrast sensitivity similar to previous work; however, we saw that all displays enabled users to achieve the maximum possible acuity with at least moderate levels of contrast. The second measurement was the perception of color. Objective measurements showed a distortion of color, notably in the blue region of color space. We devised a color matching task to quantify the distortion of color perception, finding that the displays themselves were poor at showing colors in the blue region of color space and that the perceptual distortion of such colors was even greater than the objective distortion. We noted significantly different distortions and variability between displays.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/Methodology; H.1.2 [Models and Principles]: User/Machine Systems—Human factors

1 INTRODUCTION

Upon first looking through an augmented reality (AR) display, most users notice perceptual differences between the graphical and real portions of the environment. Three differences relate directly to important measures of human perception: limited display resolution, lower contrast, and shifted colors. The performance of the display in these areas affects perceptual measures of visual acuity, contrast sensitivity, and color perception. Thus it becomes necessary to know whether a display enables users to correctly perceive the mixed environment and perform higher-level tasks.

Head-worn AR displays have a number of aspects that affect the user's perception: optical elements, the display technology, and surrounding visual context. The first and last affect the user's perception of both the graphics and the real world; the second affects only the presentation of the graphics. *Optical see-through* displays use half-silvered mirrors or similar optical combiners to merge the graphics with the user's view. This means that (in all but a few research systems) the graphics will always appear semi-transparent and thus their appearance will be affected by the background object. This affects the contrast and color perception directly, but not the resolution. Displays that overlay graphics on video signals (colloquially known as *video see-through*, despite the inaccuracy with respect to the physics of the technology) do not suffer from this confusion, but are limited in their resolution to that of the camera that acquires the video signal. The camera and lens settings will also affect the contrast, focus, and color quality of the real imagery, and the properties of the optics will affect these properties for both real and graphical imagery. We do not consider other displays for mixed or augmented reality, such as projectors or hand-held devices, though similar issues may be investigated.

1.1 Visual Perception Measures

These three perceptual differences may be quantified. In fact, the measure of visual capability of which most people first think corresponds to the resolution. *Visual acuity* is defined as the ability of an observer to discriminate fine details in the visual field. It measures the smallest stimulus the observer can resolve. Normal acuity is approximately one minute of arc at a distance of 20 feet [10]. The typical method of measuring visual acuity is with a chart of targets of different sizes which the subject must identify by name or direction, such as letters or shapes.

Contrast sensitivity describes the observer's ability to discern luminance differences in an image. This has been accepted as part of a comprehensive approach to describing visual capabilities [3]. Contrast is frequently expressed by the Michelson definition:

$$C = (L_{\max} - L_{\min}) \big/ (L_{\max} + L_{\min}) \ ,$$

where $L_{\max}$ and $L_{\min}$ are, respectively, the maximum and minimum luminances in the image. Many eye charts for measuring visual acuity may be adapted to testing contrast sensitivity by changing the foreground and background shades rather than target size.

The combined effect of contrast sensitivity and visual acuity is encapsulated in the contrast sensitivity function (CSF), which plots a curve that demarcates discernable and non-discernable regions of the 2D domain defined by the visual frequency (inverse of the visual angle subtended by a target feature) and contrast.

Color perception results from a complex set of retinal responses to light. The three types of cones (red, green, and blue) in the retina respond to different wavelengths of light, creating the effect people interpret as color. CIE L*a*b* decomposes color into a luminance channel (L) and two hue channels. The a axis moves from green to red; the b axis moves from blue to yellow. This description of color closely matches one model of how the human visual system processes wavelengths of light into color. This space is (nearly) perceptually uniform, meaning that distance comparisons in different regions of the space are valid.

1.2 Visual Perception in Head-worn Displays

Video overlay AR systems limit the user to the resolution (spatial and color) and field of view (FOV) of the camera, modulated by the display quality. Visual acuity and color perception with the real

*e-mail: mark.livingston@nrl.navy.mil

IEEE Virtual Reality 2009
14-18 March, Lafayette, Louisiana, USA
U.S. Government work not protected by U.S. copyright

world have been tested in such systems [1]. Visual acuity through the camera was degraded, but no quantitative data were reported. Success rate on a Dvorine pseudo-isochromatic color test for color blindness dropped from 97.3% to 91.3%, remained at that level during testing, and rose to 96.7% in a post-test. Color identification dropped from 98.9% accuracy to 62.2% accuracy. Some adaptation occurred; after completion of the experimental task, color identification rose to 70.0% accuracy while still wearing the AR display. Accurate (100.0%) color perception returned after removing the display. No details were given on what constituted accurate color identification.

A test of four optical see-through AR displays [11] investigated the smallest real targets visible from one meter with the display off and with the display showing a blank screen. The latter condition implies the display emits some light and, in the case the Sony Glasstron PLM-50, enables a filter that reduces transmittance of light from the environment. Two binocular displays showed differences in these two conditions. The Glasstron (33° measured horizontal FOV) allowed users to see 1 mm targets with no power (filter off) but only 6 mm targets with power (and filter) on; I-glasses enabled (25°) 0.5 mm and 3 mm. MicroOptical Corp˙ Clip-On (10°) and EyeGlass (17°) both allowed users to see 0.5 mm targets.

Snellen eye charts were included in the AR Performance Assessment Battery [5]; twenty subjects were tested. A Sony Glasstron (SVGA, 27° horizontal FOV) yielded 20/40 scores; a Microvision Nomad (SVGA, $\approx 20°$ horizontal FOV) yielded 20/30 and 20/40 scores. Another Sony Glasstron (LDI-D100B) caused eight users with normal or corrected-to-normal vision (i.e. 20/20 or better) to drop to 20/30 or worse looking through the optics of the display. All users scored 20/30 on a graphical chart [7].

Evaluation of a head-mounted projection display (HMPD) was done with a modified Landolt-C acuity test [2]. Users identified the open side in a square under three levels of light. The display limited subjects to a resolution of 4.1 arcminutes, or a Snellen score of 20/82 for all lighting levels. The type of retro-reflective material used for the HMPD affected performance with low contrast targets.

