Building a Seeing Machine

Yunfeng Li
Tadamasa Sawada
Zygmunt Pizlo
Department of Psychological Sciences
Purdue University
703 Third Street
West Lafayette, IN 47907
765-494-6930
yunfenglee14@gmail.com, tada.masa.sawada@gmail.com, pizlo@psych.purdue.edu

Keywords: Figure-ground organization, 3D shape recovery

ABSTRACT: *It is now commonly acknowledged that intelligent robots will not be able to interact with humans unless they can see as well as we do. Until very recently emulating human visual abilities was considered an insoluble problem. The two main functions of human vision that allow us to operate in everyday life are Figure-Ground Organization and 3D shape recovery. We present the very first computational models of these two functions, compare the performance of the models to the performance of human subjects and implement these models in a robot.*

1. Introduction

The modern scientific study of human visual perception started with the work of Gestalt Psychologists who completely changed the way perceptual psychologists viewed the underlying mechanisms (Wertheimer, 1923/1958; Koffka, 1935). Gestalt Psychologists pointed out that the percept of the visual world is always a combination of the information provided by the retinal image(s) and an *a priori* simplicity principle. By doing this they elaborated the ideas that had been put forth by Ernst Mach (1906/1959). Gestaltists emphasized the importance of perceptual constancies in vision. Consider shape constancy, as their primary example. Shape constancy refers to the fact that the percept of the shape of a given 3D object is constant despite changes in the shape of the object's 2D retinal image, caused by changes in the 3D viewing direction. Perceptual constancies lead to a veridical perception of the world's permanent characteristics. By veridical it is meant that we see things the way they are out there. This paper addresses two fundamental aspects of veridical perception: Figure-Ground Organization and shape constancy.

2. Human Vision as an Inverse Problem

Figure-Ground Organization (FGO) refers to the task of identifying how many objects there are in front of an observer and where they are. Figure 1(a) shows an example. It is easy to see 5 pieces of furniture in the center of the floor. The 6th object is substantially occluded, but with some effort it can also be seen on the back right. The fact that FGO is typically solved by the human visual system so effortlessly is deceptive. From a computational point of view FGO is extremely difficult. The main reason is that FGO is, as most important visual functions, and ill-posed "inverse problem." The classification of problems into direct or forward vs. inverse is due to Tikhonov (Tikhonov & Arsenin, 1977; see also Poggio et al., 1985, for the introduction of this classification into vision science). A forward problem in vision refers to producing a 2D retinal image of a 3D scene. This forward problem is easy (well-posed) because for a given 3D scene and a given viewpoint the 2D image is uniquely specified. In fact, in real imaging systems such as a robot camera or a human eye, it is the laws of optics that "solve" the forward problem (produce the image).

(a) (b)

Figure 1. A pair of stereo images for an indoor scene with six pieces of furniture. The 6th object is in the back right of the scene. It is substantially occluded by the chair in front of it.

An inverse problem in vision refers to inferring the 3D scene from a 2D retinal image. This problem is difficult (ill-posed) because there are always infinitely many possible 3D interpretations (solutions) for a given 2D retinal image. One way to see this is to realize that the object points can be moved on their projecting lines arbitrarily without changing the 2D image. The fact that the human visual system almost always arrives at a

single and correct 3D interpretation is truly amazing. The early enthusiasm of the machine vision community in the 1950s and 1960s to solve the vision problems and emulate or even surpass the abilities of the human visual system ended with a complete failure (but see successful examples of 3D interpretation of line drawings, such as the work of Guzman, 1968; Clowes, 1971, Huffman, 1971, Waltz, 1972, and Marr, 1977). As a result, the computer vision community abandoned the real 3D problem about 30 years ago and switched to the task of extracting 2D statistical features from 2D images. The 3D problem has been considered insoluble despite (i) the effort of hundreds, if not thousands of computer vision laboratories around the world, (ii) increasingly better computational and mathematical methods for signal processing and machine learning, and (iii) increasingly faster computers with growing computational power. Interestingly, many psychologists studying human vision followed the lead of the computer vision community and concluded that the 3D problem cannot be solved either by a scientist or by the human visual system. This resulted in a popularity of "multiple view theories" of 3D shape and scene perception, in which it is assumed that our percept of a 3D world consists of 2D representations. This is like claiming that the Earth is flat.

