Vision Research 126 (2016) 264–277

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Gestalt-like constraints produce veridical (Euclidean) percepts of 3D indoor scenes

TaeKyu Kwon [a,*], Yunfeng Li [a], Tadamasa Sawada [b], Zygmunt Pizlo [a]

[a] Department of Psychological Sciences, Purdue University, USA
[b] School of Psychology, National Research University Higher School of Economics, Moscow, Russia

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 12 August 2015
Accepted 15 September 2015
Available online 3 November 2015

Keywords:
3D scene
Visual space
Veridical vision
Inverse problems
A priori constraints
Triangle task

ABSTRACT

This study, which was influenced a lot by Gestalt ideas, extends our prior work on the role of *a priori* constraints in the veridical perception of 3D shapes to the perception of 3D scenes. Our experiments tested how human subjects perceive the layout of a naturally-illuminated indoor scene that contains common symmetrical 3D objects standing on a horizontal floor. In one task, the subject was asked to draw a top view of a scene that was viewed either monocularly or binocularly. The top views the subjects reconstructed were configured accurately except for their overall size. These size errors varied from trial to trial, and were shown most-likely to result from the presence of a response bias. There was little, if any, evidence of systematic distortions of the subjects' perceived visual space, the kind of distortions that have been reported in numerous experiments run under very unnatural conditions. This shown, we proceeded to use Foley's (Vision Research 12 (1972) 323–332) isosceles right triangle experiment to test the intrinsic geometry of visual space directly. This was done with natural viewing, with the impoverished viewing conditions Foley had used, as well as with a number of intermediate viewing conditions. Our subjects produced very accurate triangles when the viewing conditions were natural, but their performance deteriorated systematically as the viewing conditions were progressively impoverished. Their perception of visual space became more compressed as their natural visual environment was degraded. Once this was shown, we developed a computational model that emulated the most salient features of our psychophysical results. We concluded that human observers see 3D scenes veridically when they view natural 3D objects within natural 3D environments.

1. Introduction

There is a long tradition of applying Gestalt ideas to the interpretation of 2D retinal images (see Wagemans, Elder et al., 2012; Wagemans, Feldman et al., 2012, for recent reviews). In most of this work, two-dimensional Rules of Perceptual Organization were studied in an effort to discover how ambiguities inherent in our 2D retinal images are resolved by our visual system. Much less effort has been devoted to 3D shapes and 3D scenes. Hochberg and McAlister (1953) were an exception. They took the first important step over 60 years ago, when they showed that a simplicity principle operated in the perception of 3D shape. Unfortunately, this important work stimulated little interest in the human vision community at that time. There were only a few follow up studies, including Attneave and Frost (1969) and Perkins (1972, 1976).

Recently, one of us (Pizlo, 2008) picked this problem up when he published a book devoted entirely to the history of research on 3D shape perception. This book emphasized the important role *a priori* simplicity constraints (aka priors) had in the perception of 3D shape. *A priori* constraints[1] are essential because vision in general, and 3D vision in particular, is an ill-posed inverse problem (Pizlo, 2001). By "inverse problem", we mean that the task of producing accurate interpretations of the 3D geometrical and physical properties of objects "out there" requires an *inference* based on 2D retinal image(s). By "ill-posed", we mean that there are many possible interpretations, but only one of which is correct. The only way we know to solve such problems (i.e., to produce accurate interpretations) is to impose constraints on the family of possible interpretations. These constraints represent the visual system's knowledge of the physical world.

This knowledge can be acquired in two quite different ways; namely it could come from our personal experience with objects

* Corresponding author at: Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, USA.
E-mail address: kwont@purdue.edu (T. Kwon).

[1] The use of what we call "*a priori* constraints" is analogous to what the Gestalt Psychologists called "autochthonous organizing processes".

http://dx.doi.org/10.1016/j.visres.2015.09.011
0042-6989/

and scenes accumulated over our lifespan, or from our innate intuitions about regularities in the physical world. Emphasizing experience with concrete objects and scenes is represented by the empiristic school in perception. This school has traditionally viewed perception as a *recognition* task driven by Pavlovian stimulus–response learning (Helmholtz, 1867). The Gestalt Psychologists were nativists. They emphasized the importance of innate intuitions about regularities characterizing objects and scenes. More specifically, they viewed perception as a *reconstruction* task driven by *abstract regularities* they called *Prägnanz, or a simplicity principle* (Koffka, 1935; Wertheimer, 1923) (see also Leeuwenberg, 1969, for his use of a simplicity principle in his Structural Information Theory).[2] Note that models based on learning and familiarity tend to be fairly simple because they call for little more than using look-up tables to perform template matching. Performing 3D reconstruction based on abstract regularities as priors, an approach derived from the Gestalt Psychologists' approach to vision, requires quite elaborate models that actually solve the inverse problem.

But, once we accept that the visual system solves the inverse problem of recovering a 3D shape based on 2D images, the concept of a "veridical" interpretation must be brought into the discussion. By "veridical", we mean that the perceptual representation of an object "out there" agrees with (is identical to) this object. Why do we need this concept here? This concept is needed because the very essence of solving inverse problems in science, medicine and engineering is to produce, or guess, *the* correct solution, the truth. This is the case when someone is required to detect a cancer in a chest X-ray, predict the outcome of a presidential election, estimate the location of the center of an Earthquake or recover a 3D shape from a single 2D image.

Now one must ask whether it is possible for visual representations of objects to be veridical. The answer seems uncertain with a number of the perceptual correlates of physical attributes, such as the color of a surface. Color simply does not exist in the physical world. There are only wavelengths of light and reflectances of surfaces in the world "out there". Color is produced by the human visual system. There is only one characteristic of an object that permits certainty about veridicality, namely shape. Shape is unique precisely because it is easy to verify the veridicality of its perceptual representation. Why is it easy? It is easy because shape refers to relational characteristics, spatial similarities, called "symmetries", inherent in real 3D objects. When you look at an animal, say a horse, and you see that this horse is mirror-symmetrical, your percept is veridical, that is, the symmetry perceived agrees with the geometrical symmetry "out there." If I show you a square and you say that you see a quadrilateral with four axes of mirror-symmetry, I know that you see a square and that your percept is veridical. So, the veridicality of shape can be verified, and our recent experimental results show that human observers really do see the shapes of 3D objects veridically. Furthermore, we succeeded in developing a computational model that can recover shapes veridically, too (Li, Sawada, Shi, Kwon, & Pizlo, 2011).[3]

The reader is surely aware of the numerous studies of shape perception performed with unnatural objects and impoverished viewing conditions where the percept was observed to be far from veridical (see Pizlo, 2008, for a review of these studies). But note that such failures would be of interest *only* if a computational model could correctly solve the inverse problem of 3D shape recovery with the same unnatural objects and impoverished viewing conditions. If it could, one would be entitled to say that the visual space of the human observer is distorted; a conclusion often proposed in the past. But, if the recovery is impossible from a computational point of view, the failure of the observer to solve the same recovery problem accurately does not allow one to draw any strong conclusions about the nature of observer's visual space. The failure of the observer in such cases is more likely to represent the difficulty of the computational task, rather than some property of the underlying perceptual mechanisms. It follows that computational models are absolutely essential when studying inverse problems in vision. Ideally, computational models should be developed before a specific psychophysical experiment is conceived. A great deal of effort can be wasted if this is not done.

We first demonstrated how this kind of approach works in our studies of the veridicality of 3D shape perception by monocular and binocular observers. We started by formulating a computational model of monocular 3D shape recovery that was based on only 3 constraints, namely, mirror-symmetry, compactness and planarity. This model was able to recover a 3D shape nearly perfectly. It only showed systematic errors in the recovered aspect ratios, and these only when the viewing directions were close to "degenerate" (a degenerate viewing direction is either parallel or orthogonal to the plane of symmetry of the object). It was impossible to eliminate the systematic errors made by this model, so, it came as no surprise when nearly identical errors were observed in our monocular human observers when they were tested with the same viewing directions (Li, Pizlo, & Steinman, 2009). This permitted us to say with confidence that we understood what was going on. Namely, systematic errors in perceived aspect ratios of mirror-symmetrical objects, when they are viewed from directions close to degenerate, is a natural consequence of the operation of several *a priori* constraints that are absolutely essential for seeing 3D shapes nearly perfectly from a single 2D retinal image. Without the contribution of these constraints, there would not even be a 3D shape percept. Without our computational model, it would not have been possible to understand or explain why Biederman and Gerhardstein (1993) had observed nearly perfect shape constancy with objects composed of symmetrical parts they called "geons", while Rock and DiVita (1987) observed a complete failure of shape constancy with amorphous bent wires.

Shortly after collecting these data on monocular 3D shape recovery, we were able to conclude, without running any new experiments or simulations, that binocular shape recovery is as good as monocular shape recovery, but not better (Pizlo, Li, & Steinman, 2008). We did not expect that binocular shape recovery would be much better than monocular recovery simply because the voluminous literature on binocular vision showed that it is neither accurate nor precise (Howard & Rogers, 1995). But later, after actually testing our own human subjects, we discovered that binocular 3D shape recovery is actually *perfectly* veridical when tested under appropriate conditions (Li, Sawada et al., 2011). Here, the computational model of binocular 3D shape recovery was formulated *after* the subjects were tested, so a model explaining a given 3D visual function fully may be made before or after the psychophysical experiments are performed. The order is not critical, but the psychophysical experiments must be accompanied by a comprehensive modeling effort. We took our own advice when we conducted the present study.

