Mental geometry of perceiving 3D size in pictures

Akihito Maruya
Graduate Center for Vision Research, State University of New York, New York, NY, USA

Qasim Zaidi
Graduate Center for Vision Research, State University of New York, New York, NY, USA

We show that the classical problem of three-dimensional (3D) size perception in obliquely viewed pictures can be understood by comparing human performance to the optimal geometric solution. A photograph seen from the camera position, can form the same retinal projection as the physical 3D scene, but retinal projections of sizes and shapes are distorted in oblique viewing. For real scenes, we previously showed that size and shape inconstancy result despite observers using the correct geometric back-transform, because some retinal images evoke misestimates of object slant or viewing elevation. Now, we examine how observers estimate 3D sizes in oblique views of pictures of objects lying on the ground in different poses. Compared to estimates for real scenes, in oblique views of pictures, sizes were seriously underestimated for objects at frontoparallel poses, but there was almost no change for objects perceived as pointing toward the viewer. The inverse of the function relating projected length to pose, camera elevation and viewing azimuth, gives the optimal correction factor for inferring correct 3D lengths if the elevation and azimuth are estimated accurately. Empirical correction functions had similar shapes to optimal, but lower amplitude. Measurements revealed that observers systematically underestimated viewing azimuth, similar to the frontoparallel bias for object pose perception. A model that adds underestimation of viewing azimuth to the geometrical back-transform, provided good fits to estimated 3D lengths from oblique views. These results add to accumulating evidence that observers use internalized projective geometry to perceive sizes, shapes, and poses in 3D scenes and their pictures.

Introduction

A major part of scene understanding for locomotion, hunting, gathering, shelter, and even aesthetic appreciation, consists of judging poses, sizes, and shapes of objects. Projective geometry determines retinal and camera images of objects, so a brain or machine could calculate accurate poses and sizes if it could invert the projection. The inversion back from two dimensions (2D) to three dimensions (3D) can generate unique estimates of poses and sizes of a known object if viewing parameters are ascertained independently. Building on previous work (Koch, Baig & Zaidi, 2018; Maruya & Zaidi, 2020), we tackle the classical problem of the perception of object sizes (and shape aspect ratios) from different views of pictures of 3D scenes (Boring, 1964; Gombrich, 1972; Perkins, 1973; Hagen, 1974, 1976; Rosinski, Mulholland, Degelman, & Farber, 1980; Wallach & Marshall, 1986; Cutting, 1987; Niall & Macnamara, 1990; Yang & Kubovy, 1999; Vishwanath, Girshick, & Banks, 2005; Todorović, 2008) and identify the geometric operations that are used for estimation. The results help to understand not just picture perception, but also the mental geometry that is essential for understanding 3D scenes.

We use an inverse projective geometry approach because it was successful at explaining 3D pose perception in real and pictured scenes. With this approach, Koch, Baig, & Zaidi (2018) showed that human observers are very accurate in judging 3D poses of objects on the ground, albeit with a frontoparallel bias, and their results can be explained by a model using the back-transform of the projective geometry function that maps 3D poses to retinal orientations, with just one free parameter for the bias. There was exceedingly close correspondence across observers (The mean pair-wise correlation was 0.9934 with a standard deviation of 0.0037), and individual results all fit the same geometric back-transform function, suggesting strongly that object-pose estimation, which has been a critical function for millions of years of evolution, could be based on internalized geometric knowledge. Pictorial representation, on the other hand, is less than 40,000 years old, so we tested the possibility that observers estimate 3D poses in photographs by using the same geometrical operations as for real scenes, despite retinal images being further distorted by projection to oblique viewpoints. The model based on back-projections from retinal orientations, predicts rigid rotation of perceived poses with respect to the observer, so that objects pointing at an observer in

Citation: Maruya, A., & Zaidi, Q. (2020). Mental geometry of perceiving 3D size in pictures. *Journal of Vision*, *20*(10):4, 1–16, https://doi.org/10.1167/jov.20.10.4.

https://doi.org/10.1167/jov.20.10.4 Received April 6, 2020; published October 2, 2020 ISSN 1534-7362

the real scene should also be seen as pointing at the observer in obliquely viewed photographs of the scene. Pose estimates corresponded almost perfectly ($R^2 > 0.99$) with the rotation prediction, thus providing support for the use of these geometrical operations. These results also confirmed that the "pointing at you" pictorial phenomena can be explained simply in terms of back-projections from vertical retinal orientations without invoking any previously postulated pictorial spaces or operations (Kennedy, 1974; Ward, 1976; DeLoache, Pierroutsakos, Uttal, Rosengren, & Gottlieb, 1998; Yang & Kubovy, 1999; Niederée & Heyer, 2003; Vishwanath et al., 2005; Koenderink, van Doom, Kappers, & Todd, 2004; Koenderink, van Doorn, & Wagemans, 2011; van Doorn, Koenderink, Leyssen, & Wagemans, 2012; Pagel, 2017). This explanation fits with Cézanne's observation in his letter to Émile Bernard dated April 15, 1904: "Lines parallel to the horizon give breadth, a section of nature, or if you prefer, of the spectacle spread before our eyes ... Lines perpendicular to that horizon give depth."

The inverse projective geometry approach has also been successful at explaining 3D size perception in real scenes. Maruya and Zaidi (2020) examined 3D size perception in real scenes as a function of the pose of the object. Unlike the usual experimental method of varying distance from the observer to study size constancy (Brunswik, 1944; Gilinsky, 1951; Gilinsky, 1955; Carlson, 1960; Norman, Todd, Perotti, & Tittle, 1996; Beusmans, 1998; Ross & Plug, 1998; Loomis, Da Silva, Fujita, & Fukusima, 1992; Loomis & Philbeck 1999; Loomis, Philbeck, & Zahorik, 2002), experimentally varying object pose changes the shape projected on the retina, and so requires quite different compensations for projective distortions. Observers compensated for projective shortening as a function of 3D pose, but not sufficiently for objects pointing towards or away from the observer, which are the poses that project to the shortest retinal sizes. Modeling the empirical correction as a function of the optimal correction, revealed that perceived sizes in 3D scenes are inconstant despite observers using the correct geometric back-transform, because the retinal image evokes a slant elevation illusion that reduces the compensation. This illusion of slant elevation angle for stimuli on the ground, becomes an illusion of slant azimuth angle if the scene is rotated in the fronto-parallel plane by 90° along the line of sight, so for generality we call it the slant illusion.

The illusory rotation that accompanies oblique viewing of pictures (Koch et al., 2018), makes poses invariant to viewing azimuth, but does lead to perceived variation in sizes and shapes, especially aspect ratios. Consequently, in this study, we test whether 3D size estimates in oblique views of pictures still follow from the geometric back-transform, and if there are simple factors that can explain estimation errors. Whereas projection from a real scene to the retina reduces sizes most along the axis pointing to the observer, viewing the picture obliquely (or equivalently rotating the picture around the vertical axis) reduces sizes most along the frontoparallel axis. To distinguish this screen slant from object slant, we will call it the screen azimuth (angle). We were particularly interested in any role played by the perceived azimuth of the picture with respect to the observer, since that had no effect on 3D pose estimation from pictures, as the pose estimation depended only on retinal orientations. We found that 3D sizes at frontoparallel poses are seriously underestimated in oblique views compared to the frontal view. By contrast, there was almost no change for objects perceived as pointing to or from the viewer. We were able to model observers' corrections for size as a function of pose, by using the optimal geometric back-transform with a multiplicative parameter for the systematic underestimation of the azimuth of the display. The underestimation was confirmed by perceived azimuth angle measurements and is similar to the frontoparallel bias for object pose perception. The excellent fit of the model shows that all observers use the correct back-transform from projective geometry, indicating the use of ingrained geometrical operations just as in pose estimation, but the perceived azimuth of the picture plays a role in size estimation unlike in pose estimation. Because aspect ratios of 3D shapes depend on relative sizes along different axes, those aspects are subject to the same picture azimuth distortions as perceived 3D sizes.

Size estimates of 3D objects at oblique views of pictures

Methods

Using Blender, we created a blue rectangular 3D parallelepiped (test stick) lying on the center of a dark ground, and a yellow vertical 3D cylinder (measuring stick) standing on the test stick. Blue parallelepipeds were presented in one of 16 poses from 0° to 360° every 22.5°, of which poses in one quadrant are shown in Figure 1 (Top). The line of sight through the center of the ground was designated the 90°–270° axis, and the line orthogonal to it as the 0°–180° axis. Parallelepiped were 10, 8, or 6 cm long with a 3 × 3 cm cross-section. Images were displayed on a 22-inch DELL SP2309W Display. Matlab and Psychtoolbox were used to display the stimuli, run the experiments, and analyze the data. The observer's viewing position was fixed by using a chinrest so that the center of the monitor was viewed with an elevation angle of 15° at a distance of 1.0 m, matching the rendering parameters of the Camera in Blender. Displayed sizes in the Blender rendered images