Let us explain in some detail the nature and the degree of the difficulty of FGO. Consider a camera image with 6 million pixels. Such cameras do exist and they provide a reasonable analogy with the human retina which contains 6 million cones. Solving FGO may require evaluating all possible partitions of the 6 million pixels and deciding which partitions most likely represent objects. Ignore for a while how a criterion for making this decision can be constructed. Even if a very good criterion were available, evaluating all possible partitions of a single 2D image cannot be practically done in any reasonable amount of time even if the fastest computers were used. The reason is that the number of all partitions of n elements (called the Bell number) grows with n faster than n!, and n! grows very fast. Any problem that requires performing a number of computations proportional to n! is considered intractable. Even for small n, n! is very large. For example, 61! is equal to 10^{81}, which is equal to the number of atoms in the universe. 6,000,000! is equal to $10^{38,000,000}$. It should be obvious that the number of all partitions of 6 million pixels cannot be analyzed. Brute force approach, based on machine learning methods will not do. It is important to realize that the human visual system solves FGO within a fraction of a second despite the fact that neurons are fairly slow. If a neuron in the human brain is compared to a transistor in a CPU of a computer, then the neurons are 6-9 orders of magnitude slower than the transistors (i.e., 1 million to 1 billion times slower). So, the algorithm used by our visual system to solve FGO must be very smart and if anyone is able to emulate it in a seeing robot, this will be a real breakthrough. Section 3 describes this algorithm and the performance of a robot using it.

Next, consider the second insoluble vision problem, 3D shape recovery. Figure 2 shows a 2D image of a 3D abstract and unfamiliar polyhedron. The reader can surely see the 3D polyhedron, but again the fact that 3D shape perception seems so effortless is deceptive.

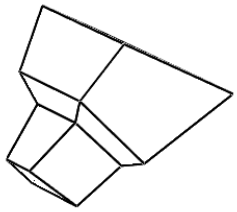


Figure 2. An image of a 3D polyhedron.

Assume that a 3D shape is represented by N points in the 3D space. These points could be points on visible surfaces of the 3D object. If the object subtends the central 20 deg of the visual field, as many as half of the 6 million cones are stimulated because most of the cones are in the central part of the human retina. So, N could be as large as 3 million. Now, as already Bishop Berkeley (1709) pointed out, each retinal point could be an image of any of the infinitely many points in the 3D space "out there", all the points being located on the line emanating from the retinal point and going through the center of perspective projection in the eye. Assume that the visual system tries to reconstruct the 3D points with spatial resolution of 1mm. Furthermore, assume that the object in front of the observer has the range in depth of 1m. This means that instead of considering infinitely many points on each projecting line, we *only* need to consider 1000 points. It follows that for a given 2D retinal image represented by 3 million points, the number of all possible 3D interpretations is $1000^{3,000,000}$ which is equal to $10^{9,000,000}$. Another astronomically large number. We know that the human visual system solves this insoluble problem in a fraction of a second despite the fact that the neurons in the brain are quite slow. How this is done will be described in Section 4. The last section will provide evaluation of what it means to have a machine that sees like us and what, if anything is missing in our effort of emulating human vision.