With this background, it becomes timely to ask whether the perception of a 3D *scene* is more like the perception of 3D *shape*, or more like the perception of color. Three dimensional scenes may or may not have mirror-symmetrical configurations of objects (imagine two chairs facing each other or standing side-by-side), but whenever similar, or identical, objects are present in the scene (now imagine several identical chairs in a haphazard configuration), all of these chairs can be described by using symmetry (invariance) under a rigid motion, a more general symmetry than mirror-symmetry. This raises the question of whether these kinds of symmetries will be

[2] Prägnanz in German means succinctness, conciseness or terseness.

[3] With other attributes, such as color, one can always verify that there is "perceptual constancy", which simply means that the identity of physical attributes is perceived veridically.

sufficient to allow veridicality to be achieved in the perception of 3D scenes. More generally, do natural 3D scenes contain a sufficient number of effective *a priori* constraints to solve the inverse problem required to recover 3D scenes accurately? Note that these are important theoretical and computational questions, and answering them *before* designing and conducting psychophysical experiments is obviously useful, and in our view, absolutely necessary. It turned out that the inverse problem of 3D scene recovery can be solved accurately, providing only that the scene contained symmetrical objects that resided on a common horizontal ground with their planes of symmetry being orthogonal to the ground because of gravity. How do we know this? We know it because we had formulated a computational model of 3D scene recovery, implemented it in a binocular robot (named "Čapek") and tested it with real scenes (Li, Sawada, Latecki, Steinman, & Pizlo, 2012). Furthermore, note that if symmetrical objects are perceived as symmetrical, and standing on a common ground, there is little room for distortions of an observer's or a model's 3D visual space. The slant of the floor cannot be underestimated perceptually because this would either destroy the right angles formed by the planes of symmetry of objects and the floor, or the symmetry of the objects would be destroyed. What is being suggested here is that the *symmetry of the objects within the physical 3D space, in conjunction with the presence and direction of gravity, calibrate the 3D visual space, making the perception of the space as well as the objects contained within it veridical.* Visual distortions of the 3D space are likely to occur when the viewing conditions are impoverished by: (i) using an empty 3D scene, or a nearly empty one, (ii) using large viewing distances, distances larger than several meters, or (iii) by eliminating visual information about the horizontal ground. There is a large body of evidence showing that several types of visual distortions of 3D space are commonplace in such cases (e.g., Foley, 1972; He, Wu, Ooi, Yarbrough, & Wu, 2004; Johnston, 1991; Li, Phillips, & Durgin, 2011; Loomis, Da Silva, Fujita, & Fukusima, 1992; Loomis, Da Silva, Philbeck, & Fukusima, 1996; Ooi & He, 2007). These distortions probably represent the fact that recovering a 3D scene veridically with impoverished viewing conditions is simply impossible. So, *the distortions reported in the literature are in the mathematics, not in the beholder's eye*. The goal of this paper is to provide empirical and computational evidence that 3D scenes can be recovered veridically under relatively natural viewing conditions. As far as the present authors know, the psychophysical studies reported here are the first to show that human observers can perceive a nearby indoor 3D scene containing several pieces of furniture, veridically. The computational model presented here is the first model that can detect nearby, real 3D objects in an indoor 3D scene; and recover the entire 3D scene nearly perfectly, as our subjects did. At this point in the model's development, its input is binocular. Extending this model to monocular input is the subject of our future work on this problem.

The remainder of this paper is organized as follows: Section 2 describes an experiment in which human subjects used monocular and binocular viewing to recover the geometry of 3D indoor scenes that contained several pieces of furniture. Sections 3 and 4 used a simpler configuration of objects to make it possible to test and falsify prior claims that the visual space of a human observer had a negative curvature. Section 5 provides a brief description of our computational model along with tests of this model with the same type of indoor scenes that were used by our subjects in experiments 1–3. The paper is concluded with a General discussion.

2. Experiment 1: Recovering 3D scenes

2.1. Subjects

Three subjects (TK, YS, and XZ) participated in the experiments. All had corrected-to-normal vision. TK was an author and XZ was a naïve subject. This experiment was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Informed consent was obtained for this experimentation from all of our human subjects.

2.2. Stimuli and conditions

The experiment was performed in a 7.92 m × 8.53 m room illuminated by overhead fluorescent lights. The walls were white and there were doors on opposite sides of the room. The floor was covered with a blue, texture-less carpet. An experimenter placed four or five pieces of children's furniture, including chairs (32 cm × 29 cm × 67 cm), tables (59 cm × 38 cm × 51 cm), bookshelves (47 cm × 30 cm × 62 cm), and garbage bins (33 cm × 33 cm × 78 cm) at haphazardly-chosen positions within the room. They were positioned before each trial to form a naturalistic indoor scene without any occlusions of the objects from the subject's viewpoint. A typical scene is shown in Fig. 1.

The positions of the furniture were measured by a motion capture system (PhaseSpace – see Fig. 2). The system has 16 pairs of cameras. Each pair consists of two orthogonally-oriented one-dimensional cameras. This system computes the 3D positions of multiple LEDs placed within a scene that are identified by their flicker rates. Their position accuracy was better than 2 cm after calibration (see Appendix for details). An LED was put at the center of each object to represent the position of each object used in this experiment.

2.3. Procedure

Each subject was tested in 40 trials: 20 trials with binocular viewing and 20 trials with monocular viewing. The subject stood in a designated position in the room where he reconstructed a top view of the scene by drawing its floor plan on a tablet computer by dragging and dropping ready-made 2D icons, representing the top views of the 3D objects (Fig. 3). The subject was instructed to use the *sizes* of the icons to judge the inter-object distances on the computer screen. The sizes of the icons were scaled to their physical size.[4] The frame of the tablet monitor, however, was not a scaled version of the room. As pointed out by an anonymous reviewer, the aspect ratio and the size of the tablet monitor could potentially influence the pattern of results.

Exposure duration was unlimited; the subject could look at the scene until the drawing of the floor plan (top view) was finished. After each trial, the experimenter put an LED on top of the center of each object in the scene and recorded the LEDs' position with the motion capture system. It took about 5 min for the subject to draw each floor plan and it took another 5 min for the experimenter to measure the objects' positions and set up a new scene.

The subject viewed the scene with both eyes in half of the trials. In the remaining half, only the right eye could see; the left eye was patched. Head motion was not restricted.

2.4. Results and discussion

The pairwise distances among all of the 3D objects, including the subject, were computed to evaluate how well the subject had reconstructed the scene. Originally, we thought that the subject's position could be treated the same way as the positions of the other objects. We found that this was not the case. Perhaps, not surprisingly, judging geometrical relations, such as pairwise distances and angles in a triangle formed by 3 chairs, is easier than judging the geometrical relations in a triangle formed by 2 chairs

[4] The subject could rotate the icons to represent the orientation of the objects but the icon's orientation data were not analyzed.

Fig. 1. A typical scene used in Experiment 1. Four or five pieces of children's furniture were placed on a texture-free floor. The specific layout of the pieces was varied from trial to trial.

and one's self. We said not surprisingly because the spatial relationships among objects in a stationary scene do not depend on the position of the observer. But, the spatial relationships between the observer and the rest of the stationary scene change constantly whenever the observer's position changes. Clearly, the coordinate system "attached" to the scene is likely to be more useful for establishing and maintaining space constancy than the coordinate system "attached" to the observer, himself. This is why the subject's position was excluded from these analyses (an analysis that did include the subject's position will be reported later). The actual Euclidean distances among the centers of the objects were obtained by the motion capture system that detected the positions of LEDs attached to the center of each object. The distances reconstructed by the subjects were obtained from the drawings they made on the tablet computer. The mean squared error (MSE) for pairwise distances was estimated to evaluate the accuracy of the subject's floor plan. The mean squared error is defined as follows:

$$\mathrm{MSE} = E\left[\left(\frac{d' - d}{d}\right)^2\right] \tag{1}$$

where d' is a reconstructed distance and d is an actual distance. The MSE is equal to the sum of the variance of normalized distances (d'/d) and the squared Bias:

$$\mathrm{MSE} = \mathrm{VAR}(d'/d) + \mathrm{Bias}^2 \tag{2}$$

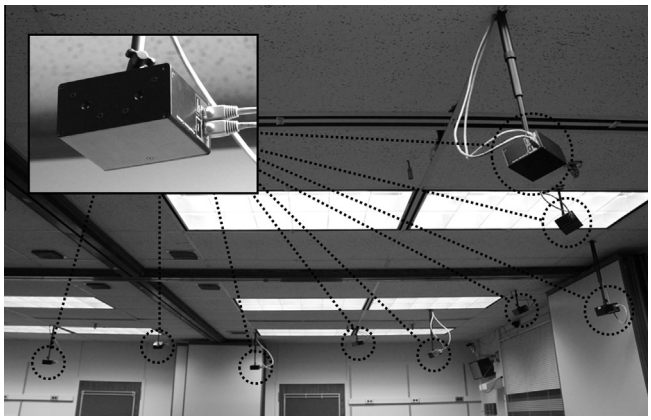

Fig. 2. Motion capture system (*PhaseSpace*). The 16 pairs of cameras attached to the ceiling measured the 3D positions of LEDs placed within the scene.

where $\mathrm{Bias} = E\left[\frac{d'-d}{d}\right]$. Taking the square root of MSE and VAR yields the root mean square (RMS) error and standard deviation (STD) which have the unit of % of actual distance:

$$\mathrm{RMS} = \sqrt{\mathrm{MSE}} \tag{3}$$

$$\mathrm{STD} = \sqrt{\mathrm{VAR}} \tag{4}$$

Table 1 shows these errors calculated for each subject.[5]

The RMS of these pairwise-distances ranged from 13% to 31%. The Bias was quite large in subjects YS and XZ. Both systematically overestimated the distance. TK was different. His Bias, computed over all of his 20 monocular and 20 binocular trials, was close to zero. Note that this does not necessarily mean that his Bias was close to zero in his individual trials, simply that it averaged out. An analysis of the nature of each subject's Bias is quite important because it can shed light on whether the source of the Bias is perceptual processing or a response bias. Here, we are using the conventional distinction between a perceptual process and a response bias as it is used in Signal Detection Theory (Green & Swets, 1966).