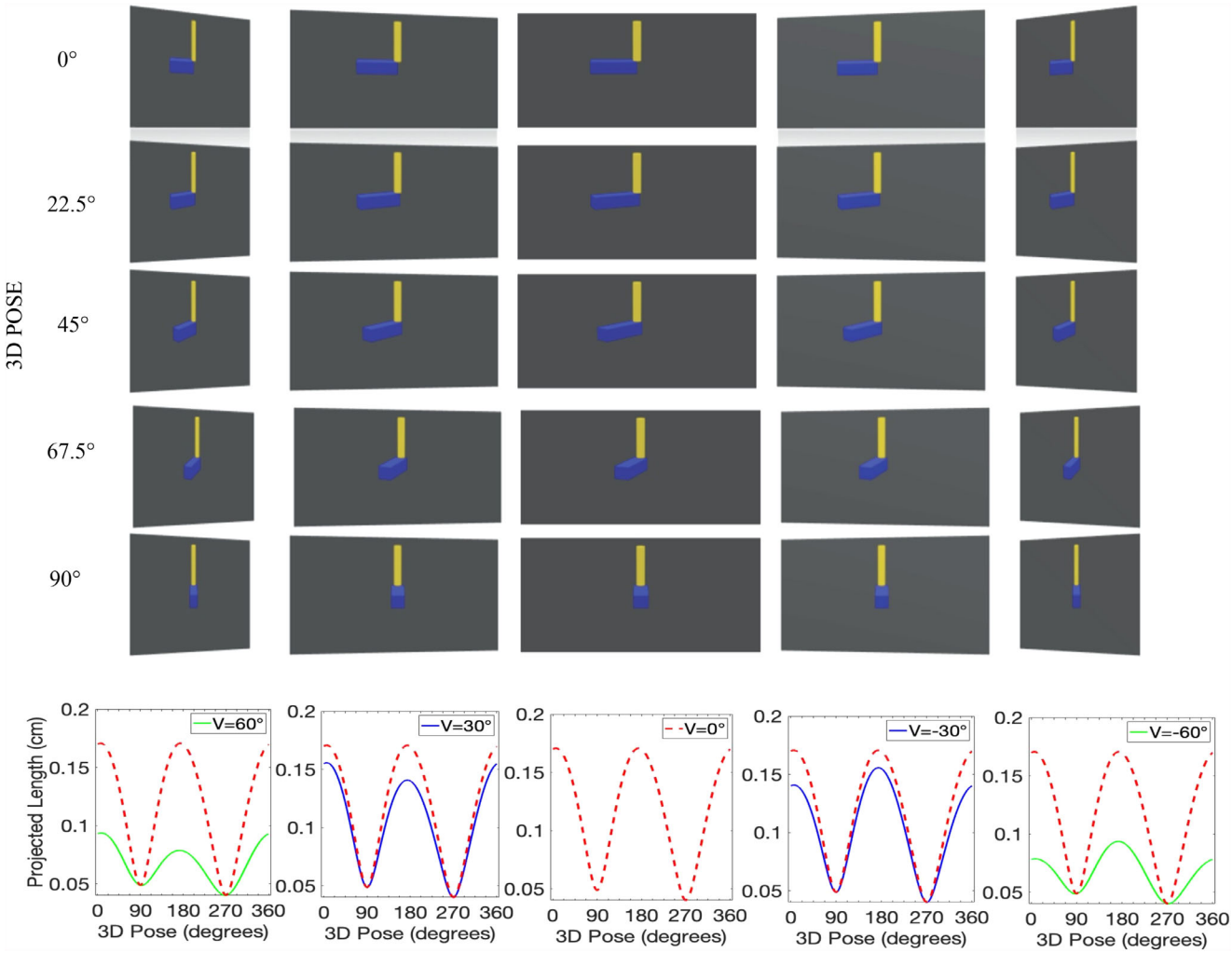

Figure 1. Projected images and lengths of Experiment 1 stimuli (Top) Blue rectangular 3D parallelepiped (test stick) of fixed length lying on the center of a dark ground, and a yellow vertical 3D cylinder (measuring stick) of adjustable length standing on the test stick. Blue parallelepipeds were presented in one of 16 poses from 0° to 360° every 22.5°, of which poses in one quadrant are shown. The line of sight through the center of the ground is designated the 90°–270° axis, and the line orthogonal to it as the 0°–180° axis. Parallelepiped lengths were 10, 8, or 6 cm with a 3- × 3-cm cross-section. Images were displayed on a 22-inch display. The monitor was viewed with an elevation angle of 15° at a distance of 1.0 m, from five viewpoints: frontal (0°), and slanted at azimuths of ±30° and ±60° around a vertical axis. Observe that projected length is shortest for the 90° pose in the 0° viewpoint, but it changes little across viewpoints, whereas the length for the 0° pose varies most across viewpoints. (Bottom) Derived lengths of retinal projections of the parallelepipeds are shown as a function of pose for the five views of the display screen (color coded according to inset and comparted to the frontoparallel view). The derivation is detailed in the Appendix and illustrated in Figure A1.

were calibrated against exact geometrical derivations to ensure accuracy of the simulations (see Maruya & Zaidi (2020)). Measurements were made from five viewpoints, with the screen in frontal (0°) position, and slanted at azimuths of ±30° and ±60° around a vertical axis. Derived lengths of retinal projections of the parallelepipeds are shown as a function of pose for frontoparallel and oblique views of the display screen in Figure 1 (Bottom). The derivation is detailed in the Appendix, and illustrated in Figure A1. For a parallelepiped of length (L_{3D}), the projected length on the retina (L_r) changes with pose (Ω) as a distorted sinusoid affected by the values for viewing elevation=Φ_c, viewing azimuth (equivalent to display azimuth) = Φ_v, focal length of camera = f_c, focal length of eye = f_v, distance from the object = d_c, and distance from the center of the display = d_v:

$$L_r = \frac{L_{3D} f_v f_c \sqrt{(\cos(\Omega)\cos(\phi_v))^2 + (\sin(\Omega)\sin(\Phi_c))^2}}{d_v (d_c - L_{3D}\sin(\Omega)\cos(\Phi_c)) - L_{3D} f_c \cos(\Omega)\sin(\phi_v)}. \quad (1)$$

However, for the vertically oriented cylinder of physical length L_{3Dm}, the projected length on the retina L_{mr} stays invariant with object pose and display

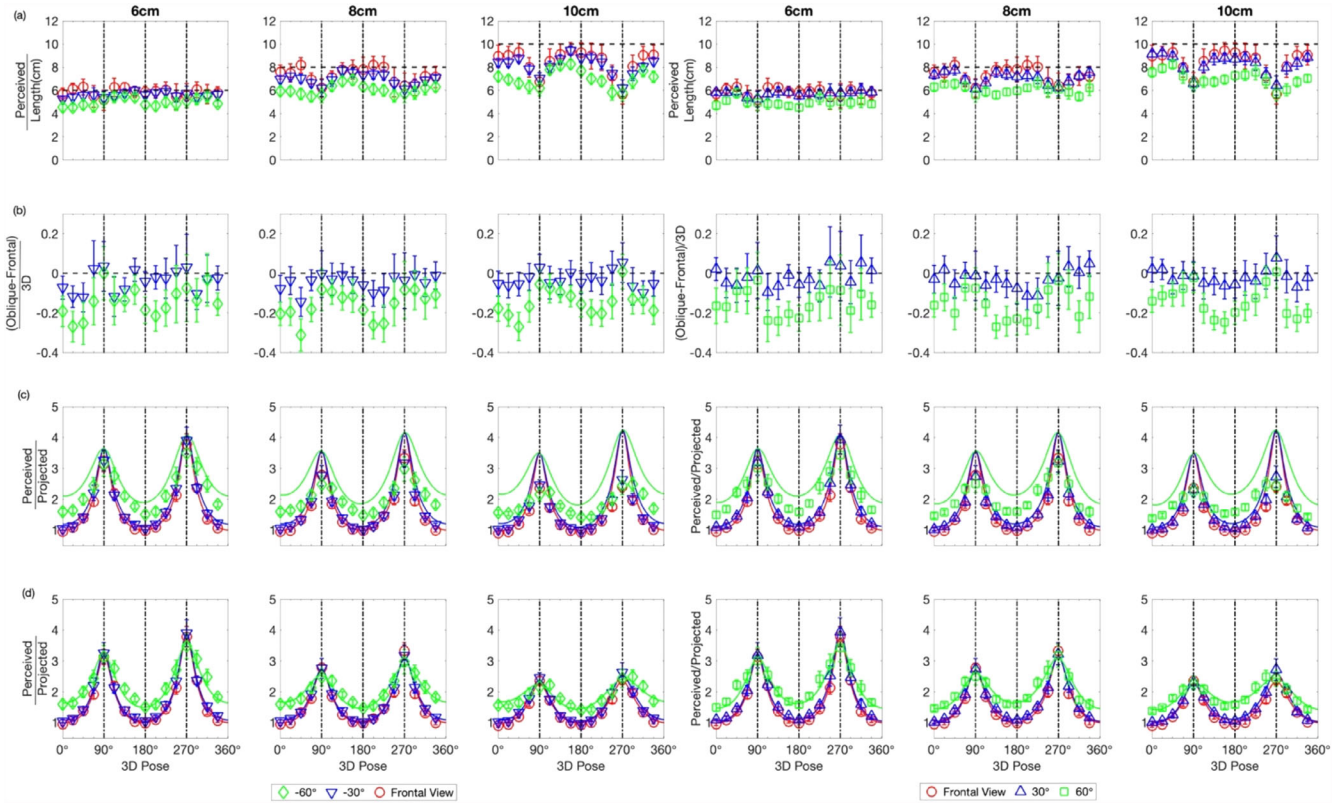

Figure 2. Average perceived 3D lengths (six observers). Each column represents estimated lengths of a parallelepiped of the indicated physical length. Left three panels are for negative oblique viewing azimuths and right for positive. (a) Perceived length across 3D pose. Horizontal dashed lines indicate physical 3D length. Underestimation of perceived length increases systematically with increased azimuth of the display. (b) Perceived 3D length for oblique view minus perceived length for frontal view divided by 3D physical length. Horizontal dashed lines indicate no difference in perceived length from frontal view. Error bars are the 95% confidence intervals. (c) Optimal correction factor (solid line) and empirical correction factor (symbols) across 3D pose, color coded for display azimuth. Higher optimal correction is required at the frontoparallel poses as display azimuth increases but is fairly constant for poses around the line of sight. (d) A model using the optimal geometrical back-transform, but incorporating underestimation of the slant of the display, fits the correction of perceived length over the projected length.

azimuth, because the 3D object and the picture plane are both rotated around the vertical axis of the cylinder:

$$L_{mr} = \frac{L_{3DM} f_v f_c \cos(\Phi_c)}{d_v (d_c - L_{3Dm} \sin(\Phi_c))}. \quad (2)$$

Observers were instructed to equate the physical lengths of the two limbs by pushing buttons to adjust the height of the measuring stick between 2.75 and 12 cm. There were no time limits. Randomly ordered trials across all three sizes and all 16 poses were run in blocks for each of the five azimuth of the display and repeated in three sets. Observers were allowed to take a break between sets. Note that because d_v plays the same role for L_r and L_{mr}, any slight variations in distance between eyes and screen will have no effect on the relative size estimates.

