Complexity of mental geometry for 3D pose perception

By
Crystal Guo

Thesis
In partial satisfaction of the requirements for the degree of
Master of Science in Vision Science
State University of New York
College of Optometry
May 2022

Approved by Master’s Thesis Committee:

[Signature]

Graduate Research Advisor: Qasim Zaidi

Committee Chair: Jose-Manuel Alonso

[Signature] [Signature]

Committee Member: Harold Sedgwick
Committee Member: Robert McPeek

Associate Dean’s Signature: Stewart Bloomfield, PhD
Associate Dean for Graduate Studies and Research
Complexity of mental geometry for 3D pose perception

Crystal Guo1, Akihito Maruya1, & Qasim Zaidi1*

1Graduate Center for Vision Research, State University of New York, 33 West 42nd St, New York, NY 10036.
cguo@sunyopt.edu
amaruya@sunyopt.edu
qz@sunyopt.edu *Corresponding author

AUTHOR CONTRIBUTIONS
QZ designed the study. CG programmed and ran the experiments. CG, AM, and QZ analyzed and modeled the results. CG & QZ wrote the paper.
Abstract

Biological visual systems rely on pose estimation of three-dimensional (3D) objects to understand and navigate the surrounding environment, but the neural computations and mechanisms for inferring 3D poses from 2D retinal images are only partially understood, especially for conditions where stereo information is insufficient. We previously presented evidence that humans use the geometrical back-transform from retinal images to infer the poses of 3D objects lying centered on the ground. This model explained the almost veridical estimation of poses in real scenes and the illusory rotation of poses in obliquely viewed pictures, including the pointing at you phenomenon. Here we test this model for 3D objects in more varied configurations and find that it needs to be augmented. Five observers estimated poses of inclined, floating, or off-center 3D sticks in each of 16 different poses displayed on a monitor viewed straight or obliquely. Pose estimates in scenes and pictures showed remarkable accuracy and agreement between observers, but with a systematic fronto-parallel bias for oblique poses. When one end of an object is on the ground while the other is inclined up, the projected retinal orientation changes substantially as a function of inclination, so the back-transform derived from the object's projection to the retina is not unique unless the angle of inclination is known. We show that observers' pose estimates can be explained by the back-transform from retinal orientation only if it is derived for close to the correct inclination. The same back-transform explanation applies to obliquely viewed pictures. There is less change in retinal orientations when objects are floated or placed off-center but pose estimates can be explained by the same model, making it more likely that observers use internalized perspective geometry to make 3D pose inferences while actively incorporating inferences about other aspects of object placement.

Keywords
- Pose estimation
- Retinal orientation
- Projective geometry
- Picture perception
1. Introduction

The human brain rapidly makes countless 3D pose estimations. Biological and machine vision systems must accurately perceive object poses to understand the world around them. For example, rapid pose estimation is key in contact sports to judge players’ positions and avoid traumatic collisions, and in locomotion over complex terrain (Fig. 1). The pose estimation problem has been extensively researched in machine vision, especially for the cases of poses of human bodies, heads, and hands, which are critically needed for designing human-computer interfaces and other applications (Cao et al., 2021; Newell et al., 2016; Toshev & Szegedy, 2014; Yang & Ramanan, 2011; Yao et al., 2011). These engineering approaches necessarily need to solve the problem in its complete generality and have used multiple techniques from template matching to deep neural networks to attack this difficult problem. The problem has not been addressed to the same extent for biological or human visual systems, so there is scant information on performance, and almost none on neural bases of the performance. Machine vision algorithms for pose estimation necessarily have to incorporate scene understanding and object recognition. Humans are extremely good at rapid scene analysis and object recognition, so human psychophysics presents an opportunity to study pose estimation separately from other tasks and to break the problem into sub-tasks requiring simpler to more complex strategies. This could enable building increasingly general models by comparing performance to optimal models as the task is made more multifaceted.

