Published in final edited form as:
SMPTE Motion Imaging J. 2012 ; 121(4): 24–43. doi:10.5594/j18173.

Stereoscopy and the Human Visual System

Martin S. Banks, **Jenny C. A. Read**, **Robert S. Allison**, and **Simon J. Watt**

Abstract

Stereoscopic displays have become important for many applications, including operation of remote devices, medical imaging, surgery, scientific visualization, and computer-assisted design. But the most significant and exciting development is the incorporation of stereo technology into entertainment: specifically, cinema, television, and video games. In these applications for stereo, three-dimensional (3D) imagery should create a faithful impression of the 3D structure of the scene being portrayed. In addition, the viewer should be comfortable and not leave the experience with eye fatigue or a headache. Finally, the presentation of the stereo images should not create temporal artifacts like flicker or motion judder. This paper reviews current research on stereo human vision and how it informs us about how best to create and present stereo 3D imagery. The paper is divided into four parts: (1) getting the geometry right, (2) depth cue interactions in stereo 3D media, (3) focusing and fixating on stereo images, and (4) how temporal presentation protocols affect flicker, motion artifacts, and depth distortion.

GETTING THE GEOMETRY RIGHT

What are we trying to do when we present stereo displays? Are we trying to recreate the scene as a physically present viewer would have seen it, or simply give a good depth percept? How should we capture and display the images to achieve each of these? Vision science does not yet have answers regarding what makes a good depth percept, but in this section we aim to cover the geometrical constraints and lay out what is currently known about how the brain responds to violations of those constraints.

Puppet Theater

Figure 1 depicts a three-dimensional (3D) display reproducing a visual scene as a miniature model—a puppet theater, if you will—in front of the viewer. Conventional stereoscopic displays cannot recreate the optical wavefronts of a real visual scene. For example, the images are all presented on the same physical screen; therefore, they cannot reproduce the varying accommodative demand of real objects at different distances. In addition, they cannot provide the appropriate motion parallax as the viewer's head moves left and right. However, they can in principle, reproduce the exact binocular disparities of a real scene. In many instances, this is an impossible or inappropriate goal. However, we argue that it is important to understand the underlying geometrical constraints of the "puppet theater" to understand what we are doing when we violate the constraints. Thus, it is a helpful exercise to consider what we need to reproduce the disparities that would be created by a set of physical objects seen by the viewer.

Presented at the 2011 SMPTE International Conference on Stereoscopic 3D for Media and Entertainment, 21–22 June 2011.

Epipolar Geometry and Vertical Disparity

The images we need to create depend on how they will be displayed. In the real world, a point in space and the projection centers of the two eyes define a plane; this is the so-called epipolar plane. To recreate this situation on a stereo 3D (S3D) display, we have to recreate such an epipolar plane. Assume the images will be displayed on a screen frontoparallel to the viewer so that horizontal lines on the screen are parallel to the line joining the two eyes (Figs. 1 and 2). This is approximately the case for home viewing of S3D television (TV). The first constraint in this situation is that to simulate real objects in the puppet theater, there must be no vertical parallax on the screen; otherwise, the points on the display screen seen by the left and right eyes will not lie on an epipolar plane. (We use the convention that *parallax* refers to separation on the screen and *disparity* refers to separation on the retina.) Figure 2 illustrates why. Irrespective of where the eyes are looking (provided that the viewer's head does not tilt to the side), the rays joining each eye to a single object in space intersect the screen at points that are displaced horizontally on the screen; that is, epipolar-plane geometry is preserved. Thus, to simulate objects physically present in front of the viewer, the left and right images must be presented on the display screen with no vertical separation.

What Happens When We Get the Geometry Wrong?

If the stereo images on the display do contain vertical parallax, they are not consistent with a physically present object. Vertical parallax can be introduced by obvious problems, such as misalignments of the cameras during filming or misalignments of the images during presentation, but they can also be introduced by more subtle issues, such as filming with converged cameras ("toe-in").[1] These two sources of vertical parallax cause a change in the vertical disparities at the viewer's retinas and are likely to affect the 3D percept.

Vertical Disparities Arising from Misalignments

The eyes move partly to minimize retinal disparities. For example, vergence eye movements work to ensure that the lines of sight of the two eyes intersect at a desired point in space. Horizontal vergence (convergence or divergence) is triggered by horizontal disparities. If the eyes are vertically misaligned, the lines of sight do not intersect in space. Instead, there is a constant vertical offset between the two eyes' images (Fig. 3 (a)). The human visual system contains self-correcting mechanisms designed to detect such a vertical offset and correct for it by moving the eyes back into alignment.[2–4] The eye movement that accomplishes this is a vertical vergence.

In stereo displays, small vertical misalignments of the images activate vertical vergence. This could occur, for example, if the cameras are misaligned because one camera is rotated about the axis joining the centers of the two cameras or, in a cinema, if the projectors are offset vertically. In these instances, the viewers automatically diverge their eyes vertically so as to remove the offset between the retinal images. This happens automatically, so inexperienced viewers are usually not consciously aware of it. It is likely to cause fatigue and eyestrain if it persists.

Passive stereo displays in which the left and right images are presented on alternate pixel rows could introduce a constant vertical disparity corresponding to 1 pixel—if the left and right images were captured from vertically aligned cameras and then presented with an offset (Fig. 4 (a)). If instead the images are captured at twice the vertical resolution of each eye's image, with the left and right images pulled from the odd and even pixel rows, respectively, there is no overall vertical disparity but just slightly different sampling (Fig. 4 (b)). In any case, the vertical disparity corresponding to 1 pixel viewed at 3 picture heights is only about a minute of arc, which is probably too small to cause eyestrain.

A similar situation occurs if the images are misaligned by being rotated about an axis perpendicular to the screen. Again, the brain automatically seeks to null out rotations of up to a few degrees by rotating the eyes about the lines of sight, an eye movement known as *cyclovergence* (Fig. 3 (b)).[5] This also produces discomfort, fatigue, and eyestrain.

Vertical Disparities Arising from Viewing Geometry—The human visual system (and human stereographers) works hard to avoid misalignments. But even if both eyes (or both cameras) are perfectly aligned, vertical disparities between the retinal (or filmed) images can still occur. Figure 5 shows two cameras converged on a square structure in front of them. Because each camera is viewing the square obliquely, its image on the film is a trapezoid because of keystoning. The corners of the square are thus in different vertical positions on the two films, and this violates epipolarplane geometry (Fig. 2). A viewer looking at the stereo display of these trapezoids receives a pattern of vertical disparities that is inconsistent with the original scene.

One can deduce the relative alignment of the cameras from the pattern of vertical disparities.[6] The exact position of objects in space can then be estimated by backprojecting from the retinal images. The visual system uses the pattern of vertical disparities across the retina to interpret and scale the information available from horizontal disparities.[7–9] For this reason, vertical disparities in stereo displays may not just degrade the 3D experience but also produce systematic distortions in depth perception.

One well-known example is the induced effect.[10] In this illusion, a vertical magnification of one eye’s image relative to the other causes a perception that the whole screen is slightly slanted, that is, rotated about a vertical axis. This is thought to be because similar vertical magnification occurs naturally when we view a surface obliquely.

Estimating Convergence from Vertical Disparity—For the purpose of S3D displays, a pertinent example concerns vertical disparities associated with convergence. For the brain to interpret 3D information correctly, it must estimate the current convergence angle with which it is viewing the world. This is because, as Fig. 6 shows, a given retinal disparity can specify vastly different depth estimates depending on whether the eyes are converging on a point close to or far from the viewer.

The brain has several sources of information about convergence. Some of these are independent of the visual content, for example, sensory information from the eye muscles. However, the pattern of vertical disparities also provides a purely retinal source of information. Consider the example in Fig. 5. The equal and opposite keystoning in the two images instantly tells us that these images must have been acquired by converged cameras. The larger the keystoning, the more converged the cameras.

An extensive vision science literature examines humans’ ability to use these cues. This shows that humans use both retinal and extraretinal information about eye position.[11–13] As we expect from a well-engineered system, more weight is placed on whichever cue is most reliable. Generally, the visual system places more weight on the retinal information, relying on the physical convergence angle only when the retinal images are less informative.[11,14] For example, because the vertical disparity introduced by convergence is larger at the edge of the visual field, less weight is given to the retinal information when it is available only in the center of the visual field.[13]

These vertical disparities can have substantial effects on the experience of depth. In one experiment, the same horizontal disparity (10 arcmin) resulted in a perceived depth difference of 5 cm when the vertical disparity pattern indicated viewing at infinity but only 3 cm when the vertical disparity pattern indicated viewing at 28 cm—although in both cases, the physical viewing distance was 57 cm.[9]

Effects of Filming with Converged Cameras—Epipolar-plane geometry is relevant to the vexing issue of whether stereo content should be shot with camera axes parallel or converged (toe-in). Some stereographers have argued that cameras should converge on the subject of interest in filming because the eyes converge in natural viewing. While there are good reasons for filming toe-in, this particular justification is not correct. It depends on the fallacy that cameras during filming are equivalent to eyes during viewing. This would be the case only if the images recorded during filming were presented directly to the audience's retinas, without distortion. Instead, the images recorded during filming are presented on a screen that is usually roughly frontoparallel to the interocular axis (Fig. 2). The images displayed on the screen are thus viewed obliquely by each eye, introducing keystoning at the retinas. As described in Fig. 5, the retinal images therefore contain vertical disparities even if there is no vertical parallax on the screen. If the images displayed on the screen have vertical parallax because they were captured with converged cameras, this adds to the vertical disparity introduced by the viewer's own convergence. The resulting vertical disparity indicates that the viewer's eyes are more converged than they really are. As we have seen, this could potentially reduce the amount of perceived depth for a given horizontal disparity.

To correctly simulate physical objects, one should film with the camera axes parallel, as shown in Fig. 7. To display the resulting images, one should shift them horizontally so that objects meant to have the same simulated distance as the screen distance have zero horizontal parallax on the screen. Provided that the viewer keeps the interocular axis horizontal and parallel to the screen, this ensures that all objects have correct horizontal and vertical disparity on the retina, independent of the viewer's convergence angle.

Back-of-the-Envelope Calculations—To get a feel for how serious these effects might be, consider some back-of-the-envelope calculations. For convergence on the midline (i.e., looking straight ahead, not to the left or right), vertical disparity is independent of scene structure and simply scales with convergence angle. To close approximation, the retinal vertical disparity at different points in the visual field is given by the following equation[8]:

$$[\text{retinal vertical disparity}]=[\text{convergence angle}]\times 0.5^{*}\sin(2^{*}\text{elevation})\times\tan(\text{azimuth}),$$

where azimuth and elevation refer to location in the visual field. This equation is for the vertical disparity in natural viewing. That is, even if an object is displayed with zero screen parallax, it still has a vertical disparity of 7 arcmin when viewed with 1° convergence at 20° elevation and 20° azimuth. The same equation can be used to compute the on-screen vertical disparity resulting from filming toed-in. For example, what degree of toe-in is necessary to cause a 1-pixel vertical disparity? For 36mm film with a 50 mm focal length, the corners of the image are at an azimuth equal to 20° and elevation equal to 13°. If the 36mm is represented by 2048 pixels, a vertical disparity of 1 pixel is 1.2 arcmin. This can be caused by a toe-in of just 14 arcmin.

Is this enough to alter perception? Suppose that the images on the screen have a pattern of on-screen vertical parallax resulting from having been filmed toed-in:

$$[\text{on}-\text{screen vertical parallax}]=[\text{some scale factor } K]\times 0.5^{*}\sin(2^{*}\text{elevation})\times\tan(\text{azimuth}).$$

This combines with the natural vertical disparity, indicating the wrong convergence angle. The scale factor K, which has angular units, is the additional, artifactual component of the convergence estimate that would be added if the visual system worked solely on the retinal information.

Suppose the viewer is in an IMAX cinema, screen size 22 × 16 m, viewing it at a distance of one screen height: 16 m. The true convergence angle is therefore 14 arcmin. At the corner of the screen, elevation equals 27° and azimuth equals 35°. Physical objects at the corners of the screen produce a retinal vertical disparity of 2.0 arcmin just because of the geometry. Suppose the toed-in vertical parallax is such that it is just 1 cm even at the corners of the screen (clearly, it is smaller everywhere else). This means that the toe-in contributes an additional 2.1 arcmin of vertical disparity at elevation equals 27° and azimuth equals 35°. That is, the barely noticeable on-screen parallax more than doubles the vertical disparity at the retina; hence, the retinal cue to convergence is 29 arcmin instead of the physical value of 14 arcmin.

Roughly speaking, the convergence overestimate in degrees equals 180/π* [viewing distance] [on-screen vertical separation at (x,y)]/x/y.

In the preceding calculation, the viewing distance was 16 m and the vertical separation was 1 cm at $x = 11$ m and $y = 8$ m, implying a convergence angle that is too large by about 0.1°.

What implications might this convergence error have for perceived shape? Suppose the images accurately simulate a transparent sphere, with a 1 m radius, at the center of the screen. The sphere has an angular radius of 3.6°, and its front and back surfaces have a horizontal disparity of −0.93 arcmin and 0.82 arcmin, respectively. If these disparities were interpreted with the actual viewing distance of 16 m and convergence of 14 arcmin, the viewer should correctly perceive a spherical object, with a 1 m radius, 16 m away. But if the images are interpreted assuming a convergence of 29 arcmin and viewing distance of 8 m, then the on-screen parallax implies a spheroid with an aspect ratio of 2: that is, a radius of 0.5 m in the screen plane and just 0.25 m perpendicular to the plane of the screen. Thus, for the same horizontal parallax and the same viewing position, a supposedly spherical object could be perceived as flattened by a factor of 2 simply because of toed-in vertical parallax, even when this is just 1 cm at the corners of the screen.

In practice, the distortion may not be so obvious. For example, other powerful perspective and shading cues may indicate that the object is spherical. Nevertheless, these calculations suggest that small vertical parallax can potentially have a significant effect on perception.

As yet, little work has been done to investigate depth distortions caused by toed-in filming. From the vision science literature to date, we predict different effects for S3D cinema versus TV. In a cinema, the display typically occupies much of the visual field. Thus, we expect convergence estimates to be dominated by the retinal information, rather than the physical value. In this situation, the same horizontal disparities could produce measurably different depth percepts if acquired with converged camera axes versus parallel. In home viewing of 3DTV, the visual periphery is generally stimulated by objects in the room. These necessarily produce vertical disparities consistent with the viewer’s physical convergence angle, while vertical disparities within the relatively small TV screen are likely to have less effect. This means that horizontal parallax on the TV screen is likely to be converted into depth estimates using the viewer’s physical convergence angle. Thus, we expect the angle between the camera axes to have less effect on the depth perceived in this situation.

Interaxial Distance—The separation of the cameras during filming is another important topic. To exactly recreate the puppet theater, one should film with the cameras one interocular distance apart. However, stereographers regularly play with interaxial distance (i.e., the separation between the optical axes of the cameras). For example, they might start with a large interaxial distance to produce measurable parallax in a shot of distant mountains and then reduce the interaxial distance as the scene changes to a close-up of a dragon on the mountain. A remarkable recent experiment demonstrated that most observers are insensitive to changes in interaxial distance within a scene. Although we could detect the resulting changes in disparity if they occurred in isolation, when they occur within a given scene we do not perceive them, because we assume the objects stay the same size. In the words of the authors, "Humans ignore motion and stereo cues [to absolute size] in favor of a fictional stable world."[15]

Why We Don't Need to Get It Right

Ultimately, the central mystery for vision science may be why S3D TV and cinema works as well as they do. By providing an additional, highly potent depth cue, S3D content risks alerting the visual system to errors it might have forgiven in two-dimensional (2D) content. As an example, an actor's head on a cinema screen may be 10 ft high, but we do not perceive it as gigantic. We could argue that this is because a 10 ft head viewed from a distance of 30 ft subtends the same angle on the retina as a 1 ft head viewed from 3 ft. Stereo displays, however, potentially provide depth information confirming that the actor is indeed gigantic. In addition, stereo displays often depict disparities that are quite unnatural, that is, disparities that are physically impossible for any real scene to produce given the viewer's eye position or disparities that conflict with highly reliable real-world statistics (mountains are hundreds of feet high, people are around 6 ft high, etc.). This is reminiscent of the "uncanny valley" in robotics, where improving the realism of a simulated human can produce revulsion.[16]

Presumably, such conflicts are the reason a minority of people find S3D content disturbing or nauseating. However, most of us find S3D content highly compelling despite these violations of the natural order. An analogy can be drawn with the way we perceive most photographs as veridical depictions of the world. We do not usually perceive objects in photographs as distorted or straight lines as curved, even though the image on our retina is substantially different from that produced by the real scene—unless we are viewing the photograph from the exact spot the camera was located to take it.[17] It is not yet known to what extent this is a learned ability, raising the possibility that as stereo displays become more commonplace, our visual systems will become even better at interpreting them without adverse effects.