In previous experiments [6], we tested contrast sensitivity and color naming. The Sony Glasstron notably reduced the contrast sensitivity of the user compared to his or her normal vision, though it should be noted that the contrast levels in this experiment were well below the standard for optometric exams. The Microvision Nomad also reduced contrast sensitivity, but by a far smaller amount. The reduction of contrast in the Glasstron appeared to also cause some color confusion near the white point of the CIE 1931 color space, especially when looking at real-world objects through the see-through optics. Color samples near the boundaries of named regions were inconsistently labeled, with lighter colors progressively less consistent in their names. Darker colors were less salient in the graphics with a white real-world background.

1.3 Displays in the Current Experiments

One of the goals in this experiment was to compare a wide range of displays: the Sony Glasstron LDI-D100B, Microvision Nomad, the nVis nVisorST, and the video overlay Trivisio ARvision. As with previous work, another goal was to compare to users' natural vision on a real object as a baseline. To provide a real object, we used an Acer AL1916w monitor placed at the focus distance for the experiment, chosen as described below. Though some devices are capable of stereo imagery, this was not tested.

The Glasstron uses two SVGA LCDs and optics that yield a fixed focus distance of 1.2 meters. It may be used in binocular mode; however, for these experiments, it was used in bi-ocular mode. (The two eyes received an identical image.) This and the fixed distance for all viewing should have avoided problems that may occur when using binocular displays in applications that require changing focus [8]. Based on previous work, we removed the Glasstron's opacity shutter for these indoor experiments. This shutter dims the real world to enable the user to see the graphics more easily.

The Nomad is a monocular, monochromatic (red) retinal scanning display with adjustable focal distance, controlled by a hardware slider. We matched the focus distance to that of the Glasstron; it remained fixed for all users. Users looked with their dominant eye for all trials. Despite different technology, the Nomad behaves in a manner typical for optical see-through displays.

The nVisorST uses two SXGA LCDs in an optical see-through system that focuses at a distance beyond arm's length. We operated it in bi-ocular mode, though we did allow the users to adjust the physical distance between the eyes to a comfortable separation; this feature was only available on the nVisorST among our displays.

The ARvision uses a stereo pair of NTSC cameras and SVGA displays. We output a single camera into a video combiner that resampled to SVGA output and overlaid the graphics on the video. The merged image was fed into both eyes (bi-ocular mode). We focused the camera on the monitor, which gave it the same focus distance as the Glasstron.

Aside from the display devices used, AR experiments require that we consider the conditions in which users view stimuli. As with previous work, we tested multiple conditions that correspond to AR scenarios: looking at the real world (i.e. the monitor) with no AR graphics (the "see-through" condition) and at graphics shown on the displays. One must also consider the background when looking at graphics on an optical see-through display, since these graphics can not occlude the real world. Even in a video overlay display, it is worthwhile to consider the context for a color stimulus, since the color perceived may vary with the surroundings. Finally, we tested users' natural vision as a baseline for comparison.

2 Contrast Sensitivity Experiment

With the summary of displays provided above, we can begin the objective evaluation of the effective acuity of the displays. But as noted in Section 1, this tells only part of the story. To predict the limits of visibility in the display, we must also know the contrast levels of the AR displays. We report objective measurements of the contrast and the contrast sensitivity as measured by a user study.

2.1 Objective Evaluation

Table 1 summarizes the expected visual acuity using horizontal FOV measurements. We used stimuli of exact integer pixel sizes in this experiment; we did not attempt to normalize the displayed size (as measured by visual angle) of the stimuli due to the aliasing problems this creates in the stimuli (as in [6]). Each display, when in use, was affixed to a chin rest apparatus (Figure 1).

The numbers give geometric insight into the visual capability a display allows. However, to truly measure the visual capability, we must also account for the display's contrast. We aimed a StellarNet EPP2000C spectrophotometer with CR2 cosine receptor through

Display	Res	h-FOV	Pix/∘	Min/Pix	Snellen
Monitor	UXGA	18.7°	85.5	0.70	20/15
nVisorST	SXGA	48.0°	26.7	2.25	20/45
Nomad	SVGA	23.7°	33.8	1.78	20/35
Glasstron	SVGA	28.1°	28.5	2.11	20/42
ARvision					
graphics	SVGA	24.9°	32.1	1.87	20/38
real	NTSC	24.9°	25.7	2.33	20/47

Table 1: Measurements and theoretical acuity for each display. Horizontal FOV (h-FOV) measurements come from the optical bench. Pixels per degree and minutes per pixel are converted to predicted Snellen score. For the ARvision, graphics are produced in SVGA, but the NTSC camera images are stretched to SVGA.

Figure 1: The experimental scene, consisting of the chin rest and monitor, with the Glasstron positioned for testing. This image corresponds to the viewing condition "Glasstron Through." Modified Landolt-C stimuli are shown (upper left: contrast=0.2, gap=2 pix, open left; upper right: contrast=0.025, gap=5 pix, open up).

each display and measured foreground and background luminance of the stimuli. We chose only one background to reduce the time required from users. For this contrast sensitivity experiment, we decided to use a black background for the graphical stimuli (black felt on posterboard).

We selected the input contrast levels used in [6] for the real and AR graphics. The graphs of the measured contrast in both conditions (Figure 2) show the differences between the various display devices. The nVisorST and the Nomad both have quite transparent lens systems, so the Through condition contrasts are close to that for the realVision condition, in which there are no intervening displays or optics in the user's view of the stimuli. Even without its opacity shutter, the Glasstron reduces the contrast of the real world significantly compared to realVision. The ARvision has lower output contrast at low input contrast, but seems to improve as the contrast level increases. In the Black condition, the ARvision, which as a video overlay device is not noticeably affected by the background, achieves the highest contrast of any display device. The Nomad, with its laser-scanning technology, achieves the next highest contrast, followed by the nVisorST.

2.2 User Study Design

Variables and Hypotheses We updated our previous design [6] with the stimuli from [2] to avoid the aliasing problem.