3. Algorithm for Solving Figure-Ground Organization Problem

As with every ill-posed inverse problem, a successful solution critically depends on the ability to impose effective *a priori* constraints on the family of possible interpretations. This is also true with FGO. There must

be a way to restrict the number of all partitions of 6,000,000 cones to just one. What are those constraints? First, an *object always comes in one piece*, rather than in many spatially discontinued pieces. Second, *objects are typically closer to the observer, than the background*. Assume that each object subtends a solid angle of 10 by 10 deg (this area is about the size of one's hand at the arm's viewing distance). If there was no occlusion, we could stack about 200 such objects in front of the observer (assuming that the visual field is equivalent to a surface of a hemisphere, whose solid angle is 20,626 deg^2). So, if an observer is trying to find objects in front of him, he will need to examine only 200 spatial neighborhoods on the retina, rather than $10^{38,000,000}$ possible partitions. Furthermore, these 200 spatially separate neighborhoods can be analyzed simultaneously in the human visual system due to the massively parallel architecture of the visual system. Clearly, *a priori* constraints can be quite effective. The constraints of spatial contiguity and large size dramatically reduced computational complexity of the problem. Next, we will describe which constraints are needed to actually find objects in the 3D scene and in the 2D image.

Consider the case of a binocular observer, or a monocular active observer. We simulate this case using a Pekee II robot equipped with a BumbleBee stereo camera. The robot's height is about 1m and the two lenses of the stereo camera are separated by 12 cm, which is roughly twice as large as the separation between the human eyes. The robot does not use any other sensors. In particular, the robot does not use the laser range sensor for reconstructing depth map. The robot looks at an indoor scene containing children furniture like that in Figure 1. Its task is to determine the number, positions and sizes of objects. Note that the floor is highly textured and it contains shadows and specular reflections. All this makes it very difficult to detect objects in the 2D images using conventional methods of texture and contour analysis. Building on Julesz's (1971) powerful demonstrations, in which a human observer was able to solve the binocular correspondence problem with random dot stereograms, our robot begins with establishing binocular correspondence of texture points, using the Sum of Absolute Difference Correlation algorithm (SADC) (Wong, Vassiliadis, & Cotofana, 2002). This stage is not error free, of course and it cannot be the basis of reliable 3D scene reconstruction. But the robot's goal is to solve FGO, not to reconstruct the 3D scene. Because objects are always spatially contiguous and important objects tend to be large (see the previous paragraph), the objects can be detected in the scene because there are always large number of depth samples close to each other in the 3D space. Using the distance between its cameras, the robot computes a 3D depth map (scales the binocular disparities) of visible points. The next step is to improve the signal to noise ratio by using another set of *a priori* constraints.

The next set of constraints is the *a priori* knowledge that all objects rest on a *common horizontal ground* due to gravity. *The ground is at a known distance below the cameras* (this distance is called the height of the observer). Using these two constraints, it is natural to estimate, as the next step, the floor in the 3D depth map and eliminate it from further processing (Faugeras, 1993, p.209). This is quite easy because in a typical scene many 3D points are actually floor points. Furthermore, if the robot, like a human observer, knows the *orientation of its cameras relative to the gravity*, the estimation of the floor calls for nothing more than fitting a known plane to 3D points and determining which points are close to the floor or below the floor (3D points below the floor represent noise). All these points are removed from the depth map because the floor points represent background and we are interested in detecting objects, called "figures". It is important to point out that a richly textured floor is not a problem for our robot. In fact, the texture is actually helpful for establishing the binocular correspondence. In contrast, conventional algorithms cannot work with richly textured backgrounds because this makes the separation of figure from ground impossible.

After floor points are removed from the 3D depth map, only object points remain. Now, the robot "mentally" rotates the remaining 3D depth map to simulate viewing the scene from above. There are two good reasons to perform this rotation. First, *all natural objects have prominent vertical structures* such as legs, surfaces and edges. This is true about animal and human bodies, as well as architectural constructions and furniture. In the presence of vertical gravitational force, vertical legs and surfaces are mechanically more stable. Because there are many vertical structures in the natural environment, there are many texture points representing these structures that lead to our 3D depth map. When these 3D points are projected orthographically onto a horizontal surface, such as floor, there is a very strong signal indicating the presence of the objects.