Six observers with normal or corrected vision participated. Viewing was binocular because it is the natural condition for looking at pictures, and Koch et al. (2018) had not found any difference between monocular and binocular viewing for pose estimation in similar conditions. All experiments presented in this article were conducted in compliance with the protocol approved by the Institutional Review Board at the State University of New York College of Optometry and the Declaration of Helsinki, with observers giving written informed consent.

Results

Perceived 3D lengths, averaged over six observers, are plotted against 3D pose in Figure 2a, for five different azimuth angles of the display, separately for each of the three physical lengths (Individual results in Figure A2). Horizontal dashed lines indicate the physical length of the test stick, and vertical dotted lines indicate poses of 90°, 180°, and 270°. In the

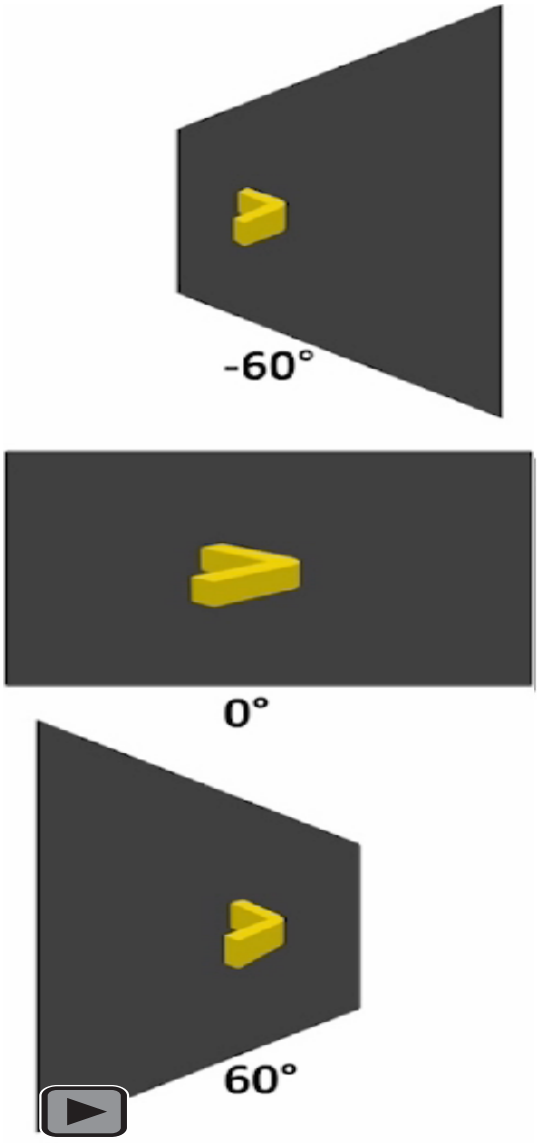

Figure 3. Dynamic demonstration of size inconstancy. A rigid object with two physically equal limbs at a right angle is rotating on the ground. When viewed from the front at a 15° elevation, the limb pointing at or away from the observer appears transitorily shorter than the orthogonal limb. Comparing the oblique views to the frontal view, the biggest change in length is transitorily for the frontoparallel limb. The effects persist for obliquely slanted pictures.

frontoparallel view, there is greater underestimation of length for poses pointing towards or away from the observer (90° and 270°), and the longer the physical length of the test stick, the greater the underestimation of length. Increasing the azimuth of the screen seriously reduced estimates of frontoparallel sizes in oblique views, although there was little effect on perceived sizes of poses pointing toward or away from the observer. The videos in Figure 3, where rigid L-shaped figures are rotated in pose in frontal and oblique views, illustrate these two effects. The magnitudes of these two effects are easily seen in Figure 2b, where the frontoparallel size estimate is subtracted from the oblique estimate, and the difference divided by the physical length, to give the fractional decrease. There are roughly equal magnitudes of underestimation as a fraction of the physical length for the three lengths, and similar patterns of underestimation as a function of object pose: the least change in estimates of 3D sizes caused by oblique azimuths of the monitor are for poses close to the line of sight (90° and 270°), whereas sizes of objects in poses close to frontoparallel (0° and 180°) were seriously underestimated in oblique views compared to the frontal view. The effect of screen azimuth also shows up in the pattern of underestimation for positive azimuths being almost mirror symmetric to the pattern for negative azimuths.

Because horizontal slants of the screen lead to the greatest length compression along the frontoparallel axis (Figure 1), the question arises whether the decrease in perceived length is explained completely by the shorter projected length, or whether the visual system does compensate partially for this compression. On the dark ground, the retinal image of the object holds the only information available for size estimation. Koch, Baig and Zaidi (2018) showed that in perception of poses in oblique views of pictures, the same observer-centered back-transform is used as for real scenes, thus leading to pose estimates that are a rigid rotation of actual scene poses by an angle equal to the viewing azimuth. Maruya and Zaidi (2020) showed that perceiving 3D sizes in real scenes also uses the optimal geometric back-transform for sizes, but estimates are suboptimal because of a slant illusion that makes longer objects appear more slanted, which leads to less correction than required. The results in Figure 2 show that observers do not make veridical estimates of 3D size. It is still possible that they use the optimal geometrical back-projection from the retinal image but misestimate one or more parameters that are part of the back-transform expression. The physical 3D length divided by the length projected on the retina (L_{3D}/L_r) gives the Optimal Length Correction index (OLC) for each pose in each viewpoint (plotted as solid curves in Figure 2c, color-coded similar to symbols for the azimuth of the screen). The symbols in Figure 2c plot the perceived lengths from Figure 2a divided by projected lengths giving the measured length corrections (MLC). OLCs are the highest at line of sight (90° and 270°) with the same values for all azimuths of the display. The largest values of MLC also correspond to the poses pointing towards or away from the observer, with similar values for across the different azimuths of the display, but they are lower than what is required for veridical estimates, especially for the longer lengths. OLCs at frontoparallel poses (0° and 180°) increase with the slant of the display. Similarly, MLCs at these poses increase with

the slant of the display, but not enough. Figure 2c enables us to reject the hypothesis that size perception in pictures uses the same back-transform as for real scenes, irrespective of the azimuth of the monitor, which makes it different from pose perception. If this hypothesis was correct, the frontal view OLC, which is also the OLC for real scenes, should fit the data with some multiplicative scaling of the red curve, but that cannot happen because OLC for 0° and 180° poses is anchored at 1.0 when display azimuth is fixed at 0°, no matter what the values for other parameters such as camera elevation, focal length or distance, whereas the MLC for 0° and 180° is higher for the obliquely viewed displays showing increasing length correction with increasing azimuth of the display. This increase is still substantially less than required by the OLC computed for the ±60° display azimuth. In general, the form of measured length correction as a function of 3D pose is similar to the optimal length correction curve, suggesting that observers may be using the optimal back-transform, but with additional multiplicative factors leading to the suboptimality. We test this hypothesis by fitting a modified back-transform model.

Model

The geometrical back-transform is obtained by inverting Equation 1 to get an expression for the estimated 3D length $\hat{L}_{3D}$:

$$\hat{L}_{3D} = \frac{L_r d_v d_c}{L_r(d_v \sin(\Omega)\cos(\Phi_c) + f_c \cos(\Omega)\sin(\phi_v)) + f_v f_c \sqrt{(\cos(\Omega)\cos(\phi_v))^2 + (\sin(\Omega)\sin(\Phi_c))^2}}. \quad (3)$$

$\hat{L}_{3D}/L_r$ gives the expression for the estimated length correction index (ELC) for each pose. The ELC will be equal to the OLC, and give veridical estimates of 3D size, only if the values used for pose Ω, viewing elevation Φ_c, viewing azimuth (equivalent to display azimuth) Φ_v, and distance from the object d_c, and distance from the center of the display d_v are accurate. If these values are not accurate, the estimated 3D length $\hat{L}_{3D}$ will be different from the veridical.