- *Pixel gap width* $\in \{ 1, 2, 3, 4, 5 \}$ *within subjects*
  (We label by pixels since visual angles are different for each display.) Hypotheses: Smaller gaps would be progressively harder to see, reflected both in increased time and decreased accuracy of the responses (consistent with practical experience [10]); all users would be accurate for all gap sizes in the realVision condition (as per the estimate in Table 1).

- *Contrast* $\in \{0.0125, 0.025, 0.05, 0.1, 0.2\}$ *within subjects*
  (The five input levels of contrast are maintained as labels for ease of reference.) Hypotheses: Lower contrast would yield slower, less accurate responses (consistent with practical experience [10]); users would perform better than guessing with their natural vision for even the lowest contrast.

- *Display condition* $\in \{$ realVision, nVisor Through, nVisor Black, Nomad Through, Nomad Black, Glasstron Through, Glasstron Black, ARvision Through, ARvision Black $\}$
  mixed design: within and between subjects
  Hypotheses: Based on Table 1 and experience with the displays, we expected that the see-through conditions would be quite easy with the Nomad and nVisorST, but quite challenging for the Glasstron and ARvision. We expected the displays to be nearly equivalent in the graphics conditions.

- *Direction* $\in \{$ up, down, left, right $\}$ *within subjects*
  Hypotheses: we expected no significant differences; mostly, this variable provides repetition of other conditions.

Each user completed a realVision condition and, to reduce time, used only two display devices with the Through and Black conditions, for five display conditions, yielding $5 \times 5 \times 5 \times 4 = 500$ trials per subject from which data was analyzed (8000 total data points).

Subjects Sixteen subjects (12 male, 4 female) completed the experiment; they ranged in age from 21 to 39, with a mean of 27.6 years. All volunteered and received no compensation. Our subjects were drawn from the science and engineering or office clerical staff at our lab. All reported being heavy computer users; half also reported being frequent players of video games. Examining the data from the realVision condition, all users were of comparable and normal (or corrected-to-normal) visual capability. Subjects did not report having any trouble learning or completing the experiment.

Experimental Task and Dependent Variables The subjects pressed one of the four arrow keys on a standard extended keyboard to indicate which of the four sides of the modified Landolt-C contained the gap. A gap was always present, and the accuracy of this response was the primary dependent variable. We also recorded the time that elapsed between the onset of the stimulus and the user's response. Each subject trained with five trials of the task on the first viewing condition they used. Between sets and halfway through each set, the user could take a break for as long as desired; users generally took just the few minutes it required to set up the next viewing condition between sets and a few moments in the midst of a set. The order of presentation of the targets was a random permutation computed at the time of the experiment. After the user responded to a stimulus, the display (monitor or head-worn display) rendered a black screen (which produced the see-through condition for the head-worn displays) for two seconds, then displayed the next stimulus. The order of display conditions was counterbalanced within each half of the user pool using a Latin square design.

2.3 Results of User Study

We analyzed each half of the subject pool separately. To verify that differences in performance were not due to differences between the groups, we considered performance on the baseline display condition, realVision. Group 1 had a mean error of 0.035 with standard deviation 0.184; Group 2 had 0.020 and 0.140. Linear regression analysis revealed that group was not a significant predictor of error scores ($\beta = 0.046, p = 0.067$), accounting for only 0.2% of the variance in error scores. In other words, even if there were a significant difference between the two groups, the effect size was tiny. Therefore, we can be confident that the two groups were not significantly different at baseline and comparisons between the displays each group used are valid.

The CSF (Figure 11, color plate) shows that the nVisorST was the best overall for graphics and see-through conditions, while the Nomad also allowed users to achieve nearly the same contrast sensitivity as realVision. The Glasstron's graphics are quite sharp, but it reduced the real world clarity; the Nomad did the inverse. The ARvision limited the clarity of both the real and virtual imagery.

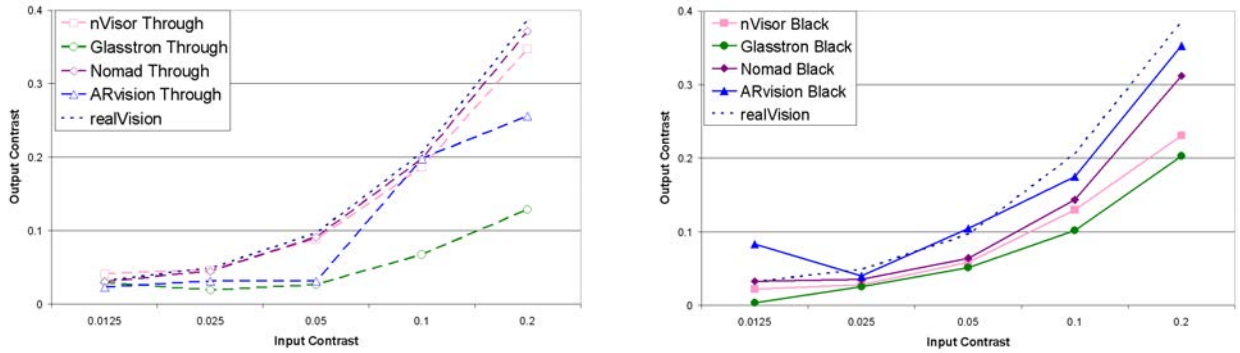

Figure 2: Contrast measurements for the display conditions used in the acuity and contrast experiment as a function of the theoretical (input) contrast. (*Left*) See-through conditions and realVision condition. (*Right*) Looking at graphics on a black background and realVision condition.

We measured the (binary) error and response time and performed a $5 \times 5 \times 5 \times 4$ within-subjects analysis of variance (ANOVA) with the `|STAT` analysis package [9] on each group of subjects. To ensure that these effects survive any non-sphericity that may violate assumptions in the ANOVA, we applied the very conservative corrective factor $\varepsilon = (n-1)^{-1}$ to the F-test [4]. Though this raised the requirements to demonstrate statistical significance, we still yielded significant results. Table 2 summarizes the main effects, their corrective factors, and the ANOVA results with the corrections.