Recall that the subjects were asked to scale the distances on the tablet by using the sizes of the icons representing the objects as a reference for their distance scale. Note that the inter-object distances were an order of magnitude larger than the actual objects themselves. This difference is similar to what obtains in a conventional magnitude estimation method when applied to size. Magnitude estimation is known to cause large variability in the data (Brunswik, 1944). It is possible that once the subject set his first distance, he used this distance as a reference to make decisions about the remaining distances. This would allow the subject to avoid comparing distances and sizes of such different magnitudes. If subjects actually did use this approach in reconstructing the scene, all distances within a given trial should share the same systematic error. This implies that we would see large random fluctuations of Bias across trials. This proved to be the case. Fig. 4 shows distributions of the intra-trial Bias for monocular and binocular viewing. The relationship between Table 1 and Fig. 4 is that the mean values of the distributions in Fig. 4 are equal to the Bias values in Table 1.

The intra-trial Bias varied from −0.2 to 0.5 (−20% to +50% of the physical distance). This variability of Bias from trial to trial contributed to large values of MSE and VAR. There are two facts that suggest that the Bias observed in individual trials was a response bias. It was not caused by a perceptual distortion of visual space. We can say this because: (i) the Bias varied a lot across trials, and (ii) the Bias had the same values in both monocular and binocular viewing.

The intra-trial Bias was removed by applying a uniform size scaling of the recovered scene in each trial (see Table 2). This optimal size scaling was done by making the recovered pairwise distances as close as possible, in the least-squares sense, to the actual distances. We next asked whether there was some additional factor representing a perceptual distortion of visual space. Traditionally, two distortions have been discussed, namely affine and projective. The presence of such distortions can be verified simply by applying affine and projective transformations to the reconstructed top view (floor plan) and then determining whether the distances among the transformed positions of the objects are closer to their true distances after this transformation. Table 2 shows the 2D affine and projective transformations used, as well as the overall size scaling that was just described. All three types

[5] One of the anonymous reviewers pointed out that we could include the PhaseSpace error in our Eq. (1). This would have reduced the estimated variance of the subjects' reconstructions. This reduction would be quite small so we decided not to include it.

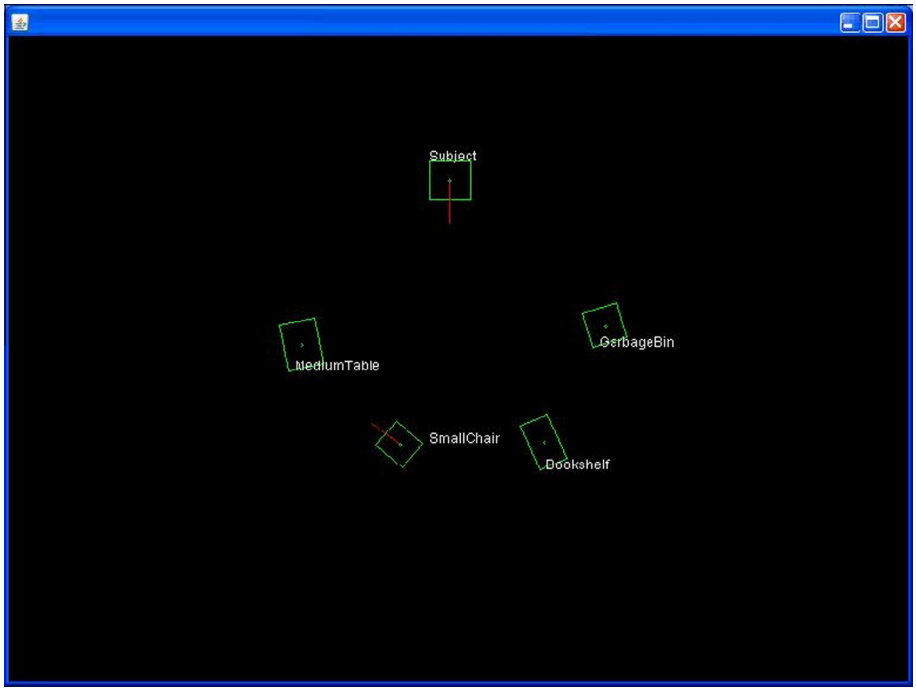

Fig. 3. Interface shown on the screen of a tablet computer. The name of each object was labeled on the lower right corner of its icon. The subject used a stylus to move the icons.

Table 1
Monocular and binocular errors of the 3 subjects' reconstruction in Experiment 1.

Subject	Viewing condition	RMS (%)	STD (d'/d) (%)	Bias (%)
TK	Monocular	20.28	19.76	4.79
TK	Binocular	12.81	11.48	−5.76
YS	Monocular	24.05	16.94	17.14
YS	Binocular	22.80	14.66	17.51
XZ	Monocular	31.37	17.45	26.10
XZ	Binocular	28.74	17.78	22.63

of transformations were applied independently to individual trials to minimize the mean squared error. Note that if there actually were perceptual distortions of visual space, the distortions will be constant across trials, and only one common distortion will be found for all trials. We applied the distortions to individual trials in order to see whether there was *any* effect because if we do not find strong effects of affine or projective distortions on a trial by trial basis, we would not have found them after all of the trials were pooled.

The three transformations described in Table 2 were applied cumulatively from size scaling to affine, and from affine to projective transformation. Simplified equations of affine and projective transformations with the most influential parameters were applied to the data to interpret these parameters more clearly. Graphs of subjects TK's, YS's and XZ's monocular and binocular standard deviations of their normalized pairwise-distances are shown in Fig. 5.[6]

[6] These graphs show errors computed from pairwise distances among the objects, when the position of the subject was not used. If the position of the subject is added to the computation of pairwise distances, the errors in Fig. 5 increase. For example, the errors after size scaling in binocular trials are twice as big. More precisely, they increase to 13%, 15% and 21% for TK, YS and XZ, respectively. This fact suggests that even though the subject can recover the 3D scene quite reliably, he has difficulty including himself in the recovered scene. This issue will be discussed further in Experiment 3.

First note that the random errors, as measured by the standard deviation of the normalized distance, were almost always smaller in binocular than in monocular viewing, but only subject TK's difference was large. Also note that the advantage of binocular viewing observed was expected because adding the second view cannot harm 3D vision; it can only help. Next, note that the errors decreased substantially when size scaling was applied. Recall that the size scaling removed the response bias from *individual* trials. This means that the standard deviations after size scaling are likely to represent the observer's actual degree of perceptual uncertainty in recovering pairwise distances in the 3D scene. The average standard deviation of the three subjects was 11.4% in monocular viewing and 8.5% in binocular viewing. These numbers are only 3–4 times greater than the Weber fraction for line-length discrimination, a task in which only one pair of approximately equal lines is presented in the frontoparallel plane (Pizlo, Rosenfeld, & Epelboim, 1995). Note that in our experiment, the subjects were faced with the task of estimating (reconstructing) the length of a line segment in a 3D space, so this task must lead to greater variability than the task of deciding which of two line segments in the retinal image is longer. It follows that the variability measured was not very large considering the computational difficulty of this 3D task. Further transformations, namely affine and projective, did not affect the errors very much, suggesting the absence of systematic distortions of 3D visual space. The standard deviation decreased by 2.3% in monocular, and by 1.8% in binocular viewing when an affine transformation was applied, and additional 0.7% and 0.9% when a projective transformation was applied. Most likely this decrease in error when affine and projective transformations were applied was caused by the addition of free parameters to the estimation rather than by a perceptual distortion of visual space. Note that the number of measurements in each trial, 6 or 10 pairwise distances when 4 or 5 objects were in the scene, was not much larger than the number of free parameters, namely, 3 when an affine, and 5 when a projective transformation was used. We can, therefore, conclude that our subjects' visual space was not distorted, which means that they saw the 3D spatial arrangement of objects within this space veridically.

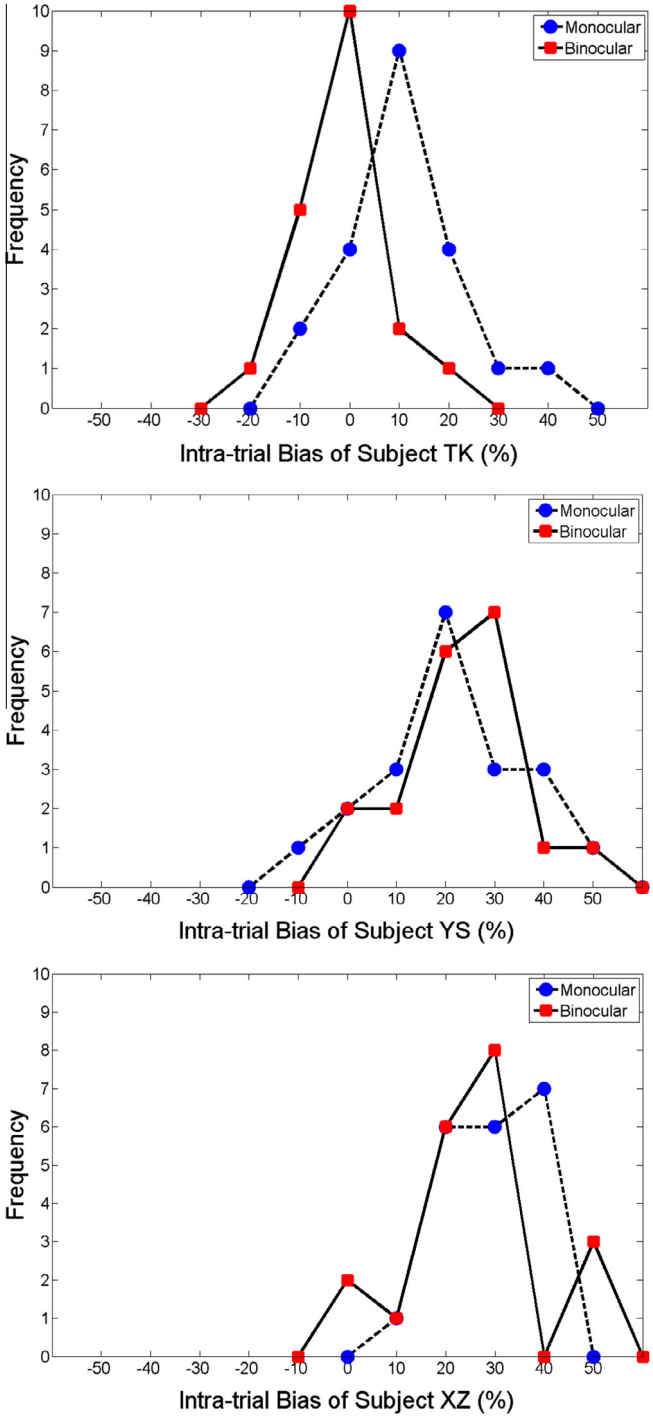

Fig. 4. Frequency polygons of the intra-trial Bias of three subjects (TK, YS and XZ) in monocular and binocular viewing.