Maruya and Zaidi (2020) showed that perceived sizes can be inaccurate in 3D scenes despite observers using the correct geometric back-transform, if the retinal image evokes a misestimate of viewing elevation. The longer the physical length of the test stick, the greater the misestimation of the elevation, and the effect is most obvious for poses pointing at or away from the observer. Their model fit the MLCs with just one free parameter that modified the estimated camera elevation (equivalent to modifying perceived slant elevation of the object). To illustrate the slant illusion, they made the video shown in the frontal view of Figure 4. When

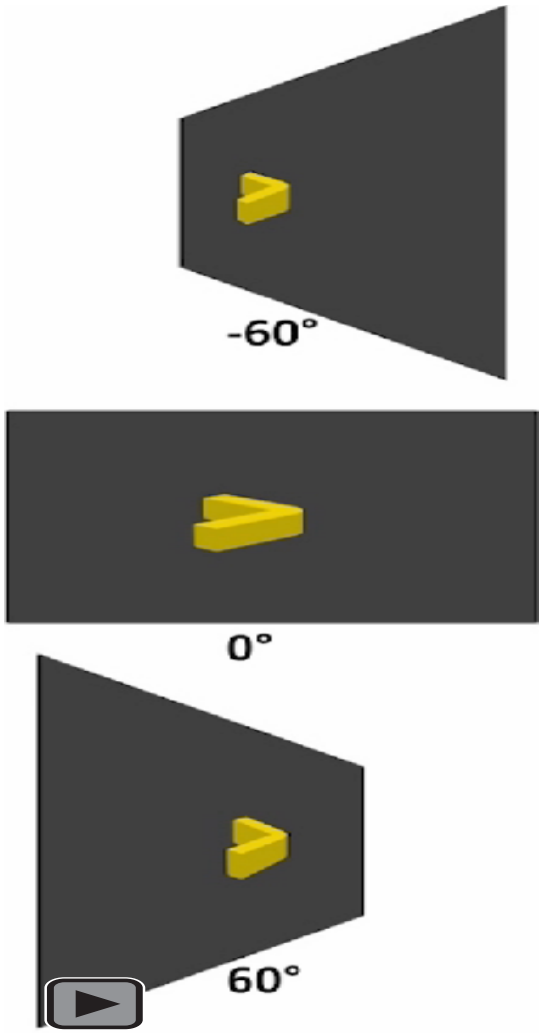

Figure 4. Dynamic demonstration of slant illusion. A rigid object with two limbs at a right angle is rotating on the ground. The length of the limb passing through the line of sight is lengthened and then shortened, to make the limbs appear equal in all poses in the view from the 15° elevation. The percept is maintained by increasing the length of the limb passing through the line of sight according to the average size estimate in Experiment 1. Instead of seeing the physical length changes, each limb seems to bounce up and down when it faces toward or away from the observer, because the pitch illusion dominates the percept in the oblique views, as well as the frontal view.

the arms of the figure from Figure 3 are dynamically adjusted to be perceptually equal in length across poses, observers perceive a change in pitch every time an arm passes through the line of site. Figure 4 shows that the illusion persists when the display is viewed obliquely, so we retain this modification of the model. Now we will try to understand the suboptimal size estimates in obliquely viewed pictures by seeing if they can be predicted by inaccurate estimates of display azimuth, and then directly measuring perceived azimuth. The

motivation for this manipulation is the frontoparallel bias in judging 3D pose (Koch et al., 2018). Figure 2c shows that higher OLCs are required at frontoparallel poses with increasing azimuth of the display, but MLCs are uniformly lower than adequate. We thus include underestimation of the azimuth of the display in the model, because ELCs are uniformly lower at the shallower azimuths of the display. The qualitative justifications for these considerations are shown in Figure A3. Under- or overestimating perceived pose shifts the peak of the ELC, which would not be consistent with the results, so we don't include misperception of pose in the model. Misperceiving viewing azimuth affects frontoparallel poses most, and essentially only for the most oblique views, which is consistent with the results. In addition, misperceiving viewing elevation (or equivalently slant elevation of the object) differentially affects size estimates of objects close to line of sight, again consistent with the results. Viewing a picture from other than the camera distance can cause perceptual distortions (Sedgwick, 1989), but it does not affect the relative size judgments in this experiment.

We thus formulated the hypothesis that observers are using an optimal back-transform, but overestimating the slant elevation of the object (or equivalently the camera elevation) which leads to underestimated lengths at poses around 90° and 270°, plus they are underestimating the slant azimuth of the display, thus correcting less than required for poses around 0° and 180° degrees. We set $f_c = 1.0\ m$ so that the image on the retina in frontoparallel viewing was identical to that from the 3D scene in real world coordinates. For f_v we tried 17, 22, and 24 mm, corresponding to best estimates from different measurement methods and found absolutely no difference up to the second digit after decimal in the secondary analyses. We tested whether adding multipliers $k_c > 1$ to the viewing elevation, and $k_v < 1$ to the azimuth of the display, in the optimal geometrical back-transform expression, could provide good fits to the MLCs:

$$\hat{L}_{3D} = \frac{L_r d_v d_c}{L_r\left(d_v \sin(\Omega)\, cos\,(k_c\Phi_c) + f_c \cos(\Omega)\sin(k_v\phi_v)\right) + f_v f_c\sqrt{(\cos(\Omega)\cos(k_v\phi_v))^2 + (\sin(\Omega)\sin(k_c\Phi_c))^2}}. \quad (4)$$

For each length and azimuth angle of the display, Figure 2d replots the empirical corrections from Figure 2c and predicted values of Equation 4 for the k_c and k_v that give the best least squares fit to the data. The model fits the results for all three lengths and azimuths of the display well, with just the two free parameters (Best fits to individual results in Figure A4). Best fitting k_c are nearly equal for frontoparallel and oblique views, and best fitting k_v correspond to perceived display azimuths 21.3° for 30° view, –19.2° for –30°, 49.8° for 60°, and –50.4° for –60°.

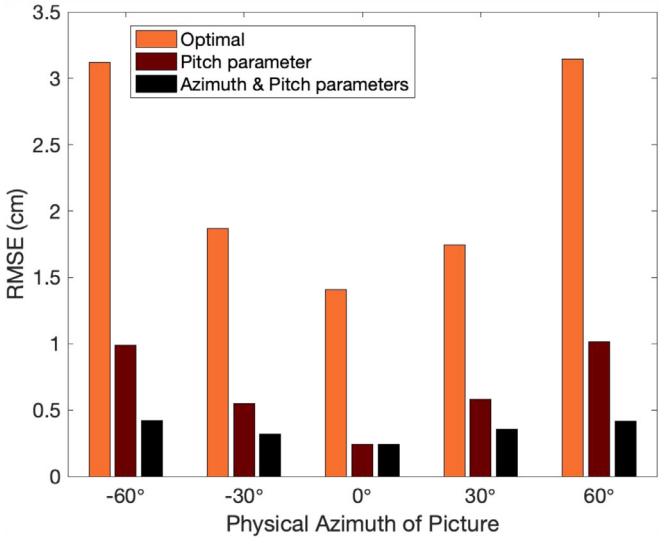

Figure 5. RMSE for optimal back-transform, compared with RMSE with added multiplier for perceived object slant elevation, and then compared to RMSE with additional multiplier for perceived screen azimuth, for five physical azimuths of the screen.

It is not as straightforward to quantify goodness of fit for non-linear functions as using R^2 for linear functions. Figure 5 illustrates the root mean square error (RMSE) for each fit. Compared to the fit of the optimal back-transform with no free parameters, adding a multiplier to vary perceived object pitch reduces RMSE by almost six times at 0° frontal view of the screen, but only a little more than three times for oblique views. Adding a multiplier to vary perceived screen azimuth further reduces RMSE by over two times at the most oblique views. Note that fitting errors for the optimal model are all in one direction, not random directions, so the reduction in RMSE magnitude is due to reducing systematic, not random, error. As expected, simulating misperceived distance or misperceived object pose by putting multipliers on the distance or pose parameters, did not reduce RMSE from the Optimal back-transform. The modeling results are thus compatible with the hypothesis that observers use the geometric back-transform, but with incorrect parameters for object and screen slant. Maruya and Zaidi (2020) previously showed that the object slant illusion was due to ambiguity between projections of increased length and slant elevation, and validated the illusion with direct measurements of perceived slant. The underestimated screen azimuth would be compatible with the general frontoparallel bias for object pose (Koch et al., 2018). To validate the screen slant inference from the best fitting model, we measured directly whether observers actually misperceived the azimuth of the display.

Display azimuth underestimation

Methods

To test whether there was actual misestimation of the azimuth of the display, we displayed a 10 cm test stimulus at 0° pose at the center of the screen, and observers were instructed to judge the 3D pose of the screen. Viewing position was identical to Experiment 1, and the scene was viewed with azimuths of −60°, −30°, 0°, +30°, or +60° in random order. Observers recorded their judgement by rotating a vector in a clock face on an iPad to the same angle as the pose of the rectangle. A horizontal iPad screen was placed close to the display screen, and observers adjusted the vector angle on a keyboard. Measurements were separated into three sets and observers were allowed to take a break between sets.

Results

The main result (Figure 6a) is that observers underestimated the azimuth of the display at oblique viewing conditions despite binocular viewing (Individual results in Figure A4). The slope of perceived to physical azimuth was 0.70, with $R^2 = 0.90$. Average perceived azimuths of the display were 23.24° for 30° view, –19.40° for –30°, 47.27° for 60°, and –49.47° for –60°. As we expected, the underestimation pattern for the obliquely viewed display surface is consistent with the frontoparallel bias for obliquely posed objects (Koch et al., 2018). The results support our hypothesis that observers may be applying a smaller correction to oblique views because there is underestimation of the azimuth of the display. Measured perceived azimuths are similar in value to those predicted by the best fitting model (Figure 6b). The slope of perceived to best-fitting azimuth was 0.90, with $R^2 = 0.88$.

