# Basic Perception in Head-worn Augmented Reality Displays

Mark A. Livingston, Joseph L. Gabbard, J. Edward Swan II, Ciara M. Sibley, and Jane H. Barrow

**Abstract** Head-worn displays have been an integral part of augmented reality since the inception of the field. However, due to numerous difficulties with designing using such unique hardware, the perceptual capabilities of users suffer when looking at either the virtual or real portions of the augmented reality. We discuss the perceptual background and a series of experiments – in the literature and in our laboratories – measuring the degradation of basic functions of the human visual system when using head-worn augmented reality displays. In particular, we look at loss of visual acuity and contrast (and how these in turn affect text legibility), distortion of perceived colors, and difficulties of fusing stereo imagery. We discuss the findings and the implications for head-worn display design.

Mark A. Livingston
Naval Research Laboratory, Washington, DC 20375, USA
e-mail: mark.livingston@nrl.navy.mil

Joseph L. Gabbard
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
e-mail: jgabbard@vt.edu

J. Edward Swan II
Mississippi State University, Starkville, MS 39762, USA
e-mail: swan@acm.org

Ciara M. Sibley
Naval Research Laboratory, Washington, DC 20375, USA
e-mail: ciara.sibley@nrl.navy.mil

Jane H. Barrow
George Mason University, Fairfax, VA 22030, USA
e-mail: jbarrow1@gmu.edu

# 1 Introduction

For many first-time users of augmented reality (AR) displays, the experience suffers compared to their expectations. While several human factors issues are responsible for this disconnect, abundant anecdotal evidence and numerous controlled laboratory studies have shown that part of the performance gap is in the low perceptual quality of the graphical presentation. Despite extensive research in producing photorealistic graphics, little work in AR has been demonstrated to have that level of visual realism. Reviewing the literature and our own experiences and research, we identified four fundamental areas in which basic perception of the virtual and real elements in the merged world may be lacking. Visual acuity captures issues of geometric resolution, limited contrast and distorted perception of colors reveal issues of color resolution and presentation. These challenges lead naturally to issues of text legibility. In many applications, depth segmentation raises issues regarding the quality of stereo imagery. In this chapter, we examine these four issues as they apply to head-worn AR displays.

Despite numerous display options for augmented reality (AR), head-worn displays are still the popular choice for industrial, medical, and military applications. Among the most important advantages they offer is hands-free viewing of the environment and adding information into the environment from the user's point of view. Such head-worn displays come closest to instantiating Sutherland's original vision [35]. However, we have documented the reduction of fundamental capabilities of the human visual system when viewing the environment through head-worn AR displays. In this chapter, we will discuss the theoretical and ecological background of these issues, experimental designs to measure the practical implications for AR users, and report published and unpublished results from experiments conducted in our respective laboratories. Further, we argue for the importance of including this type of evaluation of AR systems both early and at regular intervals during the design and evolution of an AR system.

For many years, the Naval Research Laboratory and the Office of Naval Research sponsored research on various aspects of mobile AR systems. The central application in these programs was the Battlefield Augmented Reality System (BARS$^{TM}$) [23, 25]. Development of BARS was conducted through a usability engineering paradigm [9, 12], in which technical capabilities of the system and user performance on representative tasks with that system [26] are evaluated iteratively. (See Chapters 8 and 9 for other perspectives on usability engineering and user experience applied to mobile AR.) BARS was envisioned to provide situation awareness (SA) [5] for a dismounted warfighter during military operations in urban terrain. Among the many usability engineering findings for BARS was the fundamental need to be able to read text labels presented through AR. Labels could identify streets or landmarks in the environment or be part of military standard icons giving information designed to provide SA. This was one impetus for the work on the legibility of text in head-worn AR displays discussed in this chapter.

One goal with the BARS hardware was to compare whether the AR condition was as natural as a real analog of the task was. This gave an objective basis for evaluating

the quality of AR systems. We set up a series of comparisons, the simplest of which was inspired by early efforts to test the effective visual acuity of immersive head-worn displays [16]. This led us to create an AR Snellen eye chart [27]. We found that users (all with at least normal or corrected-to-normal vision) had their visual acuity decreased by wearing the AR display and looking at a real eye chart, and similarly decreased when looking at the graphical eye chart.

An extension to the BARS application was training in basic combat skills that might be required in an urban context [23]. In the course of evaluating this application, several observations were made about the quality of the head-worn video-overlay display [1]. In particular, subjects noted difficulty seeing the real walls in the environment, perceiving depth, and that the (real) targets were too small to see. This further motivated us to examine the perceptual capabilities users were able to achieve with head-worn AR displays and is another example of the value of the iterative approach to evaluation the performance of complex AR systems.

## 2 Measures of Visual Capabilities

In this section, we discuss measures of visual capabilities relevant to head-worn AR displays. The goal of this review is to provide a brief introduction to concepts from perception for AR researchers and users. We will conclude this section with a discussion of how AR affects these basic perception factors.

### *2.1 Visual Acuity and Contrast Sensitivity*

The quantity that comes to mind for most people when asked about visual capabilities is *visual acuity*, the ability of the observer to resolve fine details in the visual field. This is quantified by determining the smallest stimulus (measured in angular size) that the observer can identify at a rate better than chance. This quantity might lead to an impression that our ability to recognize objects is a function only of the size; in truth, recognition is a function of both size and contrast. *Contrast sensitivity* describes the observer's ability to discern differences in the luminance (or color) values across the visual field; it measures the threshold of contrast required to accurately perceive the target. Since contrast is a ratio of foreground to total luminance, its value is in [0,1]. Sensitivity is the reciprocal of this threshold. For each spatial frequency, contrast sensitivity is measured; these are then connected into a curve which separates perceptible stimuli from imperceptible ones. Both axes are typically drawn with logarithmic scaling. The canonical shape of such a *contrast sensitivity function* (CSF) is shown in Fig. 1. Sensitivity forms a concave-down parabola; objects whose size and contrast fit under the curve are perceptible.

Normal visual acuity is considered to be the capability of resolving one minute of visual arc [33]; this is most often measured (e.g. by an optometrist) by reading a

Snellen chart at a standard distance of 20 ft (6 m), with the letters scaled appropriately so that the various arms, bars, and counters that differentiate letters occupy the requisite number of minutes of visual arc for 20/20 vision (one minute), 20/30 vision (1.5 minutes), etc. One can immediately see the challenge of the Snellen chart: not all letters are "equally spaced" in this design space. For example, the bar (middle, horizontal stroke) in a lower case 'e' separates it from the lowercase 'c', which is in turn separated by the counter (opening) from the lowercase 'o'. An "equally spaced" Snellen chart must ensure that the bar of the 'e' and the counter of the 'c' are equal in width. Another common example is that the length of the lower arm of a capital 'E' must be exactly one minute of arc to separate it from a capital 'F', but this implies that the separation from a capital 'L' will be different. A similar design issue is that sans-serif fonts should be used.

In response to this type of difficulty and to standardize across countries with different alphabets, one may instead use a rolling 'E' chart. This chart shows a capital letter 'E' in one of four orientations – i.e. with the arms pointing up, right, down, or left. But this again is not quite equal in all directions, unless the letter is perfectly square. The Landolt 'C' chart allows a slight improvement on this; it orients a capital 'C' with the opening on the top, right, bottom, or left of the figure. Again, a chart designer should take care to use a rotationally symmetric figure, even if the letter is not normally printed in such a fashion. This will prevent the observer from gaining a clue based on the aspect ratio of the figure. Another option for similar charts is to use sine-wave gratings. A small 2D image with a wave in one dimension and a constant value in the other dimension creates a figure of which one may query the observer for its orientation. The orientation of the waves can span up to (but not equal to) 180°. A chart designer would likely select four cardinal orientations, such as waves that are horizontal, vertical, or along the two 45° diagonals.

When transferring these figures to a digital imaging device, another issue arises for the chart designer. The discrete nature of the display interferes with the production of the desired figures for recognition. For example, a diagonally-oriented sine wave is guaranteed to create aliasing on a digital display. While anti-aliasing techniques can mitigate the problem, a better strategy would avoid such a problem altogether. A similar difficulty exists for any letter or derived figure that uses curved strokes. Thus the Landolt 'C' chart suffers from this difficulty as well. However, this can be mitigated by relaxing the requirement of using a known figure (such as a let-

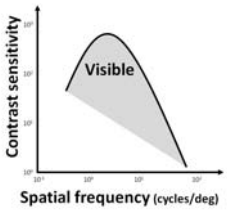


**Fig. 1** The canonical shape of the contrast sensitivity function, graphed as a function of spatial frequency for size and contrast threshold for sensitivity.

ter) in favor of simply a recognizable feature of an abstract figure. Thus a "squared" version of a 'C' that uses only horizontal and vertical strokes but leaves a small opening is a reasonable design choice [7]. A rolling 'E' could also be designed to fit these requirements, since the letter 'E' requires only vertical and horizontal strokes in any of the four cardinal orientations.

However, the design challenges do not end with the shape of the figure. The relative brightness between the foreground and the background interacts with the size of the features. So, as noted above, recognition is a function of both size and contrast. Contrast sensitivity has been accepted as part of a comprehensive approach to describing visual capabilities [14] and can be crucial in clinical evaluations for cataracts and diabetic retinopathy. The standard minimum for contrast in optometric examinations is 0.8, although it is unclear how rigidly this standard is followed in clinical practice or by what contrast definition it is to be measured. Contrast is frequently expressed by the Michelson definition:

$$C = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}},$$

where $L_{\max}$ and $L_{\min}$ are, respectively, the maximum and minimum luminances in the image. According to some, the visual system measures these extreme values from a local region around the foreground features [29]. The contrast in the image influences the observer's visual acuity score; at higher levels of contrast, the human eye is capable of detecting smaller details. Snellen charts, sine-wave gratings, rolling 'E' charts, and Landolt 'C' charts all may be adapted to provide a convenient way to measure an observer's CSF.

## *2.2 Color Perception*

The retinal responses to various wavelengths of light result in the spectrum of hues available to human *color perception*. The three types of cones (long, medium, and short wavelength – typically referred to as red, green, and blue, respectively) in the retina respond to different but overlapping ranges of wavelength of light, creating the effect that the visual system processes and interprets as color. The Commission Internationale de l'Éclairage (CIE) defined three standard primaries to describe color in 1931, leading to the CIE chromaticity diagram (Fig. 2). One problem with this diagram was that distance in the space did not have a uniform meaning to the perception of color. That is, a single distance could be a perceptually small difference in one region of the space while at the same time being a perceptually large difference in another region of the space. This led eventually to the CIE 1976 ($L^*a^*b^*$) color space (CIELAB). L denotes a luminance channel; a and b are chrominance channels (Figure 3). The a axis moves from green to red; the b axis moves from blue to yellow. This description of color closely matches the opponent process theory [15] of how the human visual system process wavelengths of light into color.