DEPTH CUE INTERACTIONS IN STEREOSCOPIC 3D MEDIA

Adding binocular disparity enriches media with a vivid sense of depth, solidity, and space. However, the traditional pictorial depth cues—shading, shadows, blur, aerial perspective (haze or smoke), linear perspective, texture gradient, occlusion, and so on—used to provide a sense of depth, space, and texture are still present. In S3D displays, as in 2D displays, these cues are important and active, as are the cues not normally provided by either 2D or S3D displays, such as accommodation and motion parallax because of head motion. These are not subsidiary or secondary cues replaced by binocular disparity when it is available; rather, they continue to contribute to the qualitative sense of three-dimensionality and the quantitative depth experienced with two eyes, as well as one. However, in S3D media (as in the real world), these multiple sources often provide incomplete, imprecise, ambiguous, and even contradictory depth information. The visual system has the challenge of reconstructing a coherent 3D percept from these myriad and changing sensory signals.

Variety and Ambiguity of Stereoscopic Percepts

Stereopsis has two inherent ambiguities that are important for cue interactions. The first occurs because the image in one eye must be matched with that in the other. This correspondence problem can be nontrivial especially with repetitive textures. It has been a major challenge for computer stereo vision. In contrast, the human visual system seems to solve this problem effectively and effortlessly.[18] In addition to this capacity, most S3D media are rich and varied, making this the lesser of the ambiguities for our purposes. The other ambiguity is that retinal disparity does not directly specify depth. As described earlier in this paper, the amount of depth corresponding to a given disparity depends on distance and to a lesser extent on direction. In the absence of good information for distance, a given disparity can correspond to a large range of possible depths. Horizontal disparity does not provide this distance information, and binocular information from vergence or vertical disparity is limited to close range and has limited accuracy.

Stereopsis can support the perception of a 3D world in many respects, including discriminating a difference in depth, ordering objects in depth, judging slant or curvature, obtaining shape and relief, judging speed or direction of motion in depth, recovering surface properties, and obtaining accurate measures of depth between objects. Depending on the nature of the task, the ambiguities of stereopsis become more or less important. For example, to determine whether one object is placed in front of another does not require calibration for viewing distance, but estimating the size of the gap between them does.

In the visual appreciation of S3D film and other content, these perceptions are complex and multifaceted. Depth ordering and segregation help reduce clutter and separate subject from background, recovering shape and relief provides volume and depth, recovering binocular highlights gives a sense of gloss and luster, and so on. In S3D media, cue integration and combination need to be considered on all these levels, because they occur simultaneously and often seamlessly.

Ambiguity, Reliability, and Accuracy

The problem of vision is to “invert” the imaging process and recover the 3D world. But information is lost in the many-to-one transformation inherent in perspective projection. A given monocular image is compatible with multiple real scenes (Fig. 8); one of the possible scenes is that we are simply viewing a 2D image on a plane, which is the case in painting, film, and TV. Not all possible interpretations are equally likely. The structure and regularities of the world greatly constrain the problem. A long tradition in perception holds that depth perception relies on recreating the most likely 3D world consistent with the retinal image or images. Helmholtz called this process “unconscious inference.”[19] Modern variants on this idea codify this probabilistic interpretation of sensory signals in ways that, as we show, seem natural if not obvious to engineers.

Just as stereopsis supports many types of 3D judgments and suffers from ambiguities, monocular cues vary in the degree that they support such judgments. For example, occlusion is one of the least ambiguous depth cues. If one object blocks the view of another, it must be between the latter object and the viewer. One object may be cut away so that it looks like it blocks another, or both “objects” could be paint on a canvas. But occlusion is mostly an unambiguous source of relative depth information. Occlusion tells us nothing about the amount of depth between the two objects. We can imagine how occlusion and stereopsis might interact to determine depth order, but it is less clear how that would work for depth magnitude. That is, the two cues are not commensurate: they are apples and oranges that cannot be directly compared or combined (but see Burge et al.[20]).

Even when two cues are commensurate, they are not always comparable in terms of precision, reliability, or range. For example, the interocular distance in humans is much larger than the pupil aperture, so stereopsis is a more precise relative depth signal than blur.[21,22] Finally, the cues can differ in accuracy and provide biased estimates of depth or other 3D properties. For instance, shape estimates from shading are usually consistent with the assumption of a light source located above the viewer. When this assumption is not correct, shape from shading can provide biased estimates.

Cue Integration, Cue Combination, and Cue Conflict

Cue combination refers to the combination of sensory information to derive a percept of an object, feature, or scene. Cue integration refers more narrowly to combining multiple sources of commensurate information, that is, about the depth, shape, velocity, or some other aspect of an object.

Cue conflict occurs when two or more cues provide different and incompatible information. This is often thought of in terms of cue integration but can apply to other cue combination scenarios. Cue conflict can take place within binocular cues (e.g., vergence and stereopsis), or between stereopsis and other cues.

S3D media almost always produce a cue conflict. Many of these conflicts come with the technology, such as the conflict with accommodation, which always indicates S3D objects are at the screen plane, not where we portray them. Other conflicts are caused by the nature of the medium. For instance, scaling of images that arises from choice of lens or how display size affects depth from disparity (stereopsis) and perspective differently. Sitting off-center in the theater also has a different effect on depth from disparity and perspective. Finally, some natural conflicts simply arise from incompatible solutions to the ambiguities of vision. For instance, unusual lighting direction can make shape from shading incompatible with shape from disparity.

Conceptual and Computational Models of Cue Combination

There are many conceptual models of how cues can be perceptually combined including the following[23]:

1. *Cue dominance or vetoing*, where one cue determines the percept. A familiar example is ventriloquism, where the sound is “captured” by the visual input. In S3D, occlusion cues can veto depth from disparity at window violations.

2. *Summation and averaging*, which are additive interactions. These can be generalized to rather complex nonlinear interactions referred to as cooperative interactions.[24]

3. *Disambiguation*, where one cue disambiguates another (or they mutually disambiguate each other). For instance, the sign of blur is ambiguous, and a given amount of blur can result from focus in front of or beyond an object; stereopsis and other depth signals could disambiguate. Information from other depth cues can also disambiguate which interpretation of a perspective image should be favored, or confirm and stabilize a bistable perception.

4. *Calibration and adaptation*, which occurs when one cue provides information necessary to interpret another. For instance, motion or perspective can provide the distance signal necessary to obtain depth magnitude from stereopsis.

5. *Dissociation.* To be integrated, the cues should be *bound* together to apply to a common object feature or location. In contrast, dissociation of cues refers to interactions in which the cues are applied differentially, either interpreting them as

arising from different objects or applying them to different aspects of the same object.

Cue integration is the *fusing* together of redundant (typically commensurate) information, usually involving vetoing or averaging processes. The computational and behavioral literature[25] has distinguished between weak and strong fusion models. In strong fusion models, sensory inputs are combined without constraining how the information is combined. There is considerable anatomical and psychophysical evidence for modularity in the visual system. To a significant extent, various depth cues such as shading, stereopsis, and perspective may be processed independently to arrive at depth estimates for points in the scene. Models that combine the outputs of such depth modules are referred to as weak fusion models. Landy et al.[26] recognized the difficulty of integrating incommensurate cues and proposed the model of "modified weak fusion," in which outputs of depth modules are combined linearly but limited nonlinear interaction is allowed to "promote" cues so that they have a common measurement scale and can be combined. This promotion may include calibration, scaling, and other effects driven by secondary cues. Although simple, this idea of a linear combination of quasi-independent depth modules has been quite successful in practice.

The solution that often arises is familiar to engineers, particularly those trained in communications theory. The brain must arrive at the optimal or most probable percept α consistent with the set of depth estimates **x** (i.e., maximize the conditional probability of $P(\alpha \mid \mathbf{x})$). Unsurprisingly, classical techniques such as maximum likelihood estimators (MLEs)[27] have been applied successfully. While theoretically limited, such linear estimators and more general Bayesian estimators have proved surprisingly successful in describing quantitatively how cues interact in the laboratory. The basic MLE solution predicts that observers should weight the depth cues according to their reliability (Figure 9), which is the inverse of their variance $\left(1/\sigma^2_{\text{cue}}\right)$ (Fig. 9).[26] For example, with depth estimates *D* from two cues, we obtain the following:

$$D_{\text{optimal}} = w_1 \cdot D_{\text{cue1}} + w_2 \cdot D_{\text{cue2}}$$

$$w_1 = \frac{\frac{1}{\sigma^2_{\text{cue1}}}}{\frac{1}{\sigma^2_{\text{cue1}}} + \frac{1}{\sigma^2_{\text{cue2}}}} \text{ and } w_2 = \frac{\frac{1}{\sigma^2_{\text{cue2}}}}{\frac{1}{\sigma^2_{\text{cue1}}} + \frac{1}{\sigma^2_{\text{cue2}}}} \frac{1}{\sigma^2_{\text{optimal}}} = \frac{1}{\sigma^2_{\text{cue1}}} + \frac{1}{\sigma^2_{\text{cue2}}}$$

If the cues have equal reliabilities, their weights are both 1/2 (averaging) and the reliability of the combined estimate increases by a factor of 2.

More generally, cue integration can be formulated to take into account the likelihood of various perceptions and changes in the reliability of cues with distance, slant, and other factors.[28] Recent research has turned to issues of dependencies among the cues and robustness when they disagree (described later).

Cue Trading—If we can successfully express depth perception as a weighted depth cue combination, as described previously, an obvious question arises: To what extent we can trade one cue for another? Could we reduce interaxial distance (i.e., camera separation) and hence disparity for visual comfort while turning up the perspective, motion parallax or shading to compensate?[29] To a certain extent, this is possible and even mandatory if cues are combined at a perceptual level, with the viewer having access to only the final percept.[30] However, there are limits to the degree to which this can be accomplished and automated:

- MLE and other weighted averages of depth cues are only appropriate if the modules provide (noisy) estimates of the same value. If one cue is suspected to be strongly biased or inaccurate, the visual system should discount it. By analogy, if you made 10 measures of a parameter and 9 measured 50 ± 2 units, you would consider a 10th measurement of 500 to be likely probably caused by error and would therefore not average it with the others. It has been proposed that the visual system is similarly robust when cues are discrepant, for instance, vetoing unreliable cues.[26,31,32] However, linear cue integration sometimes seems to occur even when cues are discrepant.[33]
- The weights adopted can vary by viewer, even if all have good stereopsis. For instance, in judgments of the slant of surfaces, some observers preferentially weight perspective and others disparity. [34,35]
- The weights assigned can vary with the type of task, previous experience, type of scene, and location in the image.
- Van Ee et al. claim that discrepant depth cues can result in alternation of discrepant perceptions over time rather than stable cue integration,[36] though Girshick and Banks failed to replicate this finding.[35]

Thus, cue trading is a complex, scene-dependent, and often idiosyncratic process.

Cue Conflict Examples

Depth Sign: Cue conflicts in depth sign (in front versus behind) or depth order are often considered especially strong conflicts. The standard example of this in S3D film is window violation. Occlusion cues indicate the frame edge is in front of the stereoscopic imagery that is portrayed in front of it. Cue dominance may be perceived, with the occlusion cue pinning the surface to the edge of the screen. In other cases, strange and uncomfortable cue dissociations can be perceived. Similar issues arise with depth-sign errors in automated 2D-to-3D conversion.

Depth Magnitude: As described previously, if cue conflicts are modest, there tends to be trading or weighted averaging of cues to depth magnitude. Thus, other cues such as perspective and shading act to modulate the depth from disparity. Many of these conflicts are a consequence of differential effects of rig and projection parameters (e.g., focal length, interaxial distance, depth of field, screen distance, and screen size) on different depth cues. In many cases, cue integration can affect other aspects besides depth, such as apparent size.

Slant: The orientation of surfaces in depth, or slant, is important for shape and object recognition. The relationship between slant specified by perspective-based cues (e.g., texture gradients) and disparity-gradient cues varies with focal length and magnification on the one hand and rig parameters such as interaxial distance on the other. Studies have shown that observers weight perspective and disparity to arrive at an estimate of surface slant.[34,37] When surface slant from perspective and disparity differ greatly, observers tend to rely on one cue vetoing the other, but the cue preference is idiosyncratic and does not necessarily favor the most reliable.[35,38]

Misalignment of the stereo rig can also produce slant distortions (see the earlier description). Rotational misalignment of the images about the z-axis produces horizontal disparity patterns consistent with the scene being slanted in depth about a horizontal axis. Similarly, size miscalibration (e.g., because of a difference in focal length in the two cameras) produces disparities consistent with slant about a vertical axis. Another distortion caused by keystone distortions arising from toed-in convergence of the cameras. This predicts perceived curvature of stereoscopic space.[39,40] While undesirable, these distortions are most noticeable when monocular cues are weak and strong perspective can attenuate or eliminate them.

Qualitative Depth and Appearance: Interaction of stereopsis with shading and lighting in an S3D context is important. Much work needs to be done here, but it seems that lighting for depth can enhance the sense of volume and space in S3D content. Similarly, beyond geometrical properties of stereopsis like slant, stereopsis can influence perception of material properties like transparency.[41] Many stereographers feel that specular highlights should be avoided at all costs. In everyday experience, however, binocular differences in intensity produce perceptions of luster that support the perception of surface gloss.[42] An effect of lighting on perceived depth is provided in Fig. 10. In S3D content, however, one can obtain intensity disparities that are not associated with surface glossiness and thus can conflict with monocular information on shininess. For instance, if the beam splitter in a mirror rig is polarization sensitive (i.e., preferentially reflects one polarization state while transmitting the other), the two images can have large differences in intensity for reflecting surfaces like water and glass. These artifacts are caused by beam-splitter characteristics rather than the interocular difference in vantage point, so they are difficult for the brain to interpret ecologically (e.g., the entire surface of a pond might be bright in one eye but not the other). By their nature, specularities are highly directional, and are hence constrained, phenomena. Binocular specular highlights are informative, but fairly small changes can make them geometrically implausible. We might be more sensitive to incorrect binocular specularities than to other cue conflicts.

Tolerance to Cue Conflict—A key concern is the tolerance of the typical observer to these cue conflicts. How much can we tolerate? When problematic, how much does it bother us? Unfortunately, particularly in the context of rich cinematic content, these are still open questions. Cue conflict has been linked to simulator sickness effects and degraded perception. We understand in certain situations how cue conflict can cause issues (e.g., see the section on vergence and accommodation conflict). Most of these data have come from either nonspecific image-quality and comfort surveys or laboratory experiments. Generalizing these results to a viewer watching rich and varied content for a full-length motion picture is important but not straightforward.

Motion pictures, S3D or not, are not normally viewed from the seat equivalent to the center of the perspective projection. Banks et al.[43] has shown that we do not experience the distortion predicted from perspective geometry as we view the image off-axis. This is expected from our ability to watch TV. As we move our head, the perception is consistent with a flat picture, and we have presumably learned a type of constancy in which we interpret the image as essentially a projection normal to the plane. In S3D content, the screen plane is shattered and we see vivid depth. Banks et al. found that when seated off-axis while viewing a simple S3D scene, observers saw the scene according to the stereo geometry. One can demonstrate this by translating the head side to side while viewing an S3D display. The scene appears to rotate with the head translates and distorts as the 3D world morphs to be appropriate with the current viewpoint. Banks et al. found essentially none of the constancy effect they found for 2D images with their simple hinged surface stimulus. It remains to be determined whether stronger perspective information could produce partial perspective constancy in rich media like S3D films or whether the difference in the viewer's amount of experience with 2D and S3D media plays a role.

FOCUSING AND FIXATING ON STEREOSCOPIC IMAGES: WHAT WE KNOW AND NEED TO KNOW

Technological advances have improved stereo media since previous 3D fads. Nonetheless, problems of discomfort and fatigue (and poor stereoscopic depth perception) remain prevalent. For example, in a recent large-scale survey ($n > 7000$) by the Russian movie website Kinopoisk.ru, 36% of respondents reported experiencing headaches or eye tiredness while watching S3D movies.[44] As S3D viewing enters the mainstream and becomes a daily activity for the general population, there is a need to better understand how the human visual system responds to stereoscopic media if safe and effective content is to be developed.