The second reason to perform "mental rotation" to simulate looking at the scene from above is related to the fact that most *objects reside on a common horizontal ground plane*. Exceptions are lamps hanging from the ceiling or a book lying on a desk. This means that if the 3D scene is actually viewed from above there will not be many partial occlusions of some objects by others. Partial occlusions are a rule rather than exception in ordinary visual images like that in Figure 1. Occlusions are common because some objects are

farther away than others from the observer. If a camera were mounted on the ceiling, there will not be many occlusions. Instead of mounting a physical camera on the ceiling (which is not practical), the robot "mentally" rotates the visible 3D points to simulate looking from above.

The result of the orthographic projection of the 3D points in our depth map on the horizontal plane is shown in Figure 3. When the viewing distance is large, the depth error from binocular disparity becomes large, too. Therefore, in this example, the algorithm only detects the objects that are within 4 meters in front of its cameras.

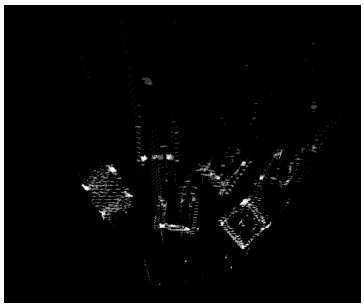


Figure 3. Orthographic projection of 3D points representing objects in Figure 1 on the horizontal plane.

Now is the time to identify individual objects in the 3D scene because these objects correspond to clusters of points. Any of the standard clustering methods can be used, but additional *a priori* constraints can greatly improve the result. For example, we expect *rectangular objects* whose *size is within some range*. We assume that the strong signal in the orthographic projection is caused by the vertical structures of the object. As a result, the projection of the 3D points onto the floor represents the shapes of the orthographic projection of the individual objects. So, we fit rectangles to the projection and estimate the position, size and orientation of each rectangle. The result of such fitting is shown in Figure 4. The green boxes show the fitted rectangles and the red boxes show the ground truth.

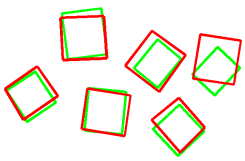


Figure 4. The green rectangles were fitted to the points in Figure 3. The red boxes show the ground truth.

In the case of the severely occluded object (the rightmost stand in the second row), our algorithm detects it and computes its position successfully, although it fails to compute its orientation accurately. This failure results from small amount information about this object. At this stage we can claim that our *robot solved FGO in the 3D representation*, on the floor. This solution can be used for planning visual navigation in the scene.

We want to emphasize one new and very important aspect of the 3D FGO. As you can see from Figure 4, the robot produces a spatially global map of its environment (the floor plan) from a single viewing position. Specifically, the robot recovers the invisible spaces behind the objects. As a result, the robot knows how much space is between objects regardless whether this space is directly visible or not. This is essential because it allows the robot to plan its navigation path in the scene even before it starts to move. This contrasts with conventional SLAM (Simultaneous Localization And Mapping) methods where the robot has to explore the environment and to reconstruct the visible surfaces from different places in advance to build the map and, at the same time to localize itself (Durrant-Whyte, Bailey, 2006). In dynamic environments, building spatially global map must be "instantaneous". Otherwise, by the time the map is produced, the environment would have changed. But solving 3D FGO is not the end of processing. More can be accomplished with these results.

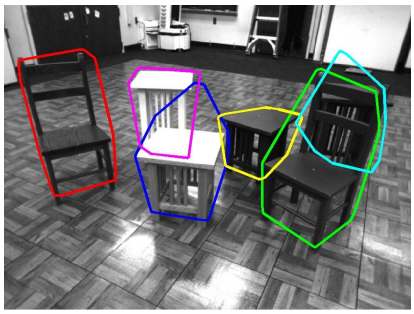


Figure 5. Detected objects in the image are circumscribed by color polygons.