The model says that three relative measurements are acquired: one differentiating black and white (the luminance channel), one differentiating between red and green, and one differentiating between blue and yellow (the latter of which is itself derived from red and green). This space is (nearly) perceptually uniform, with distortions of perceptual difference thought to be approximately 1.6:1 – i.e. about a 60% change in distance might be interpreted as perceptually identical. While far from perfect, it represents a significant improvement over the estimated 1000:1 ratio of perceptually similar distances that are present in the 1931 CIE chromaticity diagram. One curious observation about CIELAB is that the colors that may be specific well exceed normal human visual capacity for differentiating colors, which in turn typically surpasses the monitor gamut for a display device (or printer). We note in passing other color specifications, such as CIE Luv (a similar model to CIELAB, where L is lightness and *u* and *v* are chromaticity coordinates; hue-saturation-value (HSV) is another standard color space, built on a single chroma coordinate (hue), a saturation value that gives distance from achromatic value, and a brightness coordinate.

Color names may be ascribed to the regions of a color space; however, here again, problems in the 1931 CIE chromaticity diagram become obvious. The size of the regions is far from uniform. Individuals will exhibit significant differences in where the boundaries should be drawn between subjects and perhaps even within subjects, depending on any number of factors, including lighting and other physical issues, and even mood or other subjective issues. In addition, many of the "standard" names seen in Figure 2 are far from being commonly-used terms to describe color. There are, however, color names that are consistently used. Of these, eight are chromatic and have one-word English names: red, green, blue, yellow, purple, orange, brown, and pink. (Achromatic terms black, gray, and white complete the list of basic color terms.) These colors were found to be maximally discriminable and unambiguously named, even across cultures [34]. Thus color naming can be a valid task that indicates color perception.

Color vision testing can be done in several ways. The most common method is the Ishihara pseudoisochromatic test [17], which uses a series of color plates. These 38 plates show mosaic images; the tiles of the mosaic have irregular but smooth shapes and appear in one of two colors, with varying size, brightness, and saturation to prevent determination of the figure without color discrimination. One color serves as the background, while the other color tiles draw a numeral to be recognized or a curved lines to be traced. These two hues are chosen such that people with normal color vision will differentiate them; however, people with color vision abnormalities will be unable to differentiate them. With a series of plates that test color combinations known to cause confusion for the various forms of color vision deficiencies, one can measure color vision capabilities. Versions of the test that use fourteen or 24 plates are also common. The similar Dvorine pseudoisochromatic color test is based on the same conceptual design. Most figures consisted of a numeral or numerals, with three sizes of dots that did not vary in intensity. Other figures again used the path tracing task of Ishihara. Both sets of color plates (and other, similar sets) provide a test that is easy to administer and easy to take. However, the light source ought to properly match the intended source, and the colors must be carefully

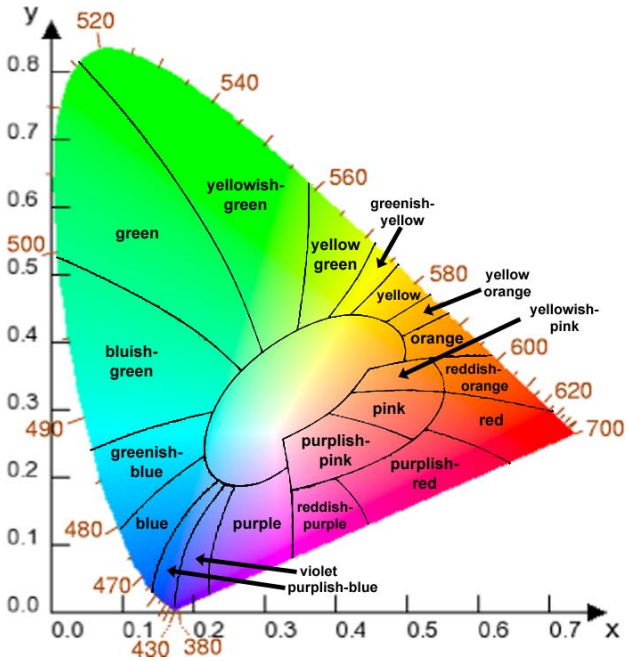


**Fig. 2** Names may be superimposed on the regions of a color space, such as the 1931 CIE chromaticity diagram seen here. However, some of these names are far from common usage, and individuals will exhibit differences in the names.

selected and may need to be varied for each individual. Finally, no accurate scoring criteria for classifying the type or extent of a defect are available.

Another strategy for color vision testing is to use a color arrangement test, of which the Farnsworth-Munsell test [6] is most relevant to our discussion below. A set of 100 colored tiles are placed before the subject, with an additional tile serving as a reference tile. (Later variations cut the number to fifteen or other values.) The idea of the color arrangement test is that the colors can be naturally ordered by someone with normal color vision. In the typical procedure, a blue shade is identified as a reference color, and the subject is instructed to place next to it the most similar of the remaining tiles; this step is iteratively repeated with the most recently placed tile becoming the reference tile. A subject with normal color vision will trace out a smooth, approximately circular path in either the CIE 1931 chromaticity diagram or in CIELAB. (A large difference in hue between the reference tile and the intended last tile ensures that subjects with normal color vision will correctly identify the first tile to place next to the reference.) A subject with normal color vision will move around the circle, whereas a user with deficient color vision will make some characteristic steps across the circle rather than follow the circumference. The advantage of this method is that any color deficiency can be detected, rather than the finite pre-planned anomalies detectable by color plates. This test can also be evaluated numerically by using the angles of the segments between the ordered tiles [37].

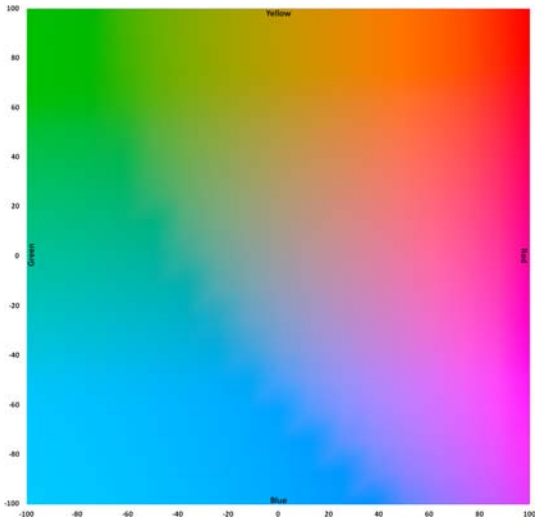


**Fig. 3** The CIELAB space aimed to both mimic the physiological measurements of the eye and form a perceptually uniform color space. This slice at $L = 65$ shows the $a$ axis running from green (left) to red (right) and the $b$ axis running from blue (bottom) to yellow (top). The origin of the $ab$ plane is the gray point in the center of this plot. The portion of the space shown here ranges from -100 to 100 in both $a$ and $b$ and corresponds to the portion drawn in later graphs.

However, the light source must still match the intended source, and the abstract ordering task requires more mature thinking, as well as manual dexterity. Thus this test in not appropriate for very young subjects.

## *2.3 Stereoacuity*

Humans (and other primates) have overlap between the visual fields of their two eyes, creating the ability to interpret 3D geometry from the 2D retinal measurements and the offset between the eyes. This yields two distinct depth cues, angle of convergence between the eyes to fixate on an object, and binocular disparity of the object in the two retinal images. Convergence angle has rather limited value, helping primarily at near field[1] distances [3]. However, binocular disparity is a much stronger cue at near field distances (enabling one or two orders of magnitude greater depth precision) and extends well beyond the near field, perhaps to several hundreds of meters from the observer. Despite this capability, the utility of stereo mechanisms in head-worn displays has not been studied extensively. This is in part due to a lack of hardware features; until recently, displays that permitted adjustment of interpupillary distance (IPD) were limited to a few research systems and the occasional

[1] The near field is most often defined as either that which is within arm's length or the slightly more generous but standardized length of two meters. Other definitions may extend the range.

commercial offering. By comparison, it has long been trivial in rendering software to create arbitrarily-separated images for left and right eyes in binocular displays. Incorrect IPD distorts distance perception; if the IPD is set at the population mean of 65 mm, then an observer with a larger IPD would think the object were farther away than it really is; an observer with smaller IPD would think it were closer. Similarly, errors in the judged depth of nearby virtual objects have been measured as a function of changes in binocular vergence.

The ability to discern a difference in depth from stereo vision is known as *stereoacuity*. Many aspects of stereo perception are thought to exhibit individual differences. The range of normal human IPD is wide relative to the absolute size; the smallest "normal" value is approximately 53mm, while the largest normal value is approximately 73mm. Additionally, the distance in front of a user's eyesat which a head-worn display may rest varies (much as different people wear eyeglasses at different points on the bridge of the nose); this distance may affect how the IPD should be set. Finally, some people (estimated as high as 20% [38] and as low as 2-5% [2]) are *stereo-blind*, and thus receive no depth information from binocular disparity (though they often compensate well enough to obscure this). So while it is a measurable quantity, a test to evaluate the effect of head-worn AR displays should account for these individual differences or screen out stereo-blind subjects. Stereoacuity has been measured for stereoscopic depth at as little as three arcseconds. This would at first seem to be in conflict with the one arcminute for normal visual acuity; it is clear that the human visual system has an extraordinary ability to accurately reconstruct the world through stereo vision.

## *2.4 Effects of AR Displays*

AR displays alter these fundamental measures of perception in meaningful ways. The user's capabilities interact with the hardware in ways that are not necessarily intuitive or obvious to AR system designers. We revisit these basic capabilities, discussing how AR displays interact with each. We will use the terms *optical see-through* and *video overlay* to describe the two major variants of head-worn AR displays. Other displays, such as head-mounted projective displays (HMPDs) and hand-held displays are mentioned only briefly. For each metric, it is important to understand how the AR user's perception of both the surrounding real environment and the virtual entities may be affected by the display. This can be very different for optical see-through displays versus video overlay displays. We also introduce our terminology of a display *device* implying the entire system that mediates and/or relays images to the observer's eyes. Devices incorporate display *elements*, usually liquid crystal displays (LCDs) or – in more recent display devices – organic light emitting diodes (OLEDs). Display *optics* include various lenses and mirrors (which may be partially-silvered and thus translucent). All these components and their arrangement can affect the fundamental measures of perception.