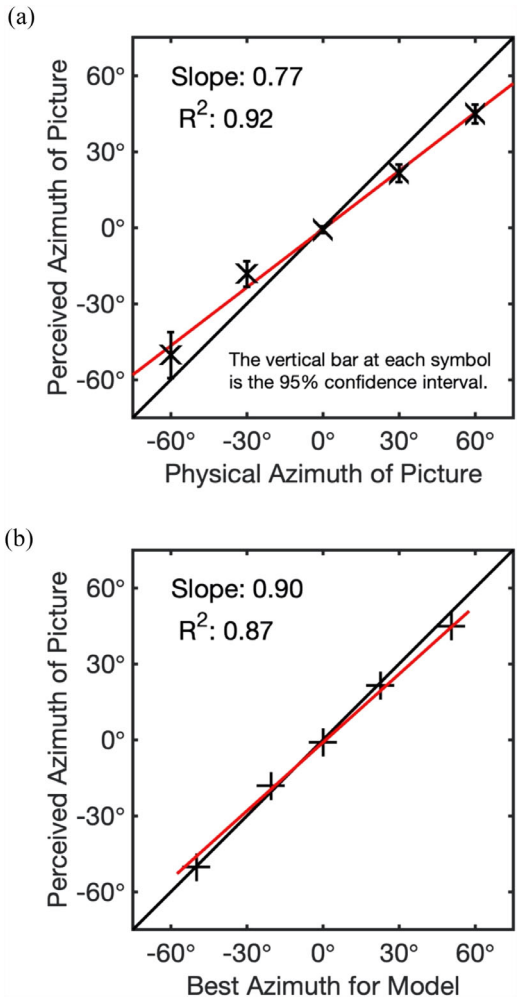

Figure 6. Display azimuth underestimation. (a) Measured azimuth versus physical azimuth of the screen, showing underestimation in every oblique viewing condition, reflected by slope of 0.77. (b) Measured azimuth versus best fitting azimuth parameter for the model, showing concordance increase of slope to 0.90.

Discussion

Many authors have argued that picture perception is different from perception of real scenes, and pictorial space is different from real space (Kennedy, 1974; Ward, 1976; DeLoache et al., 1998; Yang & Kubovy, 1999; Niederée and Heyer, 2003; Vishwanath et al., 2005; Koenderink et al., 2004; Koenderink et al., 2011; van Doorn et al., 2012; Pagel, 2017). Our approach to this issue is empirical and conceptual, as we examine visual tasks that are common to perceiving real and pictured scenes, and identify geometric operations that are involved in both perceptions. In judging poses of 3D objects, we found that observers judged poses in pictures by applying the same back-transform to retinal projections that they did for real scenes, and this predicted the illusory rotation of scenes to match the observer's viewing azimuth. Correcting for the slant of the picture would not predict the illusory rotation, so for the case of object pose we see no reason to invoke special processes for picture perception. This study asks whether that is also true for judging 3D sizes in pictures. Although poses in pictures were invariant to viewing azimuth, distortions of lengths and proportions were obvious. Display slant reduces perceived sizes of frontal objects, and makes parallelepipeds pointing at the observer look narrower, which can be seen by comparing oblique to frontal views in Figures 1 and 4. Consequently, judgments of size and proportion in pictures, which incorporate inferences about the picture's slant, are qualitatively different from judgments of poses, and thus provide new considerations in picture perception.

The main empirical contribution of this article is to measure the perceived size of 3D objects at different poses depicted on a planar picture rotated around its vertical axis. Pose variation and display slant both shorten the projected length of 3D objects, but along orthogonal dimensions, pose variations along the line of sight, and display slant along the frontoparallel axis. Our results are consistent with the hypothesis that the visual system uses the geometric back-transform to overcome these distortions but falls short because it overestimates slant elevation for poses along the line of sight and underestimates the azimuth angle of the display. The same factors lead to perceived changes in aspect ratios of shapes, so that objects pointing toward the observer look narrower in oblique views of pictures.

The frontoparallel bias for the screen seems similar to the frontoparallel bias for 3D objects (Koch et al., 2018). This bias has been reported previously for horizontal slant of planar surfaces (Erkelens, 2013) and vertical slant of oblique objects in natural scenes (Kim & Burge, 2018). Could it be based on adaptation to orientation statistics of natural scenes? Measured statistics of 2D orientations in pictures of naturalistic scenes (Hansen & Essock, 2004; Howe & Purves, 2005; Girshick, Landy, & Simoncelli, 2011) show the highest frequency for horizontal orientations, which are known from geometry to be projections of frontoparallel 3D orientations of objects lying on the ground. If prior beliefs of 3D horizontal slants are based on frequency of occurrence, then in Bayesian inference, this prior probability will shift perceived poses toward the frontoparallel. In addition, a bias towards seeing horizontal 2D orientations (Ding et al., 2017) could arise from anisotropy in populations of cortical orientation selective cells that are largest in number for horizontal orientations and most sharply tuned for horizontal orientations (Cohen & Zaidi, 2007).

The main theoretical contribution of this study is to link perception of 3D sizes in scenes and pictures to the mental use of projective geometry. Corrections of sizes from projective distortions as a function of pose form a curve that has the same shape as that predicted by the optimal back-transform, and the optimal correction expression fits measured estimates with just two free parameters multiplying the viewing elevation by greater than 1.0 and the viewing azimuth by less than 1.0. Thus our model that incorporates observers' misestimates of object slant and display slant can explain the inconstancy of relative size for different poses, object sizes, and viewpoint azimuths, suggesting that the mental use of projective geometry is common to all observers.

These empirical and theoretical results make it tractable to probe the neural mechanisms that achieve this task. Judging poses and sizes of objects is important for both scene understanding and navigation, and for object perception and manipulation. Cortical areas involved are thus likely to include object-selective occipitotemporal regions, including the lateral occipital cortex and posterior fusiform, especially sites where neuronal responses are not viewpoint invariant, as well as regions in the intraparietal sulcus and frontal cortex implicated in tool use and action planning (Yildirim et al., 2019), plus parahippocampal, retrosplenial, and occipital place areas (Epstein & Baker, 2019). In a process framework, circuits for causal generative models could incorporate knowledge from projective geometry, as could circuits for planning, possibly involving dynamic belief updates with recognition models. In such a model, the brain could first estimate distance, pose, azimuth, elevation, and then run the back-transform to obtain the size, but pose, size, and elevation are all computed from the same retinal image of the object, where they interact, so a simultaneous estimation process may be more likely. On the other hand, the human brain has an enormous number of neurons, so although template based grand-mother cells are an inelegant formal model, they need to be taken seriously for decoding (Zaidi & Conway, 2019), especially since experiments with deep learning networks suggest that the brain could easily learn hidden units that could directly translate 2D orientations and sizes in retinal images to 3D poses and sizes in real scenes (Eslami et al., 2018).

Animals and humans have constant exposure to perspective projection through image-forming eyes. Therefore, whether brains have learned to exploit projective geometry to understand real scenes is a longstanding question, e.g., Plato's dialogue Meno. We have now shown that humans use optimal projective geometry back-transforms from retinal images to estimate 3D pose and size in real scenes and continue to use the back-transforms to estimate 3D pose and size in pictures, despite the extra distortions created by oblique views of pictures (Koch et al 2018; Maruya & Zaidi, 2020). Despite the specific neural mechanisms for perceiving object pose and size still to be systematically identified, our results do provide accumulating evidence that human brains have internalized particular aspects of projective geometry through evolution or learning.

Keywords: mental geometry, picture perception, 3D size perception, 3D shape perception, projective geometry

Acknowledgments

Supported by National Institutes of Health Grants EY13312 and EY07556.

Commercial relationships: none.
Corresponding author: Qasim Zaidi.
Email: qz@sunyopt.edu.
Address: SUNY College of Optometry, 33 West 42nd St., New York, NY 10036, USA.

References

Beusmans, J. M. (1998). Optic flow and the metric of the visual ground plane. *Vision Research, 38*(8), 1153–1170.

Boring, E. G. (1964). Size-constancy in a picture. *The American Journal of Psychology, 77*(3), 494–498.

Brunswik, E. (1944). Distal focusing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs, 56*(1), i–49.

Carlson, V. R. (1960). Overestimation in size-constancy judgments. *The American Journal of Psychology, 73*(2), 199–213.

Cohen, E. H., & Zaidi, Q. (2007). Fundamental failures of shape constancy resulting from cortical anisotropy. *Journal of Neuroscience, 27*(46), 12540–12545.

Cutting, J. E. (1987). Rigidity in cinema seen from the front row, side aisle. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 323.

DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengren, K. S., & Gottlieb, A. (1998). Grasping the nature of pictures. *Psychological Science, 9*(3), 205–210.

Ding, S., Cueva, C. J., Tsodyks, M., & Qian, N. (2017). Visual perception as retrospective Bayesian decoding from high-to low-level features. *Proceedings of the National Academy of Sciences, 114*(43), E9115–E9124.

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science, 5*, 373–397.

Erkelens, C. J. (2013). Virtual slant explains perceived slant, distortion, and motion in pictorial scenes. *Perception, 42*(3), 253–270.

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., & Garnelo, M., ...Reichert, D. P. (2018). Neural scene representation and rendering. *Science, 360*(6394), 1204–1210.

Gilinsky, A. S. (1951). Perceived size and distance in visual space. *Psychological Review, 58*(6), 460.

Gilinsky, A. S. (1955). The effect of attitude upon the perception of size. *The American Journal of Psychology, 68*(2), 173–192.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*(7), 926–932.