Vergence–Accommodation Conflicts

For several reasons, viewing stereo media can have unpleasant effects on viewers.[45,46] Arguably the most important is the unnatural stimulus to the eye's focusing response. When we look at objects that are nearer or farther away, our eyes make two distinct oculomotor responses. The muscles in our eyes change the shape of the lens to try to focus the image on the retinas, a process called *accommodation*. At the same time, we rotate our two eyes equal and opposite amounts to try to bring the object of interest to the center of each retina, referred to as *vergence*. In natural viewing, we accommodate and converge to the same distance. Stereo cinema and TV systems, however, present images on a single, fixed image plane (the screen), so viewers must often make vergence eye movements to one distance (e.g., to an object nearer than the screen) while accommodating at a different distance (the screen surface; Fig. 11). Thus, there is mismatch between the stimulus to accommodation and the stimulus to vergence; this is the *vergence–accommodation conflict*.

Accommodation and vergence do not operate independently but are synergistically coupled. Under natural-viewing conditions, each response makes the other quicker and more precise. The “decoupling” of accommodation and vergence required by stereo media is difficult and effortful for many people and has been shown to cause discomfort and fatigue and to degrade the perception of depth.[47–54]

There are other causes of aversive symptoms in stereo media, including (1) misalignments or misscaling of the two eyes' images that arises from differences in the optics in pairs of stereo cameras or inaccuracies in camera rigs; (2) unnatural binocular disparities because of, for example, camera toe-in; and (3) cross-talk, or “ghosting,” where imperfect separation of the two eyes' images results in the left eye's image being partially visible to the right eye, and vice versa. Tractable solutions exist for these problems, however. Modern display technologies (active liquid-crystal shutter glasses; polarizing or chromatic filters on the projector, TV, or glasses; and line-by-line pattern polarization on some TVs) have all but eliminated ghosting. Perhaps more significantly, the switch to digital film (and displays) means that distortions and misalignments of the left and right eyes' images can be fixed in post-production using stereo image-processing software. In contrast, the vergence–accommodation conflict is fundamental to all existing stereo cinema and TV systems. Some researchers have developed multifocal-plane displays[47,55–58] that can successfully eliminate the conflict.[57,59] However, these displays do not permit multiple viewers, or even multiple single viewpoints, and thus do not offer a practical solution for cinema and TV.

If vergence–accommodation conflicts in cinema and TV cannot be eliminated, we must instead understand and quantify the exact conditions that cause aversive side effects. This allows guidelines to be developed for the amount of variation in stereo depth that is acceptable, the timescale over which variations can occur, and whether some sections of the population are more affected than others.

It has long been suspected that vergence–accommodation conflicts cause fatigue and discomfort, but it has only recently been confirmed empirically. Most studies of fatigue and discomfort compare effects of viewing stereo images with viewing normal 2D images. This approach is problematic for two reasons. First, 2D viewing typically differs from stereo viewing in several ways that can, themselves, cause aversive symptoms (including ghosting, motion judder from the temporal properties of the stereo system, and incorrectly aligned stereo images). Determining the effects of the vergence– accommodation conflict unambiguously therefore requires that the conflict be manipulated while keeping all other stimulus properties constant. Second, increased discomfort from viewing S3D media, compared to 2D viewing, could result not from the vergence–accommodation conflict per se but simply from the requirement to make vergence eye movements to fixate objects nearer and farther away. To rule out this possibility, the eye movements must also be equivalent in the two conditions. Thus, conventional stereo viewing should also be compared to equivalent real-world viewing conditions in which accommodation and vergence demands are varied together.

Hoffman et al.[49] used a multi-focal-plane display to create real-world variations in accommodation and vergence while holding all other stimulus properties constant. They compared viewers' reports of fatigue and discomfort in two viewing conditions: (1) conventional stereo display conditions, in which the stereoscopic depth of points in the images varied but the accommodation distance (screen distance) was fixed, and (2) real-world conditions, in which the accommodative distance varied with the variations in stereoscopic depth. In the conventional-display condition, viewers reported significantly higher levels of symptoms related to visual fatigue, indicating that vergence– accommodation conflicts can cause these aversive symptoms. They also showed that vergence– accommodation conflicts degrade depth perception, causing a reduction in the ability to discriminate fine detail in stereoscopic depth (stereoacuity) and increased time to fuse stereo images (see also Akeley et al.[47] and Watt et al.[53]).

Decoupling Vergence and Accommodation Responses—The eyes must focus and converge reasonably accurately; otherwise, the resulting perceptual experience will be poor. The accommodation error must be within the eye's depth of focus—approximately ±0.25 diopters (D)[60,61]—for the image to appear clear and sharp. In addition, the vergence error must be within Panum's fusion area (0.25° to 0.5°, or 0.07 to 0.14 D); otherwise, stereoscopic fusion does not occur, resulting in double vision (diplopia). The coupling of the accommodation and vergence systems means that these two responses cannot be varied independently, so with large conflicts in the stimuli to accommodation and vergence, stereo images are likely to appear blurred, diplopic, or both. It is therefore critical to understand the range within which accommodation and vergence responses can be decoupled without causing aversive side effects. Most of what we know about this comes from ophthalmological studies, designed to establish limits for prism and lens prescriptions for spectacles.[62] This work has given rise to two important concepts: the zone of clear, single binocular vision (ZCSBV) and Percival's zone of comfort (ZoC). The ZCSBV describes the extent to which accommodation and vergence responses can be decoupled while maintaining a clear, single binocular percept. It describes the maximum attainable decoupling of accommodation and vergence responses, but fatigue, discomfort, or both can be induced with much less decoupling. Based on experiments with prescribing spectacles, Percival[63] suggested that the middle third of the ZCSBV represented the range of vergence–accommodation postures that could be achieved without causing discomfort. This is referred to as Percival's ZoC (Fig. 12).

Stereoscopic ZoC—Percival's zone is useful conceptually, but it may be of only limited value in describing the ZoC for stereo displays. Vergence–accommodation conflicts resulting from lens or prism corrections in spectacles are likely to be easier for the system to adapt to (by adapting the vergence–accommodation coupling) because (1) they introduce a fixed offset between the stimuli to vergence and accommodation, whereas in stereo viewing the conflict constantly changes, and (2) spectacles are worn continuously, so people are exposed to a constant conflict for long durations, while stereo viewing occurs for relatively short durations. Thus, it is important to measure the ZoC for stereo viewing in a relevant context.

To our knowledge, only one study has attempted to map the ZoC for stereo displays while appropriately isolating the vergence–accommodation conflict. Shibata et al.[64] used an adaptive optics multifocal-plane display[56] and recorded subjective ratings of discomfort (with questionnaires) as a function of (1) the viewing distance and (2) the sign of the conflict (stereo objects nearer to versus farther from the display surface). They found effects of both factors. A given vergence–accommodation conflict resulted in overall slightly higher ratings of fatigue, discomfort, or both at far viewing distances than at near distances. The sign of conflict also had a small but significant effect that interacted with viewing distance. At near distances, fatigue and discomfort ratings to a given conflict magnitude were greater for objects nearer than the screen, and at far distances they were greater for objects farther than the screen. Interestingly, this asymmetry was related to the individual's phoria. Phoria is the vergence position adopted by the eyes when there is no stimulus to vergence but accommodation is stimulated. Thus, a person's phoria can be thought of as the extent to which that individual's accommodation and vergence responses are naturally decoupled at different accommodation distances.[62] Although there are significant individual differences in phoria, the typical pattern is to converge farther than the accommodation distance at near distances and nearer than the accommodation distance at far distances[65] (Fig. 12). Thus, we might expect, as Shibata et al.[64] found, that it is most demanding to converge nearer than the screen distance at near viewing distances and farther than the screen at far viewing distances.

Figure 12 plots the ZCSBV, and Percival's zone, as estimated from the literature by Shibata et al.[64] It also plots Shibata et al.'s estimate of the stereoscopic ZoC, based on their questionnaire data. This estimate is approximate because it is based on noisy questionnaire data and relatively few measurements (fatigue ratings to just one conflict magnitude for each distance and sign of conflict), but it nonetheless represents the current best guess of the shape of the ZoC.

Screen Distance and the ZoC—Figure 12 (a) plots the various zones in units of diopters—the reciprocal of distance in meters. Using diopters is appropriate, because the amount of blur in the retinal image is proportional to defocus in diopters, not physical distance. Changes in vergence angle have a similar relationship with physical distance. Thus, a given change in the dioptric distance to a stimulus requires approximately the same change in accommodation, vergence, or both independent of the overall distance to the stimulus. Perhaps unsurprisingly then, the width of the comfort zone is quite similar in diopters for screens positioned at different distances. This has important implications for the width of the ZoC in physical distance for different viewing situations. Figure 12 (b) plots Shibata et al.'s ZoC estimate as a function of physical distance.[64] In meters, the width of the comfort zone is small at near viewing distances. It is much larger at far viewing distances, but it is still possible to exceed the ZoC at TV and even cinema viewing distances (by presenting objects too near to the viewer). Thus, the often-made assumption that vergence–accommodation conflicts do not matter at far viewing distances is not true.

ZoC in Cinematography—The TV and movie industry is aware that large vergence–accommodation conflicts are problematic, but there is no commonly agreed rule to deal with them. Widespread practice appears to be to control the maximum amount of horizontal disparity as a proportion of screen width so that screen parallax (the horizontal separation between the left and the right eye's image points on the screen) is within 2% to 3% of the screen width for objects nearer than the screen and 1% to 2% of screen width for objects farther than the screen.[46,66]

This rule of thumb has practical value to filmmakers, because the range of on-screen parallax resulting from a given scene and camera configuration can be examined readily (i.e., on the film set) by overlaying each eye's image on a standard monitor. This rule is fundamentally incorrect, however, because it does not take into account the size of the screen that the content will be displayed on or the viewing distance. The disparities at the viewer's eyes (and therefore the vergence–accommodation conflict) depend on the differences in angular direction of image points at the two eyes, so they vary considerably if the same on-screen parallax—specified in pixels, or as a proportion of screen size—is viewed on a small versus large screen or at a near versus far viewing distance. In practice, the consequences of using this incorrect rule may not be catastrophic, because we tend to view large screens at farther distances than we view small screens (i.e., the screen size, measured in visual angle, does not vary dramatically).[67] But importantly, the asymmetry in tolerance to stereo depths nearer and farther than the screen[64] varies with viewing distance. Clearly, this has implications for how content should be optimized for different viewing situations, including scaling movies down to TV format; even if the screen has constant angular size, different on-screen parallax limits may be needed for near (computer or TV) and far (cinema) viewing.

What We Need to Know to Specify General Guidelines

Existing studies demonstrate that the underlying concept of a ZoC for accommodation and vergence responses is valid and useful. They fall short, however, of the specific knowledge required for comprehensive guidelines on producing S3D content.

Factors predicting an individual's susceptibility to aversive symptoms remain largely unknown. Large individual differences in a range of ophthalmological variables could conceivably affect a person's susceptibility to discomfort from vergence–accommodation conflicts. For instance, people's ability to decouple accommodation and vergence responses differs significantly, as do their phorias and their ability to accommodate to different distances.[62] Large-scale population studies are required to establish the relationships between these ophthalmological variables and aversive symptoms during stereo viewing. If there are indeed large differences in individuals' ZoCs, the placement of content in stereo depth may need to be conservative to remain acceptable to the majority of people.

The viewer's age is likely to be a particularly important factor. There is a belief in the stereo industry that older viewers are more affected by vergence–accommodation conflicts than younger viewers. For instance, "oculo-motor exercising [decoupling accommodation and vergence] can be painful and can increase in difficulty with age. Kids would just not care, when elderly persons may be unable to practice it."[46] However, the opposite is probably true. The ability to vary accommodation state decreases significantly with age, so under natural viewing, older adults experience vergence–accommodation conflicts most of the time (because they cannot vary their accommodation response with vergence when looking nearer and farther). Indeed, their oculomotor responses more closely resemble those required for viewing stereo media: changing vergence while accommodating to a fixed distance. Consistent with this, Yang et al.[68] recently found that people age 24 to 34 years reported more discomfort than people age 45 and over when viewing the same S3D content.

It also remains to be determined whether there are any short- or long-term effects of prolonged, repeated exposure to the unnatural stimulus presented by stereoscopic displays. In adults, accommodation–vergence coupling is quite adaptable,[69] so there is a possibility that accommodation function may take some time to return to normal following prolonged viewing of S3D media. Moreover, as the S3D industry continues to develop, our use of stereo media will change from an occasional activity to an everyday one. The introduction of stereo computer games, in particular, exposes viewers to vergence–accommodation conflicts regularly for potentially long periods. We may need to be particularly cautious about long-term effects of vergence–accommodation conflicts on younger children, because their visual systems are still developing.[70] We know of no specific causes for concern at this time, but the research required to identify relevant issues has not yet been done. It is reasonable to assume that vergence–accommodation coupling exists because it is beneficial, so we should be cautious when systematically disrupting its natural operation. The ZoC could be measured in children in the same way it has been measured in adults. Clearly, however, it would not be acceptable to carry out the long-term experimental studies that would be required to understand any potential long-term effects (although, ironically, young people may expose themselves to such a regime voluntarily). Thus, clinical, research, and industry communities should remain alert to the development of unwanted symptoms in users of stereo media.

TEMPORAL PRESENTATION PROTOCOLS: FLICKER, MOTION ARTIFACTS, AND DEPTH DISTORTIONS

Temporal Protocols in Stereo Displays

It is clearly desirable to be able to present flicker-free image content without noticeable motion artifacts or distortions of perceived depth. Here, we investigate how the means of presenting stereo images over time affects the visibility of flicker, motion, and depth.

S3D displays generally use one conventional 2D display to present different images to the left and right eyes. Because S3D displays are so similar to conventional nonstereo displays, many of the standards, protocols, technical analyses, and artistic effects that have been developed for nonstereo displays also apply to S3D. However, important differences between nonstereo and stereo displays can produce artifacts unique to stereo presentation.

There are a variety of ways to present different images to the two eyes. The field-sequential approach presents images to the left and right eyes in temporal alternation (e.g., RealD and Dolby). Among field-sequential approaches, there are several ways to present the alternating images in time, including multiple-flash methods. In addition to field-sequential approaches, one can present images to the two eyes simultaneously by using multiple projectors (IMAX), wavelength-multiplexing techniques (Infitec and anaglyph), or spatial multiplexing (micropol) on one 2D display. Figure 13 schematizes some protocols. Column 6 shows the RealD and Dolby approach. The IMAX approach is similar to column 1.

Spatiotemporal Frequencies—To examine how various temporal presentation methods affect the viewer's perceptual experience with stereo displays, it is useful to examine the temporal and spatial frequencies created by these methods. We begin by considering stroboscopic presentation of a moving object presented to one eye.[71] To create the appearance of a high-contrast vertical line moving smoothly at speed s, we present a sequence of brief snapshots of the line at time intervals Δt, with each view displaced by $\Delta x = s\Delta t$. The temporal presentation rate t_p is the reciprocal of the time between presentations: $t_p = 1/\Delta t$.

Figure 14 (a) depicts a real stimulus moving at speed s and the stroboscopic version of that stimulus. Spatial position is plotted as a function of time. By using the Fourier transform, we determine the temporal and spatial frequencies in these two stimuli. These are depicted in Fig. 14 (b). Spatial frequency is plotted as a function of temporal frequency. The Fourier transform of the real moving stimulus is the black line; it has a slope of $-1/s$. The transform of the stroboscopic stimulus is represented by the black and green lines. The black line is the same line as for the real stimulus. The green lines are *aliases*: artifacts created by the stroboscopic presentation. Their slopes are $-1/s$, and they are separated horizontally by t_p. Thus, the spatiotemporal frequencies of the stroboscopic stimulus contain a signal component (the black line) plus a series of aliases (the green ones). As the speed of the stimulus s increases, the slope of the signal and aliases decreases. As the presentation rate t_p increases, the separation between the aliases increases. When the aliases are visible to a viewer, the percept contains flicker, motion artifacts, or both. When the aliases are not visible, the percept is nonflickering and smooth.

To assess the visibility of the aliases, we consider what human viewers can and cannot see. The system's sensitivity to different temporal and spatial frequencies is described by the spatiotemporal contrast sensitivity function (CSF). Figure 15 plots the CSF for a typical viewer under room-light conditions. This function has been called the *window of visibility* because it characterizes the spatiotemporal stimuli that can be seen, as opposed to the ones that cannot be seen. The CSF is represented in Fig. 14 by the white ellipse. The dimensions of the ellipse's principal axes are the highest visible temporal frequency *cff* and the highest visible spatial frequency. Aliases falling within the ellipse are generally visible, and those falling outside are not. We can now see how a stroboscopic stimulus could appear identical to a smoothly moving real stimulus. If the two are moving at the same speed and the stroboscopic presentation t_p is fast enough, the aliases would fall outside the window of visibility, and the stroboscopic and real stimuli could not be discriminated. Similarly, the stroboscopic and real stimuli could be discriminated when the aliases fall within the window of visibility: the stroboscopic stimulus would exhibit flicker, motion artifacts, or both.