### 2.4.1 Visual Acuity

It is a relatively straightforward calculation to convert the resolution in pixels and field of view (FOV) of a head-worn display into an estimate of visual acuity. But it is important to note that simply measuring the angular resolution of a head-worn AR display is not sufficient to characterize the visual acuity or the broader experience a user will have when wearing the display. While it would theoretically place a limit on the performance a user could achieve, the human visual system is quite exceptional at filling in and interpolating information. Thus if one uses a Snellen eye chart, the visual system may be able to infer the identity of recognizable shapes (i.e. letters) without having complete information. Abstract figures and shapes, or rotationally symmetric figures and shapes, may overcome this confounding factor in acquiring measurements.

There is a fundamental difference between perceiving virtual objects and perceiving the real environment as it relates to visual acuity. Optical see-through displays theoretically should not affect the visual acuity with regard to the real environment. If acuity is measured as a purely geometric assessment, this is the case. However, the measurement using a Snellen chart is of legibility, which is a recognition task. If the chart also tests contrast, then the optics play a critical role. Given the reality of some optical see-through displays, this is an inherent characteristic of the test, as described below. The visual acuity of the user in perceiving virtual objects will be affected by the angular resolution of the display elements, even in an optical see-through display.

Video overlay AR displays (whether head-worn or hand-held) present both the real and virtual portions of the scene on finite-resolution display elements, which means that both are subject to reduced visual acuity. Further, the limit of the visual acuity may be lowered further by the camera that acquires the real environment. Some commercial video overlay AR displays have incorporated cameras that had lower resolution than the display elements, so this is a practical consideration. We note that HMPDs offer graphics at a resolution determined by the projector's angular resolution. Legibility may be further diminished by the shape of the retro-reflective surface; any deviation from flat will likely distort the shapes and inhibit recognizing a letter or understanding an abstract shape's orientation. Similarly, spatial AR systems offer a visual acuity that is a function of not only the display resolution, but also the distance that a user stands from the display surface.

### 2.4.2 Contrast Sensitivity

The same cases apply to contrast: real environment versus virtual objects, optical see-through versus video overlay. Contrast can be difficult to measure for optical see-through displays, given the uncontrolled nature of the real environment. One may wish to consider the contrast ratio of two objects in the real environment as seen through the optical elements, the contrast within the graphics as relayed by those optics from the display devices, or the contrast between a virtual object and

an object in the uncontrolled real environment. This last measurement can be quite challenging to acquire. The contrast a user perceives through an AR display depends on the display device and the optical elements that present the image to the user. Some displays significantly diminished the brightness of the real world so that the graphics did not need to be as bright. The Sony Glasstron used an electronic, global mask behind the semi-transparent mirror that reflected the graphical display into the eyes and through which light from the real environment passed. Thus the mask reduced brightness and contrast in the real world, in the hope of increasing the contrast between the real world and virtual objects. Mobile AR applications have an especially challenging task; sunlight is far brighter than any AR display (and than the user would want the AR display to be), so the mask or a similar filter is a critical element of a successful optical see-through display for outdoor use.

Video overlay AR displays have the same cases of real and virtual imagery to consider, but since all are presented on the same display elements and relayed through the same optics, the optical paths are not quite as different as in optical see-through. Contrast in the real environment will be limited by the dynamic range of the camera before the image is sent to the display elements, much in the way that the Glasstron mask or optical filters were meant to do for optical see-through displays. Although video overlay offers the possibility of matching the brightness of the real and virtual objects, this can be quite challenging to measure the real environment in real-time and reproduce the appearance. HMPDs and spatial AR often (but not always) require dark rooms, reducing the contrast in the real environment.

### 2.4.3 Color Perception

Just as a computer monitor and a printer have a gamut of colors that they can produce, so too do the display elements in AR displays (both optical see-through and video overlay) have a gamut of colors. Optical AR displays suffer from the partial transparency of graphics over an unknown real-world background; the combination can significantly change the color perceived by the user. A methodology to acquire measurements of this hue shift and measurements for an optical AR display are presented in Section 3.3. The perception of the color of real objects can be distorted by the electronic mask described above; in addition to reducing the apparent contrast, it may introduce a consistent tint to the colors and the context in which they are seen. By virtue of dimming the entire world, colors will appear quite different. Because of the contextual nature of color, the dimming and any hue shift can affect the virtual objects on the display surface as well as the real objects behind the display surface. Clear optics in theory can avoid these problems.

Color is highly contextual, and thus knowledge of the background and surrounding (2D) visual field, as is available in video overlay AR displays, can be extremely helpful in selecting colors for discriminability on the display. Video overlay AR displays are heavily dependent on the camera and display element for the richness of the color presented to the user. Between the cameras' limited range and the display elements' gamut, there can be a complex chain of modulation between a real object

and the eye of the user. The virtual objects would have a shorter, simpler path to the eye, but they will still be affected by the context. Of course, if the video image passes through the graphics memory, then it can be inspected for values and the effect of the context can be mitigated by "pre-distorting" the colors of virtual objects according to an inverse function of the distortion. Not all commercial video overlay AR displays pass the camera image through the graphics processor, however.

#### 2.4.4 Stereoacuity

Numerous challenges for proper perception of stereo imagery have been noted for head-worn AR displays, of both optical see-through and video overlay varieties. Few displays offer even some of the adjustments that are optimal for comfortable perception of stereo imagery: adjustment of the IPD, adapting the vergence angle, and alignment of the left-eye and right-eye displays. The rendering software should also adjust the IPD to match the display (which preferably will match the user). It can correct the alignment of the displays for the two eyes (as discussed in the next section). Software correction of the vergence angle is possible, although it potentially introduces perspective distortion into the imagery if the hardware does not match the angle. This latter capability is rare in head-worn display hardware.

Of course, hand-held displays generally ignore these issues, but binocular disparity is a powerful cue for depth at large distances (perhaps to several hundred meters), so the 2011 emergence of auto-stereo displays for mobile phones should encourage developers to consider these effects for hand-held displays as well.

## 3 Measurements of Visual Capabilities in AR

There have been a modest number of experiments to measure visual capabilities with head-worn AR displays. The general observations made at the end of Section 1 motivated or were discovered by the studies described in this section. We summarize these experiments and software (in approximate chronological order within the four experimental goals) and discuss some practical difficulties in collecting these measurements with AR head-worn displays.

### *3.1 Visual Acuity*

A test of four optical see-through AR displays [40] investigated the smallest real targets visible from one meter with the display off and with the display showing a blank screen. The latter condition implied that the display emitted some light and, in the case the Sony Glasstron PLM-50, enabled the mask that reduced transmittance of the light entering from the environment. Two binocular displays showed differences

in these two conditions. The Glasstron (33° measured horizontal FOV, NTSC resolution) allowed 1 mm targets (3.4 arcmin, or 20/69 Snellen score) with no power (mask off) but only 6 mm targets (20.6 arcmin, 20/412 Snellen) with power (and mask) on. Virtual I-O I-glasses (25°, NTSC) enabled users to see 0.5 mm targets (1.7 arcmin, 20/34 Snellen) without power and 3 mm targets (10.3 arcmin, 20/206 Snellen) with power. A MicroOptical Corp. Clip-On CO-1 (10°, QVGA) and MicroOptical Integrated EyeGlass (17°, VGA) both allowed users to see 0.5 mm targets (1.7 arcmin, 20/34 Snellen), although a poorly positioned CO-1 was found to limit users to 12 mm (41.25 arcmin, 20/825 Snellen) targets.

The Augmented Reality Performance Assessment Battery (ARPAB) [18] included visual acuity tests for AR displays. It was used as a pre- and post-test of a distance estimation task in optical see-through AR with 20 subjects. With natural or corrected vision, subjects were found to range from 20/13 to 20/30 in visual acuity. The AR pre-test with a Sony Glasstron[2] (SVGA, 27° horizontal FOV) yielded 20/40 Snellen scores for 18 subjects; one subject scored 20/30, and one scored 20/50. After the distance estimation task, all subjects scored 20/40 on visual acuity with the Glasstron. The pre-test with a Microvision Nomad[3] (SVGA, $\approx 21°$ horizontal FOV) yielded mostly 20/30 and 20/40 scores (precise distribution not given), with one subject scoring 20/50. Only three participants scored differently on the post-test: one from 20/50 to 20/40, one from 20/40 to 20/30, and one from 20/30 to 20/50.

A Sony Glasstron LDI-D100B caused eight users with normal or corrected-to-normal vision (i.e. 20/20 or better) to drop at least one step measured with a Snellen chart ($\approx 20/30$) looking through the optics of the display at the same real-world optometric chart [28]. All users scored 20/30 looking at a graphical chart. This test was extended to a 2D contrast sensitivity measure using a sine-wave chart [20]. The 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. Notably, looking through the Glasstron at the real target was significantly worse than looking at graphical targets in the Glasstron. This Glasstron model used a similar LCD mask as described for the PLM-50. The maximum transparency was specified as 80%, and the loss of brightness of the real world was seen in the increased contrast needed to see real targets. The Nomad 1000 also reduced contrast sensitivity both looking through the display at real targets and looking at graphical targets, but by a far smaller amount.

Video overlay AR systems mediate the real world through a camera, which limits the user to its spatial (and color) resolution. In testing a training application [4], subjects using video overlay with a camera[4] mounted to a Virtual Research V8 head-worn display (48° horizontal by 36° vertical, VGA) were found to have degraded visual acuity, but no quantitative data were reported.

---

[2] Model not reported, but given the year and reported specifications, most likely an LDI-D100B or similar

[3] Model not reported, but given the year and reported resolution, most likely a Nomad 1000

[4] The authors report using "an Auto Gain Control (AGC) and Electronic Light Control (ELC) Panasonic camera," with an FOV "compatible with the field-of-view of the HMD," but do not give precise specifications or models.