Figure 1. Depiction of pose estimation in real life scenarios. The picture on the left is an image of a group of people playing ultimate frisbee, in which human limb pose estimation becomes important to predict movements and avoid collisions (“Ultimate frisbee” by Stefano Zocca, licensed under Unsplash: https://unsplash.com/photos/dG-t92hPCKA). The picture on the right is an image of a steep set of steps in which object pose estimation is critical for locomotion (“The Climb” by George Hiles, licensed under Unsplash: https://unsplash.com/photos/qUlnWhV2LE).

Following this strategy, Koch, Baig & Zaidi (2018) studied the simple case of perceived poses of a parallelepiped lying centered on a horizontal ground. They found that all observers were very accurate in judging poses, but they all had a slight fronto-parallel bias for oblique poses. Given that the perspective mapping from the pose on the ground to orientation on the retina is 2D to 2D and
Figure 2. Projected images of stimuli. Blue rectangular 3D parallelepiped (test stick) of fixed length with one end lying on the center of a yellow disc and the other end inclined at 45° (Top); test stick of fixed length floating 7.5 cm above the center of a yellow disc with a supporting vertical yellow cylinder (Middle); test stick of fixed length displaced 12 cm to the left of the center of a yellow disc (Bottom). Blue parallelepipeds were presented in one of 16 poses from 0° to 360° spaced 22.5° apart. The line of sight through the center of the ground is designated by the 90°–270° axis, and the line orthogonal to it as the 0°–180° axis. Parallelepiped lengths were 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 the frontal viewpoint (0°).
the analytic expression is invertible, they reasoned that for the most accurate estimates of 3D poses from the information contained in retinal images, a visual system should apply the geometrical back-transform from known retinal orientations to unknown object poses. For every observer, this model explained their accuracy with just one free parameter for the fronto-parallel bias. The critical test was provided by pose estimates made while viewing pictures of the same scenes but moving the viewpoint to an oblique angle. Pose estimates made by observers were rotated by exactly the angle of oblique viewing, so that objects seen before as pointing towards the observer were still seen as pointing towards the observer, even though the observer’s position had changed, a phenomenon termed “pointing out of the picture” (Koenderink et al., 2004). They showed that this illusory rotation could be completely explained by assuming that observers were using the back-transform as for real scenes, with pose being inferred in viewer centered coordinates, thus providing strong support to the idea that observers use internalized perspective geometry to infer poses in the world relative to their position.

Our goal for this study is to generalize pose estimation by a critical step. We still used the simple parallelepipeds but now they could be inclined with one end elevated, floating above and parallel to the ground, or on the ground but placed off-center. To test the role of naturalistic cues in these positions, we tested configurations with and without shadows cast by the objects. We compared performance to the optimal back-transform function derived for each configuration. As before, we also tested if 3D pose estimates in oblique views of pictures could be explained by observers using the scene back-transform.

2. Materials and Methods

Using Blender, 3D scenes with single 6 cm x 3 cm x 3 cm blue rectangular parallelepipeds were created for each of three conditions: inclined, floating, and off-center (Fig. 2). The camera was positioned 1 meter from the center of the scene at an elevation of 15°. The ground plane was delineated by a yellow disc.

In the inclined condition, one end of the blue stick was placed at the center of a 12 cm yellow disc on the ground plane, while the other was elevated to create an inclination of 45° angle from the ground plane.

In the floating condition, the blue stick was held up by a vertical, yellow cylinder that was positioned in the center of a 12 cm yellow disc on the ground plane. The height of the yellow cylinder was 6 cm; consequently, the center of rotation for the blue stick was 7.5 cm above the center of the scene.

In the off-center condition, the blue stick was presented either 12 cm to the left or to the right of the center of the scene. A yellow disc with a diameter of 36 cm was positioned on the ground plane, its center, delineated by a red cross, coinciding with the center of the scene.

In the floating and inclined stick scenes, two conditions were presented: scenes with shadows and scenes without. The scenes without shadow were created by adjusting the lamp type to ‘Hemi’ with energy level 4 and including only diffuse reflections. Scenes with shadow were created with a ‘Point’ lamp at
energy 2, also including only diffuse reflections. The lighting conditions for the off-center condition corresponded to the settings used in the floating and inclined scenes with shadows. For all scenes, the camera and lamp positions remained constant and the surround was completely dark.