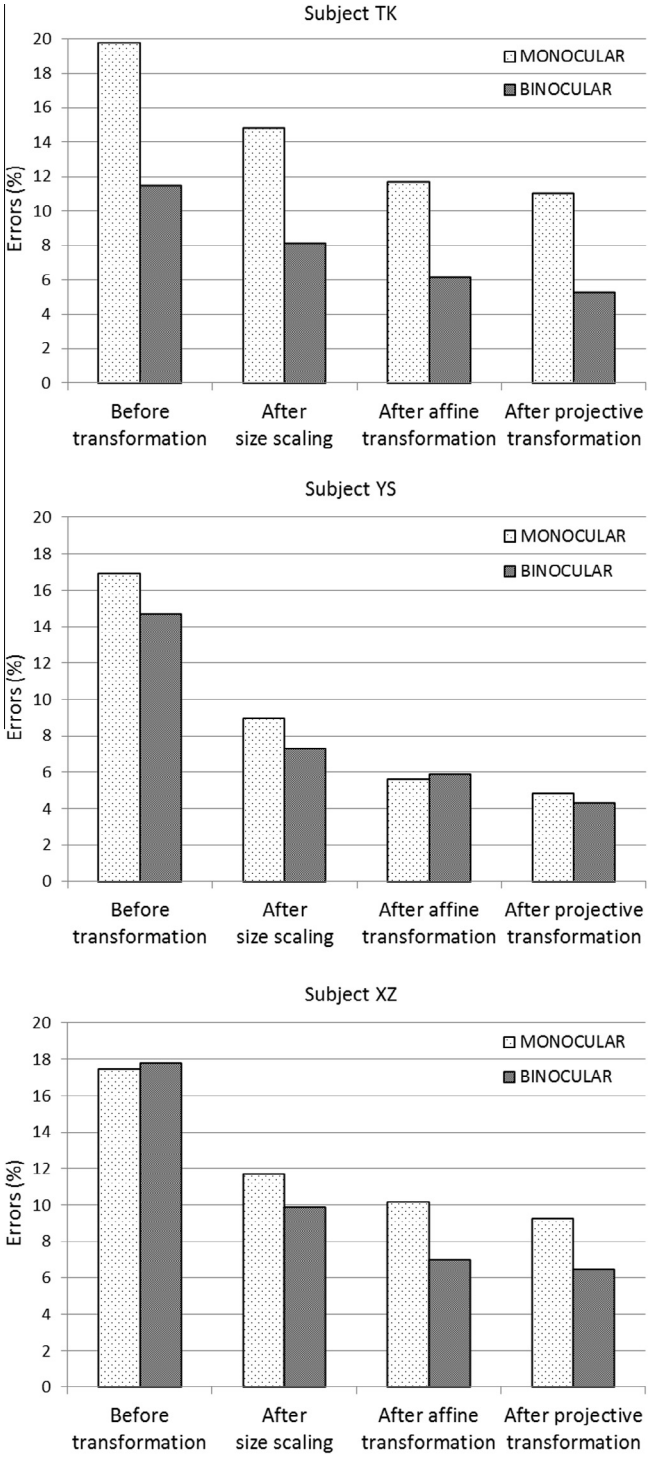

Fig. 5. Standard deviation of the normalized pairwise-distances of the 3 subjects' reconstruction in Experiment 1. The ordinate shows errors (%) represented as the standard deviation of the normalized pairwise-distances, and the abscissa shows the transformations applied to the reconstructed positions. The light and dark gray bars indicate monocular and binocular viewing conditions.

Table 2
Transformations and corresponding equations applied to the reconstructed positions.

Transformation	Equation
Size scaling (with one parameter)	$x' = ax$ $y' = ay$
Affine transformation (with two parameters)	$x'' = bx'$ $y'' = cy'$
Projective transformation (with two parameters)	$x''' = \frac{x''}{dx''+ey''+1}$ $y''' = \frac{y''}{dx''+ey''+1}$

We pointed out above that our subjects had difficulty including themselves in the reconstructed scene. Including the distances of the objects from the subject increased the reconstruction errors. Did this simply mean that systematic errors are always made when egocentric distances are estimated? Probably not because He et al. (2004) showed that perceived egocentric distances, as measured by blind walking, are quite accurate and precise. He et al.'s result,

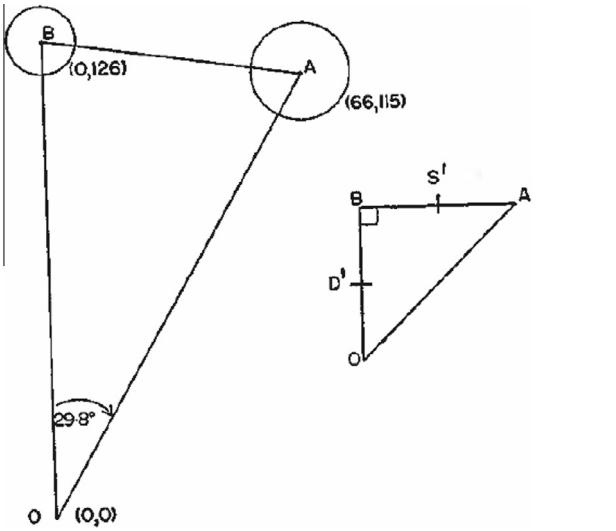

Fig. 6. Illustration of Foley's (1972) experiment. The triangle on the left is the triangle adjusted by the subject. The triangle on the right shows the triangle that the subject had been asked to construct in which OB = BA and the angle OBA = 90°. "O" indicates the subject's position. The circles contain 75% of the settings made by the subjects.

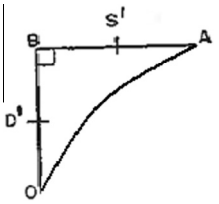

Fig. 7. If the triangle is perceived as an isosceles right triangle with the angle at O being perceived as 30°, the visual space must be curved. Specifically, when this space is curved, the sum of angles in this triangle is less than 180°.

together with our results on 3D scene recovery, suggests that our subjects had difficulty combining the ego- and exocentric coordinate systems into one. Experiment 3 addressed this question directly and showed that they did.

3. Experiment 2: Triangle production with natural viewing

The results of Experiment 1 suggested that visual space is Euclidean, but the relationship between physical locations and perceived locations does not describe the intrinsic geometry of visual space. It is possible that a subject's perception of the room is subject to some distortions, but his reconstruction on the computer's screen does not reveal them because the subject's perception of the room, represented on the computer's screen, is distorted in the same way as the subject's perception of the room, itself. The intrinsic geometry of the visual space can be evaluated by analyzing the relations among the perceived lengths and angles. This was done by using Foley's (1972) isosceles right triangle task. Foley's experiment and results will be described next.

In Foley's experiment, the subject adjusted two points of light at eye-level to form a perceived isosceles right triangle, including himself at the vertex "O" of one of the acute angles (see Fig. 6). The main result was that the adjusted distance AB is roughly one half of the viewing distance OB. This means that the triangle OAB "out there" is not isosceles, despite the fact that it is perceived as such. More importantly, the angle AOB in the constructed triangle is perceived accurately as 30°, despite the fact that it should have been perceived as 45° if the right triangle were perceived as isosceles and if the visual space were Euclidean. From this result, Foley concluded that binocular visual space is non-Euclidean (see Fig. 7). Specifically, visual space is curved and the sign of this curvature is negative. In such a space, the sum of the angles in a triangle is less than 180°.

Note that Foley performed his experiment with extremely impoverished stimuli: 2 points of light at the eye level presented in complete darkness with the subject's head immobilized on a bite-board. His conditions were very different from natural viewing conditions, so there is little reason to believe that Foley's results will generalize to visual perception in everyday life. We will examine this issue in our next experiment, but we will evaluate the reasoning that led Foley to a non-Euclidean model before describing our experiment examining the generality of his results. Foley's theoretical justification of his non-Euclidean model is not convincing. Foley assumed that the results of his experiment required the operation of a binocular visual system. This is certainly true with respect to the comparison of distances: AB vs. OB, but the judgment of angle AOB does not require binocular information. It can be done, and most likely was done, by using monocular information. So, Foley's experiment does not allow his conclusion that binocular visual space is non-Euclidean with a negative curvature because he mixed two quite different types of judgments: one binocular and the other monocular. These judgments are very different because the monocular task was very easy. It called for nothing more than "measuring" distances on the retina. The binocular task was much more difficult because it required inferring 3D distances on the basis of 2D information on the retina. Such an inference is an ill-posed Inverse Problem, the kind of problem that cannot be solved veridically without making use of *a priori* constraints (Pizlo, 2001). So, even before we did our experiment, we had a good reason to believe that Foley's theoretical claim about visual space being curved was unwarranted. But, despite the methodological flaw in how he drew his conclusion, his data demonstrating a large perceptual compression of depth can be evaluated independently from his claims about the curvature of visual space. We decided to do just this next. We re-examined Foley's triangle experiment under relatively-natural viewing conditions.