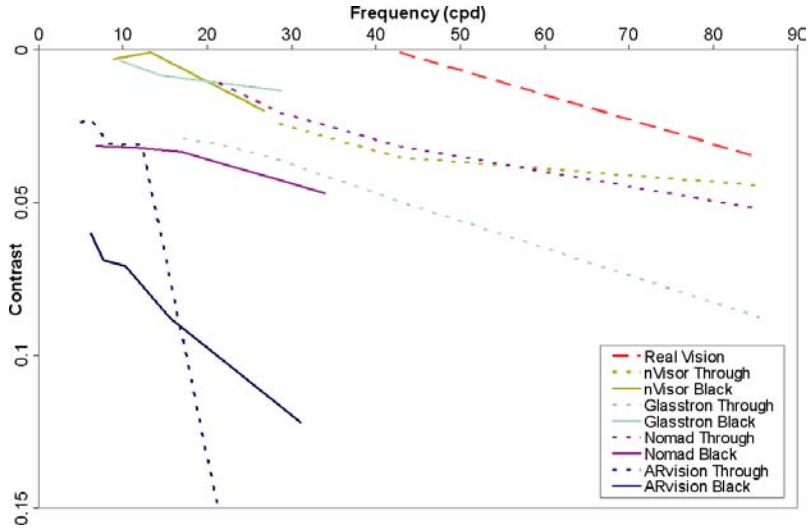


**Fig. 4** The sampled CSF for four head-worn AR displays. The size of the target decreases as frequency (measured by cycles per degree) increases to the right. The contrast is typically graphed with lower contrast at the top of the vertical axis. In this test, the nVisorST performed well in both the see-through and graphical cases, the Glasstron provided sharp graphics, the Nomad provided clear see-through, while the ARvision struggled in both cases.

Evaluation of a custom-built HMPD (52°, VGA) was done with a modified Landolt-C acuity test [7]. Users identified the location (up, down, left, or right) of an opening in a square (as described above) under three levels of light. The study found that the resolution of the display limited subjects to a visual acuity of 4.1 arcminutes, or a Snellen score of 20/82 for all lighting levels. The type of retro-reflective material (needed for the HMPD) affected performance with low contrast targets.

Bringing together the design for measuring 2D contrast sensitivity and the modified Landolt-C task provided the ability to measure a portion of the contrast sensitivity function for head-worn displays. Figure 4 shows the sampled CSF for four head-worn displays. One observation from the graph is that the varying resolutions and measured contrasts of the displays make direct comparisons somewhat difficult. The line for each condition extends from the size of one pixel at the left; this is the smallest target for which testing was possible. The right extent denotes the size of a five-pixel target. The contrasts tested were all well below the standard for an optometric exam, and the heights of the various lines – which denote the contrast necessary to see a given size of target – are best judged relative to each other. The exception to this, however, is the natural vision condition; the monitor on which this test was performed could not exceed the capacity of the human visual system.

The optical see-through nVis nVisorST performed the best overall, with both the graphics and see-through cases yielding the closest performance to users' natural vision. The Glasstron LDI-D100B enabled high performance with the graphics. Again, the difficulty seeing the real world was due to the mask, so that greater con-

trast was needed to see real targets. The Nomad 1000 saw the reverse conditions; its clear optics enabled users to see roughly as small a target as any head-worn display condition. However, the monochrome red graphics did not enable the users to see virtual targets as well as the real world. Finally, the video overlay Trivisio ARvision display struggled in both the the graphics condition, where the display quality is the limiting factor, and in the see-through condition, where the camera's resolution and contrast also influence the results.

## *3.2 Text Legibility*

Leykin and Tuceryan [19] trained automatic classifiers to assess the readability of text labels over textured backgrounds. They collected training data from six human subjects, who rated the readability of 100 text-over-texture images on an eight-point Likert scale. They then tested several standard machine learning algorithms, finding that a support vector machine was the best, but that overall performance was limited by the large number of outliers (20%) in the training data. Because the system analyzed the texture images and the features of the text itself, this method was considered appropriate only for AR systems that have video feedback (which usually is only the case for video overlay AR systems, but may be implemented more widely, as evidenced by the next experiment). Also, they limited their training and test data to grayscale presentation of text and texture.

Gabbard et al. [10] conducted a controlled study with an outdoor AR system. They studied six background textures commonly found in outdoor environments and six drawing styles, measuring response time and error for the task of identifying and reading a single numeral presented in a text sequence consisting mostly of (English uppercase) letters. Eighteen users wore a Glasstron PLM A55 bi-ocular optical see-through display (NTSC) to read text at three distances: beyond, on, and in front of a real background object (which displayed the texture). The display's focal distance was fixed (by the hardware) at two meters; the backgrounds were (respectively) at one, two, and four meters for the three distances. They found that subjects were fastest with a red brick background, perhaps because the text strings could be "framed" within a single brick. A "billboard" drawing style using blue text with a white background in the graphics rendering (which is the most opaque that the optical see-through display can be) yielded the least error among drawing styles. It was followed by a static green text color with no background in the graphics rendering. Dynamic drawing styles did not fare well, and distance of the text relative to the background real object did not show main effects.

In a follow-up study [11] using a Glasstron LDI-100B, the independent variables for the drawing of text were divided into text color (white, red, green, and cyan), drawing style (none, billboard, drop shadow, and outline), and active drawing style algorithm (maximum HSV complement, maximum brightness contrast). On each trial, users identified a letter that was repeated consecutively in one text block, and then counted occurrences of that letter in a second text block. Participants were most

accurate with the building texture as the background, and least accurate using the red brick texture. A similar trend was found for response time. Text color showed no main effect on either error or response time, which was attributed to limited luminance capabilities of the Glasstron and the positive effects of the active text drawing styles. The billboard drawing style yielded lower accuracy and slower responses. Among the active styles they found a small but significant main effect; participants were most accurate when reading text drawn with maximum brightness contrast.

Renkewitz et al. [32] devised an experiment to find the optimal font size for displaying text in a head-worn display for outdoor AR applications. They used a video overlay AR system presented on an unknown HMD with SVGA resolution. They found that for a concrete wall texture, blue fonts needed a size of 8.3 points and red fonts needed a size of 8.1 points. For a bush texture, blue fonts needed 8.6 points and red fonts 9.0 points. Response times rapidly decreased until the font size was increased to 15 points, with slight gains at greater sizes. Thus they concluded that while 9 points was readable, it was not sufficient for fast reading. Selecting two seconds as the maximum acceptable reading time, they recommended fonts sizes of (concrete, blue) 23 points, (concrete, red) 42 points, (bush, blue) 34 points, and (bush, red) 29 points, equivalent to 0.92°, 1.68°, 1.36°, and 1.18° (respectively).

Labels that overlap in an AR view become difficult to read; however, stereoscopic segmentation can perceptually separate multiple, overlapping labels [31]. Seventeen subjects read an airplane call sign in AR and, on a subsequent AR display, selected the object with that label. As the amount of overlap increased, response times increased; similarly, as the amount of overlap decreased, the error rate decreased. The viewing conditions varied disparity: ordered with depth, random with respect to depth, and constant with respect to depth. There was a significant interaction between the viewing condition and the amount of overlap. The ordered condition led to faster responses when the amount of overlap was high, implying that the targets in close proximity to other objects benefited from the ordering of disparity with depth. Dynamic scenes (including moving labels) yielded lower error rates than static scenes. An earlier experiment [30] found that correct vertical separation based on depth yielded lower error than no vertical separation and inverted vertical separation. Correct vertical separation meant that the closest target, which was lowest in the visual field, had the lowest label in the visual field. Inverted separation meant that the closest object was lowest in the visual field, but its label was highest among the labels. These two conditions did not show a significant difference in response time, and both were faster than no vertical separation.

## *3.3 Color Perception*

In designing ARQuake [36], color selection was recognized as an important consideration. The dark colors of the original game were not conducive to display on a see-through AR display. Therefore, nine colors were tested (red, green, blue, cyan, magenta, yellow, purple, pink, and orange), each at four intensities and in four light-

ing conditions (standing in shade or sunlight, crossed with looking into shade or sunlight). Users assigned a Likert rating (1-10) for visibility and opaqueness. Nine color/intensity combinations were found to have a minimum score of six and mean score of at least seven in each lighting condition: three intensities of purple, two of blue, two of yellow, and two of green.

As noted above, video AR systems limit the user to the color resolution of the camera, modulated by the display's color gamut. A training system was used for testing color perception [4]. 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 accuracy in color perception.

One can also quantify the perception of color through head-worn AR displays using a color naming task. The reduction of contrast in the Glasstron noted for the visual acuity and contrast sensitivity test 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 [20]. 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.

Switching to a color matching task [22] gave objectivity and much greater precision in the data about color perception. Users were asked to control the chroma values ($a$ and $b$ of CIELAB) with a two-dimensional trackball. An initially gray target patch would change hue, and users could move through the color space in order to match the color of a reference patch. Colored bars around the target helped remind users which way to move through the space. Another helpful change was conceiving of the task in a perceptually uniform color space, such as CIELAB. With this experimental design, color error could be expressed as $\Delta E$, a distance in color space which has been well-studied in perceptual literature for perceptually uniform color spaces such as CIELAB and CIE Luv. Setting up the matching task so that users could not receive unintended assistance by looking around the display required some careful arrangement of physical barriers. However, the result was a rich set of data.