The robot has an estimate of each object's height, which means that a 3D circumscribing box can be formed for each object. We then project the occluding boundaries of these boxes to one of the 2D perspective images that were used to solve the 3D FGO. The result is shown in Figure 5. The color curves in the image are estimated convex hulls of 2D projections of original objects. Some of these curves partially overlap as they should because the objects themselves partially overlap. These color curves represent the *solution of FGO in the 2D image*.

Figure 6 is the diagram of the algorithm and it shows how to identify individual objects from a pair of stereo images. The algorithm was tested on a DELL T5500

computer, and on average, it can acquire and process about 4 pictures (with the resolution of 512x384 pixels) per second.

Once we know which regions in the 2D image represent individual objects, we can proceed with extracting important 2D contours that represent essential aspects of the 3D shape. How a 3D shape can be recovered from a single 2D perspective image will be described next.

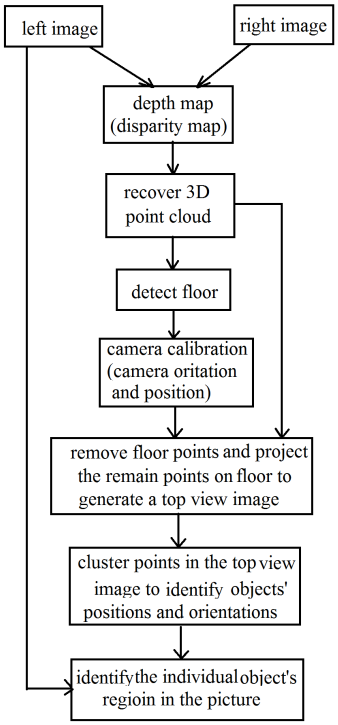


Figure 6: The diagram of the algorithm.

4. Algorithm for recovering a 3D shape from a single 2D image

Unlike all prior algorithms and models for 3D shape reconstruction, we begin by asking about the nature of *a priori* constraints, rather than about the nature of visual data. Constraints proved so useful in solving FGO that the reader should be confident that they will be essential in 3D shape recovery, as well. Computational complexity of 3D shape reconstruction is so large (see Section 2) that without *a priori* constraints it just does not seem possible to reconstruct 3D shapes and scenes accurately and reliably. This observation has been supported by enormous amount of results, both theoretical and empirical. In human vision, researchers have been testing for dozens of years the observer's ability to judge depth relations among points and surfaces, as well as the ability to judge 3D orientation of surfaces. Most of these experiments were done with amorphous stimuli that precluded the visual system from using *a priori* constraints. This was done on purpose; the researchers wanted to study 3D perception not "contaminated" with *a priori* constraints. These experiments universally showed that perception of depth and surface orientations is very poor: there are large systematic errors across viewing conditions, across observers and even across replications of the same experiment with the same observer. These results should have provided a warning sign that visual data without *a priori* constraints is not the way to go. But the researchers wanted to study the most general type of stimulus devoid of familiarities or constraints of any kind. But shapes of natural objects are not devoid of regularities; shapes of all animals are mirror symmetrical. Their parts, as pointed out by Biederman (1987), conform to translational symmetry (Biederman called the family of shapes that are used to represent parts, "geons"). Finally, flowers often represent rotational symmetry. Once we acknowledge that most, if not all important natural object are symmetrical, then the symmetry *a priori* constraint is no longer a "contamination" of one's experiment, but it becomes its essential part. Using Brunswik's (1956) terminology, a stimulus in one's laboratory experiment must be "ecologically valid." How far can one go with a symmetry constraint? The answer is, all the way. The symmetry constraint can often reduce the enormous family of possible 3D interpretations to a unique and the correct one.