Six types of scenes were presented: an inclined stick with shadow, an inclined stick without shadow, a floating stick with shadow, a floating stick without shadow, an off-center stick on the left, and an off-center stick on the right. In all six conditions, the blue stick was rotated by 22.5° for 16 poses spanning 360° (Fig 2). Additionally, each pose for each condition was repeated 3 times, totaling 48 images per scene, or 96 images per condition. Altogether, each subject made 288 judgments for all three conditions in one session. The scenes were displayed on a 22-inch DELL SP2309W Display monitor. Subjects viewed the vertical monitor from a distance of 1 meter with their heads placed on a chin rest such that the viewing level was 15° above the center of the scene. The stimulus display, response vector, and data analyses used MATLAB.

Observers recorded their judgement by rotating a vector in a clock face on a horizontal 11” tablet to match the perceived pose of the blue stick in the 3D scene. Observers viewed this response touchscreen from an elevation of 70°. The experiment was blocked by the three conditions in random order, and images were displayed in random order as well. There was unlimited time to make judgments as well as unlimited break time between each condition. Every observer completed the study in one session.

Five observers with normal or corrected to normal vision participated in the study. All observers provided their written, informed consent to participate in the study. The study was approved by the SUNY College of Optometry Institutional Review Board in accordance with the Declaration of Helsinki.
3. Results

A

![Graph A](image1)

B

![Graph B](image2)

C

![Graph C](image3)

**Figure 3.** Perceived 3D pose against physical 3D pose. Left three columns show results for the three observer viewing angles: -60°, 0° or frontal, and +60°. Rows represent different stimulus conditions: A) inclined condition, B) floating condition, and C) off-center condition. Each large blue dot is the average estimation for a given pose across all observers, each small blue dot represents a single pose estimation, and the black line is the unit diagonal. Right column shows perceived 3D pose for stimuli with shadow vs. stimuli without shadow, and stimuli with left stick vs. stimuli with right stick. R² values were 0.98888, 0.99082, and 0.99131 for each of the conditions respectively. Each blue point is one pose estimate by an observer and the black line is the line of unity from which R² was measured.

The results for 5 observers are presented in Figure 3, with the average as large blue dots, and individual observer’s responses as small blue dots. Individual observers’ data are presented separately in Figures A1, A2, and A3 in the Appendix. Perceived 3D pose plotted against the physical 3D pose for the 0° viewing angle shows that observers judge pose almost veridically for all three configurations, except for a slight fronto-parallel bias where oblique poses are seen at shallower angles than veridical. The fronto-parallel bias is more pronounced for poses that are pointing towards the observer, i.e. poses from 0°
to 180°. The results for oblique viewing angles of -60° and +60° away from the frontal viewing angle show that observers are very accurately estimating poses, except that the perceived pose is shifted up or down by an angle equal to the viewing angle, implying a rigid illusory rotation of the scene.

The panels left column show that there is very little difference in observer estimates whether the stimuli are presented with shadow or without, as in the inclined and floating conditions, and whether the stimuli are presented with the stick on the left or the right, as in the off-center condition. Therefore, the inclusion of shadows in the stimuli does not seem to alter the observers' perception of pose. Similarly, for off-center sticks, displacing the origin of the stick by the same number of units to the right or to the left of the center does not affect observer perception. For the analyses, we combined the data across shadow and offset conditions for each stimulus configuration.
4. Analyses