Our study went out of its way to produce the kind of environment and natural viewing conditions we humans encounter almost every day. Also, unlike everyone else who has studied the geometry of 3D vision, we recognized that 3D vision is a difficult inverse problem, whose solution depends critically on the operation of *a priori* constraints. Note that once the inverse problem approach is adopted, it is essential to emulate natural conditions as much as possible to guarantee that these constraints can operate effectively. If the stimuli or viewing conditions are impoverished, some or all of the *a priori* constraints may have become ineffective, making 3D judgments biased as well as variable, not because visual space is non-Euclidean, but because the experimental conditions were not ecologically-valid (Brunswik, 1956). Ecological-validity was approximated in this experiment by using a triangle task with 3 natural objects, rather than 2 points of light, standing on the floor with the room illuminated. This experiment clearly showed that our 3D visual space is Euclidean. Our third, and final, experiment worked out which aspects of natural viewing are essential for vision to be veridical.

3.1. Subject

Three subjects (TK, YS and JC) participated in this experiment. All three had corrected-to-normal vision. TK and YS had also participated in Experiment 1.

3.2. Stimuli

Three ordinary children's chairs, standing naturally on the floor, served as the 3 stimuli that formed the triangle. This is completely different from what Foley did. In his experiment, the subject served as 1 of the 3 objects forming the triangle. The motion capture system identified the exact position of each chair by detecting a LED mounted at its center. These LEDs were used as the reference points for the distance judgments made by both the subject and the motion capture system. Two of the chairs (O, B) were placed in front of the subject within a viewing distance of 4 m, and the subject was required to adjust the position of the remaining chair (A) so as to construct an isosceles right triangle (see Fig. 8). Note that "O" in our experiment, refers to one of the three objects, not to the subject as it did in Foley's experiment.

3.3. Procedure

Each subject was tested in 20 trials during which he set the position of A so that the segment AB appeared perpendicular to OB, and the distance AB appeared equal to OB. The physical changes of the position of the chair moved by the subject were actually carried out by the experimenter in response to verbal commands from the subject.

In half of the 20 trials, the subject viewed the scene binocularly. The right eye viewed the scene in the other half, the left eye was occluded. The subject's head was supported on a chin-rest to minimize motion parallax cues produced by moving the head.

After completing his adjustment of the position of the object A, the subject was asked to report verbally whether the angle AOB was equal to 45°. The positions of LEDs on all 3 chairs were measured by the motion capture system at the end of each trial.

3.4. Results and discussion

The adjusted ratio, AB/OB, averaged across the three subjects, was 0.89 in monocular and 0.87 in binocular viewing. Fig. 8 shows the physical layout that each subject perceived as an isosceles right triangle. The subjects always reported that the angle AOB was 45°. The actual average angle AOB shown in these Figures was computed from the three distances among the chairs. On average, the angle AOB was less than, but close to, 45°. These results, which are completely different from Foley's (1972), allow us to conclude that in natural viewing conditions, visual space is Euclidean in both monocular and binocular viewing.

4. Experiment 3: Triangle production task with varying viewing conditions

We could not know at this point which experimental factor(s) was responsible for the differences in our and Foley's experimental results because the stimulus and the viewing conditions differed in more than one way (see Table 3). There are at least three possible factors, namely: (i) the illumination level (his dark and our highly illuminated room), (ii) the placement of the stimuli (ours standing on the floor and his at eye level), and (iii) the number and nature of the stimuli. Foley used the subject as 1 of the 3 vertices and tiny lights for the other 2 vertices of his isosceles triangle. Our 3 vertices were ordinary children's chairs standing naturally on the floor. We both used binocular viewing. In our Experiment 3, the head of the subject was not restricted in any way, but the head in Foley's was immobilized by a bite-board.

So, an experiment designed to find out why our results were different from Foley's had to manipulate 2^3 variables, which meant that we needed 8 experimental conditions to find out which factors were responsible for the distortion of visual space reported by Foley (1972).

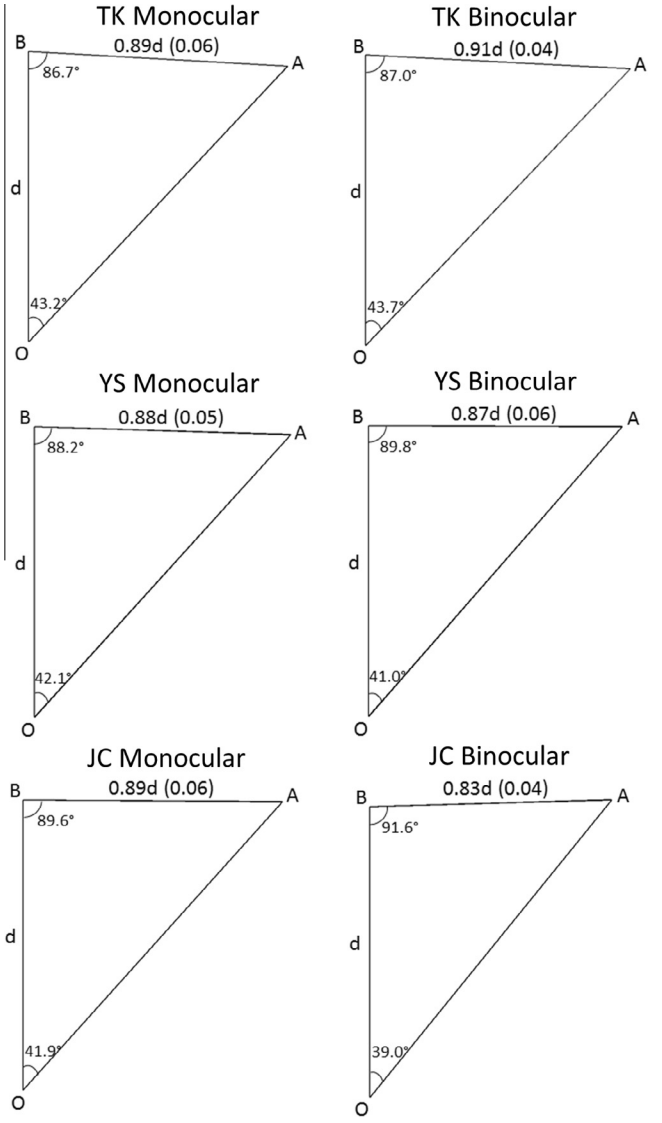

Fig. 8. Average physical arrangement that was perceived as isosceles right triangles by the three subjects in Experiment 2. The length of OB was set as *d* and the length of AB was represented as a fraction of *d*. Angles were computed from the measured lengths. The length and the angles were averaged across the 10 trials for each subject and viewing condition. The number in the parentheses indicates the standard deviation of the ratio AB/OB.

Table 3
Differences between Foley's (1972) experiment and our Experiment 2.

	Foley (1972)	Our Experiment 2
Lighting condition	Dark room	Illuminated room
Level of stimuli	Eye level	Knee level
Number and type of stimuli	Two points of light	Three pieces of furniture

4.1. Subjects

The 3 subjects with corrected to normal vision (TK, YS and XZ), tested in Experiment 1, participated in this experiment.

4.2. Stimuli

Children's free-standing coat racks (height = 111 cm – eye level) and low rectangular stands (height = 46 cm – knee level) were

placed on the floor in different positions to form the triangles. Their locations were determined by a LED placed in their center as done previously. LEDs also served once again as the reference points for distance judgments. In the 3-objects condition, object O was put in front of the subject at a distance between 1 m and 2 m and object B was aligned with object O and with the subject. The distance between B and O was set between 1 m and 2 m. Initially, object A was placed behind object B, and was moved so as to construct an isosceles right triangle. The subject sat on a chair with his head free to move. The height of the chair was adjusted to make the subject's eyes approximately at the same height as the LEDs mounted on the coat racks. This was done to make the trials with coat racks similar to the conditions used by Foley who presented his stimuli at his subject's eye level.

4.3. Procedure

Each subject was tested with 3 trials in each of the 8 conditions. The subject sat on a chair and kept his eyes closed while the experimenter set up the scene. Once it was set up, the subject opened his eyes and directed the experimenter to move the position of object A so that the segment AB appeared perpendicular to OB, and the distance AB appeared equal to OB. The experimenter recorded the positions of the LEDs with the motion capture system after the subject was satisfied with the position of A.

The subject was given unlimited time to view the scene binocularly and his head was free to move. The subject was not asked to judge the perceived magnitude of the angle AOB. The angle AOB was computed from measurements made by the motion capture system.

4.4. Results and discussion

Table 4 shows the average of the aspect ratios and the average angles of the constructed triangles that the subjects perceived as an isosceles right triangle in each of the conditions. Note that if the subject constructs the isosceles right triangle accurately, the angle AOB would be 45° and the ratio AB/OB should be 1. When the room was illuminated (lighting condition: bright) and when three objects were used at knee and eye-level, the ratio AB/OB, averaged across the three subjects was 0.94. So, our estimated depth compression with natural viewing of a 3D scene is only 6%. This estimated systematic error is comparable to the standard deviation of such judgments (see Experiment 2). We can, therefore, conclude that there is almost no evidence of distortions in 3D scene perception when natural scenes are viewed naturally. The systematic errors we measured with natural viewing, 6%, are no way near to the 50% errors measured by Foley and everyone else with unnatural viewing. It is clear that *human observers see things the way they are* "*out there*".