We discussed stroboscopic presentation in Fig. 14 because such presentation determines the spatiotemporal frequencies of the aliases and those frequencies remain for other protocols that are actually used in S3D displays. Other presentation methods (sample and hold, multiflash, etc.) do not change the pattern of aliases; they only change their amplitudes.

Flicker Visibility

We now consider when flicker is visible in an S3D display. We define visible flicker as perceived fluctuations in the brightness of the stimulus. We assume that flicker is perceived when aliases such as those in Fig. 14 (b) encroach the window of visibility near a spatial frequency of zero (i.e., along the temporal-frequency axis).

In field-sequential stereo displays (e.g., active shutter glasses or passive glasses with active switching in front of the projector), the monocular images consist of presentation intervals alternating with dark intervals. In some cases, each presented image of the moving stimulus is a new one. We refer to this as a *single-flash protocol*; it is schematized in Fig. 16 (a). There is also a *double-flash protocol*, in which the images are presented twice before updating, and a triple-flash protocol, in which they are presented three times before updating. We use f to represent the number of such flashes in a protocol. Those protocols are also schematized in the left column of Fig. 16 and in Fig. 13. Multiflashing is similar to the double and triple shuttering that is done with film-based movie projectors. The double- and triple-flash protocols are used to reduce the visibility of flicker (RealD and Dolby use triple flash, and IMAX uses field-simultaneous double flash). We refer to the rate at which new images are presented as the *capture rate* t_c (or $1/t_c$, where t_c is the time between image updating). We refer to the rate at which images, updated or not, are delivered to an eye as the presentation rate t_p (or $1/t_p$). Thus, $t_c = t_p/f$ (or $t_c = ft_p$).

The Fourier transform for the single-flash, field-sequential protocol is shown in Fig. 16 (b). With the insertion of dark frames, the amplitude of the aliases at a temporal frequency of t_c is rather high. As a result, flicker should be quite visible, whether the stimulus is moving or not, unless a high presentation rate is used. The transforms for the double- and triple-flash protocols are also shown in Fig. 16 (b). The frequencies of the aliases are the same in the single- and double-flash protocols, but their amplitudes go to zero at t_c in double flash and at $2t_c$ in single flash. In the tripleflash protocol, the aliases are again the same, but their amplitudes go to zero at t_c and again at $2t_c$, remaining small in-between. The first alias with nonzero amplitude along the temporal-frequency axis occurs at a temporal frequency of t_p ($1/t_p$), which is the presentation rate. Thus, we predict that presentation rate, not capture rate, determines flicker visibility. This prediction was confirmed in a perceptual experiment.[73] Because presentation rate should be the primary determinant, we should be able to reduce flicker visibility for a fixed capture rate by using multiflash protocols. Specifically, flicker should be less visible in the triple-flash than in the double-flash protocol and less visible in the double-flash than in the single-flash protocol. This prediction has been shown to be correct.[73]

Stereo processing in the visual system is sluggish[74]; therefore, the visual system is less sensitive to rapidly changing disparities than to time-varying luminance signals. From this observation, we predict little if any difference in flicker visibility between stereo and nonstereo presentations, provided that the temporal protocols are the same. This prediction is basically correct.[73]

Motion Artifacts

We now turn to the visibility of motion artifacts. These artifacts include judder (jerky or unsmooth motion appearance), edge banding (more than one edge seen at the edge of a moving stimulus), and motion blur (perceived blur at a moving edge). The analysis of motion artifacts is somewhat more complicated than the one for flicker because with a given capture rate, multiflash protocols do not change the spatiotemporal frequency of the aliases. Instead, they differentially attenuate the amplitudes of the aliases at certain temporal frequencies. Thus, the visibility of motion artifacts is determined by the spatiotemporal frequencies and amplitudes of the aliases.

Viewers typically track a moving stimulus with smooth-pursuit eye movements that keep the stimulus on the fovea, and this affects what motion artifacts look like. With smooth pursuit, the image of a smoothly moving stimulus becomes fixed on the retina; that is, for a real object moving smoothly at speed s relative to the observer, and an eye tracking at the same speed, the retinal speed of the stimulus is zero. With a digitally displayed stimulus moving at the same speed, the only temporally varying signal on the retina is created by the difference between smoothly moving and discretely moving images. Each image presentation of duration t_p displaces across the retina by $\Delta x = -st_p$. Thus, significant displacement can occur with high stimulus speeds and low frame rates, thereby blurring the stimulus on the retina ("motion blur").[75,76]

From the analysis of temporal and spatial frequencies, we can make a number of predictions about the visibility of motion artifacts. First, the visibility of motion artifacts should increase with increasing stimulus speed and decrease with increasing capture rate. More specifically, combinations of speed and capture rate that yield a constant ratio (s/t_c) should have approximately equivalent motion artifacts. Hoffman et al.[73] tested this prediction and found that it is essentially correct. Second, although speed and capture rate should be the primary determinants of motion artifacts, multiflash protocols for a fixed capture rate should produce more visible motion artifacts. This too has been tested empirically and found to be correct. [73] Finally, edge banding should be determined by the number of flashes in multiflash protocols: two bands being perceived with double flash, three with triple flash, and so on. This prediction has been borne out empirically.[73,76,77]

Distortions of Perceived Depth

A temporal delay to one eye's input can cause a moving object to appear displaced in depth.[78,79] Many protocols in Fig. 13 introduce such a delay to one eye. If the delay alters the visual system's estimate of the disparity over time, this would in turn produce distortion in the depth percept. Consider, for example, the C_{sim}/P_{alt-1X} protocol (fourth column in Fig. 13). The solid horizontal line in the lower panel of the figure represents the correct disparity over time; that is, the disparity that would occur with the presentation of a moving real object in the plane of the screen. To compute disparity, the visual system must match images in one eye with images in the other. But the images in this protocol are presented to the two eyes at different times, so nonsimultaneous images must be matched. If each image in one eye is matched with the succeeding image in the other eye, the estimated disparities would be the green dots in the lower panel of the figure. For every two successive matches (three images), one disparity estimate is equal to the correct value and one is greater. As a result, the time-average disparity is biased relative to the correct value, and this should cause a change in perceived depth: a perceptual distortion. The difference between the time-average disparity and the correct disparity depends on the protocol: largest with single flash (Fig. 13, column 4) and smallest with triple flash (column 6). For this reason, the largest distortions should occur with single-flash protocols and the smallest with triple-flash protocols. The magnitude of the distortions should also depend on speed, because the difference between the time-average disparity and the correct disparity is proportional to speed. We refer to the distortions predicted from the average disparity over time as the *time-average model*.

The most frequently used protocols employ simultaneous capture and alternating presentation in which one eye's image is delayed. Figure 17 plots the predictions and data from such protocols with different capture rates and numbers of flashes. The predictions from the time-average model are the dashed lines. Experimental data in which the magnitude of the depth distortion was measured are represented by the colored symbols. As expected, the size of the distortion increases as stimulus speed increases. With 75 Hz capture, the distortion increases up to the fastest speed tested. With 25 Hz capture, the distortion levels off around 3°/sec and then decreases at yet higher speeds. We conclude that perceived depth distortions occur, as predicted by the time-average model, when capture and presentation synchrony are not matched. The model's predictions are accurate at slow speeds, but smaller distortions than those predicted are observed at fast speeds. The prediction failure at fast speeds is the consequence of a temporal disparity-gradient limit.[73]

Thus, distortions of perceived depth occur with moving objects in some stereo presentation protocols because they delay the input to one eye relative to the other eye. As a consequence, objects moving in one direction can be perceived as closer and objects moving in the opposite direction can be perceived as farther than they are meant to be. Such distortions can be readily observed in stereo TV and cinema. For example, in S3D broadcasts of World Cup soccer in 2010, a ball kicked along the ground appeared to recede in depth when moving in one direction (paradoxically seeming to go beneath the playing field) and appeared to come closer in depth when moving in the opposite direction. This speed-dependent effect can be quite disturbing, so it is clearly useful to understand its cause and how to potentially minimize or eliminate it.

Summary of Temporal Protocols

The work described here adds to the theoretical and empirical foundation for determining what display parameters are likely to yield noticeable flicker, motion artifacts, and depth distortions. From this foundation, we can make effective decisions about how to minimize or even eliminate these undesirable effects.

CONCLUSION

Stereoscopic displays are being used ever more frequently, particularly for entertainment. In the various applications, the S3D imagery should create a faithful impression of the structure of the scene being portrayed. Temporal artifacts like flicker and motion judder should be minimal. Moreover, the viewer should be comfortable and not experience eye fatigue or headache. This paper reviewed current research on stereo human vision and how it informs us about how best to create and present S3D imagery. Several issues were discussed including when and why it is important to get the projection geometry correct, how interactions with other depth cues can enhance or degrade the viewer’s experience, why eye fatigue and discomfort occurs with S3D imagery and how to minimize such adverse effects. Why presentation protocols affect the visibility of flicker, motion artifacts, and depth distortions and how to minimize those problems was also discussed. We hope this will be useful to practitioners in creating the best experiences for viewers.

References

1. Allison RS. Analysis of the Influence of Vertical Disparities Arising in Toed-In Stereoscopic Cameras. J. Imag. Sci. Tech. 2007; 51:317–327.
2. Allison RS, Howard IP, Fang X. Depth Selectivity of Vertical Fusional Mechanisms. Vision Res. 2000; 40:2985–2998. [PubMed: 11000396]
3. Howard IP, Allison RS, Zacher JE. The Dynamics of Vertical Vergence. Exp. Brain Res. 1997; 116:153–159. [PubMed: 9305824]
4. Howard IP, Fang X, Allison RS, Zacher JE. Effects of Stimulus Size and Eccentricity on Horizontal and Vertical Vergence. Exp. Brain Res. 2000; 130:124–132. [PubMed: 10672465]
5. Rogers BJ, Bradshaw MF. Disparity Minimisation, Cyclovergence, and the Validity of Nonius Lines as a Technique for Measuring Torsional Alignment. Perception. 1999; 28:127–141. [PubMed: 10615455]
6. Longuet-Higgins HC. A Computer Algorithm for Reconstructing a Scene from Two Projections. Nature. 1981; 293:133–135.
7. Gårding J, Porrill J, Mayhew JE, Frisby JP. Stereopsis, Vertical Disparity and Relief Transformations. Vision Res. 1995; 35:703–722. [PubMed: 7900308]
8. Read JCA, Phillipson GP, Glennerster A. Latitude and Longitude Vertical Disparities. J. Vision. 2009; 9(13):11–37.
9. Rogers BJ, Bradshaw MF. Vertical Disparities, Differential Perspective and Binocular Stereopsis. Nature. 1993; 361:253–255. [PubMed: 8423851]
10. Ogle KN. Induced Size Effect I: A New Phenomenon in Binocular Vision Associated with the Relative Size of the Images in the Two Eyes. Archives of Ophthalmology. 1938; 20:604.
11. Backus BT, Banks MS, van Ee R, Crowell JA. Horizontal and Vertical Disparity, Eye Position, and Stereoscopic Slant Perception. Vision Res. 1999; 39:1143–1170. [PubMed: 10343832]
12. Banks MS, Backus BT. Extra-retinal and Perspective Cues Cause the Small Range of the Induced Effect. Vision Res. 1998; 38:187–194. [PubMed: 9536348]
13. Bradshaw MF, Glennerster A, Rogers BJ. The Effect of Display Size on Disparity Scaling from Differential Perspective and Vergence Cues. Vision Res. 1996; 36:1255–1264. [PubMed: 8711905]
14. Backus BT, Banks MS. Estimator Reliability and Distance Scaling in stereoscopic Slant Perception. Perception. 1999; 28:217–242. [PubMed: 10615462]

15. Glennerster A, Tcheang L, Gilson SJ, Fitzgibbon AW, Parker AJ. Humans Ignore Motion and Stereo Cues in Favor of a Fictional Stable World. Curr. Biol. 2006; 16:428–432. [PubMed: 16488879]
16. Geller T. Overcoming the Uncanny Valley. IEEE Comput. Graph. Appl. 2008; 28:11–17. [PubMed: 18663813]
17. Vishwanath D, Girshick AR, Banks MS. Why Pictures Look Right When Viewed from the Wrong Place. Nat. Neurosci. 2005; 8:1401–1410. [PubMed: 16172600]
18. Julesz B. Binocular Depth Perception of Computer Generated Patterns. Bell Syst. Tech. J. 1960; 39:1125–1162.
19. Helmholtz, H. Physiological Optics. English Translation 1962 by J. P. C. Southall from the 3rd German Edition of Handbuch Der Physiologischen Optik. Vos. New York: Hamburg. Dover; 1909.
20. Burge J, Fowlkes CC, Banks MS. Natural-Scene Statistics Predict How the Figure–Ground Cue of Convexity Affects Human Depth Perception. J. Neurosci. 2010; 30:7269–7280. [PubMed: 20505093]
21. Mather G, Smith DR. Depth Cue Integration: Stereopsis and Image Blur. Vision Res. 2000; 40:3501–3506. [PubMed: 11115677]
22. Held RT, Cooper EA, O'Brien JF, Banks MS. Using Blur to Affect Perceived Distance and Size. ACM Trans. On Graphics. 2010; 29:1–16.
23. Howard, IP.; Rogers, BJ. Depth Perception. Vol. Vol. 2. Toronto: Porteous; 2002.
24. Bülthoff HH, Mallot HA. Integration of Depth Modules: Stereo and Shading. J. Opt. Soc. Am. A. 1988; 5:1749–1758. [PubMed: 3204438]
25. Clark, JJ.; Yuille, AL. Data Fusion for Sensory Information Processing Systems. Boston: Cluwer; 1990.
26. Landy MS, Maloney LT, Johnston EB, Young M. Measurement and Modeling of Depth Cue Combination: In Defense of Weak Fusion. Vision Res. 1995; 35:389–412. [PubMed: 7892735]
27. Fisher RA. Theory of Statistical Estimation. Math. Proc. Camb. Phil. Soc. 1925; 22:700–725.
28. Knill DC. Learning Bayesian Priors for Depth Perception. J. Vision. 2007; 7:1.
29. Siegel M, Nagata S. Just Enough Reality: Comfortable 3-D Viewing via Microstereopsis. IEEE Trans. Circ. Syst. Video. Tech. 2000; 10:387–396.
30. Hillis JM, Ernst MO, Banks MS, Landy MS. Combining Sensory Information: Mandatory Fusion Within, but Not Between, Senses. Science. 2002; 298:1627–1630. [PubMed: 12446912]
31. Knill DC. Robust Cue Integration: A Bayesian Model and Evidence from Cue-Conflict Studies with Stereoscopic and Figure Cues to Slant. J. Vision. 2007; 7:1–24.
32. Meese TS, Holmes DJ. Performance Data Indicate Summation for Pictorial Depth-Cues in Slanted Surfaces. Spat. Vis. 2004; 17:127–151. [PubMed: 15078016]
33. Muller CM, Brenner E, Smeets JB. Testing a Counter-intuitive Prediction of Optimal Cue Combination. Vision Res. 2009; 49:134–139. [PubMed: 18983869]
34. Allison RS, Howard IP, Rogers BJ, Bridge H. Temporal Aspects of Slant and Inclination Perception. Perception. 1998; 27:1287–1304. [PubMed: 10505175]
35. Girshick AR, Banks MS. Probabilistic Combination of Slant Information: Weighted Averaging and Robustness as Optimal Percepts. J. Vision. 2009; 9:1–20.
36. van Ee R, Adams WJ, Mamassian P. Bayesian Modeling of Cue Interaction: Bistability in Stereoscopic Slant Perception. J. Opt. Soc. Am. A Opt. Image Sci. Vis. 2003; 20:1398–1406. [PubMed: 12868644]
37. Hillis JM, Watt S, Landy MS, Banks MS. Slant from Texture and Disparity Cues: Optimal Cue Combination. J. Vision. 2004; 4:967–992.
38. Allison RS, Howard IP. Temporal Dependencies in Resolving Monocular and Binocular Cue Conflict in Slant Perception. Vision Res. 2000; 40:1869–1885. [PubMed: 10837832]
39. Woods AJ, Docherty T, Koch R. Merritt JO, Fisher SS. Image Distortions in Stereoscopic Video Systems. Proc. of SPIE, Stereoscopic Displays and Applications IV. 1993
40. Held RT, Banks MS. Misperceptions in Stereoscopic Displays: A Vision Science Perspective. Proc. Of APGV08. 2008; 5:23–31.

41. Tsirlin I, Allison RS, Wilcox LM. Stereoscopic Transparency: Constraints on the Perception of Multiple Surfaces. J. Vision. 2008; 8:1–10.