Figure 4. (A) (Left) A depiction of the projection from the 3D scene to the picture plane. (Right) Top-down view of the ground plane, above a frontal view of the picture plane. The brown line is the projected image of a stick at an oblique pose. These plots show how a circle on the ground plane would become vertically compressed in the picture plane, and how the orientation of a stick extending from the center of the circle on the ground plane becomes compressed towards the horizontal axis in the picture plane. (B) (Left) A depiction of the projection from the picture plane to the retinal plane for oblique viewpoints. (Right) The same frontal view of the picture plane as in (A) and a frontal view of the retinal plane (resulting from oblique viewing). The retinal plane becomes horizontally compressed relative to the picture plane giving the illusion of vertical elongation and a tilting of the ground plane. (C) Directed vector on horizontal touch screen used to report perceived 3D pose, set by touching a location along the blue circle and fine tuned using keyboard arrows. (D) Top-down view of the response coordinate space (0° to 180° axis is fronto-parallel to the observer), and 3 locations from which the observer viewed the screen.
Information about the pose is contained in the retinal orientation of the image of the stick, and the best any system could do at pose estimation is to use the back-transform of the correct projection function to estimate the pose from the retinal image. We used perspective projection (Fig. 4A) to derive the equations for 2D retinal orientation as a function of 3D pose (Koch et al., 2018) as detailed in the Appendix. \( \theta_R \) is the orientation of the stick’s image on the retina, \( \Omega_T \) is the physical 3D pose of the stick on the ground, \( \phi_C \) is the camera elevation, which is fixed at 15°, and \( d_c \) is distance to camera, which is fixed at 1 meter.

For sticks with an angle of inclination \( \phi_E \) from the ground plane:

\[
\theta_R = \tan^{-1}\left( \tan(\Omega_T) \sin(\phi_C) - \frac{\tan(\phi_E) \cos(\phi_C)}{\cos(\Omega_T)} \right)
\]

Eq. (1)

For sticks floating \( H \) units above the center of the scene:

\[
\theta_R = \tan^{-1}(\tan(\Omega_T)(\sin(\phi_C) - (H/d_c)))
\]

Eq. (2)

For off-center sticks on the ground plane extending from \((X_0, Z_0)\), where the origin is at the center of the camera’s line of sight:

\[
\theta_R = \tan^{-1}\left( \frac{\tan(\Omega_T) \sin(\phi_C) d_c}{d_c + \cos(\phi_C)(Z_0 - X_0 \tan(\Omega_T))} \right)
\]

Eq. (3)

To test pose estimation in scenes depicted in pictures viewed from oblique viewing angles, observers were instructed to perform the same experiment while viewing the screen with azimuth angles of \(-60^\circ\) or \(+60^\circ\). The viewing conditions were blocked by azimuth and presented in random order. We derived the following equations for retinal projections from oblique viewing angles, given by \( \phi_V \) (Fig. 4B):

For sticks with an angle of inclination \( \phi_E \) from the ground plane:

\[
\theta_R = \tan^{-1}\left( \frac{\tan(\Omega_T) \sin(\phi_C)}{\cos \phi_V} - \frac{\tan(\phi_E) \cos(\phi_C)}{\cos(\Omega_T) \cos(\phi_V)} \right)
\]

Eq. (4)
For sticks floating H units above the center of the scene:

$$\theta_R = \tan^{-1}(\tan(\Omega_T) \left( \sin(\phi_c) - \frac{H \cos^2(\phi_c)}{\cos(\phi_c) d_c - H \sin(\phi_c)} \right) - \left( \frac{H \cos(\phi_c) \sin(\phi_c) f_c}{d_p \cos(\phi_p) (d_c - H \sin(\phi_c))} \right))$$

Eq. (5)

For off-center sticks on the ground plane extending from \((X_0, Z_0)\), where the origin is at the center of the camera’s line of sight:

$$\theta_R = \tan^{-1}\left( \frac{-\sin(\Omega_T) \sin(\phi_c) (1 + \frac{X_0 \cos(\phi_c)}{d_c - Z_0 \cos(\phi_c)})}{\cos(\phi_p) \frac{X_0 \sin(\Omega_T) \sin(\phi_c) + Z_0 \cos(\Omega_T) \cos(\phi_c) + \frac{X_0 Z_0 (\cos(\phi_c) - \sin(\phi_c))}{d_c - Z_0 \cos(\phi_c)}}{d_p \cos(\phi_p) (d_c - H \sin(\phi_c))} \right)$$

Eq. (6)

Since some of these expressions are not analytically solvable for \(\Omega_T\) as a function of \(\theta_R\), we calculated the back-transforms from plots of the functions set for the configuration parameter. We found previously that the fronto-parallel bias could be explained by adding a parameter \(1/K\) to the term inside the \(\tan^{-1}\) for objects lying on the ground (Koch, Baig & Zaidi, 2018), so when fitting the back-transforms to the data we did the same. Values of \(K\) are given in figure captions.
4.1 Analysis: Inclined objects