Display Condition The CSF foreshadows the main effect of the displays. The nVisor Through and Nomad Through conditions enabled nearly the same performance as realVision. Our hypothesis was correct; the optics are the most transparent and thus enabled the best performance. The Glasstron Through condition had extremely low contrast in the objective measurements; this manifests itself here. The dimness of the real world may have actually helped the users in the Glasstron Black condition, however. Interestingly, the monochrome graphics of Nomad Black also presented some difficulty for the users. Users performed poorly with the ARvision in both Through and Black conditions. This would seem to be the result of the low camera resolution for the Through and the difficulty subjects reported keeping their eyes in the exit pupil of the display for both Through and Black. It is worth noting that none of the AR display conditions – neither looking through the displays nor looking at graphics on the displays – enabled the users to equal their performance with their natural vision.

The main effect of display on response time (Figure 3) is similar. Users were fastest in realVision and nearly as fast looking through the (clear) optics of the nVisorST and Nomad displays. The Glasstron Black condition was also quite fast, while Glasstron Through and both ARvision conditions, likely for the reasons above, slowed users down.

Contrast Not surprisingly, contrast had a main effect on the error rate. The users' overall error rate was negatively related to the input contrast within both groups (Figure 4), and this effect was quite consistent across the display conditions. However, as noted in [6], the output contrast is not linear throughout the intensity range available, so more investigation may reveal further insights. The main effect of contrast on time shows an approximately exponential decrease in time required with increasing contrast (Figure 5).

Gap size As expected, smaller openings were progressively harder to see (Figure 6). Once again, since gap size is not the same for the displays, the CSF is more descriptive. Subjects were faster as the gap size increased (Figure 7); however, we see diminishing returns at sizes greater than three pixels.

Gap direction The direction of the gap – up, down, left, or right – showed a trend in the first group and a significant main effect in the second group. The up and down directions were slightly

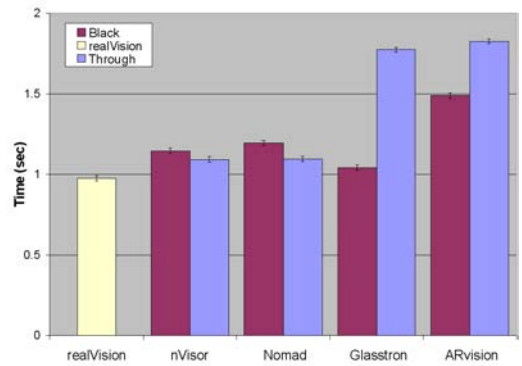

Figure 3: The main effect of display on time shows that users were fastest with their natural vision and nearly as fast looking through the nVisorST and Nomad displays or at Glasstron graphics. Subjects were generally slower with the remaining display conditions.

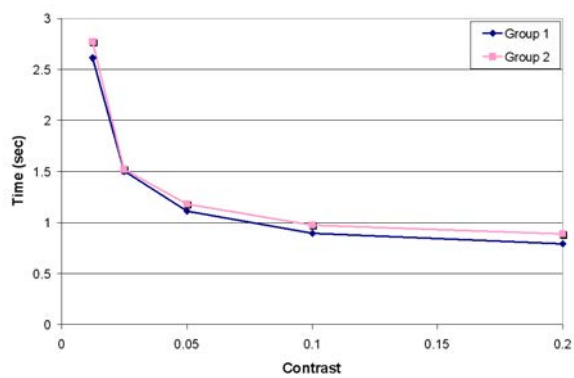

Figure 5: The main effect of contrast on time shows that users were progressively faster with increasing contrast.

more difficult for the users and caused them to be slightly slower. In the second group, there was a significant interaction between display and gap direction – F(3,21)=6.014, p=0.004. In the ARvision, pixels are larger vertically, so gaps on the left and right sides were in fact slightly larger and thus easier to see.

Afterimages We do not see any evidence of afterimages in our data. The potential afterimages constitute 13%-34% of the errors for the display conditions, 15%-24% for gap sizes, 21%-28% for contrasts, and 18%-24% for gap directions. The sine-wave patterns in [6] covered several degrees of visual angle; our gaps covered much smaller angles. Smaller size implies lower signal response in the visual system; this in turn lowers afterimage potential, so we are not surprised by the lack of afterimage effects.

	Effect	ε	n	d	Finding – Group 1		Finding - Group 2	
Main Effects on Error	Display condition	0.25	4	28	$F(1, 7) = 86.257$	$p = 0.000$	$F(1, 7) = 22.240$	$p = 0.002$
	Contrast	0.25	4	28	$F(1, 7) = 152.953$	$p = 0.000$	$F(1, 7) = 90.719$	$p = 0.000$
	Gap size	0.25	4	28	$F(1, 7) = 73.683$	$p = 0.000$	$F(1, 7) = 55.684$	$p = 0.000$
	Gap direction	0.33	3	21	$F(1, 7) = 4.276$	$p = 0.077$	$F(1, 7) = 8.844$	$p = 0.021$
Main Effects on Time	Display condition	0.25	4	28	$F(1, 7) = 22.425$	$p = 0.002$	$F(1, 7) = 18.923$	$p = 0.003$
	Contrast	0.25	4	28	$F(1, 7) = 196.262$	$p = 0.000$	$F(1, 7) = 236.271$	$p = 0.000$
	Gap size	0.25	4	28	$F(1, 7) = 184.508$	$p = 0.000$	$F(1, 7) = 127.130$	$p = 0.000$
	Gap direction	0.33	3	21	$F(1, 7) = 24.767$	$p = 0.002$	$F(1, 7) = 17.037$	$p = 0.004$

Table 2: Statistically significant effects in the two groups in the acuity/contrast experiment. Main effects are found in all four of the independent variables for both error and time, though gap direction was only significant for Group 2. ε is the (very conservative) corrective factor for the F-test; n and d are the numerator and denominator for the F-test. The finding for each group shows the corrected p values.