When only two objects were used in a bright room, the reconstructed ratios AB/OB were smaller; they ranged between 0.75 and 0.87. When the room was dark, performance was worse overall. With 3 objects in the dark, the reconstructed ratio AB/OB ranged between 0.76 and 0.93, and when only 2 objects were shown in a dark room, performance was the worst: the reconstructed ratio AB/OB ranged from 0.59 to 0.76. Our naïve subject, XZ, when tested with two objects at the eye-level in a dark room (the conditions used by Foley) produced triangles almost identical to those reported by Foley (1972) with the ratio AB/OB being close to 0.6.

It is easy to understand why a brightly illuminated room could lead to more veridical perception than an almost completely dark room. In a brightly illuminated room, there is much more visual context available to which effective *a priori* constraints, known to be used for the veridical perception of 3D objects and 3D scenes, can be applied. They include the direction of gravity, the orientation of the horizontal floor, as well as the symmetry of the 3D objects within the room. It is, however, harder to explain why constructing a triangle with 3 objects is more accurate than constructing one with 2 objects and oneself. It seems that incorporating one's own position when recovering positions of objects is more difficult than recovering positions among three external objects; this issue was mentioned in Experiment 1. Recall that with two objects and oneself, the subject has to compare both egocentric and exocentric distances. With three external objects, the entire task can be performed by comparing only exocentric distances. Egocentric distance estimation depends on the subject's viewpoint and involves a viewer-centered representation, but exocentric distance estimation can be independent from the viewpoint which makes it similar to a scene-centered representation. It has already been demonstrated that egocentric distance estimations, as measured by blind walking to an object, are accurate (He et al., 2004). This suggests that subjects may simply not be able to use both types of distances in a common perceptual coordinate system despite the fact that they can reconstruct both. A scene-centered coordinate system may simply dominate 3D scene perception because it is likely to be used for achieving space constancy.

5. Model

Following the approach we used in our work on the veridical recovery of 3D shape described in the Introduction, we started formulating and testing models of 3D scene recovery at the same time that we started designing the psychophysical experiments described just above. The most recent version of the model, which was published last year, will be described briefly here (Pizlo, Li, Sawada, & Steinman, 2014). This computational model was implemented in a robot, called "Čapek". Čapek is equipped with a stereo-camera (*BumbleBee* made by Point Grey Research). The camera's height above the floor was about 1 m and its two lenses were separated by 12 cm.

Table 4
The average angles AOB and the average ratios AB/OB for the subjects in Experiment 3. Angles AOB were computed by using the measured lengths.

Lighting condition	Level of eye	Number of stimuli	TK		YS		XZ	
			AOB	AB/OB	AOB	AB/OB	AOB	AB/OB
Dark	Eye level	2	39.1°	0.73	37.1°	0.76	29.37°	0.59
Dark	Floor level	2	38.0°	0.75	34.1°	0.66	33.18°	0.65
Dark	Eye level	3	43.0°	0.79	39.8°	0.93	37.15°	0.78
Dark	Floor level	3	40.1°	0.82	40.1°	0.89	34.42°	0.76
Bright	Eye level	2	44.3°	0.85	39.5°	0.79	37.91°	0.75
Bright	Floor level	2	42.1°	0.81	41.5°	0.87	38.56°	0.79
Bright	Eye level	3	47.1°	0.98	42.6°	0.97	42.37°	0.91
Bright	Floor level	3	43.4°	0.89	43.7°	0.99	41.60°	0.89

Fig. 9 illustrates typical results obtained by this robot using our computational model for 3D scene recovery. This particular scene contains 7 pieces of furniture that produce partial occlusions of each other. The ground is a highly textured dance floor. The complexity of the ground's surface is enhanced considerably by specular reflections and shadows. The partial occlusions, together with the complex ground surface, make figure-ground organization challenging from a computational point of view. A reader looking at Fig. 9 (left) will have no difficulty finding objects in front of the background despite its complexity and the partial occlusions in this scene. A human observer has no difficulty producing veridical figure-ground organizations in such scenes. This always seems easy and effortless, but this is far from easy for a computational model. Our model is probably the very first to be able to solve such a difficult figure-ground organization problem with real images, as well as it does. Note that the scene shown in Fig. 9 is much more complex than the scenes used to test both the model and the subjects in the present study. Recall that in the present study, the floor was covered by a homogeneous carpet that produced no specular reflections and the objects were arranged without any occlusions (see Fig. 1). Fig. 9 makes it clear that these differences did not matter much at all: The veridicality of the Figure-Ground Organization in Fig. 9 is quite good. How was this done?

Čapek started by solving the binocular correspondence problem for texture points, using a standard algorithm that came with the *Bumblebee* camera. It then computed a 3D depth-map of the visible points. Čapek then applied an *a priori* constraint that says that all objects reside on a common horizontal ground. With this in place, the robot estimated the orientation and position of the floor by finding the plane that can fit the largest number of points in the depth-map. This provided Čapek with an estimate of the position of the horizontal ground plane and the vertical direction of gravity. These estimates were very good: there was no systematic error and the standard deviation of the floor orientation estimate was less than half a degree. These estimates are similar to the known visual abilities of human subjects presented with such tasks. The points representing the floor are then eliminated, and the remaining points, which represent the objects, are projected orthographically onto the horizontal plane. The result of this projection is called a "top view" representation. Čapek then identifies individual objects in the 3D scene by solving a clustering problem in this representation and fitting rectangles to the individual clusters (a rectangle is characterized by 5 parameters, that is, coordinates of the center, 2D orientation plus width and length). Solving the clustering problem in a top-view representation is much more reliable than solving this problem in a 2D camera image. Objects rarely occlude each other in a top-view, but they often occlude each other in a camera view. The center of the fitted rectangle is Čapek's estimate of the object's position on the floor. Note that a rectangle is a natural figure to use in a top-view of a mirror-symmetrical object standing upright, that is, an object with its plane of symmetry vertical. One side of the rectangle is parallel to the plane of symmetry and the other side is orthogonal to it. This is true for all mirror-symmetrical objects, not only pieces of furniture, but also for dogs, horses, cows and people. The rectangles *fitted* by Čapek are shown in Fig. 9 (right), where the *estimated* rectangles are superimposed on the ground truth that was measured by the PhaseSpace system. The fit is obviously very good.

Adding the estimated height of each object to the estimated rectangle results in 3D bounding-boxes that represent a solution of the 3D figure-ground organization (3D FGO). These boxes are projected to the camera image, producing bounding-polygons. These polygons, which are shown in the center of Fig. 9, are a solution of a 2D FGO problem. The bounding-polygons partially overlap, as they should, because the objects overlapped in the camera image. Note that using 3D bounding-boxes allowed the recovery of where the objects ended on their back, invisible-sides, as well as the recovery of the invisible, empty spaces behind the objects. Clearly, Čapek was able to produce a spatially-global 3D map of the room from a single vantage point, a map that can be used to plan navigation. Simply looking at Fig. 9 (left), makes it easy for the reader to see where objects end on their back, invisible-sides, as well as how large the invisible spaces behind the objects are. So, Čapek can do what we humans do, so well. The fact that we humans can do it is not very surprising because our common sense tells us that we can all do this very well. Imagine what getting around during rush hour in a train station would be like if we could not. The new thing here is that we have the first computational model that explains *how* this can be done. This result is important because it breaks with the long tradition started by David Marr's paradigm, in which we only see the *visible* surfaces of objects (Marr, 1982; Pizlo, 2008). According to Marr, our knowledge about the back parts of the objects comes from our familiarity with individual objects, something that must be learned. Here, as well with our 3D shape recovery model, the back parts of objects are recovered on the basis of abstract geometrical and physical regularities, namely, symmetry, compactness, planarity and gravity. Our model has no familiarity with individual objects; it has no memory, and it cannot learn anything about the shapes, sizes or positions of objects. Note that this calls attention to an additional advantage inherent in testing a model like ours. One can never be sure about the role familiarity is playing in a human observer. With our model, we can be sure that it plays no role, whatsoever.

Now, we will report Čapek's test results with experimental conditions identical to those used with our human subjects. Čapek reconstructed a top view of four different scenes using its binocular camera. At this point of the model development, Čapek cannot recover 3D scenes monocularly. The top views of recovered 3D scenes showed no sign of any systematic distortions of the 3D space around it. This result was shown in Fig. 9. This result made it clear that there was no need to apply the full set of transformations, *viz.*, size scaling, affine stretching and projective distortion, as we did in Experiment 1. Čapek's reconstructions of the pair-wise distances among several of the objects used in the 4 scenes are shown in Table 5. The standard deviation of Čapek's reconstructed distances ranged between 7.4% and 10.3%, values very close to the errors of our subjects (see Fig. 5 for our binocular subjects' results after size-scaling).

Čapek was next tested in a triangle task in which 3 objects were standing on the floor and the room was illuminated. Čapek reconstructed the 3D scene using its binocular camera, reporting the position of each object, as well as the distances and angles among them. An experimenter read the distances and angles reported, and based on these readings, adjusted the position of object A in the direction that would make the departure from the reconstructed isosceles right triangle smaller. The criterion used to end each trial was that the difference between the length of the reconstructed AB and OB was less than 1 cm, and the reconstructed angle ABO was different from 90° by less than 1°. The actual position of the objects was measured again with the motion capture system after this final adjustment had been made. The model was tested in 4 trials. The average physical arrangement of the isosceles right triangle that was reconstructed by Čapek is shown in Fig. 10. Čapek's visual space is obviously Euclidean: it is not distorted. We drew the same conclusion from our experiments with human subjects. Furthermore, Čapek's variability in adjusting the AB/OB ratio was very similar to the variability demonstrated by our human subjects in Experiment 2.