Gombrich, E. H. (1972). The visual image. *Scientific American, 227*(3), 82–97.

Hagen, M. A. (1974). Picture perception: Toward a theoretical model. *Psychological Bulletin, 81*(8), 471.

Hagen, M. A. (1976). Influence of picture surface and station point on the ability to compensate for oblique view in pictorial perception. *Developmental Psychology, 12*(1), 57.

Hansen, B. C., & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision, 4*(12), 5–5.

Howe, C. Q., & Purves, D. (2005). Natural-scene geometry predicts the perception of angles and line orientation. *Proceedings of the National Academy of Sciences, 102*(4), 1228–1233.

Kennedy, J. M. (1974). *A psychology of picture perception*. San Francisco: Jossey-Bass Publishers.

Kim, S., & Burge, J. (2018). The lawful imprecision of human surface tilt estimation in natural scenes. *Elife, 7*, e31448.

Koch, E., Baig, F., & Zaidi, Q. (2018). Picture perception reveals mental geometry of 3D scene inferences. *Proceedings of the National Academy of Sciences, 115*(30), 7807–7812.

Koenderink, J. J., van Doom, A. J., Kappers, A. M., & Todd, J. T. (2004). Pointing out of the picture. *Perception, 33*(5), 513–530.

Koenderink, J. J., van Doorn, A. J., & Wagemans, J. (2011). Depth. *i-Perception, 2*(6), 541–564.

Loomis, J. M., Da Silva, J. A, Fujita, N., & Fukusima, S. S. (1992). Visual space perception and visually directed action. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 906.

Loomis, J. M., & Philbeck, J. W. (1999). Is the anisotropy of perceived 3-D shape invariant across scale?. *Perception & Psychophysics, 61*(3), 397–402.

Loomis, J. M., Philbeck, J. W., & Zahorik, P. (2002). Dissociation between location and shape in visual space. *Journal of Experimental Psychology: Human Perception and Performance, 28*(5), 1202.

Maruya, A., & Zaidi, Q. (2020). Mental geometry of three-dimensional size perception. *Journal of Vision, 20*(8), 14–14.

Niall, K. K., & Macnamara, J. (1990). Projective invariance and picture perception. *Perception, 19*(5), 637–660.

Niederée, R., & Heyer, D. (2003). The dual nature of picture perception: A challenge to current general accounts of visual perception. *Looking into pictures: An interdisciplinary approach to pictorial space,* 77–98.

Norman, J. F., Todd, J. T., Perotti, V. J., & Tittle, J. S. (1996). The visual perception of three-dimensional length. *Journal of Experimental Psychology: Human Perception and Performance, 22*(1), 173.

Pagel, R. (2017). The duality of picture perception and the robustness of perspective. *Art & Perception, 5*(3), 233–261.

Perkins, D. N. (1973). Compensating for distortion in viewing pictures obliquely. *Perception & Psychophysics, 14*(1), 13–18.

Rosinski, R. R., Mulholland, T., Degelman, D., & Farber, J. (1980). Picture perception: An analysis of visual compensation. *Perception & Psychophysics, 28*(6), 521–526.

Ross, H. E., & Plug, C. (1998). The history of size constancy and size illusions. In V. Walsh, & J. Kulikowski (Eds.), *Perceptual constancy: Why things look as they do* (p. 499–528). Cambridge, UK: Cambridge University Press.

Sedgwick, H. A. (1989) The effects of viewpoint on the virtual space of pictures. Available at https://ntrs.nasa.gov/search.jsp?R=19900013616. Last accessed July 25, 2020.

Todorović, D. (2008). Is pictorial perception robust? The effect of the observer vantage point on the perceived depth structure of linear-perspective images. *Perception, 37*(1), 106–125.

van Doorn, A. J., Koenderink, J. J., Leyssen, M. H., & Wagemans, J. (2012). Interaction of depth probes and style of depiction. *i-Perception, 3*(8), 528–540.

Vishwanath, D., Girshick, A. R., & Banks, M. S. (2005). Why pictures look right when viewed from the wrong place. *Nature neuroscience, 8*(10), 1401.

Wallach, H., & Marshall, F. J. (1986). Shape constancy in pictorial representation. *Perception & Psychophysics, 39*(4), 233–235.

Ward, J. L. (1976). The perception of pictorial space in perspective pictures. *Leonardo, 9*(4), 279–288.

Yang, T., & Kubovy, M. (1999). Weakening the robustness of perspective: Evidence for a modified theory of compensation in picture perception. *Perception & Psychophysics, 61*(3), 456–467.

Yildirim, I., Wu, J., Kanwisher, N., & Tenenbaum, J. (2019). An integrative computational architecture for object-driven cortex. *Current Opinion in Neurobiology, 55*, 73–81.

Zaidi, Q., & Conway, B. (2019). Steps towards neural decoding of colors. *Current Opinion in Behavioral Sciences, 30*, 169–177.

Appendix: Derivation of 2D retinal lengths from 3D lengths of object at poses

This projection is derived for the blue parallelepiped lying on the ground plane in 3D-Space (XYZ space, where the X-Z face is on the top surface of the parallelepiped and the Y-axis is the center of the rotation) and extending from the center of the scene $(0, 0, 0)$ to $(x, 0, z)$. The end-point $(x, 0, z)$ can be expressed in terms of the physical length (L_{3D}) and pose (Ω) of the parallelepiped:

$$(x, 0, z) = (L_{3D} \cdot \cos(\Omega), 0, L_{3D} \cdot \sin(\Omega)). \quad \text{(A1)}$$

Since the camera elevation, Φ_c, from the Z-axis is equivalent to a rotation around the X-axis, the rotation of the end-point $(x, 0, z)$ is given by:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\Phi_c) & -\sin(\Phi_c) \\ 0 & \sin(\Phi_c) & \cos(\Phi_c) \end{pmatrix} \begin{pmatrix} x \\ 0 \\ z \end{pmatrix}. \quad \text{(A2)}$$

So that:

$$\begin{aligned} x' &= x \\ y' &= -z\sin(\Phi_c). \\ z' &= z\cos(\Phi_c) \end{aligned} \quad \text{(A3)}$$

The center point stays the same, $(0, 0, 0) \rightarrow (0, 0, 0)$.

In the projection to the Picture Plane (UV-Space), the central point is mapped as $(0, 0, 0) \rightarrow (0, 0)$, and if d_c is the distance from the camera's aperture to the center of the picture plane, the coordinates of the end-point are:

$$\begin{aligned} u &= \frac{x'}{d_c - z'} \cdot f_c \\ v &= \frac{y'}{d_c - z'} \cdot f_c \end{aligned}. \quad \text{(A4)}$$

So the projected length on the Picture Plane is given by:

$$L_c = \sqrt{u^2 + v^2}. \quad \text{(A5)}$$

By substituting Equations (A1), (A3), and (A4) into (A5)

$$\begin{aligned} L_c &= \sqrt{\left(\frac{x'}{d_c - z'} f_c\right)^2 + \left(\frac{y'}{d_c - z'} f_c\right)^2} \\ &= \frac{L_{3D} f_c \sqrt{\cos^2(\Omega) + \sin^2(\Omega)\sin^2(\Phi_c)}}{d_c - L_{3D}\ \sin(\Omega)\cos(\Phi_c)} \end{aligned}. \quad \text{(A6)}$$

Finally the Picture Plane is projected to the retinal plane (RS-space). Because the Picture Plane is slanted

by ϕ_v ($\pm 30^\circ$ *or* $\pm$ 60°), while the observer's viewing position is fixed, the new coordinates are defined in 3D-space (UVW space, where the W axis is orthogonal to the frontoparallel location of the screen, adding depth to the 2D space defined by the frontoparallel picture plane. The central point is mapped as $(0, 0) \rightarrow (0, 0, 0)$. The new coordinates for the end-point are:

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = \begin{pmatrix} \cos(\phi_v) & 0 & -\sin(\phi_v) \\ 0 & 1 & 0 \\ \sin(\phi_v) & 0 & \cos(\phi_v) \end{pmatrix} \begin{pmatrix} u \\ v \\ 0 \end{pmatrix}.$$

Giving:

$$\begin{aligned} u' &= u\cos w' = u\sin(\phi_v)(\phi_v) \\ v' &= v \\ w' &= u\sin(\phi_v) \end{aligned} \quad . \quad \text{(A7)}$$

If d_v is the distance from the observer's pupil to the center of the picture plane, and f_v is the observer's focal length, the endpoint's projection from UVW-space to retinal space is:

$$\begin{aligned} r &= \tfrac{u'}{d_v - w'} \cdot f_v \\ s &= \tfrac{v'}{d_v - w'} \cdot f_v \end{aligned}. \quad \text{(A8)}$$

Therefore the projected length on the retina:

$$L_r = \sqrt{r^2 + s^2} \quad \text{(A9)}$$

is derived by substituting Equations (A1), (A3), (A4), (A7), and (A8) into (A9):

$$\begin{aligned} L_r &= \sqrt{\left(\tfrac{u'}{d_v - w'} f_v\right)^2 + \left(\tfrac{v'}{d_v - w'} f_v\right)^2} \\ L_r &= \tfrac{L_{3D} f_v f_c \sqrt{(\cos(\Omega)\cos(\phi_v))^2 + (\sin(\Omega)\sin(\Phi_c))^2}}{d_v(d_c - L_{3D}\sin(\Omega)\cos(\Phi_c)) - L_{3D} f_c \cos(\Omega)\sin(\phi_v)} \end{aligned} \quad \text{(A10)}$$