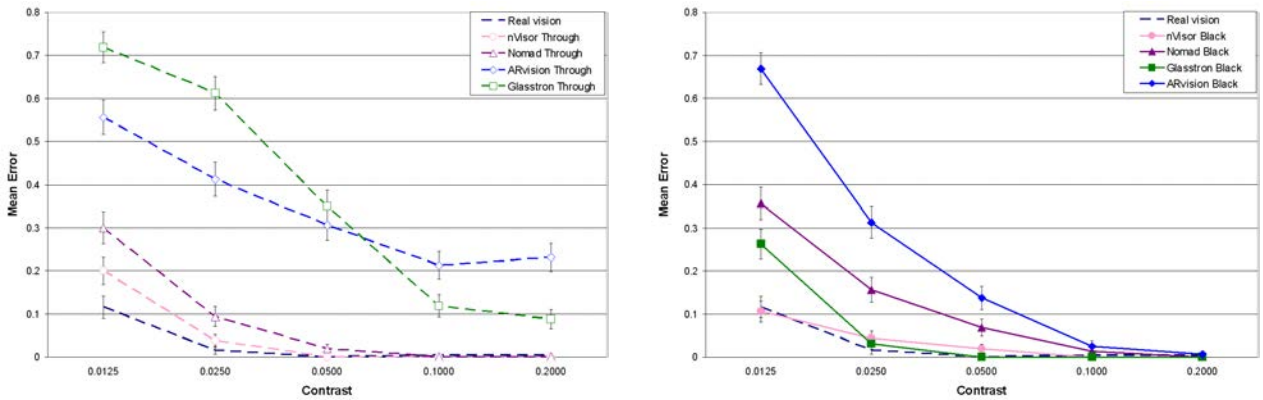

Figure 4: These two graphs show the main effect of the display condition and the main effect of the contrast, which shows a similar pattern of performance improvement for all display conditions. The left graph shows the see-through conditions along with the realVision condition, while the right graph shows the graphical conditions (black background) along with the realVision condition.

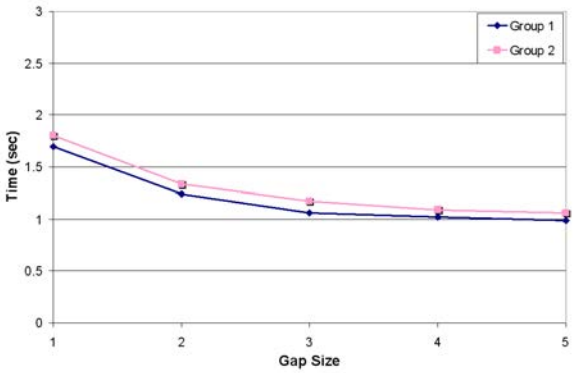

Figure 7: The main effect of gap size on time shows that users were, as expected, faster with increasing gap size, but that a diminishing return was in evidence with gaps larger than three pixels.

3 Color Perception Experiment

Initial experiments showed that users interpret color inconsistently in AR displays [1, 6]. We took objective measurements of color patches in the displays and developed a color matching task for a new user study. (The monochrome Nomad was not used in this experiment.) We expanded the graphics conditions to include both black (felt) and white (plain posterboard) backgrounds.

3.1 Objective Evaluation

Using the spectrophotometer in the same configuration, we measured CIE L*a*b* coordinates of color samples. We selected colors on a regular grid in a plane of Lab space. The plane was L=65 for the nVisorST and Glasstron, but the ARvision was washed out at this lightness; we used L=45 instead. The sampling grid was defined by the cross-product of $\{-67, -33, 0, 33, 67\}$ with itself. Removing the gray value (at the origin) gave us 24 color samples.

Figure 9 graphs the resulting measurements. The nVisorST comes the closest to matching the input sample locations. However, we note that the blue portion of the graph is not populated well by the measurements, except in the Through condition. In the White condition, the blue hues were overwhelmed by the background; the Black condition saw some distortion of the green colors. The Glasstron pulls all the samples into a small area of the graph, especially in the case of the white background. This matches our experience of having the graphics in the Glasstron overwhelmed by background illumination. Finally, the ARvision also distorts the input samples into a small area of the graph, with again the Through condition experiencing the most distortion. But this time, the measurements are pulled away from yellow into the blue region.

3.2 User Study Design

Our new study design mirrored the objective measurements, to take advantage of metric comparisons that can be made between the two types of results.

Variables and Hypotheses We provide the independent variables and hypotheses for each.

- *Display* $\in$ {nVisor, Glasstron, ARvision} *between subjects* (All displays offered 32-bit color.) Hypotheses: We expected that the ARvision would distort users' color perception most and that the nVisorST would distort it least.

- *Visual condition* $\in$ {Black, White, Through} *within subjects* (We separated the background and display in the analysis.) Hypotheses: we expected interactions with the variable Display, due to the different optics of various devices.

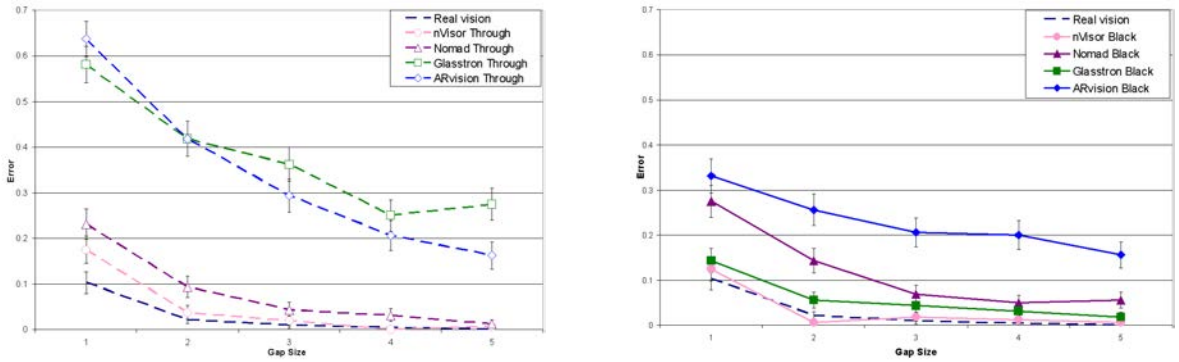

Figure 6: These two graphs also show the main effect of the display condition, but this time indirectly show the main effect of the gap size; this shows a similar pattern of performance improvement with increasing size of the gap for all the display conditions. The left graph shows the see-through conditions along with realVision, while the right graph shows the graphical conditions (black background) along with realVision.