This similarity does not *prove* that this model describes how the human visual system solves the 3D scene recovery task, but it does provide a first step towards formulating such an explanation. Note that Čapek was able to solve a complicated, naturalistic,

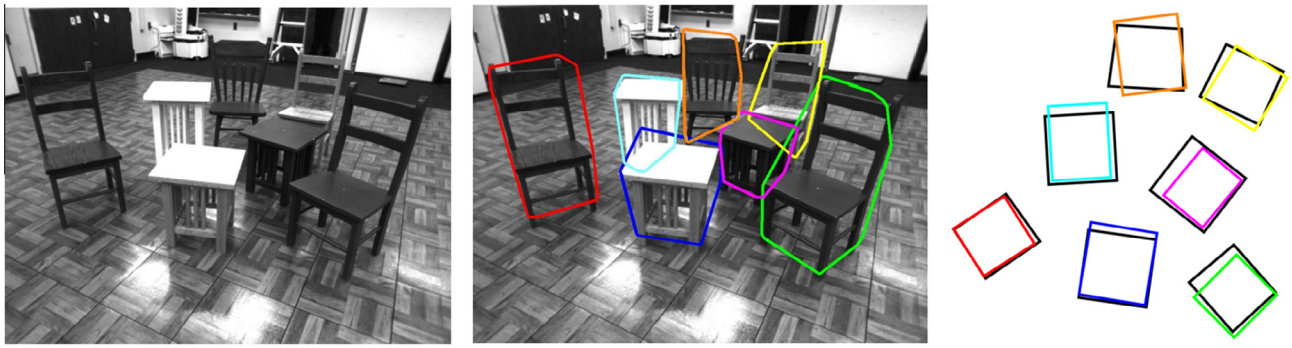

Fig. 9. Left: one of the two camera images of a 3D indoor scene acquired by our robot (Čapek). Center: the color curves indicate the regions containing individual objects. This shows that our robot solved a figure-ground organization problem on its camera image. Right: top view of the 3D scene in front of the robot. Black quadrilaterals represent the "ground truth": the true positions, sizes and orientations of the 7 pieces of furniture. Color rectangles are the robot's reconstructions.

Table 5
Robot's errors measured as the standard deviation of normalized distances when reconstructing pairwise-distances among the 4 or 5 objects in each of 4 scenes.

Model	Number of objects	STD (d'/d) (%)
Setting 1	4	10.31
Setting 2	5	9.39
Setting 3	4	7.36
Setting 4	5	7.50
Average	–	9.36

FGO problem. It was able to find real objects in a 3D scene, as well as in its 2D camera images. A number of tests showed that the model does this nearly perfectly (see Pizlo et al., 2014, for more examples). Human beings do this nearly perfectly, too. Our future work will address the effect of impoverishing the visual stimuli on the model's performance by using a dark room as well as by using two, rather than three objects. It is not clear, however, how informative these manipulations will be because, as you may recall, these manipulations caused our subjects to perceive illusions when they were tested. Traditionally, it has been assumed, often implicitly, that veridical perceptions can be produced in many different ways, which suggested that one needed to study illusions to disambiguate among several possible explanations. Our recent work on human and robot 3D vision suggests that the converse might be true. It is much more difficult to formulate multiple models of veridical perceptions than multiple models of biased and unreliable perceptions. This is precisely why we started by formulating a model of *veridical* vision.

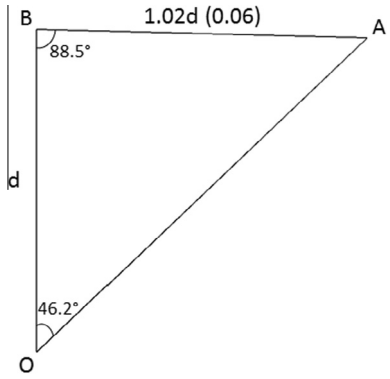

Fig. 10. The average physical arrangement that was perceived by Čapek as an isosceles right triangle. The length of OB was set as d and the length of AB was represented as the fraction of d and averaged across 4 trials. Angles were computed from the measured lengths. The number in the parentheses indicates the standard deviation of the ratio AB/OB.

6. General discussion

In the Introduction, we argued that computational modeling is a necessary tool for explaining how the visual system solves inverse problems. Inverse problems are typically ill-posed, which means that the input data available allows for more than one possible interpretation. This is always true with 3D vision. The Ames's chair demonstration is a classical example of this fact (see http://shapebook.psych.purdue.edu/1.3/). The visual system must apply effective *a priori* constraints if it is going to produce a unique 3D percept. The key word here is "effective". A single constraint may be sufficient to do the job, but more often than not, the visual system must use several constraints in order to make sure that the 3D perceptual interpretation is: (i) unique, (ii) stable, that is, the 3D percept does not change much in the presence of noise and uncertainty in the 2D retinal image, and (iii) veridical. This is a lot to ask of the visual system. But even more importantly, there is simply no way to decide how to accomplish these goals without formulating and testing computational models that allow you to find out: (1) which constraints should be applied, and (2) how they should be applied to satisfy the three criteria, listed just above. This is critical because you cannot evaluate the computational stability of a 3D interpretation without running simulations. Furthermore, these simulations must be done with realistic input images, not with the toy input images usually used in 3D shape and scene research. What all this means is that what we mean when we formulate and talk about a "model" is quite different from what other researchers, usually working within different contexts, mean by a "model". In those other cases, fitting a regression surface to data points, results in what is called a "model", and this model can be verified or falsified by collecting additional data. Formulating a model of how the visual system solves the 3D vision problem, which is what we do, is a completely "different animal" because the main challenge here is not to choose the best model from a family of already existing models, as is typically done when you decide which is the best regression function, but to come up with even a single model that can solve the recovery problem in 3D vision reasonably well. It follows, that with the types of models we are using, falsification is not your "best friend" as it often is elsewhere. In 3D vision, there is usually no model you can turn to, so there is nothing to falsify. Simply put, scientific discovery in 3D vision is not accomplished by using an ANOVA or Bayesian tests to reject some hypotheses. Discovery is accomplished by correctly guessing which cost function is actually being used by the visual system.

In this paper, we presented our "guess" about the cost function that is used for the veridical perception of 3D scenes. Our guess took the form of a computational model that takes real camera images of a real 3D scene and forms a veridical 3D interpretation

of the scene. Our model works well in the sense that its 3D recovery is as good a representation of the *ground truth* about the locations of 3D objects in the scene as our human subjects' percepts. Furthermore, this model uses constraints known to be used by the human visual system. We do not claim that this model is the ultimate, correct model. It will surely benefit from elaboration. We also will not be surprised to find out that a completely new model may be needed to emulate the human beings' vision as well as this can be done. Our next step will be to formulate a second model, with a different cost function, and/or with different constraints to find out whether it can produce equally good 3D scene recoveries. If it can, we will try to develop tests to choose between them.

Emulating human 3D vision is only one part of this endeavor. Showing that human beings perceive 3D scenes veridically under natural conditions is also important. For some obscure reason, studying failures of vision ("illusions") have attracted much more attention in the vision community than studying the veridicality of human visual perceptions. The study of illusions at the expense of veridical perceptions may simply be a product of the widespread commitment to testing null hypotheses. This strategy encourages looking for perceptual "effects", such as the effect of depth cues on judging the aspect ratio of an ellipse, or the joint effect of texture and binocular disparity on perceived slant of the surface, etc. If a percept is veridical and constant, despite changes in the viewing conditions, there are no effects! There is nothing to publish. This might explain, at least partially, why there have been very few studies designed to assess the degree of veridicality in visual perception rather than its failures.

We showed, in three psychophysical experiments, that human beings see naturalistic indoor scenes accurately, so accurately that there was no appreciable bias that could be used to justify using non-Euclidean spaces as models of the visual space, which is the perceptual representation of the physical space in which we humans live and act. The variability of our subjects' judgments was also fairly small. When these subjects were asked to compare distances among 3 objects forming a right isosceles triangle on the floor, a task similar to line-length discrimination, their standard deviations were only 4–6%, barely twice as large as it is when the subject is required to discriminate the lengths of two line-segments in the frontal plane. It seems likely that the performance we measured is as good as the very best that can be done in 3D vision. If this proves to be the case, it seems unlikely that there will be two, or more, different models that are also capable of producing performance equal to the best performance demonstrated here.

In conclusion, our results with human subjects, when taken along with our robot's performance, suggest that human 3D vision is based on the operation of three *a priori* constraints, specifically: (i) the direction of gravity, (ii) a horizontal floor upon which all objects reside, and (iii) the symmetry inherent in the 3D objects. These specific constraints were essential for our computational model to work as well as it did. More psychophysical experiments must be done before we can make an equally explicit statement about how the human visual system works. Details about how these constraints are combined with visual data in our computational model are available elsewhere (Li et al., 2012; Pizlo et al., 2014). The fact that *a priori* constraints are *essential* was anticipated by the Gestalt Psychologists over 100 years ago. When the 3D visual stimuli available to a human observer are natural, these *a priori* constraints can be fully effective, and the 3D percept will be veridical, that is, we humans will see things as they actually are "out there". But, whenever viewing conditions are very impoverished, our natural *a priori* constraints are likely to be ineffective, so our visual reconstructions must become unreliable as well as biased. It is important to keep in mind that unreliable vision with impoverished stimuli is inherent in the ill-posed nature of the visual task presented to the subject. Under natural viewing conditions, *a priori* constraints can be used by the visual system, allowing the ill-posed task to be regularized and solved veridically as a well-posed task.