There are two sets of graphs; the first (Figure 5) shows the objective color distortion measured by a StellarNet EPP2000CXR spectrophotometer with CR2 cosine receptor. This data maps the color gamut of the display device under conditions of both see-through and graphics conditions. The graphics condition was further tested with cases of a black background and a white background; with the optical see-through, the background can heavily influence the perceived color of light that enters into the user's eye through each pixel of the graphical display. In this data, the nVisorST was shown to have only modest distortion away from the yellow-green and cyan corners of CIELAB space in the see-through condition; this was credited to the clear optics of the display, which caused little distortion of color. In the case of graphics-on-black background, the ambient light in the room pushed the objectively-

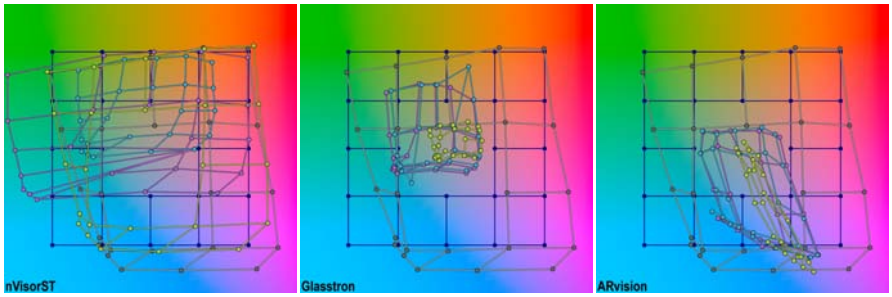


**Fig. 5** Objective measurements of color distortion in the nVisorST (*left*), Glasstron (*center*), and ARvision (*right*) as determined by spectrophotometer measurements. Three visual conditions were of interest: see-through (yellow graphs) – a misnomer for the video overlay ARvision, but still a convenient label, graphics-on-black background (magenta graphs), and graphics-on-white background (cyan graphs). The see-through data should be compared to the measured color gamut of the reference monitor that was the real-world background; this data is mapped by the gray graph. The blue reference grid gives the definitions in CIELAB space of the colors sent to the displays.

measured color out of the blue half of the color space (perhaps not surprising, since blue accounts for little perceived intensity) and pulled it towards green (the strongest component of what we perceive as intensity). The white background increased the push away from blue region of CIELAB and pushed away from the red and green slightly as well. The Glasstron, again owing in part to the reduction of contrast, pulled all colors towards the center (gray) point (reducing apparent saturation of color), but also slightly towards the yellow half of the color space. The amount of distortion was approximately the same for the graphics-on-white and graphics-on-black background conditions; the pull towards the center intensified significantly in the see-through condition. Finally, the ARvision pulled the entire graph towards the blue half of the space and also reduced the saturation of colors, although less in the magenta corner of CIELAB than regions. Again, the distortion of color was similar in the case of the graphics-on-white background condition as in the graphics-on-black background condition (subject only to the display's capabilities). Analogous to the Glasstron, the video overlay (subject to the camera and display capabilities) intensified the distortion, but this in the red-green dimension of CIELAB; it slightly increased the shift towards blue but not the distortion in the blue-yellow dimension.

The second set of graphs (Figure 6) shows the perceived colors. These are only compared to the monitor gamut of the reference display, the gray grid also shown in Figure 5. Before examining the data, we note that user responses spanned nearly the entire portion of CIELAB graphed ($a,b \in [-100,100]$). Thus the relative patterns of distortion would seem not to be an effect of attempting to match CIELAB color specifications that are outside of normal human color perception. With that observation, we can turn our attention to the data. The nVisorST caused users to perceive less strength in the cyan corner in the see-through condition. There was notable variance in the answers in this condition as well, denoted by the scale of the circular data indicators. The graphics-on-white background condition showed a similar distortion away from the cyan region of CIELAB and a similar variance. The

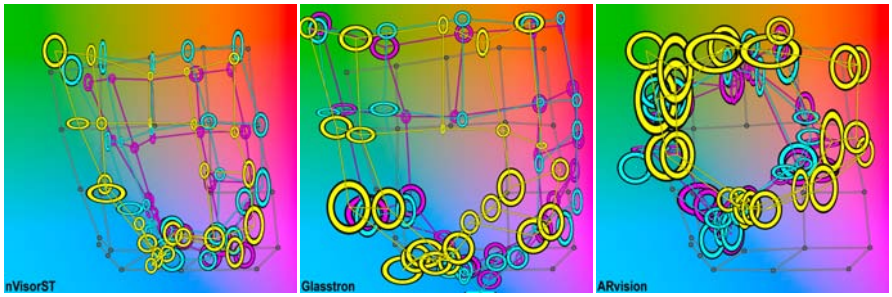


**Fig. 6** The perceptual matching showed the color distortion in the nVisorST (*left*), Glasstron (*center*), and ARvision (*right*). The same three visual conditions are graphed with the same colors as in Figure 5. All data should be compared to the measured color gamut of the reference monitor, again mapped by the gray graph. In this figure and Figure 5, the CIELAB domain is represented by the L=65 slice with a and b in [-100,100]. This portion of the domain is also depicted in Figures 7 and 8 and thus may be used to help compare the data in all four graphs.

graphics-on-black condition caused further distortion away from cyan, but a little less variation in user responses. For the Glasstron, users appeared to generally overcompensate for the reduction in saturation, perceiving colors that were much further from the center of the color space than they should have. This was accompanied by an increase in the variation between users, especially as the perceived colors moved toward the outside of CIELAB. The patterns of distortion and variation were similar in the three visual conditions of see-through, graphics-on-black-background, and graphics-on-white background. Finally, for the ARvision, the stark distortion is away from the magenta corner of CIELAB; further, colors near the center of the space appeared to be perceived as more saturated than they were. The video overlay appeared to suffer from this effect more than the other two visual conditions. There was a notable increase in individual variation for nearly all colors in all three display conditions compared to the other two head-worn displays.

Gabbard et al. [13] applied the textured background from the text legibility experiments (described above) to create a testbed for measuring the effect of blending natural light reflected off real-world backgrounds with virtual light produced by an optical see-through display. They found large perceptual shifts (Figure 7) between a "no-background" condition and brick, foliage, pavement, sidewalk, and white background conditions. The white background versus the remaining textured conditions showed the next largest set of changes. In terms of direction, the no-background condition pulled the colors towards the white point of the color space compared to the white background, which allowed the perceived colors to be distributed more over the color space. The foliage texture background pushed the colors away from the white point compared to the white background. The brick texture background pushed the colors away from the green region of the color space compared to the foliage texture background, and the sidewalk texture background pulled the colors toward to white point compared to the brick texture background.

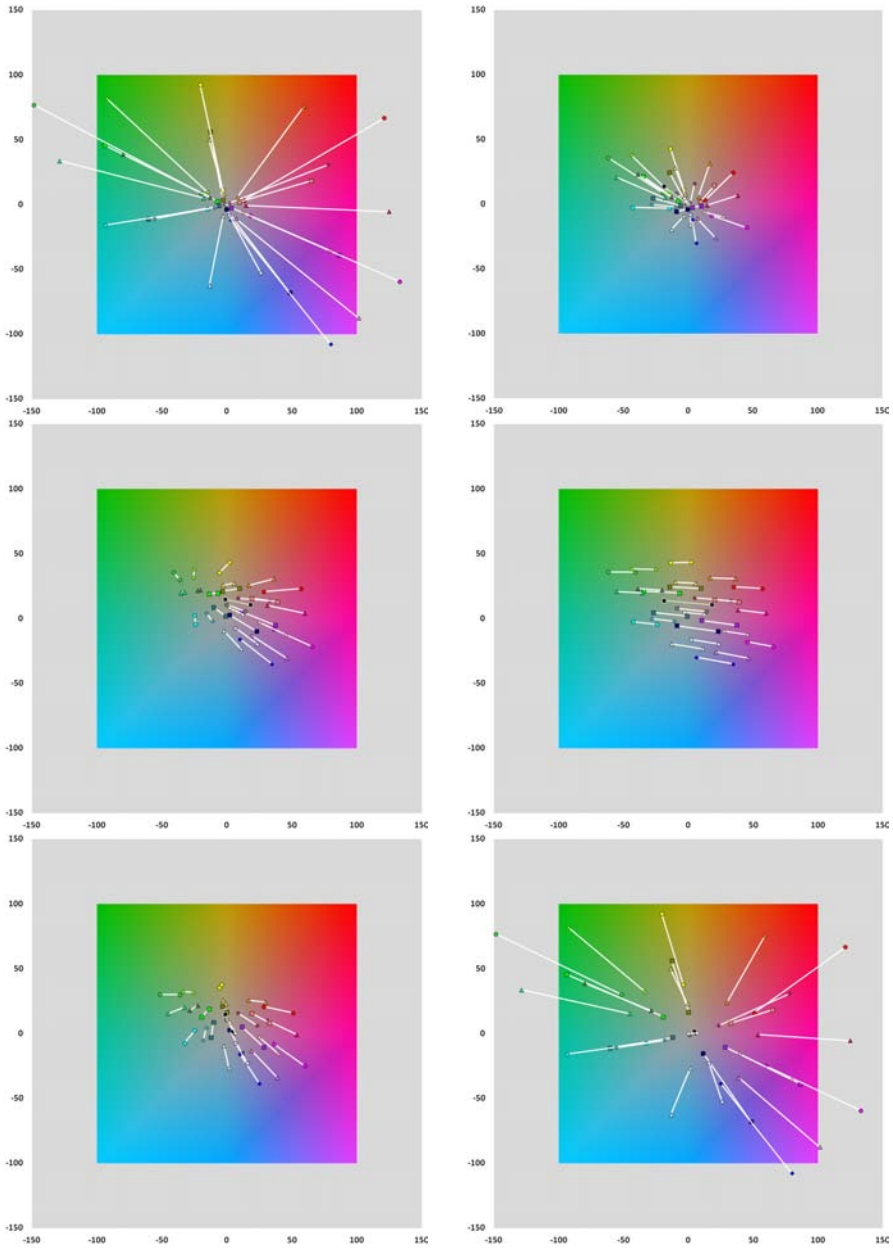

**Fig. 7** Objective chromatic changes between textured backgrounds measured in [13], converted to CIELAB; colored portion of background corresponds to the background of Figs. 5, 6, 8, and 9. *Top left*: no-background (outer) versus white background (inner). *Top right*: foliage background (outer) versus white background (inner). *Center left*: brick background (outer) versus sidewalk background (inner). *Center right*: brick background (right) versus foliage background (left). *Bottom left*: pavement background (outer) versus sidewalk background (inner). *Bottom right*: pavement background (inner) versus no background (outer).

## *3.4 New Results on Color Perception*

We present two previously unpublished studies of color perception in AR displays.

### 3.4.1 Farnsworth Color Arrangement

One simple strategy to test for distortion in color perception with AR displays is to conduct a standard color vision test with the AR display. This was done for a Sony Glasstron LDI-D100B, a Trivisio ARvision, and an nVis nVisorST. Before giving the results, it will help to define two types of color vision deficiency. The most common form of color blindness is poor red-green discrimination. The most common type of this deficiency is caused by a shift in the medium wavelength (colloquially, green) retinal receptors towards the long (red) wavelengths; subjects with this condition are called *deuteranomal*. It affects perhaps 5% of the male population and is hereditary. A more rare form (0.01% of the population) is caused by a defect in short wavelength (blue) receptors and affects the ability to differentiate blue and yellow hues; subjects with this condition are called *tritanomal*. This is also a hereditary deficiency, but shows no difference in gender. There are color vision deficiencies marked by the absence of one of the three types of cone receptors and deficiencies marked by a defect in one of the three types, as well as combinations of these, so numerous other types of color vision deficiencies exist, but these two will be sufficient to describe the results of this study.