A

View: -60 deg

View: 0 deg

View: +60 deg

Retinal Orientation

Physical 3D pose

B

View: -60 deg

K=0.97

View: 0 deg

K=0.91

View: +60 deg

K=0.97

Retinal Orientation

Predicted 3D Pose

C

View: -60 deg

K=0.97

View: 0 deg

K=0.91

View: +60 deg

K= 0.97

Retinal Orientation

Predicted 3D Pose

D

View: -60 deg

View: +60 deg

PhE = 65

Retinal Orientation

E

View: -60 deg

View: 0 deg

View: +60 deg

Perceived Indination (deg)

Pose: 45 deg Pose: 135 deg

Pose: 45 deg Pose: 135 deg

Pose: 45 deg Pose: 135 deg
Figure 5.
(A) Projected retinal orientation against the physical 3D pose for the inclined condition. Columns represent the three observer viewpoints. The green curve is the projected retinal orientation described by Eq. (4). (B) 3D pose against the retinal orientation. The red curve is the back-transform derived from Eq. (4) and represents the predicted 3D pose from the 2D retinal orientation. Each blue dot is the averaged perceived pose estimate corresponding to the retinal orientation of that pose across all observers. The fronto-parallel bias is given by K. (C) 3D pose against retinal orientation for a stick with an inclination angle of 35°, 40°, 45°, 50°, and 55°. Each blue dot is the averaged pose estimate corresponding to the retinal orientation of that pose across all observers. Each colored curve represents the predicted 3D pose from the retinal orientation for a given angle of inclination in the following order: blue for 35°, red for 40°, yellow for 45°, purple for 50°, and green for 55°. For the -60° viewing angle, the RMSE for each curve is 9.8645, 5.6321, 2.4948, 2.6502, and 5.0768 respectively. For the 0° viewing angle, the RMSE for each curve is 13.1852, 8.7633, 5.0911, 4.2393, and 7.0668 respectively. For the +60° viewing angle, the RMSE for each curve is 10.0727, 5.8579, 2.7453, 2.724, and 5.054 respectively. (D) 3D pose against retinal orientation for an inclined stick of 65° assuming the back-transform used for the 0° viewing angle. Each blue dot is the averaged pose estimate corresponding to the retinal orientation of that pose across all observers. (E) Perceived angle of inclination across 4 observers for three viewing positions depicted in boxplots. Observers estimated the angle of inclination for two poses (45° and 135°) from three viewing positions (-60°, 0°, and +60°) when the inclination angle was kept constant at 45°. The first whisker is the minimum, the box is the first and third quartile, the red line is the median, and the last whisker is the maximum. The black horizontal line denotes the true inclination angle of 45°.

The configuration that is most different from the ones analyzed by Koch, Baig, & Zaidi (2018) is the inclined case, so we analyze that first. The retinal projection curve in Fig. 5A deviates from the ones in Koch, Baig & Zaidi (2018) by not being 1 to 1 because for every pose of the inclined stick, the retinal projection is greater than 180°. This is because the retinal orientation of the inclined stick is a function of the inclination angle, and for poses pointing towards the observer, the inclination angle is greater than the camera elevation angle. Thus, observers are perceiving the inclined stick from its bottom side and this perspective causes the stick to be perceived as greater than 180° for poses pointing towards the observer, as depicted by Fig. A4.

Inverting the retinal projective curves gives the back-transform, or the predicted 3D pose from the retinal orientation, as shown by the red curves in Fig. 5B. There is a close agreement between observer estimates and the back-transform curve for the inclined condition in Fig. 5B, suggesting that observers are using the most appropriate back-transform to make their judgments. The K values are all close to 1.0 indicating only a slight adjustment accounts for the fronto-parallel bias. In order to test whether observers are perceiving the correct inclination angle, Fig. 5C depicts various back-transforms for inclination angles of 35°, 40°, 50°, and 55°. Root Mean Square Error (RMSE) values indicate that perceived pose estimates best fit the back-transform when the inclination is close to 45° for all three viewing azimuths. Therefore, it appears that observers may be accurately perceiving the slant of the picture as well as the inclination angle to use the appropriate back-transform. Figure A5 shows fits for individual observers.