Acknowledgments

This research was supported by the AFOSR (FA9550-09-1-0207), NSF (0924859) and NIH (1R01EY024666-01). The authors thank Prof. Robert M. Steinman for helping us set up the experiments and write this paper.

Appendix A

A.1. Calibration of the motion and position capture system

The coordinate system in the room was defined by drawing a 5 m by 5 m square with a grid of 1 m steps on the floor (Fig. A1). The positions of the vertices of the grid were then measured by the motion capture system and the transformation between the coordinate systems of the physical space and the motion capture system was derived.

A.2. Drawing the grid on the floor

Straight line segments on the floor were marked by first projecting a line from a laser and then stretching strings along the laser lines. We then determined the positions of the three vertices p_0, p_1, and p_2 of the square (see Fig. A2). The two line segments, p_0p_1 and p_0p_2 form a rectangular corner at p_0. The two segments were drawn roughly first to make them parallel to the walls of the room which is approximately rectangular (Fig. A2). The positions of p_1 and p_2 were marked at a distance of 5 m from p_0. The positions of p_1 and p_2 were then corrected so that p_0, p_1, and p_2 formed an isosceles right triangle. Next, a vertex p_3 of the square was determined by simple trigonometry. Once these measurements were done, markings were placed at 1 m steps on the four sides of this square. The grid-lines were drawn by connecting these markings.

A.3. Measuring the physical grid using the motion capture system

The positions of the vertices of the grid were measured with the motion capture system in order to compare the coordinate systems of the physical space and the motion capture system. LED markers (used in the motion capture system) were placed sequentially at the vertices on the floor and 60 cm above the floor. The position of the marker was measured for 10 s at 30 Hz at each position (total samples = 300). The measured positions of points on two sides of the square proved to be unstable and these positions were eliminated from subsequent analyses (see gray area in Fig. A2). The useful area that remained was a 4 m by 4 m square. Measurements made at the other positions were stable and their standard deviations were less than 2 mm. The 300 samples obtained at each useful position were averaged and used in the subsequent analyses.

Deriving the transformation between physical $[x', y', z']^t$ and measured $[x, y, z]^t$ spaces:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = S_{xyz} R_{xyz}(\theta_x, \theta_y, \theta_z) \left(\begin{bmatrix} x \\ y \\ z \end{bmatrix} - T_{xyz} \right)$$

$$R_{xyz}(\theta_x, \theta_y, \theta_z) = R_z(\theta_z) R_y(\theta_y) R_x(\theta_x)$$

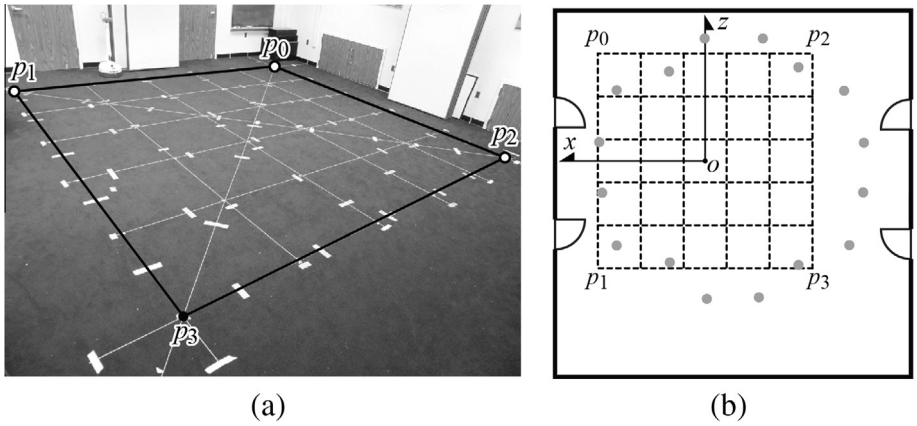

Fig. A1. (a) A photo of the experimental room during the calibration. The grid on the floor was marked with strings. (b) The top view of the room. The four corners, p_0, p_1, p_2 and p_3 form a square reference frame used for producing the grid that was used as the ground truth in the calibration. Dashed contours indicate the grid on the floor and the gray discs indicate the positions of the motion capture system's cameras attached to the ceiling. The 3D coordinate system was set so that the xz-plane coincided with the floor and the y-axis was vertical and pointing up. The origin of the coordinate system was set at the center of the grid. The photo (a) was taken from the bottom-right corner of the room as shown in (b).

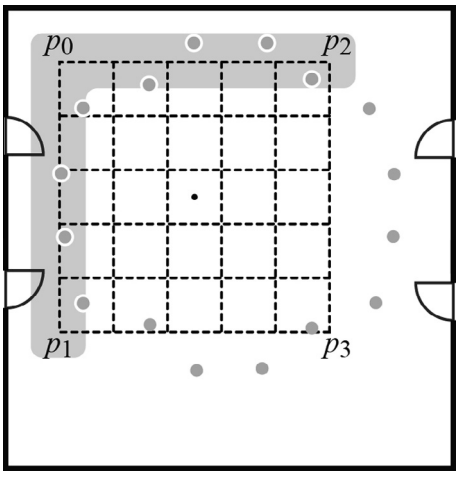

Fig. A2. A copy of Fig. A1b. The points of the grid in the shaded area could not be measured reliably. Note that all of these unreliable points were located farthest from the center of the circular array of cameras. These points could only be seen by cameras on the opposite side of the circle. The unreliable points were far away from the cameras, and the limited viewing angles of the cameras above the points prevented these points from being seen.

$$R_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix}$$

$$R_y(\theta_y) = \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{bmatrix}$$

$$R_z(\theta_z) = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $[x, y, z]^t$ is a measured position of a point, $[x', y', z']^t$ is its physical position, and θ_x, θ_y, and θ_z are degrees of rotations about the x, y and z axes respectively. In total, there were seven free parameters. These parameters were determined using the least squares method: $T_{xyz} = [12.6\text{ mm}, -4.8\text{ mm}, -9.0\text{ mm}]^t$, $\theta_x = 0.086°$, $\theta_y = 0.017°$, $\theta_z = 0.12°$, and $S_{xyz} = 1.004$. The difference between the physical position and the measured position of the point after the transformation was less than 2 cm. It means that the motion capture system could detect a position of a point in a 3D scene within the maximum error of 2 cm after the calibration.

References

Attneave, F., & Frost, R. (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception and Psychophysics, 6*(6B), 391–396.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*(6), 1162–1182.

Brunswik, E. (1944). Distal focusing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs, 56*(254).

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Foley, J. M. (1972). The size-distance relation and intrinsic geometry of visual space: Implications for processing. *Vision Research, 12*, 323–332.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

He, Z. J., Wu, B., Ooi, T. L., Yarbrough, G., & Wu, J. (2004). Judging egocentric distance on the ground: Occlusion and surface integration. *Perception, 33*, 789–806.

Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology, 46*(5), 361–364.

Howard, I. P., & Rogers, B. J. (1995). *Binocular vision and stereopsis*. New York: Oxford University Press.

Johnston, E. B. (1991). Systematic distortions of shape from stereopsis. *Vision Research, 31*(7/8), 1351–1360.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York, NY: Harcourt Brace Jovanovich.

Leeuwenberg, E. L. (1969). Quantitative specification of information in sequential patterns. *Psychological Review, 76*(2), 216–220.

Li, Z., Phillips, J., & Durgin, F. H. (2011). The underestimation of egocentric distance: Evidence from frontal matching tasks. *Attention, Perception & Psychophysics, 73*, 2205–2217.

Li, Y., Pizlo, Z., & Steinman, R. M. (2009). A computational model that recovers the 3D shape of an object from a single 2D retinal representation. *Vision Research, 49*(9), 979–991.

Li, Y., Sawada, T., Latecki, L. J., Steinman, R. M., & Pizlo, Z. (2012). A tutorial explaining a machine vision model that emulates human performance when it recovers natural 3D scenes from 2D images. *Journal of Mathematical Psychology, 56*(4), 217–231.

Li, Y., Sawada, T., Shi, Y., Kwon, T., & Pizlo, Z. (2011). A Bayesian model of binocular perception of 3D mirror symmetrical polyhedra. *Journal of Vision, 11*(4), 1–20.

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*, 906–921.

Loomis, J. M., Da Silva, J. A., Philbeck, J. W., & Fukusima, S. S. (1996). Visual perception of location and distance. *Current Directions in Psychological Science, 5*, 72–77.

Marr, D. (1982). *Vision*. New York: W.H. Freeman.

Ooi, T. L., & He, Z. J. (2007). A distance judgment function based on space perception mechanisms: Revisiting Gilinsky's (1951) equation. *Psychological Review, 114*, 441–454.

Perkins, D. N. (1972). Visual discrimination between rectangular and nonrectangular parallelepipeds. *Perception & Psychophysics, 12*, 396–400.
Perkins, D. N. (1976). How good a bet is good form? *Perception, 5*(4), 393–406.
Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research, 41*(24), 3145–3161.
Pizlo, Z. (2008). *3D shape: Its unique place in visual perception*. Cambridge, MA: MIT Press.
Pizlo, Z., Li, Y., Sawada, T., & Steinman, R. M. (2014). *Making a machine that sees like us*. NY: Oxford University Press.
Pizlo, Z., Li, Y., & Steinman, R. M. (2008). Binocular disparity only comes into play when everything else fails; a finding with broader implications than one might suppose. *Spatial Vision, 21*(6), 495–508.
Pizlo, Z., Rosenfeld, A., & Epelboim, J. (1995). An exponential pyramid model of the time course of size processing. *Vision Research, 35*, 1089–1107.
Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology, 19*, 280–293.
Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138*(6), 1172–1217.
Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin, 138*(6), 1218–1252.
Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. Laws of organization in perceptual forms. *Psycologische Forschung, 4*, 301–350.