42. Sakano Y, Ando H. Effects of Head Motion and Stereo Viewing on Perceived Glossiness. J. Vision. 2010; 10:1–14.

43. Banks MS, Held RT, Girshick AR. Perception of 3-D Layout in Stereo Displays. Information Display. 2009; 25:12–16. [PubMed: 21687822]

44. Berezin O. Digital Cinema in Russia: Status of 2D/3D DC Rollout. Is 3D Still a Driver for the Development of the Cinema Market? 3Dmedia. 2010 http://www.kinopoisk.ru.

45. Lambooij M, Ijsselsteijn W, Fortuin M, Heynderickx I. Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review. J. Imag. Sci. Tech. 2009; 53(3):1–14.

46. Mendiburu, B. 3D Movie Making: Stereoscopic Digital Cinema from Script to Screen. Oxford: Focal Press; 2009.

47. Akeley K, Watt SJ, Girshick AR, Banks MS. A Stereo Display Prototype with Multiple Focal Distances. ACM Trans. Graph. 2004; 23:804–813.

48. Emoto M, Niida T, Okano F. Repeated Vergence Adaptation Causes the Decline of Visual Functions in Watching Stereoscopic Television. J. Display Tech. 2005; 1:328–340.

49. Hoffman DL, Girshick AR, Akeley K, Banks MS. Vergence–Accommodation Conflicts Hinder Visual Performance and Cause Visual Fatigue. J. Vision. 2008; 8:1–30.

50. Ukai K. Visual Fatigue Caused by Viewing Stereoscopic Images and Mechanism of Accommodation. Proc. 1st Int. Symp. Univ. Comm. 2007:176–179.

51. Ukai K, Howarth PA. Visual Fatigue Caused by Viewing Stereoscopic Motion Images: Background, Theories and Observations. Displays. 2008; 29:106–116.

52. Wann JP, Mon-Williams M. Health Issues with Virtual Reality Displays: What We Do Know and What We Don’t. ACM SIGGRAPH Comput. Graph. 1997; 31:53–57.

53. Watt SJ, Akeley K, Ernst MO, Banks MS. Focus Cues Affect Perceived Depth. J. Vision. 2005; 5:834–862.

54. Yano S, Emoto M, Mitsuhashi T. Two Factors in Visual Fatigue Caused by Stereoscopic HDTV Images. Displays. 2004; 25:141–150.

55. Liu S, Cheng D, Hua H. An Optical See-Through Head Mounted Display with Addressable Focal Planes. IEEE Int. Symp. Mixed and Augmented Reality. 2008; 33:33–41.

56. Love GD, Hoffman DM, Hands PJW, Gao J, Kirby AK, Banks MS. High-Speed Switchable Lens Enables the Development of a Volumetric Stereoscopic Display. Optic. Express. 2009; 17:15715–15725.

57. MacKenzie KJ, Hoffman DM, Watt SJ. Accommodation to Multiple-Focal-Planes Displays: Implications for Improving Stereoscopic Displays, and for Accommodation Control. J. Vision. 2010; 10(8):22.

58. McDowall I, Bolas M. FakeSpace Labs Accommodation Display Research. 1994 unpublished report.

59. MacKenzie KJ, Dickson R, Watt SJ. Vergence and Accommodation to Multiple-Image-Plane Stereoscopic Displays: ‘Real World’ Responses with Practical Image-Plane Separations? SPIE Proc. XXII Stereoscopic Displays and Applications. 2011; 7863:1–11.

60. Campbell FW. The Depth of Field of the Human Eye. Opt. Acta. 1957; 4:157–164.

61. Charman WN, Whitefoot H. Pupil Diameter and the Depth of Field of the Human Eye as Measured by Laser Speckle. J. Mod. Optic. 1977; 24:1362–3044.

62. Scheiman, M.; Wick, B. Clinical Management of Binocular Vision: Heterophoric, Accommodative, and Eye Movement Disorders. Philadelphia: J. B. Lippincott; 1994.

63. Percival AS. The Relation of Convergence to Accommodation and Its Practical Bearing. Ophthalmol. Rev. 1892; 11:313–328.

64. Shibata T, Kim J, Hoffman DM, Banks MS. The Zone of Comfort: Predicting Visual Discomfort with Stereo Displays. J. Vision. 2011; 11:1–28.

65. Sheedy JE, Saladin JJ. Phoria, Vergence, and Fixation Disparity in Oculomotor Problems. Am. J. Optom. Physiol. Optic. 1977; 54:474–478.

66. Advanced Television Systems Committee (ATSC). ATSC Planning Team 1 Interim Report. 2011 http://www.atsc.org/PT1/PT-1-Interim-Report.pdf.

67. Ardito M. Studies of the Influence of Display Size and Picture Brightness on the Preferred Viewing Distance for HDTV Programs. SMPTE J. 1994; 103:517–522.

68. Yang S-N, Schlieski T, Selmins B, Cooper S, Doherty RA, Corriveau PJ, Sheedy JE. Individual Differences and Seating Position Affect Immersion and Symptom in Stereoscopic 3D Viewing. Technical Report, Vision Performance Institute, Pacific University. 2011

69. Schor CM, Kotulak JC. Dynamic Interactions Between Accommodation and Convergence Are Velocity Sensitive. Vision Res. 1986; 26:927–942. [PubMed: 3750876]

70. Rushton SK, Riddell PM. Developing Visual Systems and Exposure to Virtual Reality and Stereo Displays: Some Concerns and Speculations About the Demands on Accommodation and Vergence. Appl. Ergon. 1999; 30:69–78. [PubMed: 10098818]

71. Watson AB, Ahumada AJ, Farrell JE. Window of Visibility: Psychophysical Theory of Fidelity in Time-Sampled Visual Motion Displays. J. Opt. Soc. Am. 1986; 3:300–307.

72. Kelly DH. Motion and Vision. II. Stabilized Spatio-temporal Threshold Surface. J. Opt. Soc. Am. 1979; 69:1340–1349. [PubMed: 521853]

73. Hoffman DM, Karasev VI, Banks MS. Temporal Presentation Protocols: Flicker Visibility, Perceived Motion, and Perceived Depth. J. SID. 2011; 19:255–281.

74. Tyler CW. Depth Perception in Disparity Gratings. Nature. 1974; 251:140–142. [PubMed: 4420707]

75. Klompenhouwer MA, Velthoven LJ. Motion Blur Reduction for Liquid Crystal Displays: Motion-Compensated Inverse Filtering. Proc. SPIE. 2004; 5308:690.

76. Feng XF. LCD Motion-Blur Analysis, Perception, and Reduction Using Synchronized Backlight Flashing. Proc. SPIE. 2006; 6057:1–14.

77. Klompenhouwer, MA. Ph.D. thesis. Eindhoven Univ. Press; 2006. Flat Panel Display Signal Processing: Analysis and Algorithms for Improved Static and Dynamic Resolution.

78. Morgan MJ. Perception of Continuity in Stroboscopic Motion: A Temporal Frequency Analysis. Vision Res. 1979; 19:523–532. [PubMed: 483580]

79. Burr DC, Ross J. How Does Binocular Delay Give Information About Depth? Vision Res. 1979; 19:523–532. [PubMed: 483580]

Biographies

Martin Banks received a bachelor’s degree at Occidental College in 1970, majoring in psychology and minoring in physics. After a year in Germany, he entered graduate school, receiving a master’s in psychology at University of California, San Diego in 1973 and Ph.D. in developmental psychology from the University of Minnesota in 1976. He was assistant and associate professor of psychology at the University of Texas at Austin from 1976–1985. Banks moved to UC Berkeley School of Optometry in 1985 where he is professor of optometry and vision science. He was chair of Vision Science from 1995–2002. Banks has received several awards for basic and applied research on topics including human visual development, visual space perception, the development and evaluation of stereoscopic displays, and inter-sensory perception.

Jenny Read studied physics at Oxford University, graduating in 1994, and stayed on to do a doctorate in theoretical astrophysics. She began her neuroscience career in 1997 with a Wellcome Trust Training Fellowship in mathematical biology from Oxford's Laboratory of Physiology, receiving an M.Sc. in neuroscience in 1999. In 2001, she moved to the U.S. to work at the National Eye Institute at the National Institutes of Health in Maryland. Since 2005, Read has been at Newcastle University in the northeast of England on a Royal Society University Research Fellowship. Her research includes computational models of disparity encoding and psychophysical measurement of depth perception.

Robert Allison is associate professor of computer science and engineering at York University in Toronto (Canada). He received a BASc in computer engineering from the University of Waterloo in 1991. After graduation, he worked as an electrical engineer at Atlantis Aerospace in Brampton, Canada, where he designed electronics for flight training devices. Following this, he completed a MASc in electrical engineering (biomedical engineering) from the University of Toronto before obtaining a Ph.D. specializing in stereoscopic vision from York University in 1998. He was on the experimental team for the 1998 Neurolab space shuttle mission and did post-doctoral research at York University and the University of Oxford. Allison works on perception of space and self-motion in virtual environments, the measurement and analysis of eye-movements, and stereoscopic vision.

Simon Watt received a bachelor's degree in psychology from Cardiff University, Wales, in 1994, and a masters in psychology from Surrey University, England, in 1996. He completed a Ph.D. in human vision and visuo-motor control at Surrey in 2001. He was then a postdoctoral researcher at the School of Optometry, UC Berkeley, before moving to a faculty position at the School of Psychology, Bangor University, Wales, in 2004, where he is now a senior lecturer. He does basic and applied research on several topics related to 3D perception and binocular vision. In particular, he works on perceptual and ergonomic problems with current 3D display technologies. He also carries out research on depth perception, how 3D vision is used to control everyday movements such as reaching to grasp, and how the brain combines 3D information across sensory modalities such as vision and touch.

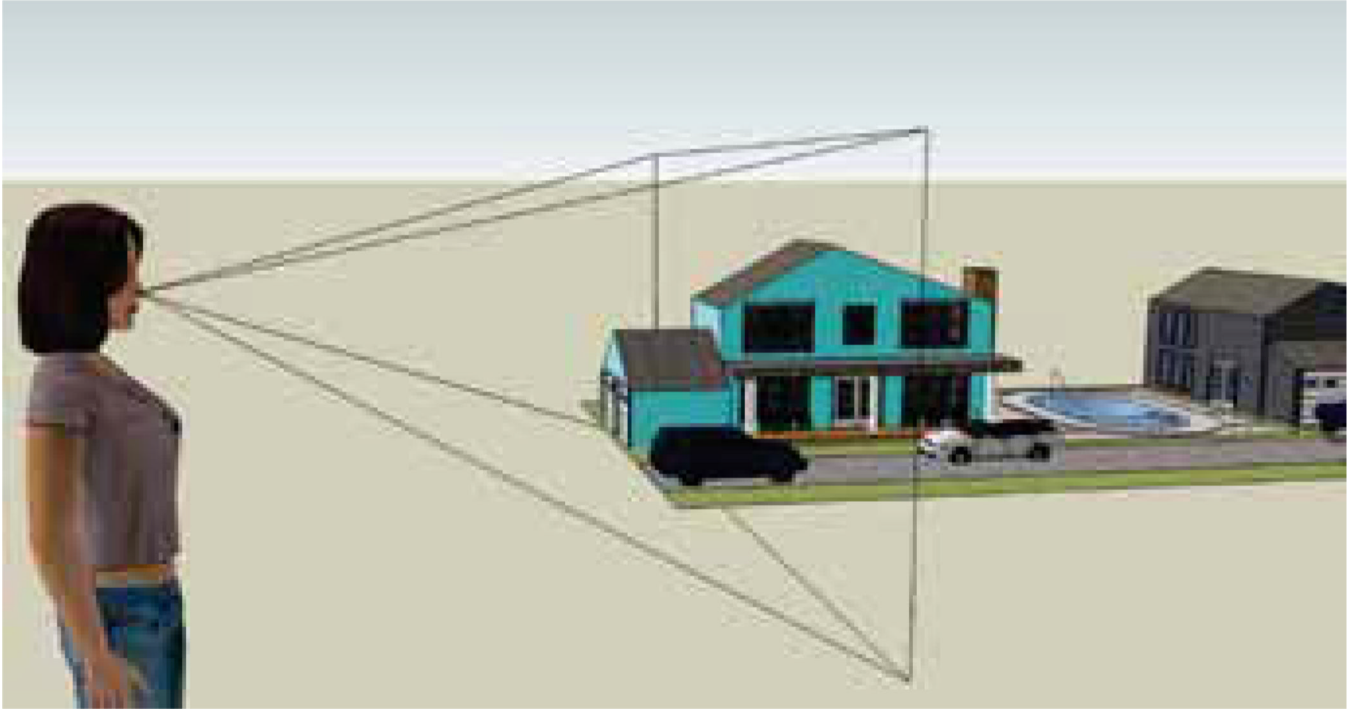

Figure 1.
A visual scene as a miniature model in front of the viewer.

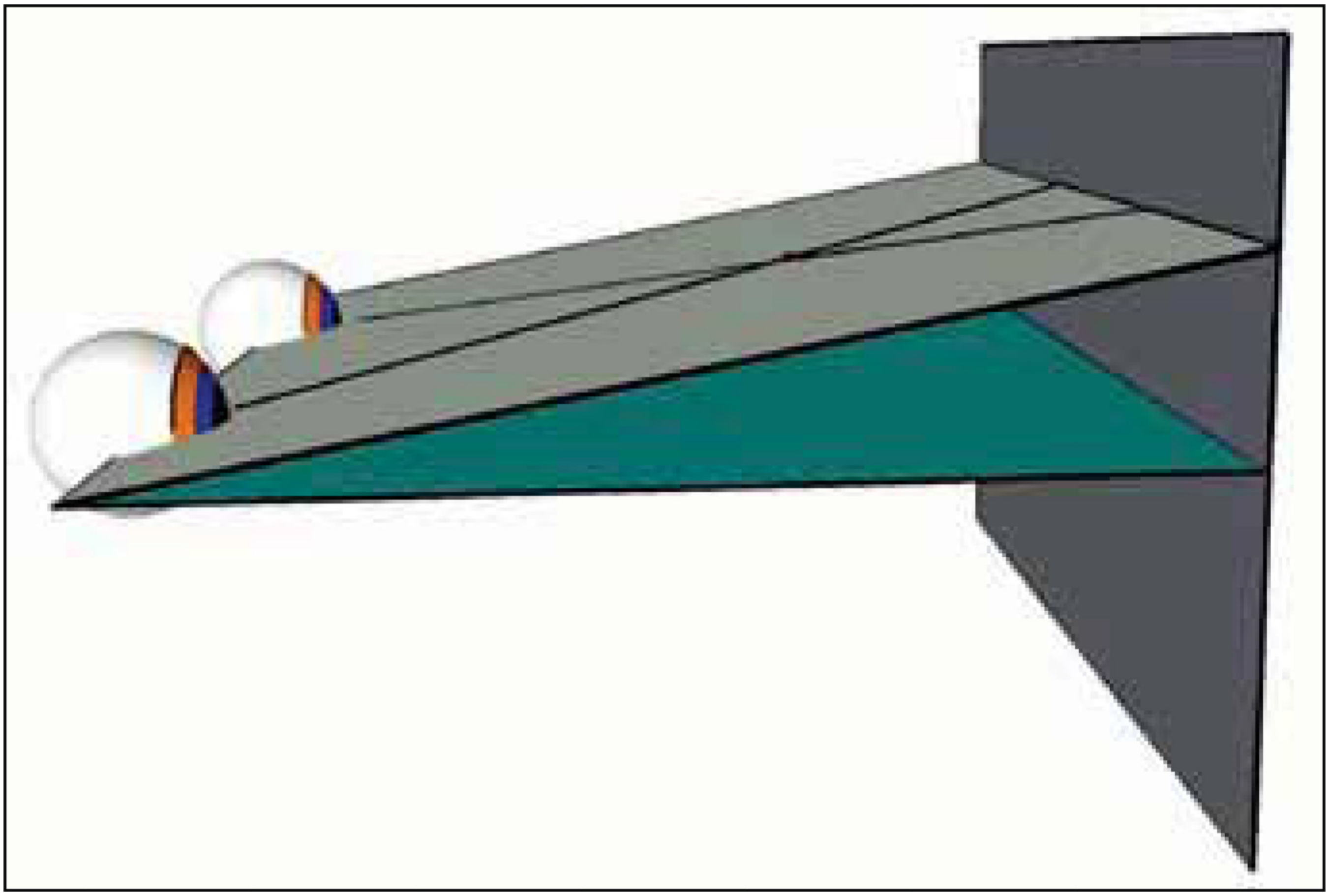

Figure 2.

An object in space and the centers of projection of the eyes define an epipolar plane. If the screen displaying the S3D content contains a vector parallel to the interocular axis, then the intersection of this plane with the screen is also parallel to the interocular axis. In the usual case, where the interocular axis is horizontal, this means that to reproduce the disparity of the real object, its two images must have zero vertical parallax. Their horizontal parallax depends on how far the simulated object is in front of or behind the screen.