Given the similarity in shape between the back-transforms for different inclination, we have to rule out the possibility that observers are using a more simplified process – rather than judging the slant of the picture, they are using
the back-transform from the frontal viewing angle but for the wrong inclination, and applying that to oblique views. Fig. 5D shows that the frontal back-transform fits the observer data closely for oblique views when the inclination angle is estimated to be 65°. Therefore, it is possible that observers do not have to judge the slant of the picture to make accurate pose estimations if it is also true that the perceived angle of inclination is steeper in the oblique views. To test this hypothesis, a short experiment was run on 4 additional observers to estimate the angle of inclination for two poses of inclined sticks, 45° and 135°, from the three viewing angles. These observers viewed 10 randomized repetitions of each pose and recorded their estimate on the same response touchscreen, oriented vertically. Fig. 5E shows that the perceived angle of inclination varies with the viewing angle as well as the pose. For example, a pose of 45° appears more steep when viewed from +60° but less steep when viewed from the -60°, and vice versa for a pose of 135°. Additionally, though certain poses may look steeper from one viewing angle, many of the estimates are significantly lower than 65°, which rules out the fit of the frontal back-transform. Therefore, it is likely that observers are accurately perceiving both the slant of the picture as well as the inclination angle in order to use the best possible back-transform to make their judgments.
4.2 Analysis: Floating objects

Figure 6.
(A) Projected retinal orientation against the physical 3D pose for the floating condition. The green curve is the projected retinal orientation described by Eq. (5). Columns represent the three observer viewpoints. The black line is the unit diagonal. (B) 3D pose against retinal orientation. The red curve is the back-transform derived from Equation (5) and represents the predicted 3D pose from the 2D retinal orientation. Each large blue dot is the averaged perceived pose estimate corresponding to the retinal orientation of that pose across all observers and the black line is the unit diagonal. The fronto-parallel bias is given by $K$. (C) 3D pose against retinal orientation for a stick floating at $H = 6.5$ cm, 7 cm, 7.5 cm, 8 cm, and 8.5 cm. Each colored curve represents the predicted 3D pose from the retinal orientation for a given amount of float in the following order: blue for 6.5 cm, red for 7.0 cm, yellow for 7.5 cm, purple for 8.0 cm, and green for 8.5 cm. The curves are almost indistinguishable.
4.3 Analysis: Off-center objects

(A) Projected retinal orientation against the physical 3D pose for the off-center condition. Columns represent the three observer viewpoints. The green curve is the projected retinal orientation described by Eq. (6). The black line is the unit diagonal. (B) 3D pose against the retinal orientation. The red curve is the back-transform derived from Eq. (6) and represents the predicted 3D pose from the 2D retinal orientation. Each blue dot is the averaged perceived pose estimate corresponding to the retinal orientation of that pose across all observers and the black line is the unit diagonal. The fronto-parallel bias is given by $K$. (C) 3D pose against retinal orientation for a stick off-center by $X = 10$ cm, 11 cm, 12 cm, 13 cm, and 14 cm. Each colored curve represents the predicted 3D pose from the retinal orientation for a given amount of off-set in the following order: blue for 10 cm, red for 11 cm, yellow for 12 cm, purple for 13 cm, and green for 14 cm. The curves are almost indistinguishable.

Figure 7.

The retinal projection curves in Fig. 6A and 7A are very similar to the ones for objects on the ground (Koch, Baig & Zaidi, 2018) and show that for the floating and off-center conditions, there is systematic modulation around the unit diagonal that becomes shallower in oblique viewing angles. For the floating and off-center conditions, the perceived poses show close agreement with the appropriate back-transform for all viewing angles in Fig. 6B and 7B. Fig 6C and 7C show that the back-transform curves for different heights and offsets are essentially identical and appear superimposed, therefore the fits cannot test for
misestimation of these parameters. These configurations show that for objects that are not inclined, the optimal back-transforms don’t vary as functions of height and offset, so pose estimates can be robust to misestimations of height or offset. Figures A6 and A7 show fits for individual observers.