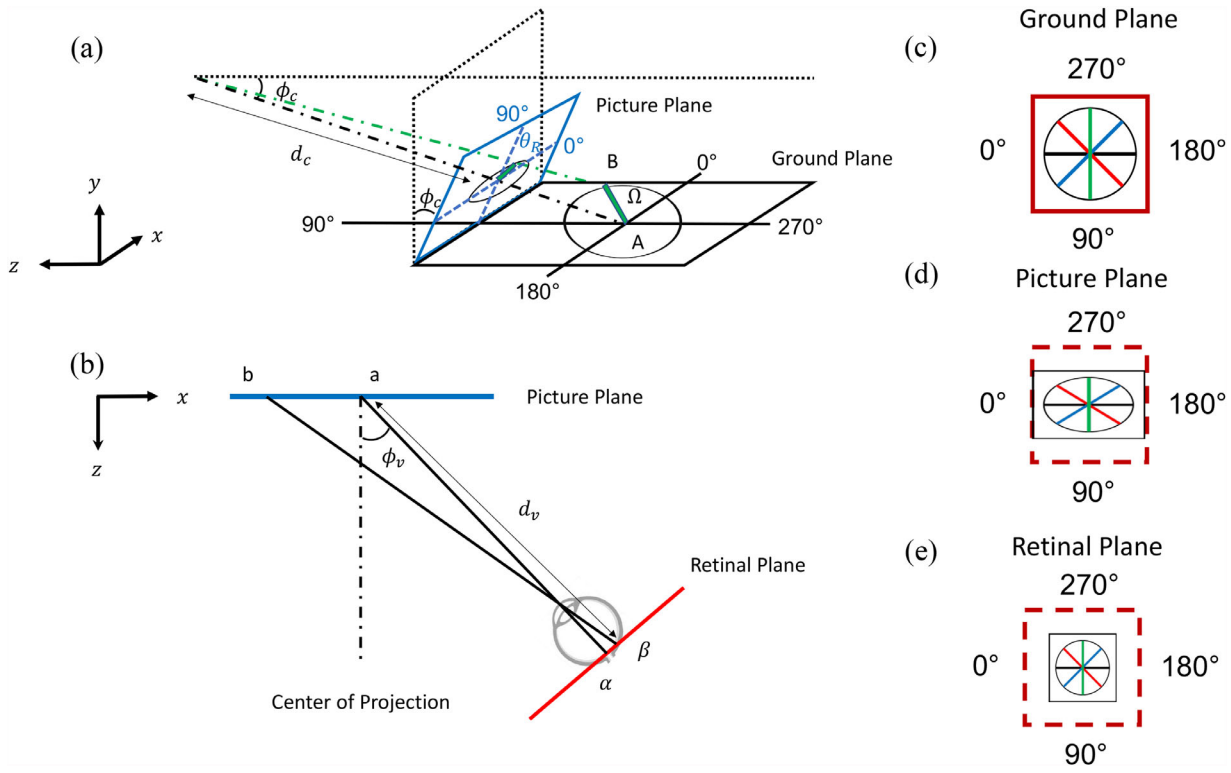

Figure A1. Derivation of projected length as a function of object pose and viewing azimuth (a) Equal lengths of objects on the ground project to different lengths on the screen as a function of pose, as a circle is compressed into an ellipse. Comparing panel (c) to (d) shows greatest reduction for line of sight poses. (b) When the screen is viewed obliquely, in the retinal projection, the ellipse is compressed towards a circle again. Comparing panel (d) to (e) shows the reduction in length is greatest for frontoparallel poses.

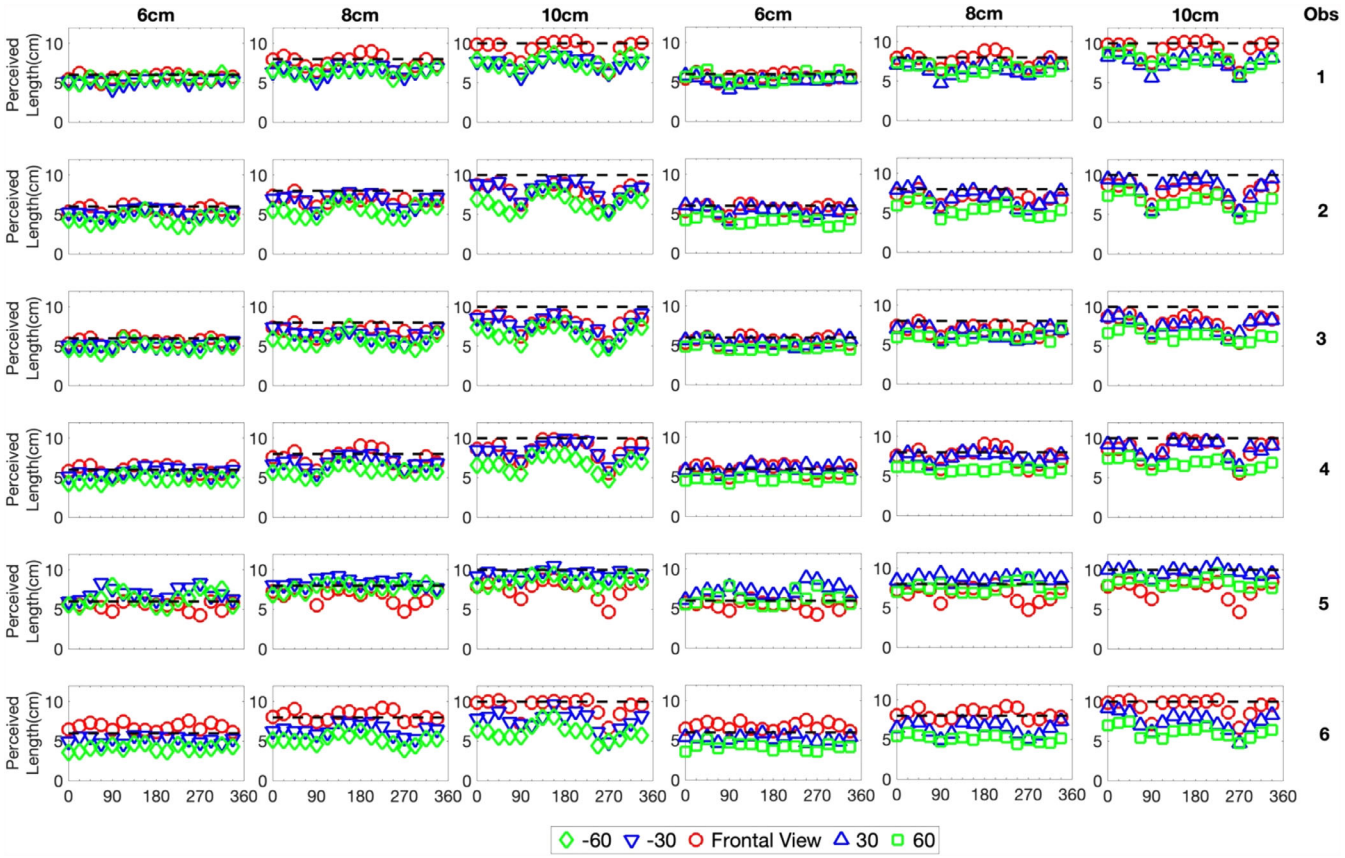

Figure A2. Perceived 3D lengths across pose and viewing azimuth for six observers. Each column represents estimated lengths of a parallelepiped of the indicated physical length. Left three panels are for negative oblique viewing azimuths and right for positive. Points show perceived length across 3D pose. Dashed lines indicate physical 3D length. Individual observer's data are plotted on separate rows. Underestimation of perceived length increases systematically with increased azimuth of the display.