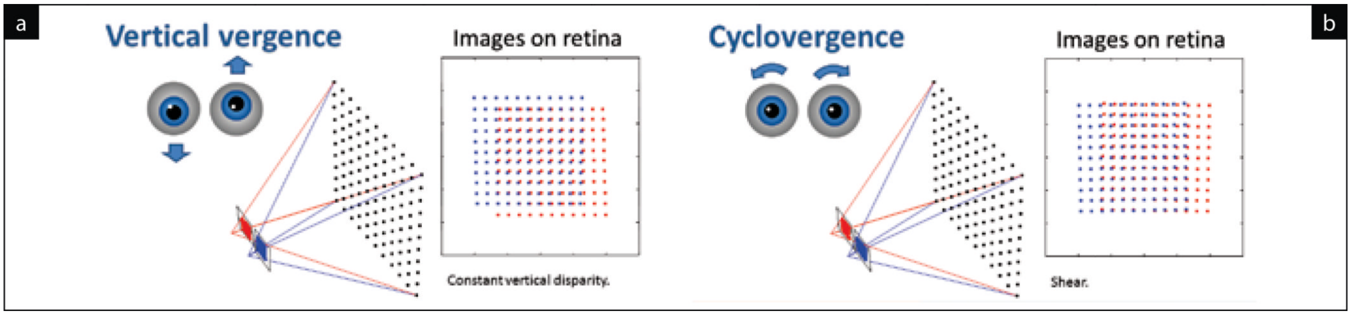

Figure 3.
Different eye postures cause characteristic patterns of vertical disparity on the retina, largely independent of the scene viewed. Here, the eyes view an array of points on a grid in space, directly in front of the viewer. The eyes are not converged, so the points have a large horizontal disparity on the retina. (a) The eyes have a vertical vergence misalignment. This introduces a constant vertical disparity across the retina. (b) The eyes are slightly cyclodiverged (rotated in opposite directions about the lines of sight). This introduces a shearlike pattern of vertical disparity across the retina.

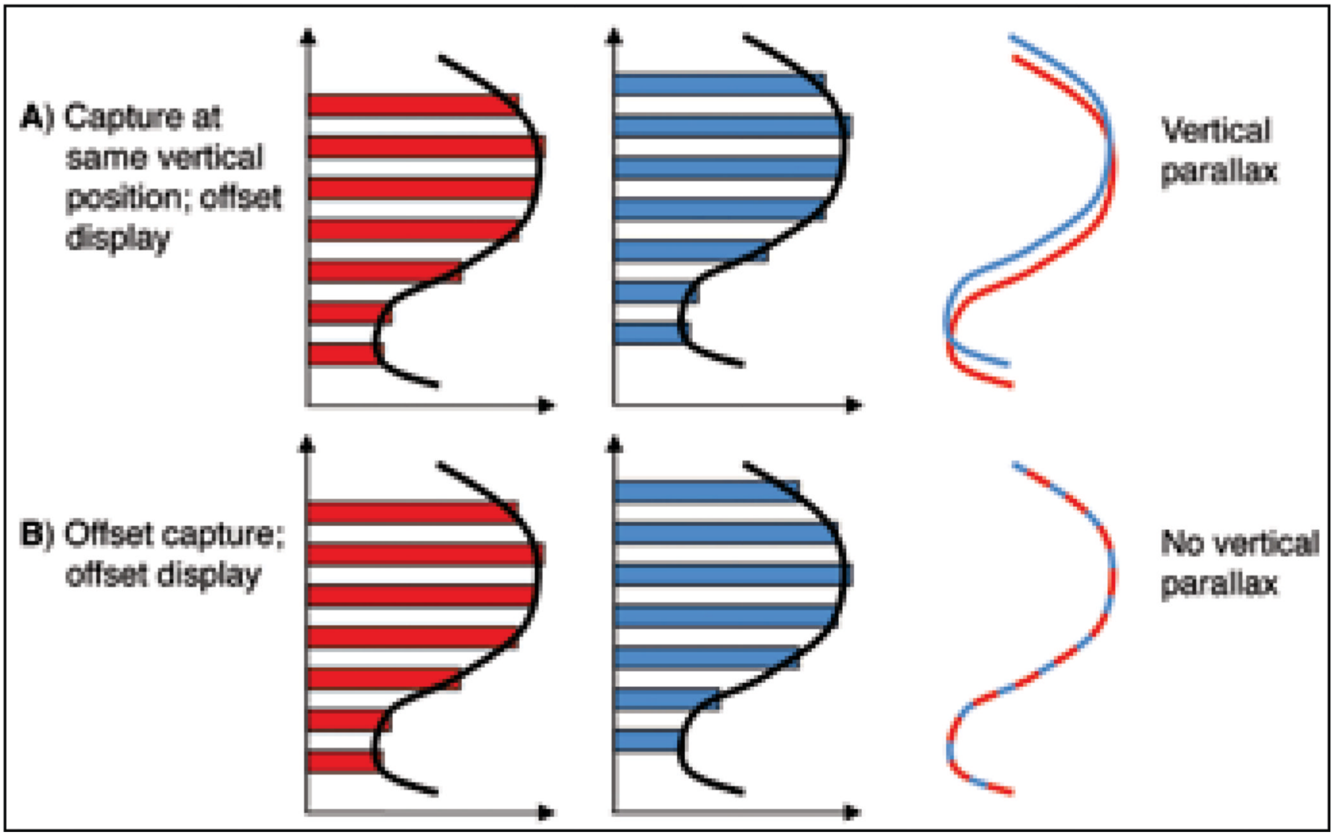
A) Capture at same vertical position; offset display
Vertical parallax
B) Offset capture; offset display
No vertical parallax

Figure 4.
Passive stereo in which left and right images that are displayed on different pixel rows (a) can introduce vertical parallax but (b) need not do so if created appropriately.

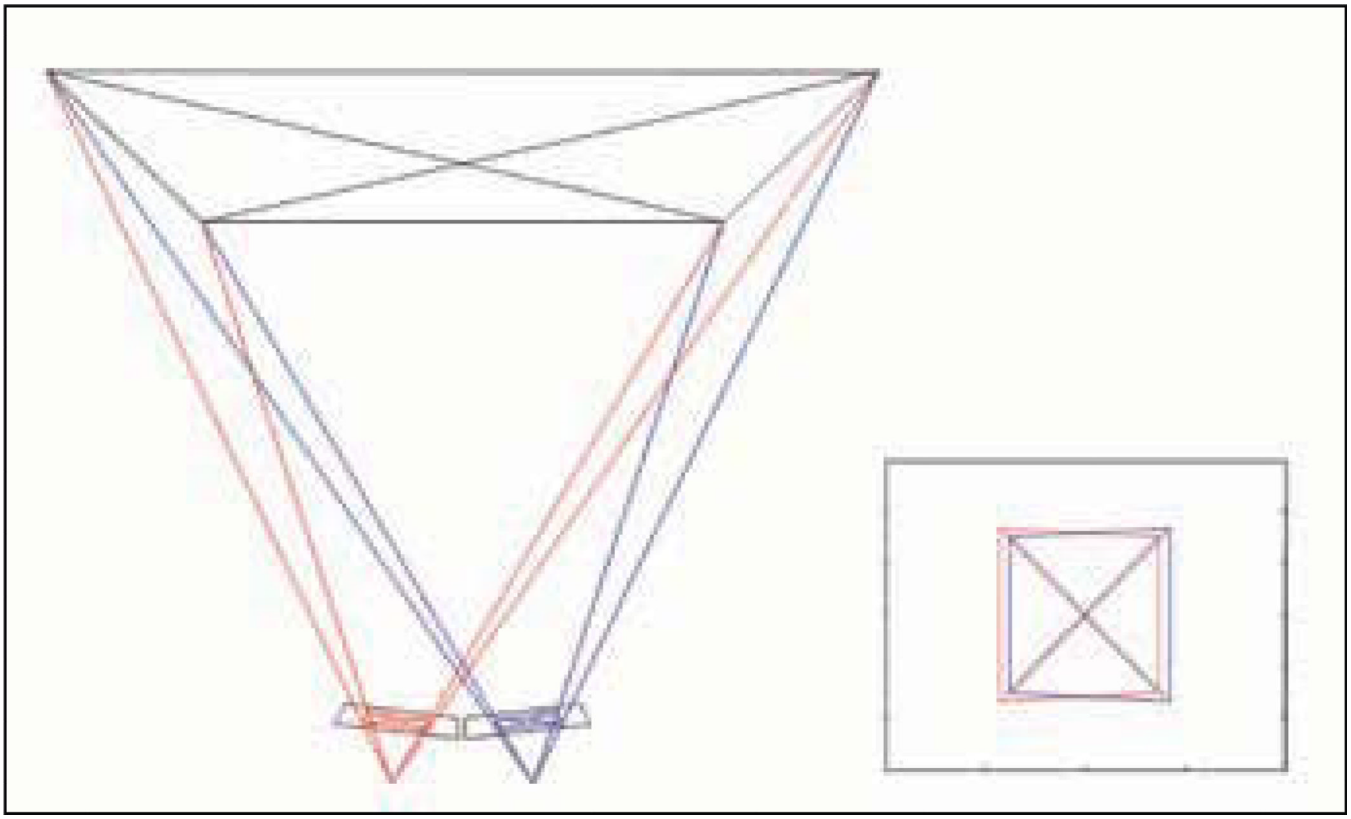

Figure 5.
Vertical parallax introduced by camera convergence.

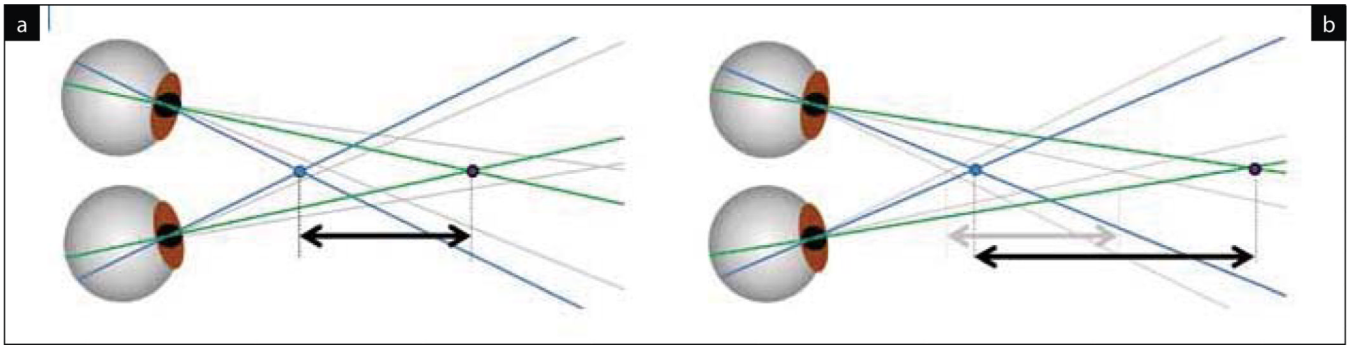

Figure 6.

Mapping from disparity to depth depends on the convergence angle. In both panels, the eyes are fixating on the purple sphere. The retinal disparity between the two spheres is the same in both panels. (a) The sphere is close, so the eyes are more strongly converged. (b) The physical distance the eyes map onto is much larger when the convergence angle is smaller.

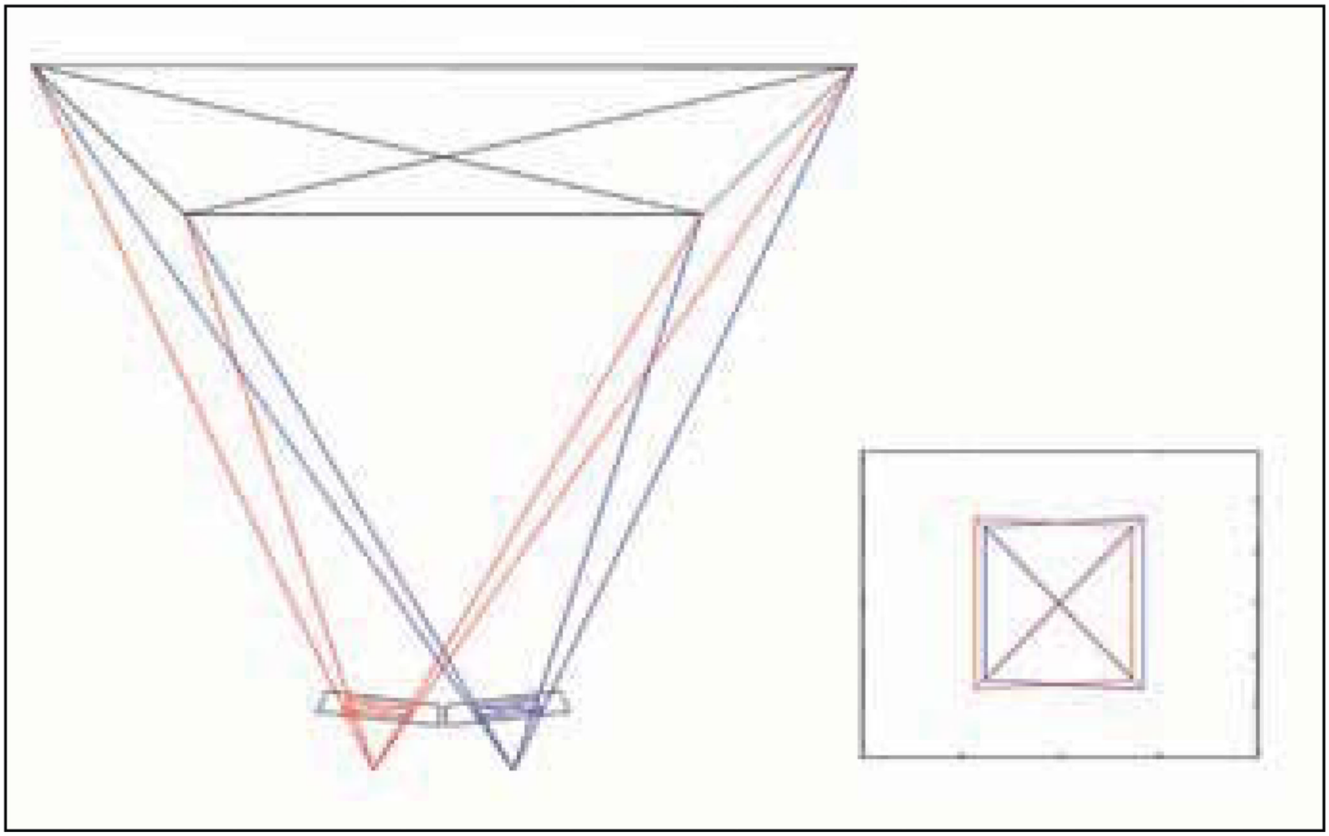

Figure 7.
Filming with parallel camera axes.

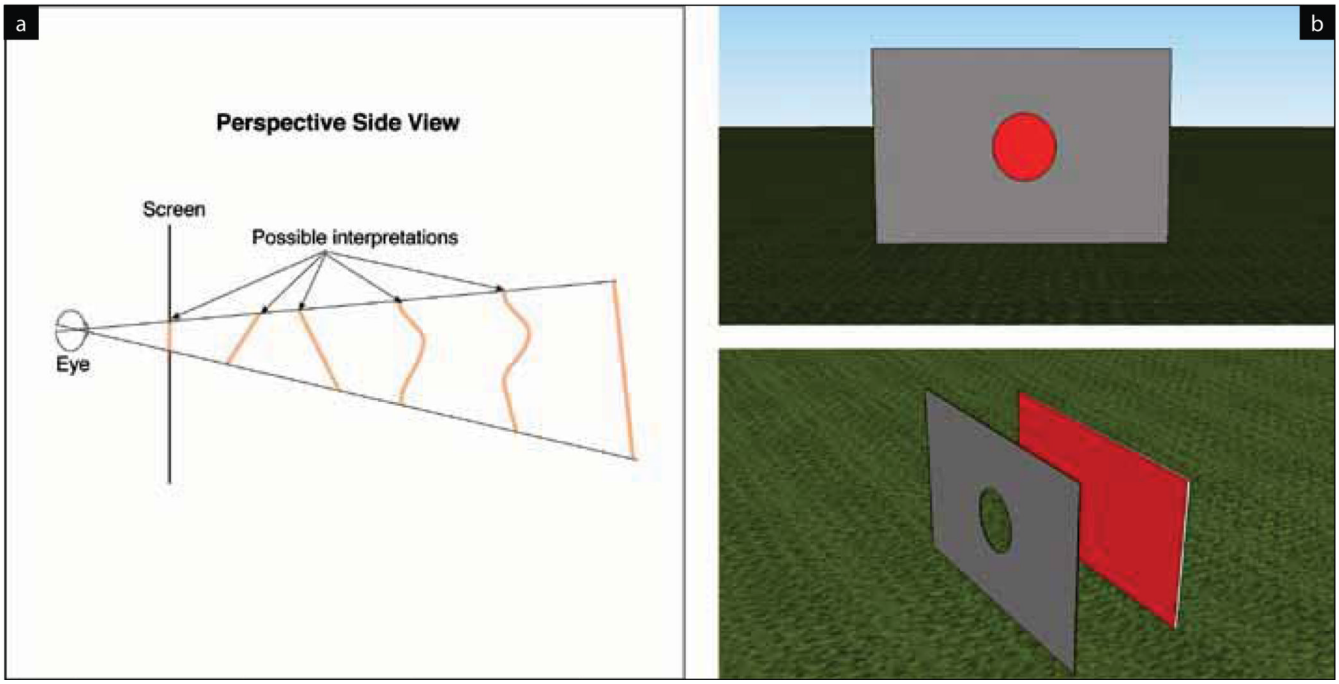

Figure 8.
Ambiguity in perspective projection. (a) A perspective projection is compatible with many real scenes, including a drawing of a 2D surface. (b) Even occlusion can be ambiguous if it is uncertain which surface is the occluder.