5. Ancillary cues

The results from Koch et. al of a single stick lying on the ground as well our data from the floating and off-center conditions show us that observers are likely using an automatic process when estimating the poses of these objects. They perceive the stimulus and can then quickly and intuitively judge the pose from its retinal orientation without needing to integrate any additional processes. This automatic process, however, cannot explain how observers are able to accurately estimate the pose of an inclined stick because the projected retinal orientation varies with inclination angle, and observers choose the back-transform from close to the physical inclination angle. It is more likely that observers are relying on ancillary cues to first assess the stick’s angle of inclination before they make their judgment.

Figure 8.
(A) Original stimulus of an inclined stick with a pose of 0°. (B) Inclined stick with no ground plane. (C) Pose estimates of inclined sticks with disc versus without disc for two observers. Each dot represents one setting.
We hypothesized that the strongest cues to 3D pose estimation for the inclined condition are: 1) the presence of the ground plane, and 2) the three-dimensionality of the stimulus stick. To test the first hypothesis, we conducted a short experiment with 2 additional subjects. We compared pose estimations of the same stick with and without the ground plane disc (Fig. 8). The results of this short experiment indicate that the presence of the ground plane is not a strong cue to determine 3D pose, as estimations with and without the ground plane are comparable. To assess the second hypothesis, we altered the stimuli to eliminate different three-dimensional cues to the shape of the stick. In keeping the shape of the stick but taking away both the shading and internal contours in Fig. 9 (A, D), whether the stick is pointing towards or way from the observer becomes unclear. Fig. 9 (B, E) incorporates internal contour lines onto a stick with no shading – these contour lines help to disambiguate the direction the stick is facing, but without shading the observers may still confuse the appearance of convexity versus concavity. Finally, Fig. 9 (C, F) shows the original stimulus with appropriate shading and contour lines. The shading of the stick allows viewers to determine the direction the stick is facing, while the contour lines allow viewers to accurately perceive the angle of inclination.

6. Discussion

The distinction between the perception of pictures versus perception of real scenes has long been an issue of debate. In our experiment, we identify geometric operations that are involved in both. We found that observers use the same back-transform to retinal projections as they do for real scenes. Our results strongly suggest that human brains use knowledge of projective geometry, the geometry of light, to make pose estimations from 3D scenes. Historical figures, such as Berkeley (1709), Plato (1976), and Kant (1781), have separately addressed the innate nature of this knowledge, while Poincaré (1905) and others
have argued that it is learned through visual observation. Through our experimentation and analysis, we show that percepts of 3D pose for non-centered objects not lying on the ground can be explained by the hypothesis that observers use the optimal geometric back-transform from retinal images, in addition to other 3D cues for more complex poses. Our hypothesis provides a simple explanation for 3D scene inference, but prefaces the need to study even more complex situations, ones that we are more likely to find in the real 3D world.
Acknowledgments

Funded in part by NIH grants EY007556 & EY013312

References


Appendix

Derivation of 2D retinal orientations from 3D object poses:

This projection is derived for objects in 3-Space (XYZ – Space) lying on the ground plane (i.e., object elevation = 0), and extending from the center of the scene (0,0,0). The camera is centered at that point. Thus each object has one endpoint at (0,0,0), and the other at (x,0,z).

We first rotation the ground plane around the x-axis to account for the camera elevation angle, $\phi_c$. The center point, (0,0,0) $\rightarrow$ (0,0,0). We compute the new coordinates of the other endpoint:

$$
\begin{pmatrix}
  x' \\
  y' \\
  z'
\end{pmatrix} = \begin{pmatrix}
  1 & 0 & 0 \\
  0 & \cos(\phi_c) & \sin(\phi_c) \\
  0 & -\sin(\phi_c) & \cos(\phi_c)
\end{pmatrix}
\begin{pmatrix}
  x \\
  y \\
  z
\end{pmatrix}
$$

(S1)

giving:

\begin{align*}
  x' &= x \\
  y' &= z \sin(\phi_c) \\
  z' &= z \cos(\phi_c)
\end{align*}

(S2)