- *Color* *within subjects*
  (We used 24 samples in CIE $L^*a^*b^*$ space, described above.) Hypotheses: We expected significant differences between colors; based on experience, we expected blue hues to suffer the most distortion. We emphasize that, for each display, intensity was equal between color samples.

Each user completed a realVision version of the task and used only one display device (with all three background visual conditions). There were no repetitions of color samples within each condition. This yielded $3 \times 24 = 72$ samples per subject (1728 data points).

Subjects Twenty-four subjects (17 male, 7 female) completed the experiment; they ranged in age from 20 to 69 (mean=31.3). All volunteered and received no compensation; several also participated in the acuity/contrast experiment. All subjects reported heavy computer use; eleven played video games frequently. All subjects passed a Farnsworth D15 color blindness test. Subjects did not report any trouble learning or completing the experiment.

Experimental Task and Dependent Variables Subjects viewed a reference and a target color patch. The target began as a neutral gray with the appropriate value for L (65 or 45, as above) for the display. Users controlled the target's color with a trackball mapped to the two hue parameters (a, b). To help them navigate through the space (with which none claimed to be familiar after concluding the experiment), bars indicating the color for each cardinal direction were placed around the target (Figure 12). To prevent the users from getting assistance from their natural vision, we prevented them from seeing the target patch unless they were looking through the display device. Users physically moved from one side of the barrier to the other, adjusting the target patch until they felt its color matched that of the reference patch, then pressed the space bar on a keyboard nearby. We recorded the hue parameters chosen by the user and the response time for each trial. We used Euclidean distance in color space between the actual hue and the user-selected hue as the dependent variable in our analysis.

3.3 Results of User Study

Display We found a main effect of display device on the error (Figure 8) – F(2,21)=3.941, p=0.035. The ARvision was clearly the most difficult display for the users in completing the task. Users were slower with the ARvision display (mean response time of 37.5 sec) than with the nVisorST (28.0 sec) or Glasstron (27.9 sec), but that this was only a trend – F(2,21)=3.084, p=0.067 – and that it was in the Black condition that ARvision users were slowest, not the Through condition.

Color Color had a main effect on error – F(23,483)=28.612, p=0.000. We graph the results for each display, separated by the visual condition, on $L^*a^*b^*$ color space (Figure 10). For the nVisorST, the cyan corner of the color space creates the most error in all three visual conditions, although one can see that the Black background did deteriorate the quality of the color matching in the yellow half of the graph compared to the White and Through conditions. There is increased variance in the blue (bottom) half of the space, and within that region, more in the cyan and magenta corners. For the Glasstron, there is a noticeable over-compensation occuring within the users' visual systems. One could understand this for the Through condition, which is seen through severely dimming optics, but this is harder to understand for the Black and White conditions. Again, there is more variance in the blue half of the graph, and slightly more in the green half than the red half. The ARvision notably led users away from the cyan and magenta corners in all visual conditions, and away from all four corners in the Black and White visual conditions. The variance is extremely high nearly everywhere.

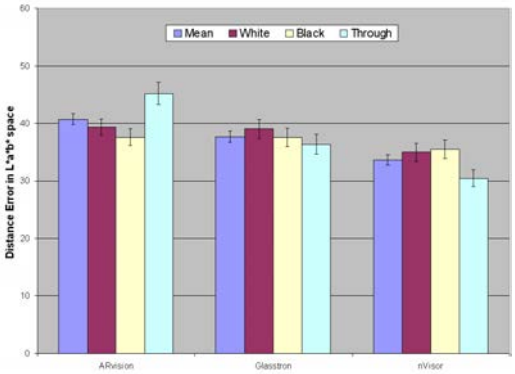

Figure 8: The main effect of display on error showed that the nVisorST yielded the best color perception; the ARvision yielded the poorest. The clear nVisorST optics benefitted users in the Through condition, while the ARvision suffered in the Through condition.

Color had a main effect on response time – F(23,483)=2.733, p=0.000. Users were slower on hues in the yellow (upper) half of the space: a greenish-blue, the most yellow shade, a pink-to-orange band, and the most red. Users were fastest with the six colors in the blue (bottom) half that were closer to green (left), where accuracy was poor. Perhaps users knew they were unable to perform the task well and thus did not try to be accurate.

4 Discussion

4.1 Contrast Sensitivity

Our experiments confirmed common experience with these displays and enabled us to quantify the performance we can expect users to

have on fundamental tasks of perception. As predicted, the clear optics of the nVisorST and Nomad were easy to see through; the dimmed real world through the Glasstron and the low-resolution camera image in the ARvision were not. On the graphical conditions, the nVisorST's high resolution certainly helped. The Nomad did not enable the sensitivity expected; users noticed a fuzziness in the display. Perhaps the variable focus mechanism was imprecise or the calibration of the eye to the exit pupil should be further investigated. Users also reported a bright spot in the middle of the display. The ARvision gave users problems with staying in the exit pupil, so the order predicted by the geometric analysis of acuity and the contrast measurements did not quite occur. Our hypotheses were correct only for the nVisorST and ARvision, for all contrasts, and for all gap sizes, as well as for all gap directions except in the ARvision display.

In particular, it is important to note that users had high error rates for single-pixel gaps in the Nomad and ARvision graphics, well above the error rate for their natural vision, whereas in the Glasstron and nVisorST, the error rates were only slightly higher than with natural vision. The Glasstron and ARvision displays did not enable users to see any of our targets very well when looking through to the real environment. Thus these displays are not recommended when fine details of the real world are important to the task. The quantitative results give minimum contrast values that should be heeded for AR visualizations and indicate that one-pixel objects are indeed visible in the graphics of most of these display conditions, but not necessarily looking through the displays into the real world. Data points under 0.375 mean error in Figures 4 and 6 indicate that users could see the stimuli and the confidence with which this may be true in general.