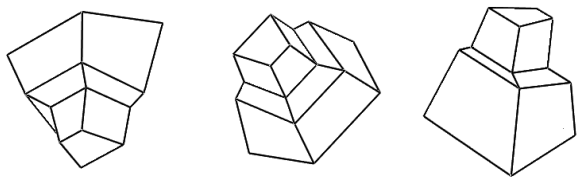
Similar efforts with similar outcomes took place in the machine vision community. The researchers became aware of such *a priori* constraints as symmetry at least as early as 1981(Kanade, 1981; Gordon, 1989) , but the use of strong *a priori* constraints has not become the main stream of research. Marr's (1982) paradigm, with its emphasis on surfaces, dominated the field. When this paradigm failed, instead of exploring the role and availability of constraints, machine vision community switched to 2D operations on 2D images. The hope was that modern machine learning methods will be able to extract invariant signatures of 3D objects. Despite some moderate progress, recognition performance of these 2D "appearance models" did not come even close to the performance of human observers. During the last decade, the present authors have provided psychophysical evidence showing that *a priori* constraints such as 3D symmetry and planarity of contours are essential in 3D shape perception (Pizlo, 2008; Pizlo et al., 2010). If these constraints cannot be applied to the family of possible 3D interpretations of a 2D image, shape constancy performance is at chance level. When these constraints can be applied, performance is close to perfect. So, now we do not have to deal with a question as to whether or not constraints should be used. They should and they are used. The real question is how to design a 3D shape

recovery model whose performance will match that of a human observer.

Consider a set of N points in a 3D space forming a mirror symmetric configuration. This means that there is a plane in 3D such that N/2 of the 3D points are mirror images of the other N/2 points with respect to this plane. Recall that in the absence of a symmetry constraint, when the task is to reconstruct the depth of N points from a single 2D perspective image, there are N free (unknown) parameters, the depth values of all N points. When these N points form a 3D mirror symmetrical configuration, there is no unknown! The 3D configuration is uniquely recovered! Recall the degree of uncertainty that we estimated in Section 2 for the case of 3D shape recovery. It should now be obvious that when a designer of a robot vision system is faced with a decision of adding additional visual data vs. an effective *a priori* constraint, she should choose the latter. This is what we did.

We showed not only how to apply a symmetry constraint to a 2D perspective image, but also to a 2D orthographic image (Li et al., 2009; see also Vetter & Poggio, 1994). Next, we showed how the human visual system combines the 3D symmetry constraint with two 2D retinal images (Li et al., 2011). This is the case of a binocular observer or an active monocular observer. Besides 3D symmetry and planarity of contours constraints, we showed that the human visual system uses 3D compactness constraint (Li et al., 2009). 3D compactness is a well known concept in mathematical physics (Polya & Szego, 1951), but it has never been used as a constraint in visual perception. A 3D compactness is defined as V^2/S^3, where V and S are the volume and the surface are of an object or of its convex hull.

Figure 7 shows three views of a 3D shape recovered by our model based on the 2D image shown in Figure 2. We want to point out two important aspects of our shape recovery model: (i) it recovers 3D shape without measuring 3D distances; 3D distances can be reconstructed after the 3D shape is recovered, and (ii) it often recovers the back invisible parts of the object as well as its front visible ones. This contrasts with Marr's (1982) 2.5D sketch. According to Marr, the visual system can inform the observer only about the visible surfaces of the 3D object. The back, invisible ones are completed by memory. Our robot can "see" the entire object. This is critical because it gives the robot knowledge of where the object ends on its back, "invisible" side. Note that so far, the symmetry correspondences for the image of a 3D shape have been established manually. Once the symmetric points are identified in the 2D image, the recovery is instantaneous.



(a) (b) (c)

Figure 7. Three views of the recovered polyhedron from the 2D image in Figure 2.[1]

To compare the recovery of our model with human performance, we measured the subjects' percept of the symmetric polyhedra, like that in Figure 2, in binocular and monocular viewing conditions. In the case of binocular condition, the subject viewed the stereo images of polyhedra through stereoscopic shutter glasses. On each trial, two objects were shown. The reference 3D shape was a stationary object shown on the left (monocularly or binocularly). The test 3D shape was shown on the right monocularly. The test shape was rotating around the vertical axis so that the subject could see many views of this shape. The slant of the symmetry plane of the reference 3D shape was 15, 30, 45, 60 or 75 degrees. The subject was asked to adjust the aspect ratio of the test 3D shape so that it matched the reference 3D shape. The rationale behind this adjustment task is related to the fact that a single 2D orthographic image of a symmetric 3D shape, determines this shape up to only one free parameter – its aspect ratio (Vetter & Poggio, 1994). Four subjects were tested and each ran 100 trials for each viewing condition (20 trials per each of the five slants).