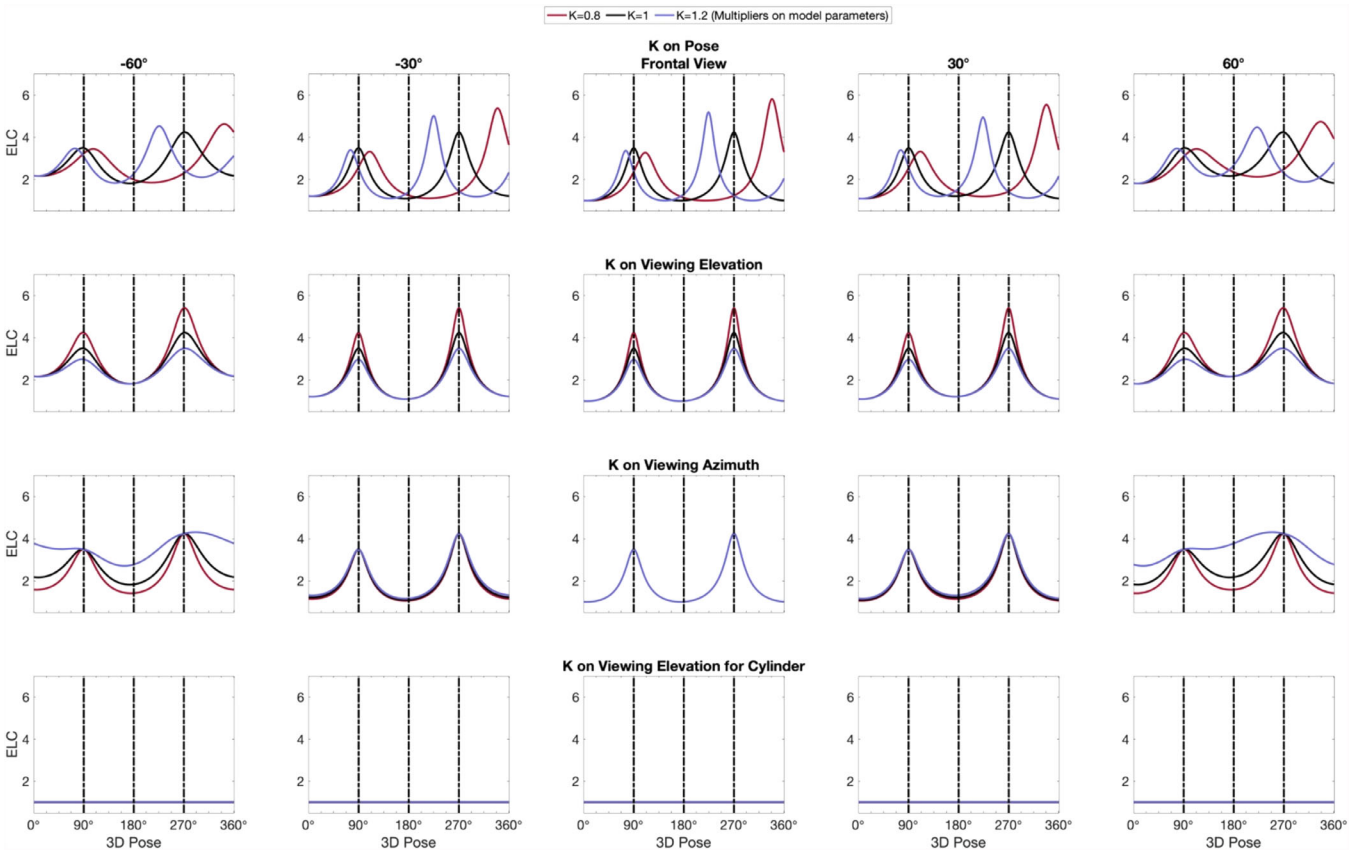

Figure A3. Effect of multipliers on object pose, viewing elevation, and viewing azimuth on predicted length correction. Curves for K = 1.0 represent the optimal length correction. Curves for K = 0.8 and 1.2 represent empirical length correction if a parameter is under- or overestimated. Under- or overestimating perceived pose shifts the peak of the ELC, which would not be consistent with the results. Misestimating viewing elevation (or equivalently slant elevation of the object) differentially affects size estimates of objects close to line of sight, consistent with the results for all viewpoints. Misestimating viewing azimuth affects frontoparallel poses most, and essentially only for the most oblique views, which is consistent with the results for slanted displays.

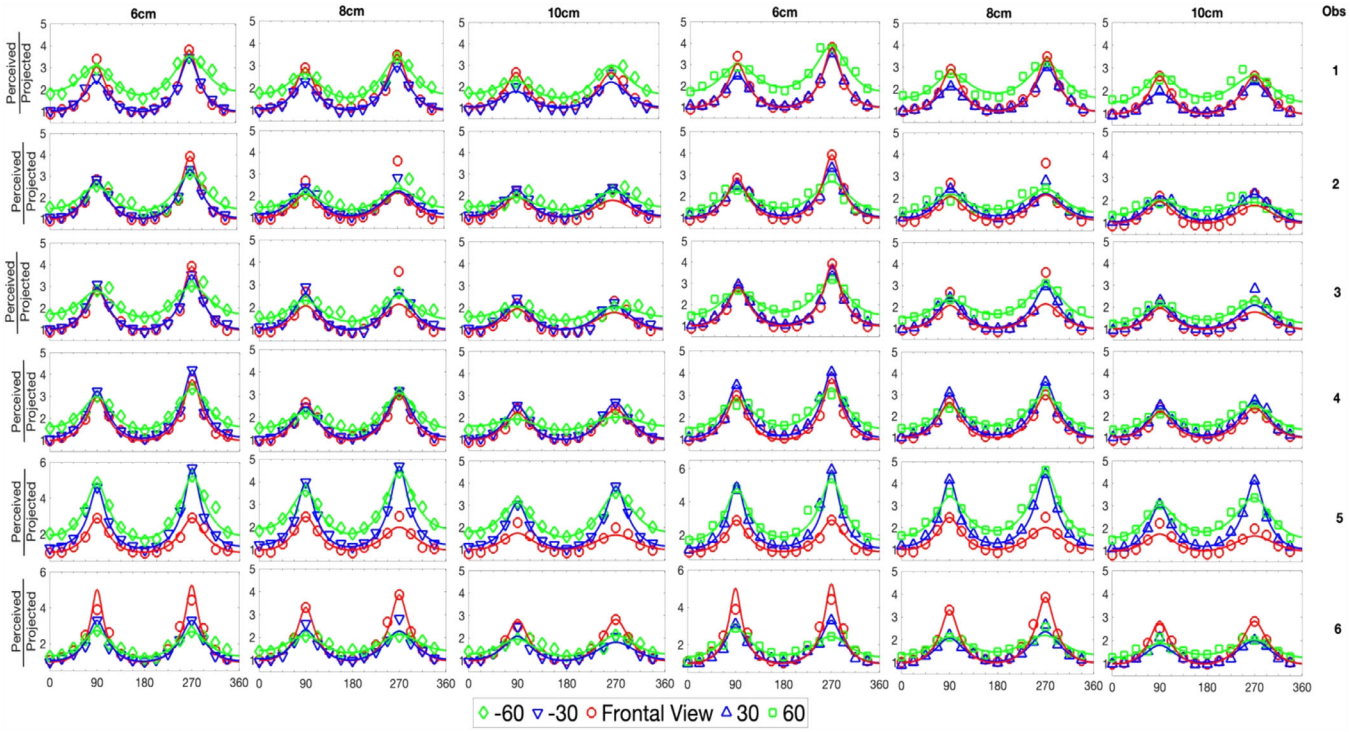

Figure A4. Model fits for 3D length estimation across pose and viewing azimuth for six observers. Each column represents estimated lengths of a parallelepiped of the indicated physical length. Left three panels are for negative oblique viewing azimuths and right for positive. Individual observer's data are plotted on separate rows. A model using the optimal geometrical back-transform, but incorporating underestimation of the azimuth of the display, fits the correction of perceived length over the projected length.

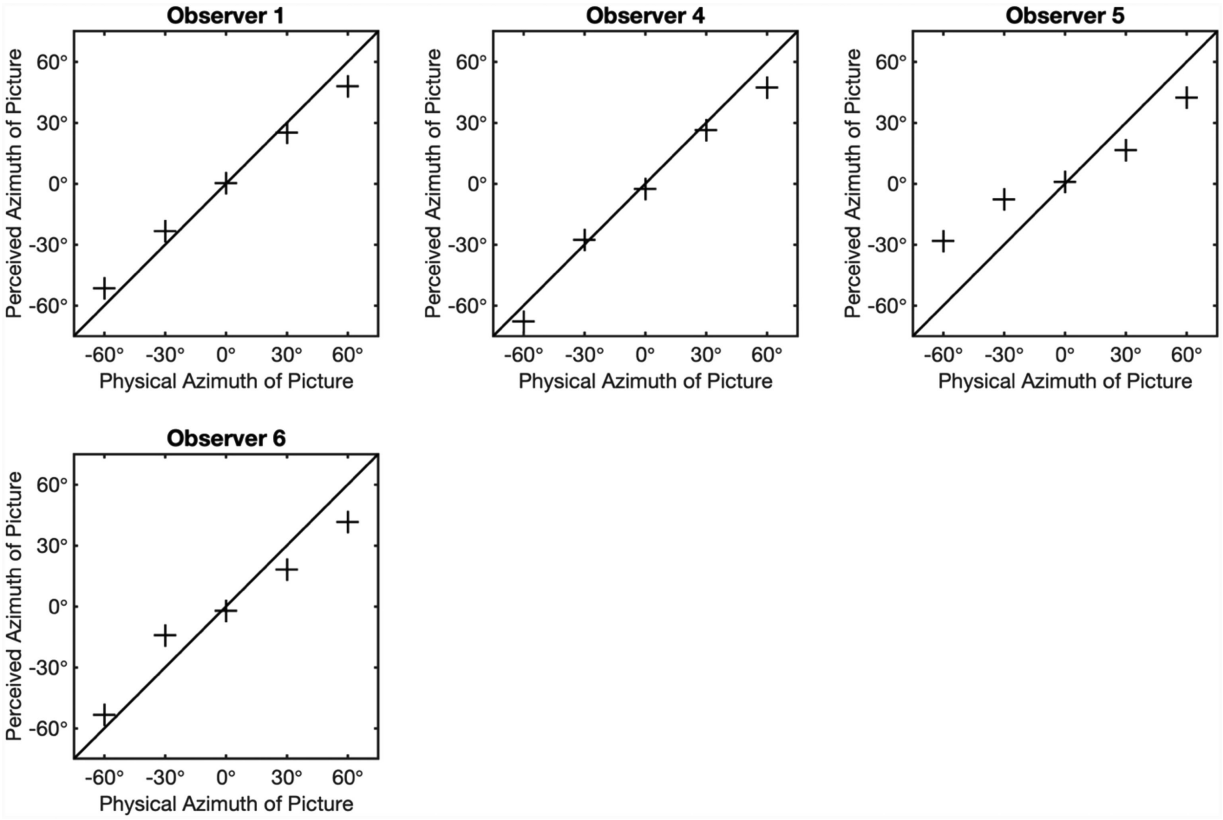

Figure A5. Perceived display azimuth for four observers. Perceived display azimuth as a function of physical azimuth of monitor for six observers. In almost every oblique viewing condition, there is underestimation of the azimuth of the display.