Twenty-four subjects (eight for each display) completed the Farnsworth D-15 color arrangement test in four modalities. As a baseline case, each subject completed the test using a standard computer monitor. The subject completed a "see-through" version of the task on the same monitor, but with his or her vision mediated by the AR display. Note that this has very different meanings for the optical see-through Glasstron and nVisorST than it does for the video overlay ARvision. Two graphical conditions rounded out the set: seeing the Farnsworth test on the AR display with a white background, and seeing it with a black background. The order of these four display conditions was counterbalanced by a Latin square (repeated twice for each display's eight subjects).

All twenty-four users passed the test in the baseline case, so we believe that all deviations from normal with the displays were due to the displays and their properties. With the nVisorST, all eight users tested as having normal color vision both looking through the display at the monitor (the see-through condition) and with the virtual graphics displayed over both black and white backgrounds. In all, there was no significant distortion of color perception with the nVisorST. Similarly, almost all users passed the test in all conditions with the Glasstron. One user made a minor error in the see-through condition, which for that user was the last condition mandated by the Latin square. This error resulted in a Confusion Index (C) of 1.39 and a Scatter Index(S) of 1.10 using the scoring method of Vingrys and King-Smith [37]. A value of 1.00 (in both indices) is considered normal, but it would require values near or above 2.00 in both to be considered to have a color vision deficiency. So

although the user would still have been judged to pass the test, there is a hint of difficulty in the Glasstron see-through condition, although given that this was the final condition tested, fatigue may also have been a factor.

Turning to the results with the video overlay ARvision, we recall that the colors perceived are a function of the camera parameters *and* the display element's color gamut. The combined effect yielded a wide array of color vision results. We emphasize that all users tested as having normal color vision using a standard computer monitor, although minor errors occurred, as noted below. In the condition equivalent to see-through for the video overlay display – i.e. where the users saw the computer monitor through the camera and head-worn display – five users tested as normal but with Confusion Indices and Scatter Indices ranging from 1.10 to 1.69, the latter of which approaches the Confusion Index for tritanomal subjects. Another subject was scored as C=2.24 and S=2.16, which is consistent with being tritanomal, although the angle of crossing the circle was consistent with normal color vision. One subject did match the tritanomal profile in all three measures, and one matched the deuteranomal profile in all three scores. The results for the ARvision with the graphical test seen over white and black backgrounds were more encouraging. Five users tested as normal with the white background, while two tested as normal, but with C and S in the range of [1.11,1.65]. One user matched the tritanomal profile. With the black background, again five users tested as having normal color vision, two users tested as normal but with C and S in [1.10,1.39], and one user matched the tritanomal profile (which was the same user as for the white background).

Thus we can see that some users were transformed by some AR display conditions into partially color-blind subjects; given that these deficiencies are known to be hereditary and rare, this speaks to the limitations of the AR display (including display element for all displays, as well as the optics of the see-through displays and the cameras of the video overlay display).

### 3.4.2 CIELAB Measurements with Chromatic Backgrounds

We extended the experiment described above using the CIELAB space to include conditions with chromatic backgrounds; data with these backgrounds was acquired only for the nVisorST display. We extended our experimental design to four new display conditions: solid color backgrounds of (respectively) red, green, blue, and yellow. For comparison purposes, subjects again completed a natural vision condition in which there was no AR display used; they merely matched what they saw on one side of the barrier to the other side of the barrier. This baseline allows us to quantify differences and natural variation in performance of the perceptual matching task. Nineteen subjects completed the experiment with 22 color samples in each of the five display conditions.

We found main effects on both dependent measures, distance ($\Delta E$) in color space and the direction of this error within the *ab* plane of CIELAB. Looking at the distance measure, we found a main effect of the display condition – $F(4,72)=80.444$, $p=0.000$. Not surprisingly, subjects were most accurate to the input colors in the nat-

| | $\Delta E$ | | Angle | | Background |
|---|---|---|---|---|---|
| Display Condition | Mean | SD | Mean | SD | Intensity |
| Natural vision | 27.8397 | 21.8763 | 0.1855 | 1.7092 | 20.5 |
| Red background | 41.2221 | 27.9827 | -0.2676 | 1.6762 | 22.9 |
| Green background | 53.7014 | 28.6392 | -0.1540 | 1.0546 | 53.7 |
| Blue background | 44.1641 | 24.9506 | 0.8739 | 1.1027 | 13.4 |
| Yellow background | 56.6131 | 29.6001 | -0.2579 | 1.4069 | 78.9 |

**Table 1** The main effect of display condition on the distance ($\Delta E$ metric in CIELAB) and angle of the error (in radians). The absolute values for $\Delta E$ are not as meaningful as the relative size of the chromatic background conditions to the natural vision condition. Regarding the angular errors, it is interesting to note that only the green, blue, and yellow backgrounds caused significant differences from the natural vision condition. The intensity of the background may be a cause for the effects observed.

ural vision condition. Since some of this error is inevitably due to an imperfect monitor gamut for the reference monitor, the more interesting finding is not the absolute error in this baseline condition, but that the four chromatic backgrounds all exhibited significantly higher error and higher variance in the responses (Table 1). Furthermore, the red and blue backgrounds exhibited significantly lower error than green and yellow. One could easily hypothesize that this may be an effect of the background intensity, although this would seem to be an incomplete explanation for the difference between the natural vision condition and the four chromatic backgrounds. There was also a significant main effect on the angle of this error – $F(4,72)=46.127$, $p=0.000$. The natural vision condition had a low mean and a high variance, which means the distribution of the direction of error was likely more similar to noise. The red, green, and yellow had a similar mean direction, although the green was more consistent (lower variance) than the other two, and the blue background induced an entirely different direction and was also somewhat more consistent about this effect than the red or yellow backgrounds. The variances are so high that it is unclear whether the mean angles for the various conditions are meaningful measures.

Figure 8 shows a plot of the mean and standard deviation for each color sample of this data in the same spirit as Figure 6. The center of the circle is at the mean location for each color sample's response, and the scale of the circle in the $a$ and $b$ dimensions is a linear function of the standard deviation of the subject responses. While these graphs showed that we tried to sample colors away from the achromatic central region, there is often distortion further away from the center than the natural vision condition. This illustrates both the increased $\Delta E$ and the patterns of angular error. In this graph, we can see that the mean angle for the blue background diverges well away from the other conditions towards the yellow region. One might expect to see such an effect, pushing users towards the opposing color in human perception. This effect is not seen consistently with the other chromatic backgrounds, although it does apparently appear for colors near the opposing color with the green background (right side of the plot) and yellow background (bottom side of the plot). In order to produce hues that would be recognized by subjects as having "typical" appearance for the nominal background colors, we did not equalize the intensity of

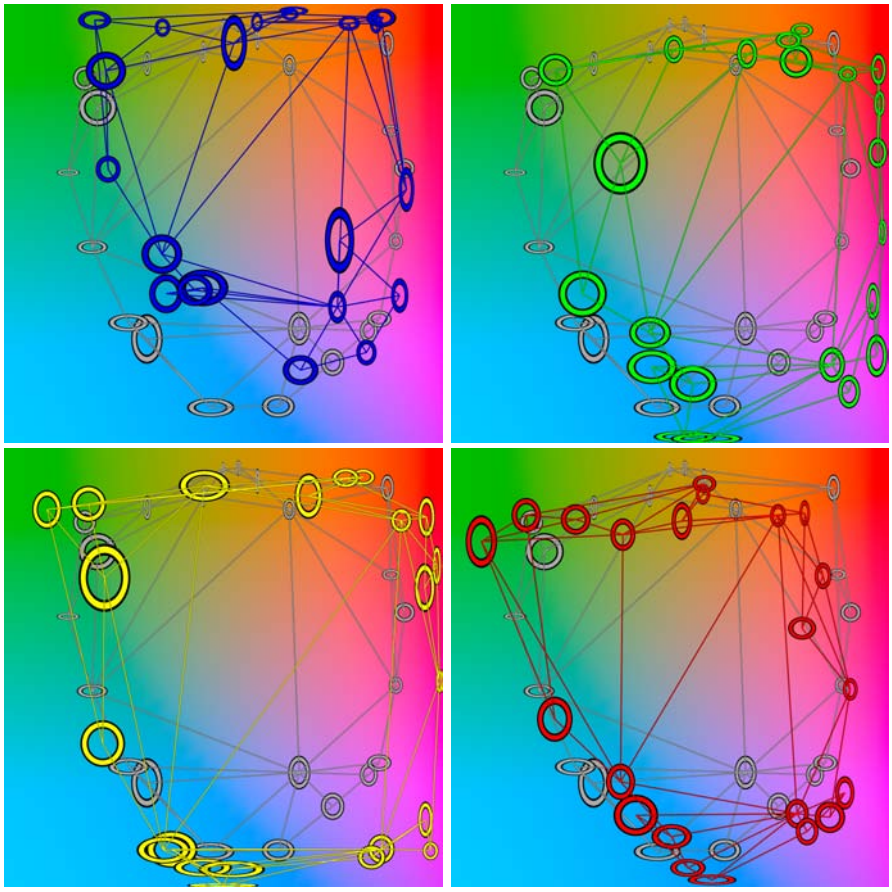

**Fig. 8** The perceptual matching showed the color distortion in the nVisorST for four canonical chromatic backgrounds of solid color: Blue (upper left), Green (upper right), Yellow (lower left), and Red (lower right). Each graph is shown on CIELAB space and includes a gray reference grid of the matching as completed with no mediation by any AR display.

the background colors. Also, we note that the intensity of black (listed as the background for the natural vision condition in Table 1) was greater than the intensity of the blue background seen through the nVisorST display optics. While it is possible the optics did cause this difference as measured by the color meter, it is curious to observe such numbers. Thus we could hypothesize that the low intensity of the blue background is the cause of the unique direction for the mean error in this condition. One could also formulate theories about the saturation of the background as the cause of any of the differences observed.

We noted a significant main effect of sample color on $\Delta E$ – $F(21,378)=23.879$, $p=0.000$ – and on angle $F(21,378)=14.362$, $p=0.000$. While the graphs indicate that various samples were moved by different distances and different directions, it is hard

to make general statements about the patterns beyond those made above. Further, there were significant interactions between the display condition and the color samples for both $\Delta E$ – F(84,1512)=4.301, p=0.000 – and on angle F(84,1512)=3.492, p=0.000. These serve to solidify our assertion that the pattern of errors is quite complex and deserves further data collection and analysis. We also measured response time; we found a significant main effect of color sample – F(21,378)=4.276, p=0.000 – but not of the display condition – F(4,72)=0.386, p=0.818. While we may have expected such an effect based purely on the starting condition for each trial of a gray field to match to a chromatic sample, there was no apparent pattern of the response time as a function of the distance of the color sample from the origin of the *ab* plane. We saw a significant interaction between display condition and color sample – F(84,1512)=1.556, p=0.001 – but defer interpretation until such time as the main effect can be explained.