Figure 8 shows the comparison of monocular and binocular performance of a human subject in 3D shape recovery to the performance of our model. The bottom line is that the model's performance is very close to the subject's performance. Specifically, in monocular viewing, there are systematic errors in recovering the aspect ratio of the 3D shape when the slant of the symmetry plane is close to 0 or 90 deg. Slants 0 and 90 deg represent "degenerate views". They are called degenerate because with such views the 3D symmetry constraint is ineffective. Note that the systematic errors of the model are very similar to the systematic errors of the subject. In binocular viewing, all systematic errors disappear. Both, the subject and the model see the 3D shapes veridically. This is the first empirical and computational demonstration of perfect 3D shape perception.

[1] Our polyhedral stimuli consisted of two boxes whose bottom faces were coplanar. So, each polyhedron had 28 edges and all of the edges were included in the object's representation, despite the fact that some of them were coplanar, like those seen in Figure 7c.

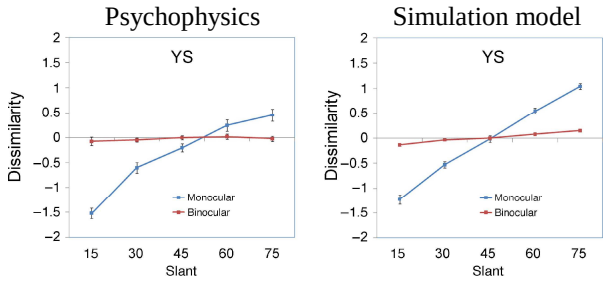


Figure 8. Performance of one subject (YS) is shown on the left. Performance of our model simulating YS's performance is shown on the right (from Li et al., 2011). The subject and the model performed a 3D shape recovery task based on the information provided by one or two images (monocular vs. binocular viewing). The horizontal axis shows the slant of the symmetry plane of the 3D reference shape. The vertical axis shows the error in the recovered aspect ratio of the test 3D shape using a log scale.

5. Conclusions

We explained two fundamental functions of the visual system: Figure-Ground Organization and 3D shape recovery. These two functions have been universally considered insoluble problems because of their inherent ill-posedness: there are simply too many possible solutions. We showed that these problems can be solved and a unique and correct interpretation produced when effective *a priori* constraints are used. We "explained" these functions in the sense that we formulated computational models and implemented them in a seeing robot. By doing this, we followed Richard Feynman's proposal "what I cannot create, I don't understand".

Is there any other problem to solve before the robots can interact with us? There is only one such problem. To apply the symmetry *a priori* constraint, the robot must be able to detect where, in the 3D object, the symmetry transformation is present. This is not trivial because the 2D image of a 3D symmetrical object is, itself asymmetrical. To detect a 3D symmetry from perspective images, one has to use perspective invariants of symmetry and formulate a robust method of using them with real images of real scenes. We have already made good progress in solving this problem. For more examples illustrating how our algorithm solves FGO problem, as well as demos of 3D scene and shape recovery of synthetic and real objects, as well as robot navigation, please refer to the following links http://web.ics.purdue.edu/~li135/Demo.html and http://web.ics.purdue.edu/~li135/Demo/People/objdet.mov.

6. References

Berkeley, G. (1709/1910). *A new theory of vision.* New York: Dutton.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94 (2),* 115-147.

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

Clowes, M.B. (1971) On seeing things. *Artificial Intelligence, 2,* 79-116.