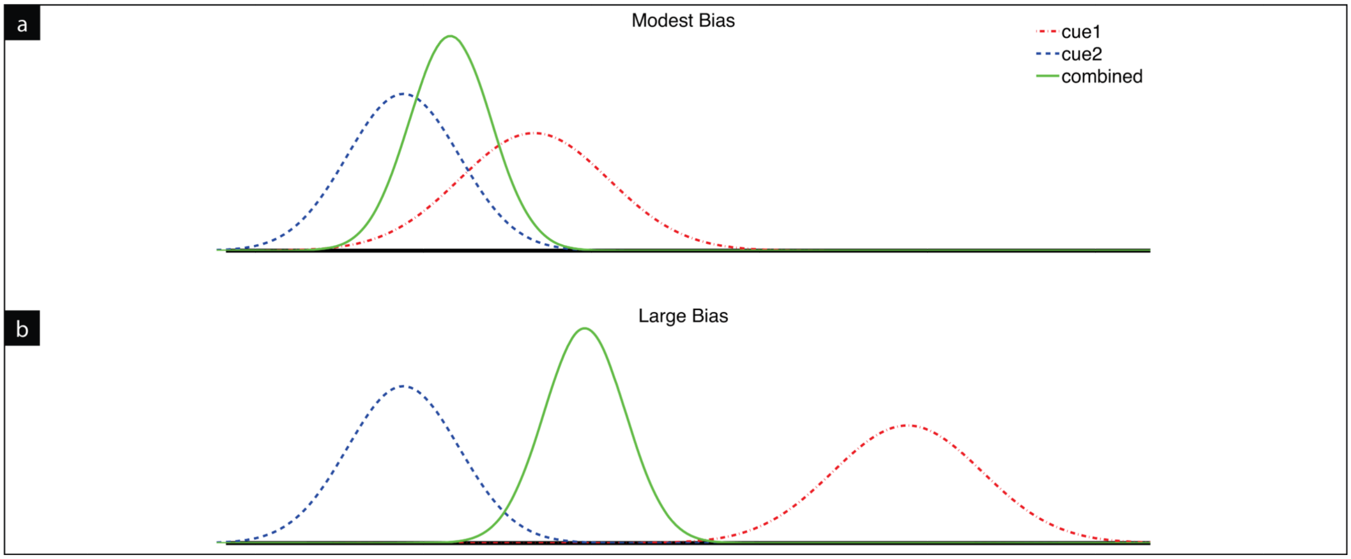

Figure 9.

Cue combination. Curves show the likelihood of a given depth value (horizontal axis) provided by two cues and the MLE combination (all normalized for unit area). (a) When the estimates of the two cues are similar, the weighted combination gives a more precise estimate. (b) When bias is large, the cue combination may not be consistent with either cue.

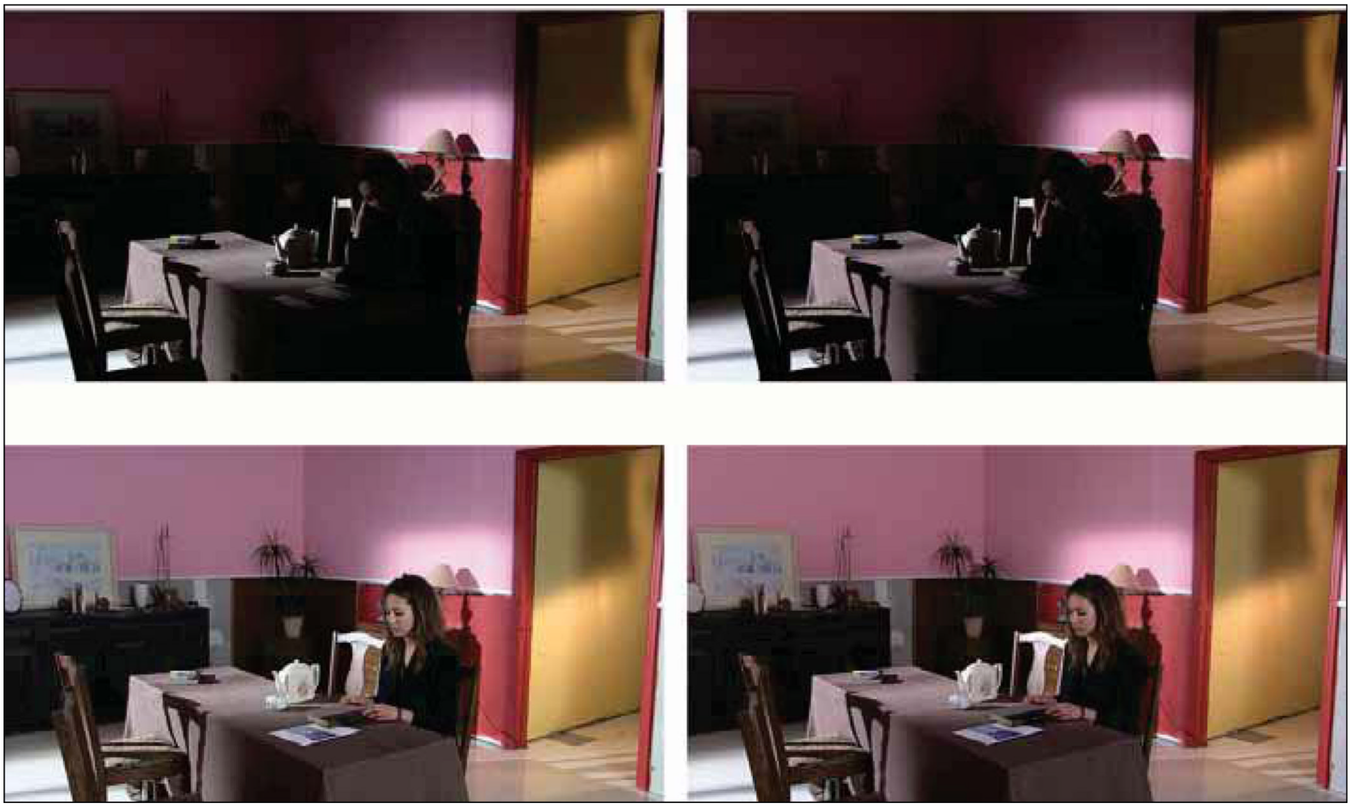

Figure 10.
Use of lighting to enhance the perception of depth. Top and bottom stereo pairs are arranged for cross-eyed fusion and have the same camera parameters. The top pair has high-contrast lighting; the bottom has flat lighting. Viewers generally report that the top pair has more depth than the bottom pair.

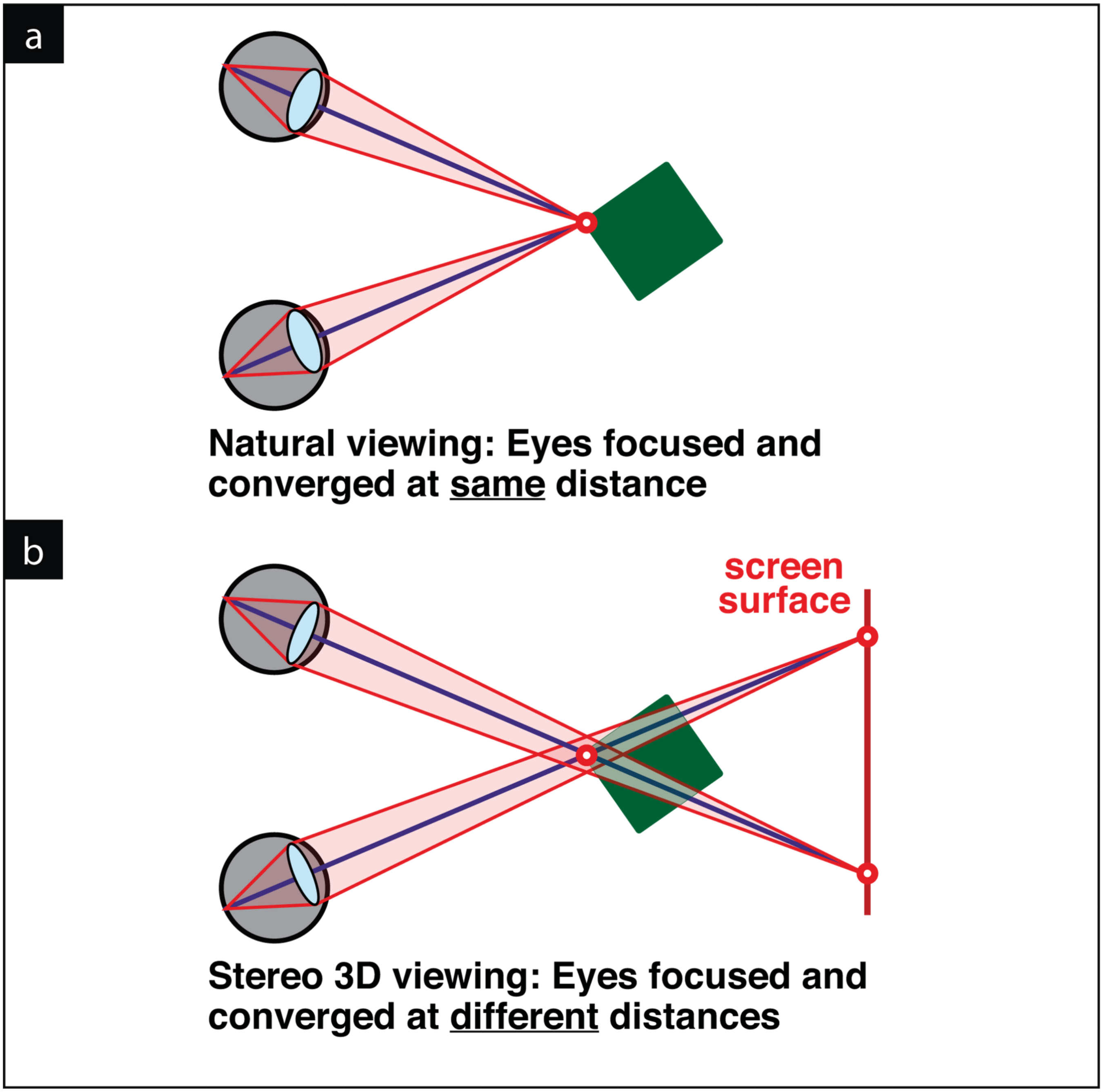
a
Natural viewing: Eyes focused and converged at same distance
b
screen surface
Stereo 3D viewing: Eyes focused and converged at different distances

Figure 11.
The vergence–accommodation conflict in stereoscopic displays. (a) In natural viewing, vergence and accommodation are to the same distance. (b) In stereo displays, these two oculomotor responses must be decoupled for the viewer to have clear, single binocular vision.

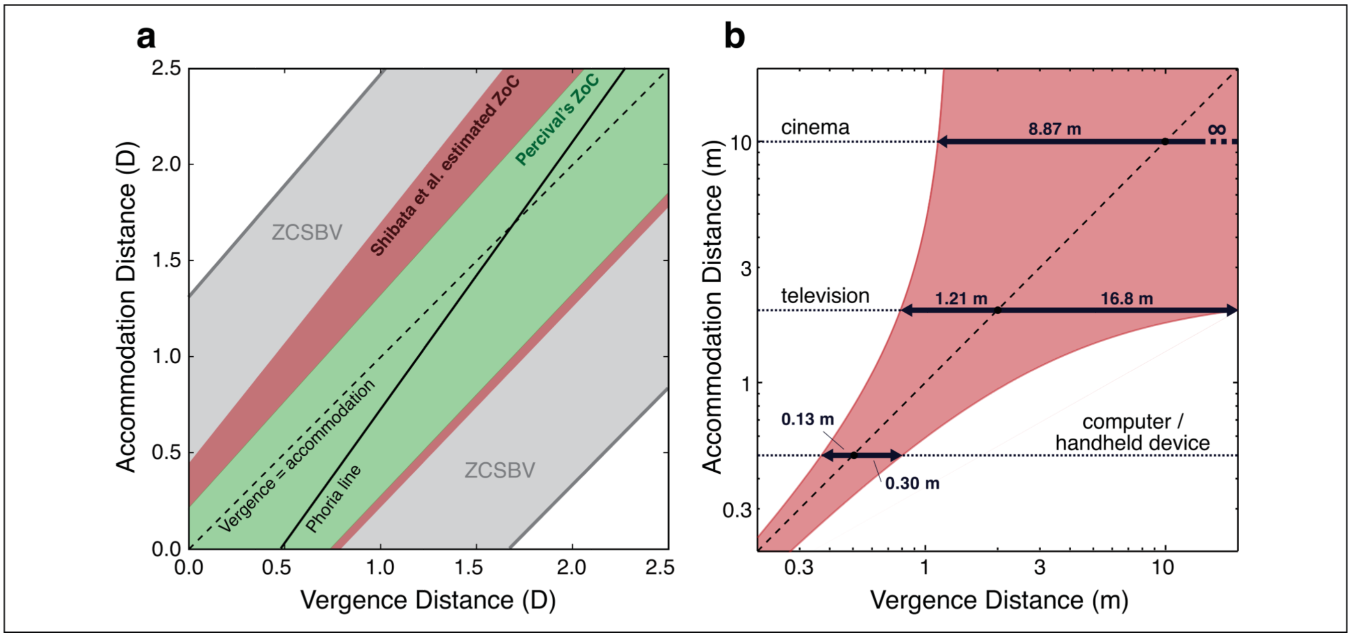

Figure 12.

ZoC. (a) Plot of accommodation distance as a function of vergence distance, both in diopters. The estimate of the ZCSBV is in gray, Percival's ZoC is in green, and the estimate of the ZoC for S3D viewing from Shibata et al.64 is in red. The phoria line for a typical viewer is also shown. (b) ZoC from Shibata et al. when plotted in units of distance rather than diopters. The horizontal lines represent the typical viewing distances for various common devices.

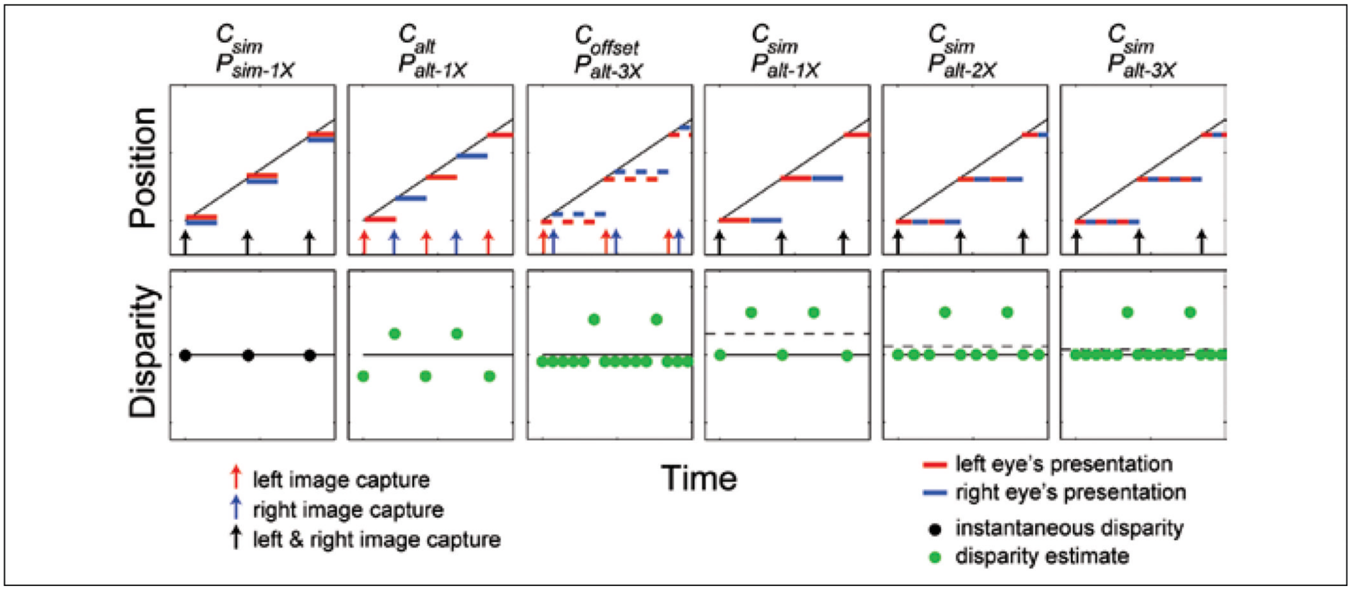

Figure 13.