4.2 Color Perception

Our improved task design enabled us to make metric rather than qualitative evaluations of the performance of the users in perceiving colors. We can clearly see the objective and subjective distortion of color space with each of the displays. Not surprisingly, there is an interaction with the background. The nVisorST clearly yielded the best matches for the objective measurements, though the distortion away from the blue region of color space is notable in the graphics. This distortion also occured with the Glasstron. The Glasstron and ARvision both compressed the objective color space to a small portion of the color space, indicating the relatively lower passage of light through their optics and electronics. Thus it was surprising that users performed as well as they did on the color matching task. Since color is highly contextual, perhaps users were able to compensate for changes in color. Our users had some reference from the experimental apparatus which may have helped. The Glasstron reduced saturation, as demonstrated by the objective measurements. However, users apparently overcompensated during subjective measurements, showing the difficult nature of making this transformation cognitively. The nVisorST compressed blue colors somewhat beyond the objective distortion and, in the graphics, green colors as well. The ARvision led users away from the cyan and magenta corners. We also note that the consistency of perception was much better with the nVisorST, which may be important for applications in which a group of users must interpret colors.

4.3 Future Work

We have already identified one improvement we would like to make in the experimental design. The nVisorST's headband was not detached for this experiment, making it hard to move one's eyes to the display. With some practice, the users adjusted to the manuever, though it may have discouraged looking at the reference patch multiple times. Thus it may be that users could actually improve on their performance in the color matching task. Other interesting avenues for future work may also be identified. Adding more displays to the experiment would be of interest; we could explore contrast levels that occur outside the middle of the displayable intensity. We could also use standard color vision tests in place of our color matching task; such a test would give clinical but less numerical results. Additionally, we could explore the analysis further than the main effects presented here; some interactions found were excluded due to space limitations. Finally, a range of intensities and backgrounds could be explored for the color matching task.

4.4 Conclusions

These experiments expand on previous work by adding new displays; notably, the nVisorST demonstrates a readily apparent improvement over our other displays in the overall perception for both graphical and real portions of an AR environment. It offers notable improvement over the Glasstron (no longer available from the manufacturer, but still commonly used and sold among researchers). The performance of the ARvision indicates the need for detailed study of how video-overlay AR displays affect the user's perception. We found that head-worn AR displays are not yet enabling users to maintain their natural abilities in visual acuity and contrast sensitivity, though the optics do not necessarily interfere with their perception of the real environment. The distortion of color perception is a concern for applications in which color is a cue, but it appears that users may be able to overcome this. Still, it appears that avoiding a few key regions of color space would be recommended until a more complete mapping can be found.

Acknowledgements

The authors wish to thank James A. Ballas, Ellen Carter, Rob Carter, Stephen R. Ellis, Carlos Font, Patrick McKnight, Jannick Rolland, J. Edward Swan II, and the anonymous subjects. Support provided by the NRL Base Program.

References

[1] R. P. Darken, J. A. Sullivan, and M. Lennerton. A chromakey augmented virtual environment for deployable training. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2003)*, Dec. 2003.

[2] C. Fidopiastis, C. Fuhrman, C. Meyer, and J. Rolland. Methodology for the iterative evaluation of prototype head-mounted displays in virtual environments: Visual acuity metrics. *Presence: Teleperators and Virtual Environments*, 14(5):550–562, Oct. 2005.

[3] A. P. Ginsburg and W. R. Hendee. *Quantification of Visual Capability*, pages 52–71. Springer-Verlag, 1992.

[4] D. C. Howell. *Statistical Methods for Psychology (Sixth Edition)*. Thomson Wadsworth, 2007.

[5] S. E. Kirkley, Jr. *Augmented Reality Performance Assessment Battery (ARPAB): Object Recognition, Distance Estimation and Size Estimation Using Optical See-through Head-worn Displays*. PhD thesis, Instructional Systems Technology, Indiana University, May 2003.

[6] M. A. Livingston. Quantification of visual capabilities using augmented reality displays. In *Proceedings of the International Symposium on Mixed and Augmented Reality*, pages 3–12, Oct. 2006.

[7] M. A. Livingston, C. A. Zanbaka, J. E. Swan II, and H. S. Smallman. Objective measures for the effectiveness of augmented reality. In *IEEE Virtual Reality 2005 (Poster Session)*, pages 287–288, Mar. 2005.

[8] M. Mon-Williams and J. P. Wann. Binocular virtual reality displays: When problems do and don't occur. *Human Factors*, 40(1):42–49, Mar. 1998.

[9] G. Perlman. Data analysis in the unix environment. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, pages 130–138. Springer-Verlag, July 1982.

[10] L. A. Riggs. Visual acuity. In *Vision and Visual Perception*, pages 321–349. John Wiley and Sons, 1965.

[11] R. L. Woods, I. Fetchenheuer, F. Vargas-Martin, and E. Peli. The impact of non-immersive head-mounted displays (HMDs) on the visual field. *J. of the Society for Information Display*, 11(1):191–198, 2003.