This type of data may be used to adapt the displayed colors so that they will be perceived as intended and appear as if matched to the lighting of the real environment [39]. Two important considerations are adaptation of the user over time and the system latency that may be introduced by pre-distorting the colors. If the latter must be done on a per-pixel basis to account for the background of each pixel, then the compensation algorithm may be an expensive rendering process. An implementation that takes full advantage of the programmable nature of the modern graphics processing unit (GPU) may alleviate this.

It is natural to compare the various backgrounds (black, white, see-through, red, green, blue, and yellow) for patterns to see what appears to behave similarly and what appears to behave differently. Figure 9 compares the data from all background conditions, displayed over CIELAB color space (Figure 3). Recall that green shades are on the left, while red shades are on the right; blue shades are at the bottom, while yellow shades are at the top. We see that the green and blue backgrounds generally caused users to shift away from that color in the matching task (though exceptions exist). The yellow and red backgrounds seemed to cause users to shift away from the achromatic center of the slice of the color space. The colored backgrounds generally caused larger errors than the achromatic and see-through conditions, which saw little consistent pattern of errors. One may perhaps see in the bottom (blue) half of the reference sample space that the shift was toward the yellow half, and vice-versa. However, this data is rather sparse, and strong, consistent patterns would appear to require more data.

## *3.5 Stereoacuity*

In order to perceive stereo images, the human visual system must fuse the input from the left and right eyes into a coherent picture of the three-dimensional world. Among the many issues created by stereo optical see-through displays is vertical disparity in the graphics relative to the real world. Such disparity will cause *diplopia* (double vision) for the user of a head-worn AR display, and even if the visual system

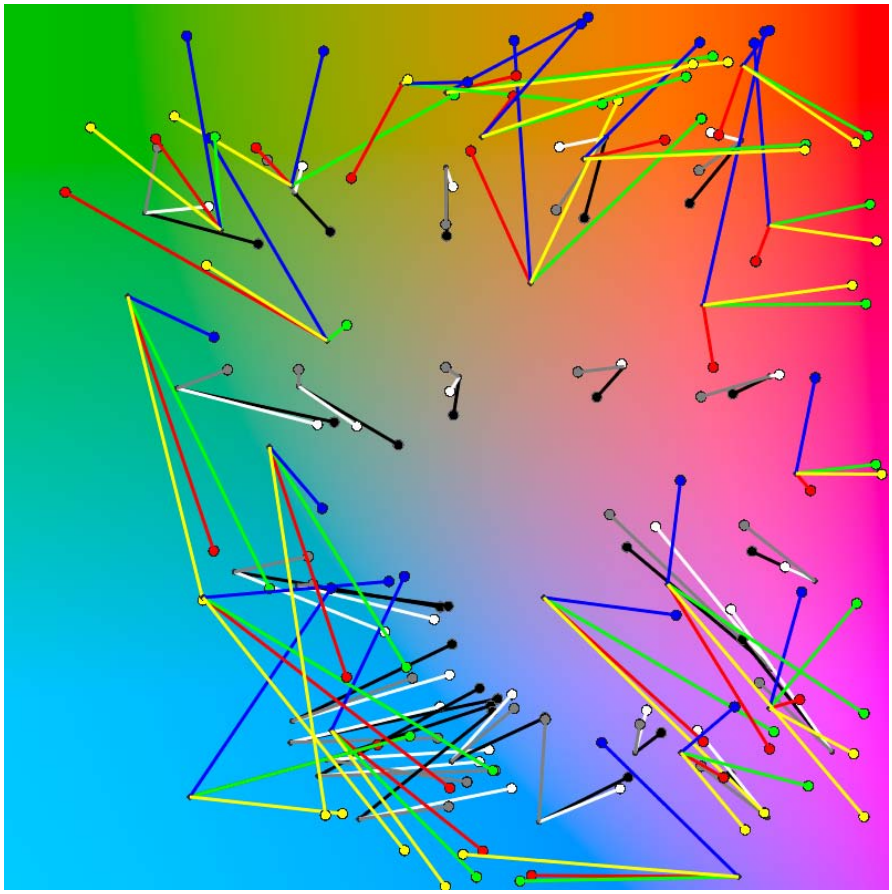

**Fig. 9** In comparing the direction and magnitude of the color matching error under the various background conditions for the two color matching experiments, we see a few patterns emerge in CIELAB color space. The reference color is denoted by small, dark gray circles with no outlines, whereas the circles with outlines represent the mean color error ($\Delta E$, represented by distance and angle in color space) in user matching. The achromatic circles with gray or black outlines indicate the black or white backgrounds according to their inner color, with the gray inner circles with black outlines indicating the see-through condition. In the first test, we used a set of 24 reference samples. The colored circles with black outlines indicate the colored backgrounds (red, green, blue, and yellow) from the second experiment, which used a new set of 22 reference samples.

manages to compensate for misalignment, eye strain and headaches are likely from extended use of a device with this disparity. This disparity was measured using nonius lines for a set of Glasstron displays [24] and corrected with three algorithms. Two modified the six degree-of-freedom offset between the eyes to include (1) a vertical component to the IPD translation or (2) a pitch rotation; (3) correction by image shift was also studied. Notable variability between different devices of the same manufacturer and model were noted, and the correction methods did not yield

equivalent visual angles, indicating the tolerance of the human visual system to adjust to errors (despite the potential for fatigue effects).

While correction of such vertical disparity is a necessary condition for proper perception of stereo, it is not sufficient for understanding the stereo capability provided by a head-worn display. By testing the depth of a virtual target to the depth of a real reference object, one can measure the stereoacuity users experience with an AR display. For two custom-built HMPDs, stereoacuity measurements were recorded in pilot tests [8]. With a $52^\circ$ diagonal FOV and $640 \times 480$ graphical resolution, five subjects were determined to have stereoacuities between 1.1 arcminutes and 6.2 arcminutes with the real reference at 80cm, between 1.4 arcminutes and 3.0 arcminutes with the reference at 150cm, and between 0.6 arcminutes and 0.8 arcminutes with the reference at 300cm. With a $42^\circ$ diagonal FOV and $800 \times 600$ graphical resolution, five subjects were determined to have stereoacuities between 1.1 arcminutes and 1.7 arcminutes with the real reference at 80cm, between 1.2 arcminutes and 2.6 arcminutes with the reference at 150cm, and between 0.4 arcminutes and 1.7 arcminutes with the reference at 300cm.

Stereoacuity also may be measured through application of a depth-matching task with horizontal disparity as the independent variable. This disparity normally gives the impression of depth to the human visual system. A test task of matching the apparent depth of a virtual target to a real reference object (provided on a monitor visible through the HMD) was applied to the nVisorST [21]. The results showed that subjects generally achieved a stereoacuity between 1.9 and 3.9 arcmin; this may be compared to the typical stereoacuity of one arcminute, although hyperacuity for stereoscopic depth can reach three arcseconds. As may be inferred from the discussion above, lower contrast and smaller size of the object decreased the stereoacuity (i.e. raised the detection threshold measured in arcmin). Regression of the mean disparity for each subject versus the subject's IPD showed an excellent linear fit, indicating that users were able to convincingly verge the real and virtual objects.

## 4 Discussion

Summarizing a diverse array of experiments such as those described above is destined to be a difficult task. But from each set of experiments, we learned important lessons about how head-worn AR displays affect basic perceptual capabilities of the human visual system. It is also important to note differences in the methods use to collect data.

First, we note the importance of evaluating the quality of seeing both the real environment and the virtual objects that the AR application places within it. This merging of real and virtual worlds is the fundamental characteristic of AR, and as such should be taken into account in any evaluation of the perceptual quality of an AR technology or application. Thus the see-through condition for optical see-through displays and the background video in video overlay displays are critical conditions to be evaluated, just as the display elements and optics that (respectively) generate

and mediate the view of the virtual objects are obvious targets of evaluations. It immediately follows that identification of the limiting factor in an optical or video AR display (display elements, optics, masks, and cameras, as applicable) is of great importance to the AR system designer. It is also worth noting that we reviewed primarily work on head-worn displays of optical see-through and video overlay types, adding comments about alternative displays in the few locations where data exists. As hand-held displays become more popular, there will be an increasing need for these types of evaluations to be conducted with hand-held displays.

The second defining feature of AR is the *registration*, or alignment, of the virtual objects to the real environment. While the geometric measurements in modeling and tracking the objects and the user are central to registration, being able to accurately discern the virtual and real objects is also a prerequisite. If a user is to be able to understand the relationship between real and virtual with a reasonable degree of precision, then it stands to reason that basic measures such as contrast sensitivity (incorporating both size and difference in brightness or color) contribute to the understanding of whether users will perceive objects to exist in the merged AR world.

A chief application of AR is to convey information about the surrounding world; as such, many useful AR applications overlay text on the real environment. If this text is to be useful, then it must be legible when presented in an AR display. While the contrast sensitivity measure will indicate this in an abstract context, reading language operates at a level above raw distinction of which (tiny) regions of an image (whether real or virtual) compose a letter; familiarity with a language breeds an ability to understand words without seeing all the letters. Thus the raw resolution required may be too strict a requirement; in the world of small mobile devices, this may be exploited to the benefit of the human visual system.

One underlying theme we detect in the conduct of the experiments described here is that there are many experimental tasks that AR can copy from the perception literature and everyday contact with specialists in aspects of perception (e.g. optometrists). Virtual eye charts can be traced to the early days of virtual environment research [16]. But the perception research literature may have superior recommendations over the clinical practices we experience in our personal lives. The measurement of contrast sensitivity versus visual acuity is one example of this. Measuring the precise distortion of color is another; this type of data is far superior in value to the results of standard color vision testing for the display manufacturer who works to improve the color quality of an AR display. Careful consideration of the task design may become more critical in shifting the emphasis of work to mobile displays; the style of use may not be as conducive to adapting optometric examinations as the style of use of head-worn displays.