Temporal protocols used in S3D displays. The columns represent different protocols. In the upper row, each panel plots the position of stimulus moving at constant speed in the plane of the screen as a function of time. Red and blue line segments represent the presentations of the images to the left and right eyes, respectively. The arrows indicate the times at which the stimulus was captured (or computed). Black arrows indicate left and right images captured simultaneously. Red and blue arrows indicate left and right images captured in alternating fashion. Black diagonal lines represent the correct positions for the left and right images as a function of time. In the lower row, each panel plots disparity as a function of time. Black horizontal lines represent the correct disparities. Black dots represent the disparities when the two eyes' images are presented simultaneously. Green dots represent the disparities that would be calculated if the left-eye image is matched to the successive right-eye image and the right-eye image is matched to the successive left-eye image. Dashed horizontal lines represent the time-average disparities that would be obtained by such matching. Wherever a horizontal line is not visible, the average disparity is the same as the correct disparity, so the two lines superimpose.

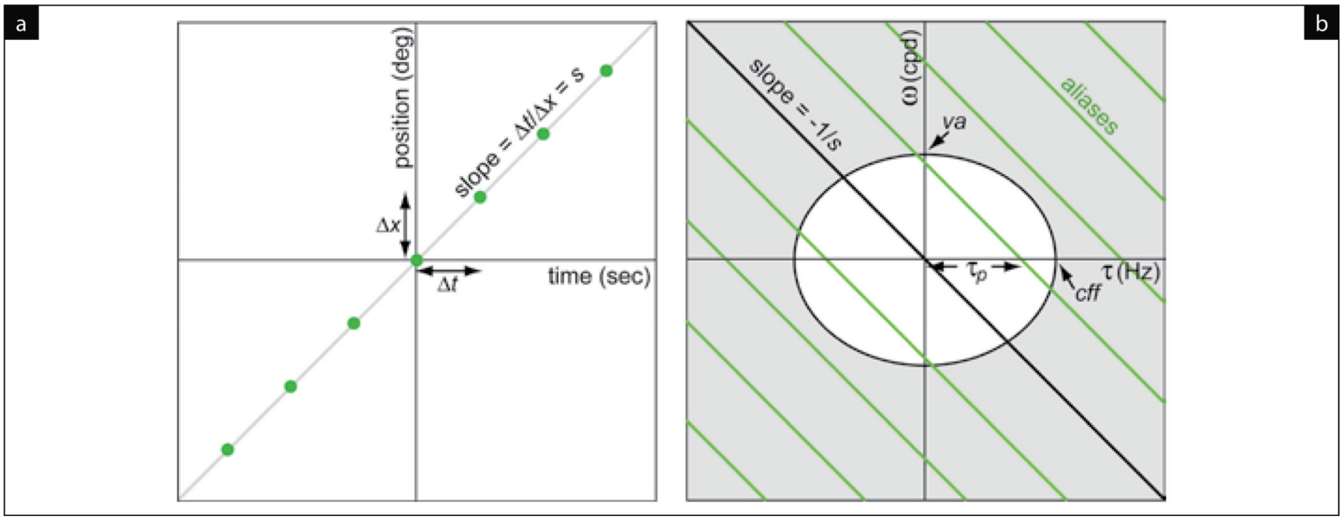

Figure 14.

Properties of a smoothly moving stimulus and a stroboscopic stimulus. (a) The gray diagonal line represents the motion of a smoothly moving vertical line on axes of time and horizontal position. The green dots represent the stroboscopic presentation of that stimulus; brief flashes occur at multiples of Δt. (b) Fourier transform (technically the amplitude spectrum) for the smoothly moving and stroboscopic stimuli plotted on axes of temporal frequency (in cycles per second or hertz) and spatial frequency (in cycles per degree). The black diagonal line represents the temporal and spatial frequencies of the smoothly moving stimulus. Green lines are the additional frequencies from the stroboscopic stimulus; they are temporal aliases separated by $\tau_p = 1/\Delta t$. The ellipse contains combinations of temporal and spatial frequency that are visible to the visual system. The highest visible temporal frequency is indicated by *cff*, and the highest visible spatial frequency is indicated by *va*. The shaded region contains combinations of temporal and spatial frequency that are not visible.

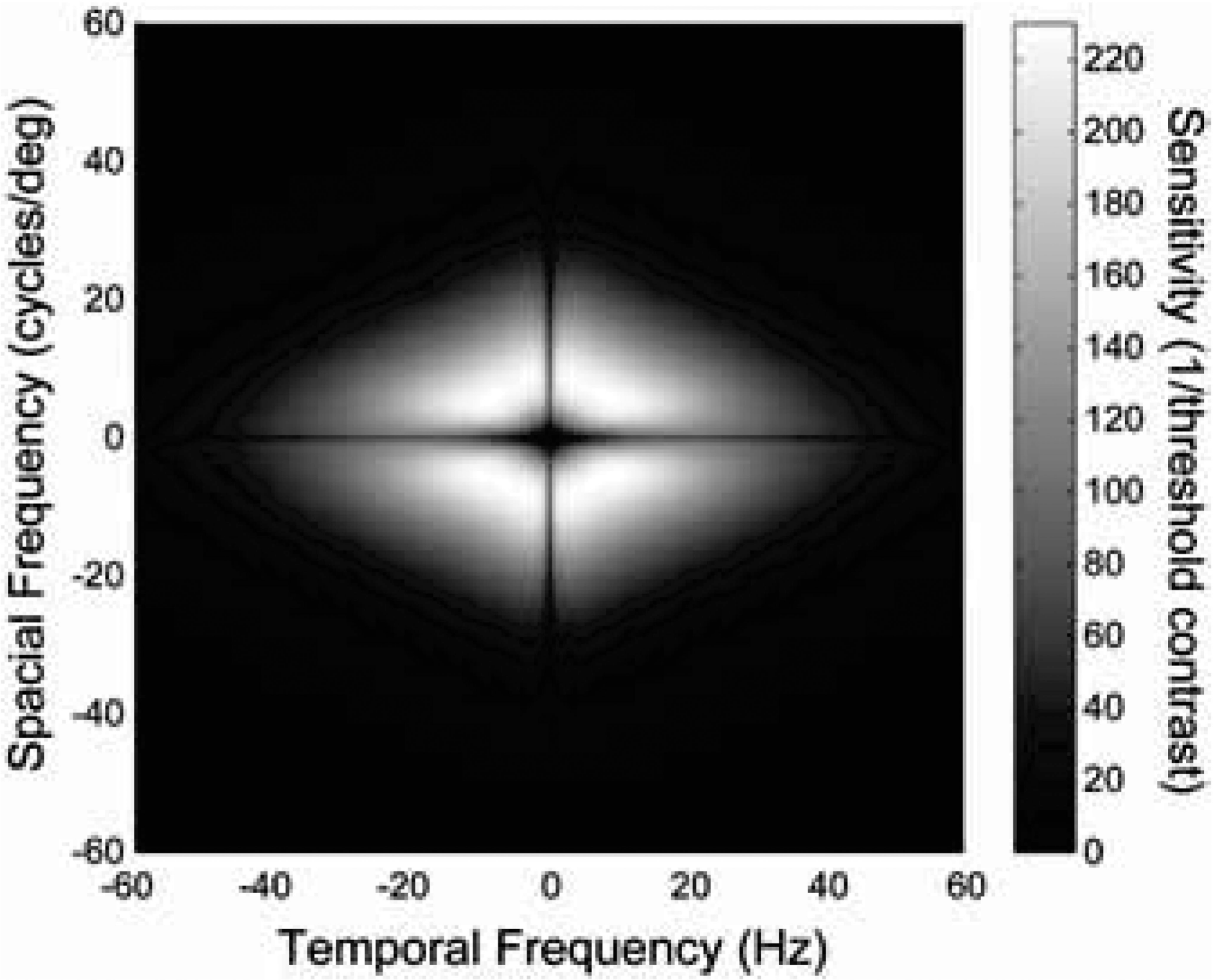

Figure 15.

The human spatiotemporal CSF. The sensitivity to a moving sinusoidal grating is plotted as a function of temporal frequency and spatial frequency. Sensitivity is the reciprocal of the contrast required to detect the stimulus and is represented by gray scale; brighter values corresponding to higher sensitivity. Adapted from Kelly.[72]

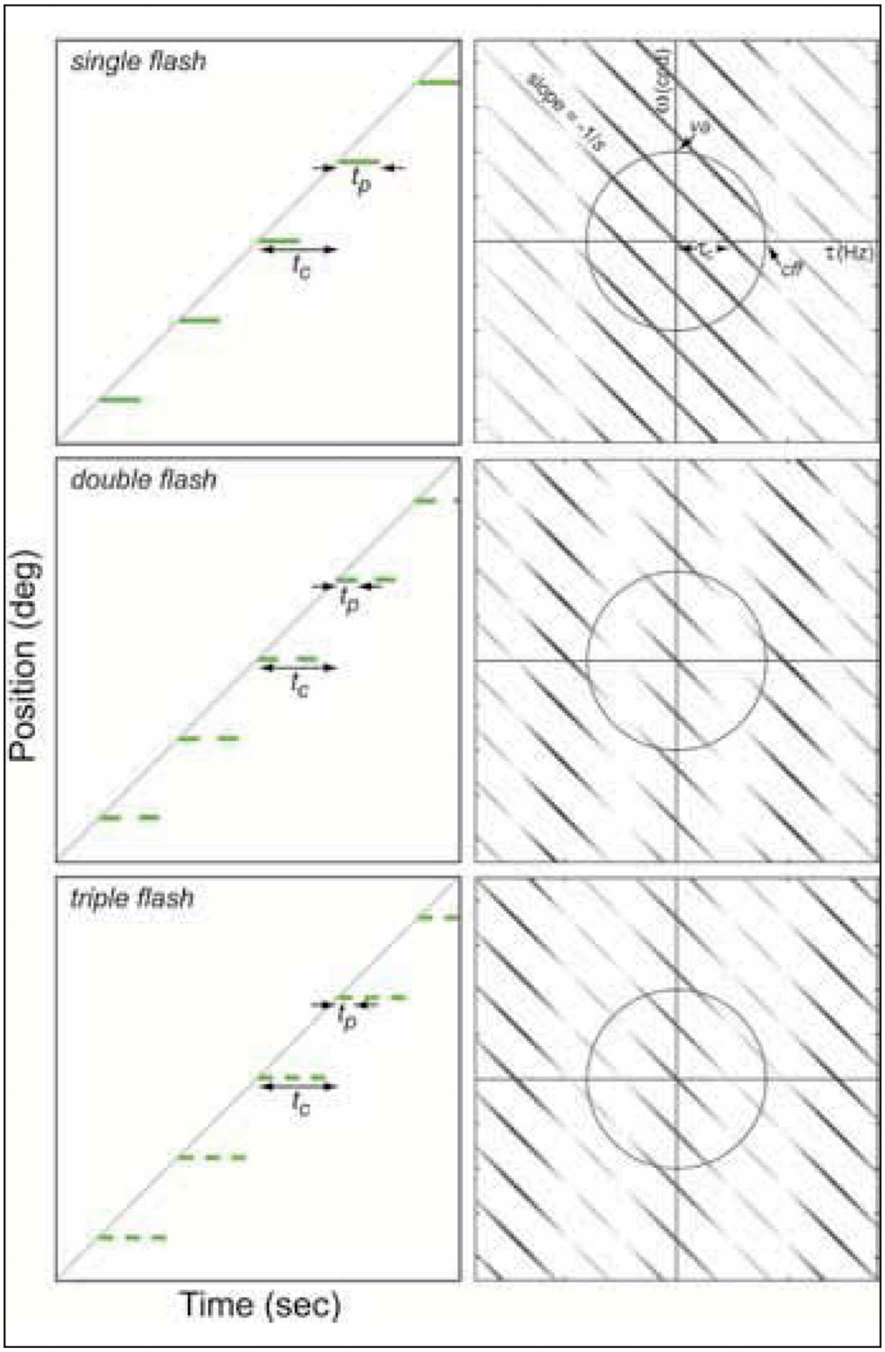

Figure 16.

Properties of stimuli presented with multiple-flash protocols. (a) Schematization of the single-, double-, and triple-flash protocols. In each case, the same images are presented during the interval tc until updated images are presented in the next interval. In multiflash protocols, the duration of each image presentation t_p is t^c/f, where f is the number of flashes. (b) Corresponding Fourier transforms of the multiflash stimuli plotted as a function of temporal and spatial frequency. The transform of a smoothly moving real stimulus is again a diagonal line with the slope $-1/s$. Amplitude is represented by gray scale, with dark values corresponding to higher amplitudes. The presentation rate t_p (or $1/t_p$) is indicated by arrows.

The aliases are separated by $t_c\,(1/t_p)$, which is also indicated by arrows. The circles represent the window of visibility.

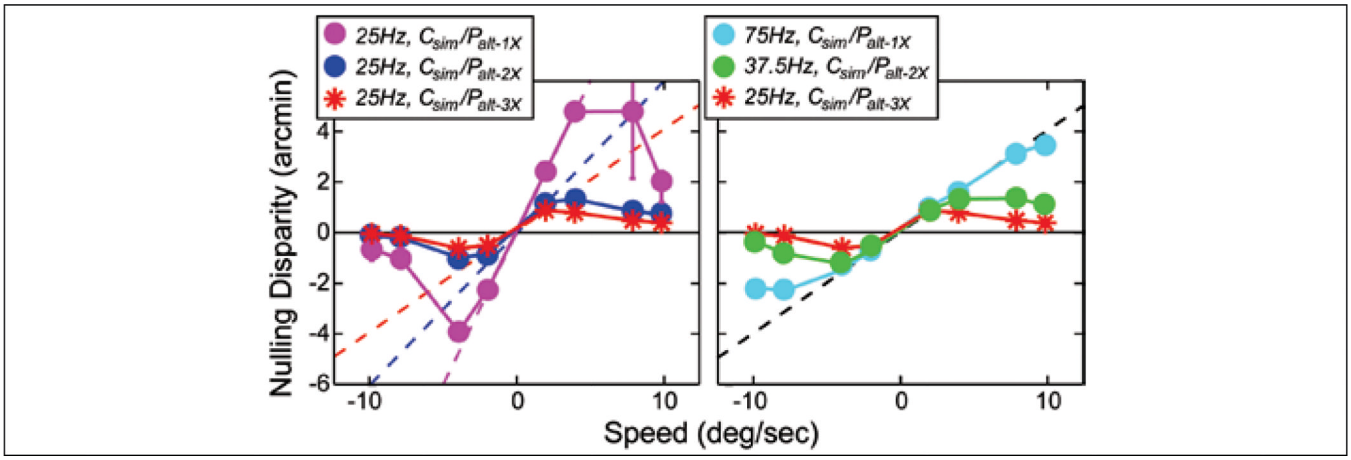

Figure 17.

Distortions of perceived depth with simultaneous capture and alternating presentation. The disparity distortion is plotted as a function of the speed of a stimulus moving in the plane of the display screen. (a) Data from protocols with a 25 Hz capture rate. Purple circles represent the data with the single-flash protocol (C_{sim}/P_{alt-1X}). Blue circles represent the data with the double-flash protocol (C_{sim}/P_{alt-2X}). Red asterisks represent the data from the tripleflash protocol (C_{sim}/P_{alt-3X}). The predictions for the time-average disparity model (lower row of Fig. 13) are the dashed lines with the colors corresponding to the appropriate temporal protocol. (b) Data from the same protocols, but with different capture rates. In each case, the presentation rate was 75 Hz, so the right eye's image was delayed relative to the left eye's image by 1/150 sec. The predictions for the time-average model are the dashed line. Cyan circles, green circles, and red asterisks are the data from the single-, double-, and triple-flash protocols, respectively.