Next, we compute the projection of each endpoint onto the Picture Plane (UV-space). We let the distance from the camera to the center of the 3D scene be $d_c$, the focal length of the camera be $f_c$. The central endpoint is mapped at (0,0,0) $\rightarrow$ (0,0). We compute the picture plane coordinates of the other endpoint:

\begin{align*}
  r &= \frac{u'}{d_c + w'f_c} \\
  s &= \frac{v'}{d_c + w'f_c}
\end{align*}

(S3)

The final projection is from the picture plane to the 2D retinal space (RS-space). In our experiment the monitor remains static, and the observers changes viewpoint by physically moving to different locations that form a semi-circle around the monitor. Because the latter is more computationally intuitive, we will frame the derivative of the projection in this way. We first compute the new coordinates in 3-Space (UVW space), where the plane W=0 is defined by the fronto-parallel location of the monitor. Moreover, we add depth to the 2D—Space defined by the fronto-parallel picture plane. The central point is mapped as, (0,0,0) $\rightarrow$ (0,0,0). Computing the new coordinate for the other endpoint we have:

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

(S4)

And:

\begin{align*}
  u' &= u \cos(\phi_v) \\
  v' &= v \\
  w' &= -u \sin(\phi_v)
\end{align*}

(S5)

We let $d_v$ equal the observer's distance to the center of the picture plane, and $f_v$ equal the observer's focal length. Projecting the endpoints from UVW-Space to 2D retinal space (RS-space), the central endpoint is mapped as, (0,0,0) $\rightarrow$ (0,0), and the other endpoint is defined by:
To geometrically recover the orientation in the retinal plane in terms of the original orientation on the 3D ground plane we use a simple trigonometric derivation and substitution from the above equations:

\[ r = \frac{u'}{d_v + w'f_v} \]  

(S6)

\[ s = \frac{v'}{d_v + w'f_v} \]

Lastly, we can take the inverse of the above projection to get the geometric back projection from the retinal orientation to 3-Space orientation:

\[ \theta_R = \tan^{-1}(\frac{\alpha}{\beta}) = \tan^{-1}\left(\frac{\frac{u'}{d_v + w'f_v}}{\frac{v'}{d_v + w'f_v}}\right) = \tan^{-1}\left(\frac{\frac{v'}{u'}}{\frac{v}{u'\cos(\phi_v)}}\right) \]

(S7)

\[ \theta_R = \tan^{-1}(\tan(\Omega_T) \cdot (\sin(\phi_c)/\cos(\phi_c))) \]

Equations 1-6 for inclined, floating, and off-centered objects were derived similarly by using the proper coordinates for the object endpoints.
Fig A1: Refers to Fig 3A. Observer 1-5. Perceived dose estimates by observer for the inclined condition. Each observer made 6 judgments for each physical pose (symbols are as in Figure 3).
Fig A2: Refers to Fig. 3B. Observer 1-5. Perceived pose estimates by observer for the floating condition. Each observer made 6 judgments for each physical pose (symbols are as in Figure 3).
Obs 1

Off-center View: -60 deg

Fig A: Refers to Fig. 3C. Observer 1-5. Perceived pose estimates by observer for the off-center condition. Each observer made 6 judgments for each physical pose (symbols are as in Figure 3).
Figure A4. Depiction of inclined stick at 8 cardinal positions.
Fig. A5: Refers to Figure 5B. Observer 1-5. Predicted 3D pose vs. retinal orientation for each observer for the inclined condition. Rows represents different observers. Symbols and curves are as in Figure 5B. Each small blue dot represents 1 pose estimate. K is the fronto-parallel bias parameter.
Fig. A6: Refers to Figure 6B. Observer 1-5. Predicted 3D pose vs. retinal orientation for each observer for the floating condition. Rows represents different observers. Symbols and curves are as in Figure 6B. Each small blue dot represents 1 pose estimate.
Fig. A7: Refers to Figure 7B. Observer 1-5. Predicted 3D pose vs. retinal orientation for each observer for the off-center condition. Rows represents different observers. Symbols and curves are as in Figure 7B. Each small blue dot represents 1 pose estimate. K is the fronto-parallel bias parameter.