Durrant-Whyte, H., Bailey, T. (2006). Simultaneous Localization and Mapping (SLAM): Part I the Essential Algorithms. *Robotics and Automation Magazine 13(2),* 99-110.

Faugeras, O. (1993). *Three-dimensional computer vision: A geometrical viewpoint.* Cambridge: MIT press.

Gordon, G. G. (1989). Shape from symmetry. In *Proceedings of SPIE, Intelligent Robots and Computer Vision VIII: Algorithms and Techniques, 1192,* 297-308.

Guzman, A. (1968) Decomposition of a visual scene into three-dimensional bodies. *Proceedings of AFIPS Conference, 33,* 291-304. Washington, DC: Thompson.

Huffman, D.A. (1971) Impossible objects as nonsense sentences. In: Meltzer, B. & Michie, D. (Eds.), *Machine Intelligence, vol. 6* (pp. 295-323) Edinburgh: Edinburgh University Press.

Julesz, B. (1971). *Foundations of cyclopean perception.* Chicago: University of Chicago Press.

Kanade, T. (1981) Recovery of the three-dimensional shape of an object from a single view. *Artificial Intelligence, 17,* 409-460.

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

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

Mach, E. (1906/1959) *The Analysis of Sensations.* New York: Dover.

Marr, D. (1977) Analysis of occluding contour. *Proceedings of the Royal Society of London, B 197,* 441-475.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Pizlo, Z. (2008). *3D shape: Its unique place in visual perception.* Cambridge: MIT press.

Pizlo, Z., Sawada, T., Li, Y., Kropatsch, W. & Steinman, R. M. (2010). New approach to the

perception of 3D shape based on veridicality, complexity, symmetry and volume. *Vision Research, 50 (1),* 1-11.

Poggio, T., Torre, V. & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317 (6035),* 314-319.

Polya, G. & Szego, G. (1951). *Isoperimetric inequalities in mathematical physics.* Princeton: Princeton University Press.

Tikhonov, A. N. & Arsenin, V. Y. (1977). *Solutions of ill-posed problems.* New York: Wiley.

Vetter, T. & Poggio, T. (1994). Symmetric 3D objects are an easy case for 2D object recognition. *Spatial Vision, 8 (4),* 443-453.

Waltz, D. (1972/1975) Understanding line drawings of scenes with shadows. In: Winston, P.H. (pp. 19-91) New York: McGraw-Hill.

Wertheimer, M. (1923/1958) Principles of perceptual organization. In: D.C. Beardslee & M. Wertheimer (Eds.) *Readings in Perception*, pp. 115-135. NY: D. van Nostrand.

Wong, S., Vassiliadis, S., & Cotofana, S. D. (2002). A sum of absolute differences implementation in fpga hardware. *Proceedings 28th EUROMICRO conference.* 183-188.

Acknowledgements

This project was supported by AFOSR, DOE, DOD, NSF.

Author Biographies

YUNFENG LI is a postdoctoral research fellow in the Department of Psychological Sciences at Purdue University. He received his Ph.D. degree in Psychology from Purdue University in 2009. His research interests are in 3D shape and scene recovery, figure-ground organization, invariants of symmetry. His work combines psychophysical experiments, computational and mathematical modeling.

TADAMASA SAWADA is a postdoctoral research fellow in Department of Psychological Sciences at Purdue University. He received a Ph.D. degree in Psychophysics from Tokyo Institute of Technology in 2006. He has been working on human perception of 3D shape, 2D and 3D symmetry, figure-ground organization and perception of 3D scenes. His work involves psychophysical experiments and computational modeling.

ZYGMUNT PIZLO is a Professor of Psychology and of Electrical and Computer Engineering at Purdue. He received his Ph.D. degree in Electronic Engineering in 1982 and his Ph.D. degree in Psychology in 1991. He served as the president and vice president of the Society for Mathematical Psychology. He published over 100 journal and conference papers on all aspects of visual perception, as well as motor control, problem solving and image quality. In 2008 he published the first monograph on human shape perception.