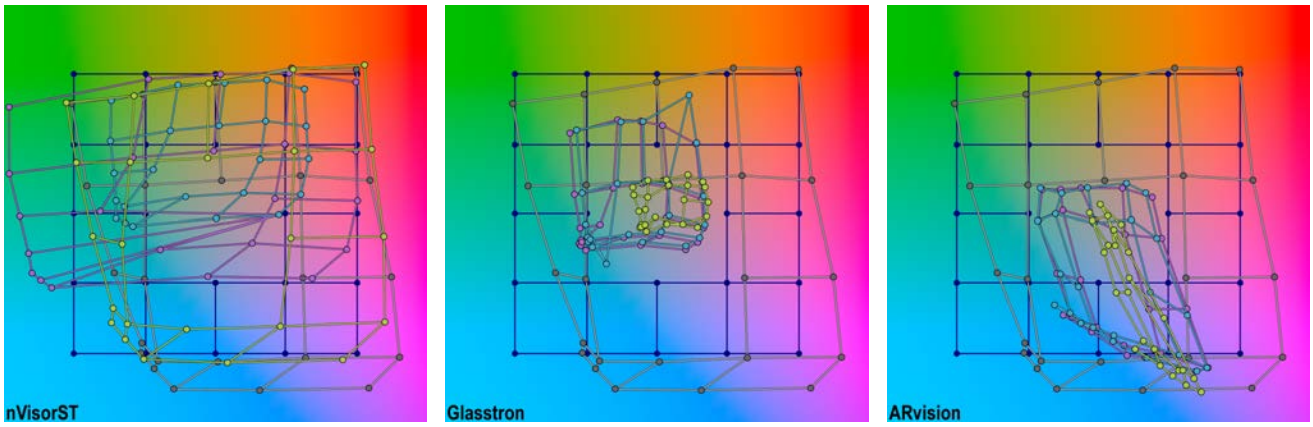

Figure 9: Measured color samples from our displays are graphed over Lab space. These graphs show the sampling grid (dark blue) and the measurements from the monitor used for the real world (gray). For each display, we measured samples in the Black (magenta points and lines), White (cyan), and Through (yellow) visual conditions. The last should be compared to the samples from the monitor (gray); deviation between these two reflect the distortion of color in the Through condition. The first two should be compared to the blue grid; this represents how colors of graphics are distorted by the display device in combination with the background. *Left*: nVisor, *Center*: Glasstron, *Right*: ARvision

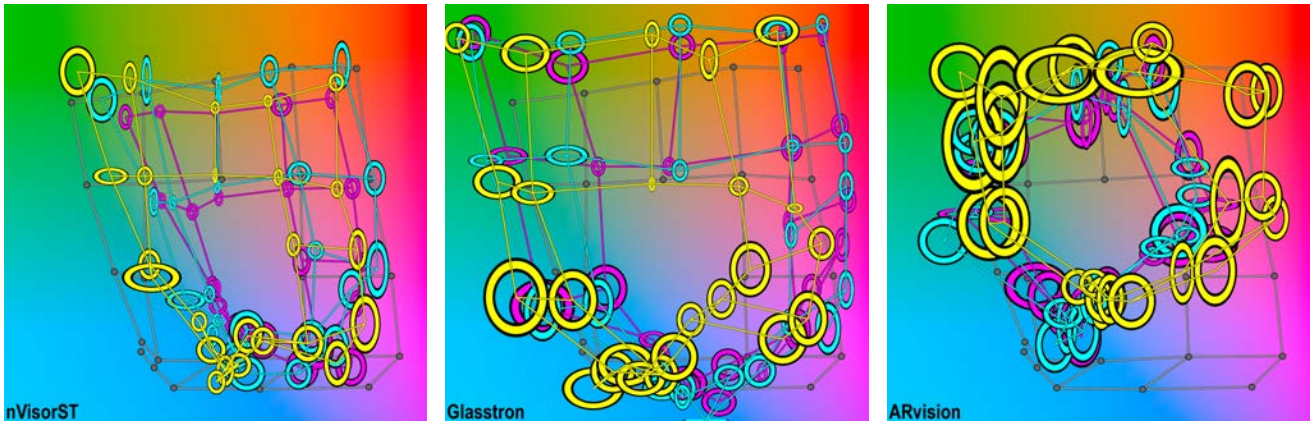

Figure 10: The main effect of the sample location in color space on the error in matching the color may be seen in these graphs. Each graph shows the measured samples from the monitor to which the users were matching colors. The users' performance varies in the Black (magenta), White (cyan), and Through (yellow) visual conditions. Each sample represents the mean location (the grid point and center of the circle) and the standard error in the two hue dimensions (radius of the error ellipse in each dimension). *Left*: nVisorST, *Center*: Glasstron, *Right*: ARvision

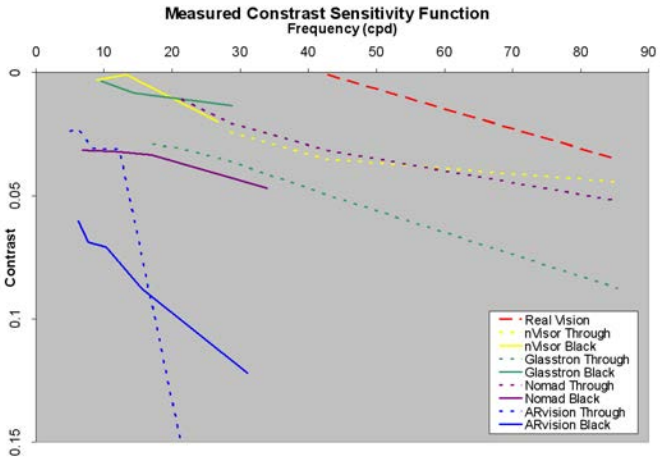

Figure 11: We interpolated and extrapolated segments of the CSFs; the visible region lies below each curve. Offsets between starting and ending positions along the frequency axis reflect the sizes of pixels (Table 1). Offsets along the contrast axis reflect measured contrasts and user performance of 62.5% accuracy, the threshold for better performance than guessing on a four-alternative forced-choice task.

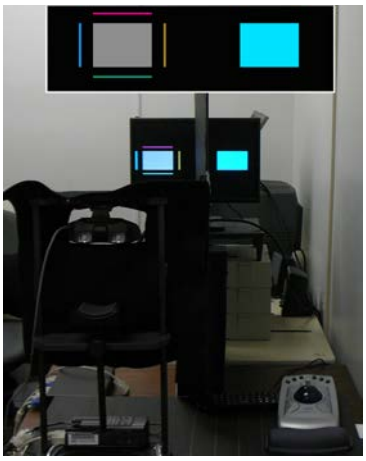

Figure 12: The color perception experiment showed side-by-side patches to be matched. The physical barrier shielded the user's view of the target (left) so that it was seen only through the AR display. With the Glasstron on the chin rest and both sides of the monitor visible, this image corresponds to the "Glasstron Through" condition. The inset shows an example initial target and a reference.