Our emphasis on an iterative evaluation process is justified by the results of most of the basic perception tests. The experiments studied showed how AR displays frequently limit the basic perceptual capabilities of the users. But users may not always be able to identify the precise nature of such problems. In the BARS application mentioned above, we could identify numerous situations in which the display of text would help the user. When we built prototype applications and showed them to users, many of the negative comments indicated the difficulty subjects had seeing

in displays, but how to address low visibility of text that is overlaid optically on top of a bright real environment with a combination of size and contrast is not likely to be prescribed by end users. At the same time, even something as simple as the time required to read text was shown to vary with the color contrast and size, which clearly has implications for the application.

Color presents an array of issues for AR displays. We reviewed data showing the color distortion that occurs with both optical and video displays. The former must compete with the uncontrolled background, and have their color gamut severely altered by the optical conditions of the display and the background. In this area, far more data is likely to be needed before detailed correction functions can be derived for the diverse set of circumstances proposed for AR applications, especially mobile applications. Even with regard to video overlay displays (which, in the form of mobile phones, are increasingly the choice for mobile applications), the issues of color contrast are not solved, although the greater control of the final image that reaches the eye has enabled greater progress. Here, both objective measurements taken with a spectrophotometer (also known as a color meter) and subjective measurements collected from human subjects contribute to our understanding and towards a solution. The contextual nature of color – in which adjacent colors and intensities affect the perception of the neighboring colors and intensities – implies the need to acquire both objective and subjective measurements, as well as an understanding of the application context.

With regard to stereoacuity, we saw two critical issues. First was the potential for improper stereo to lead to eye fatigue and headaches. While studies have also demonstrated the tolerance of the human visual system to errors in stereo displays, these factors clearly limit the use of improperly calibrated stereo displays. As AR is still considered a strong candidate for medical and manufacturing applications in which work is to be done at close distances, stereo would seem to be an important feature to correct in AR displays. The second critical issue is that the displays still limit the human visual system from applying the binocular cues for scene understanding, as evidenced by the studies conducted.

Perhaps the most important lesson to be learned from this review is the sparse amount of data that has been collected on these fundamental questions for AR display and thus AR systems. Replication and extension to new devices, filling in the gaps in data collection, and designing compensation or correction algorithms would all benefit the field. We also encourage AR researchers who conduct evaluations of AR applications to learn from the lessons taught to us by our users: AR applications may fail to meet expectations (of users and/or designers) for reasons that range from the “high-level” application soundness down to “low-level” issues of basic perception. Evaluators would be wise to conduct studies of basic perception when looking for the reasons an application fell short. Improved hardware will surely improve the results of the studies discussed here. But, as several studies showed, clever use of the limited resources can overcome the perceptual challenges and lead to greater application success.

## References


1. Brown, D.G., Stripling, R., Coyne, J.T.: Augmented reality for urban skills training. In: IEEE Virtual Reality, pp. 249–252 (2006)
2. Bruce, V., Green, P.R., Georgeson, M.A.: Visual Perception: Physiology, Psychology, and Ecology, 3rd edn. Psychology Press (1996)
3. Cutting, J.E., Vishton, P.M.: Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In: W. Epstein, S. Rogers (eds.) Handbook of Perception and Cognition, Vol. 5: Perception of Space and Motion, 2nd edn., pp. 69–117. Academic Press (1995)
4. Darken, R.P., Sullivan, J.A., Lennerton, M.: A chromakey augmented virtual environment for deployable training. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (2003)
5. Endsley, M.R.: Measurement of situation awareness in dynamic systems. Human Factors **37**(1), 65–84 (1995)
6. Farnsworth, D.: The Farnsworth-Munsell 100-hue and dichotomous tests for color vision. Journal of the Optical Society of America **33**(10), 568–578 (1943)
7. Fidopiastis, C., Fuhrman, C., Meyer, C., Rolland, J.: Methodology for the iterative evaluation of prototype head-mounted displays in virtual environments: Visual acuity metrics. Presence: Teleoperators and Virtual Environments **14**(5), 550–562 (2005)
8. Fidopiastis, C.M.: User-centered virtual environment assessment and design for cognitive rehabilitation applications. Ph.D. thesis, Dept. of Modeling and Simulation, Univ. of Central Florida (2006)
9. Gabbard, J.L., Hix, D., Swan II, J.E., Livingston, M.A., Höllerer, T.H., Julier, S.J., Brown, D., Baillot, Y.: Usability engineering for complex interactive systems development. In: Human-Systems Integration Symposium (2003)
10. Gabbard, J.L., Swan II, J.E., Hix, D.: The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. Presence: Teleoperators and Virtual Environments **15**(1), 16–32 (2006)
11. Gabbard, J.L., Swan II, J.E., Hix, D., Jung Kim, S., Fitch, G.: Active text drawing styles for outdoor augmented reality: A user-based study and design implications. In: IEEE Virtual Reality, pp. 35–42 (2007)
12. Gabbard, J.L., Swan II, J.E., Hix, D., Lanzagorta, M., Livingston, M.A., Brown, D., Julier, S.: Usability engineering: Domain analysis activities for augmented reality systems. In: The Engineering Reality of Virtual Reality (SPIE Volume 4660), pp. 445–457 (2002)
13. Gabbard, J.L., Zedlitz, J., Swan II, J.E., Winchester III, W.W.: More than meets the eye: An engineering study to empirically examine the blending of real and virtual color spaces. In: IEEE Virtual Reality, pp. 79–86 (2010)
14. Ginsburg, A.P., Hendee, W.R.: Quantification of Visual Capability, pp. 52–71. Springer-Verlag (1992)
15. Hering, E.: Outlines of a Theory of the Light Sense. Harvard University Press, Cambridge, MA (1964)
16. Holloway, R., Fuchs, H., Robinett, W.: Virtual worlds research at the University of North Carolina at Chapel Hill as of February 1992. In: Visual Computing: Integrating Computer Graphics with Computer Vision, pp. 109–128. Springer-Verlag (1992)
17. Ishihara, S.: Tests for Colour-blindness. Hongo Harukicho, Handaya, Tokyo (1917)
18. Kirkley, Jr., S.E.H.: Augmented reality performance assessment battery (ARPAB): Object recognition, distance estimation and size estimation using optical see-through head-worn displays. Ph.D. thesis, Instructional Systems Technology, Indiana University (2003)
19. Leykin, A., Tuceryan, M.: Automatic determination of text readability over textured backgrounds for augmented reality systems. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 224–230 (2004)
20. Livingston, M.A.: Quantification of visual capabilities using augmented reality displays. In: IEEE International Symposium on Mixed and Augmented Reality, pp. 3–12 (2006)

21. Livingston, M.A., Ai, Z., Decker, J.: A user study towards understanding stereo perception in head-worn augmented reality displays. In: IEEE International Symposium on Mixed and Augmented Reality (2009)
22. Livingston, M.A., Barrow, J.H., Sibley, C.M.: Quantification of contrast sensitivity and color perception using head-worn augmented reality displays. In: IEEE Virtual Reality, pp. 115–122 (2009)
23. Livingston, M.A., Brown, D., Julier, S.J., Schmidt, G.S.: Mobile augmented reality: Applications and human factors evaluations. In: NATO Human Factors and Medicine Panel Workshop on Virtual Media for Military Applications (2006)
24. Livingston, M.A., Lederer, A., Ellis, S.R., White, S.M., Feiner, S.K.: Vertical vergence calibration for augmented reality displays. In: IEEE Virtual Reality (Poster Session) (2006)
25. Livingston, M.A., Rosenblum, L.J., Julier, S.J., Brown, D., Baillot, Y., Swan II, J.E., Gabbard, J.L., Hix, D.: An augmented reality system for military operations in urban terrain. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (2002)
26. Livingston, M.A., Swan II, J.E., Julier, S.J., Baillot, Y., Brown, D., Rosenblum, L.J., Gabbard, J.L., Höllerer, T.H., Hix, D.: Evaluating system capabilities and user performance in the battlefield augmented reality system. In: Performance Metrics for Intelligent Systems Workshop (PerMIS'04) (2004)
27. Livingston, M.A., Zanbaka, C., Swan II, J.E., Smallman, H.S.: Objective measures for the effectiveness of augmented reality. In: IEEE Virtual Reality (Poster Session), pp. 287–288 (2005)
28. Livingston, M.A., Zanbaka, C.A., Swan II, J.E., Smallman, H.S.: Objective measures for the effectiveness of augmented reality. In: IEEE Virtual Reality 2005 (Poster Session), pp. 287–288 (2005)
29. Peli, E.: Contrast in complex images. Journal of the Optical Society of America A **7**(10), 2032–2040 (1990)
30. Peterson, S.D., Axholt, M., Ellis, S.R.: Label segregation by remapping stereoscopic depth in far-field augmented reality. In: IEEE International Symposium on Mixed and Augmented Reality, pp. 143–152 (2008)
31. Peterson, S.D., Axholt, M., Ellis, S.R.: Objective and subjective assessment of stereoscopically separated labels in augmented reality. Computers & Graphics **33**(1), 23–33 (2009)
32. Renkewitz, H., Kinder, V., Brandt, M., Alexander, T.: Optimal font size for head-mounted-displays in outdoor applications. In: International Conference on Information Visualisation, pp. 503–508 (2008)
33. Riggs, L.A.: Visual acuity. In: Vision and Visual Perception, pp. 321–349. John Wiley and Sons (1965)
34. Smallman, H.S., Boynton, R.M.: On the usefulness of basic colour coding in an information display. Displays **14**(3), 158–165 (1993)
35. Sutherland, I.E.: A head-mounted three-dimensional display. In: 1968 Fall Joint Computer Conference, vol. 33, pp. 757–764. Thompson Book Co. (1968)
36. Thomas, B., Close, B., Donoghue, J., Squires, J., Bondi, P.D., Piekarski, W.: First person indoor/outdoor augmented reality application: ARQuake. Personal and Ubiquitous Computing **6**(1), 75–86 (2002)
37. Vingrys, A.J., King-Smith, P.E.: A quantitative scoring technique for panel tests of color vision. Investigative Ophthalmology and Visual Science **29**(1), 50–63 (1988)
38. Ware, C.: Information Visualization: Perception for Design, 2nd edn. Morgan Kaufmann (2004)
39. Weiland, C., Braun, A.K., Heiden, W.: Colorimetric and photometric compensation for optical see-through displays. In: Intelligent and Ubiquitous Interaction Environments, Universal Access in Human-Computer Interaction, part of HCI International, LNCS Volume 5615, pp. 603–612 (2009)
40. Woods, R.L., Fetchenheuer, I., Vargas-Martín, F., Peli, E.: The impact of non-immersive head-mounted displays (HMDs) on the visual field. Journal of the Society for Information Display **11**(1), 191–198 (2003)