NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Tadamasa Sawada

A COMPUTATIONAL MODEL THAT RECOVERS DEPTH FROM STEREO-INPUT WITHOUT USING ANY OCULOMOTOR INFORMATION

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: PSYCHOLOGY
WP BRP 106/PSY/2019

T. Sawada [1]

A Computational Model that recovers depth from stereo-input without using any oculomotor information[2]

It is commonly believed that the visual system requires oculomotor information to perceive depth from binocular disparity. However, any effect of the oculomotor information on depth perception is too restricted to explain depth perception under natural viewing conditions. In this study, I describe a computational model that can recover depth from a stereo-pair of retinal images without using any oculomotor information. The model shows that, at least from a computational perspective, any oculomotor information is not necessary for perceiving depth from the stereo retinal images.

Keywords: binocular disparity; stereo vision; P3P problem; multiple view geometry

JEL Classification: Z

PsycINFO Classification Categories and Codes: 2323 Visual Perception

[1] Assistant Professor, PhD, tsawada@hse.ru (tada.masa.sawada@gmail.com), School of Psychology, National Research University Higher School of Economics.

[2] This manuscript was prepared as a result of a project "Visual perception in our everyday life" within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant № 19-04-006) and by the Russian Academic Excellence Project «5-100».

Introduction

Visually perceiving a 3D scene and the 3D shapes of objects within the scene is a difficult problem but we all perceive them veridically in our everyday life. Some have assumed that such veridical 3D perception requires some kind of extra-retinal information, usually memories of past experiences of moving ourselves, or our arms, or eyes around in our 3D world. Any effect of such memories has been questioned, however, by psychophysical studies on the effect of memory on 3D visual perception (Hochberg & Brooks, 1962; Hochberg & Hochberg, 1952; Hochberg & McAlister, 1955; Mershon & Gogel, 1975). Studies by Pizlo and his colleagues have shown that the veridical perception of the 3D shapes of familiar objects can be explained better by *a priori* constraints than by the memorized shapes of the objects (Pizlo, 2008; Pizlo, Sawada, Li, Kropatsch, & Steinman, 2010; Pizlo, Li, Sawada, & Steinman, 2014). This group has developed and tested computational models that emulate veridical human 3D shape and scene perception rather well by using only a few *a priori* constraints (priors), namely, 3D symmetry, the planarity of contours, minimum surface area, and maximum compactness. The present study addresses another classical visual problem in which oculomotor information has been assumed to be essential for the perception of 3D depth. Here, "binocular disparity", which is an important input for eliciting and controlling slow vergence eye movements and disjunctive saccades, could provide information needed for the veridical perception of depth (see Erkelens, Van der Steen, Steinman, & Collewijn, 1989; Erkelens, Steinman, & Collewijn, 1989; Collewijn, Erkelens, & Steinman, 1995; for studies of vergence eye movements). It has been widely assumed that our visual system uses binocular disparity to perceive depth. The present study will show that the perception of depth can be recovered entirely on the basis of geometrical optics. Our visual system does not need to make use of any oculomotor information. Note that my research problem, when viewed within the rubric called "Inverse Problem Theory," is a "Direct" problem because its solution does not require the use of any a priori constraints (aka priors, see Pizlo, Sawada, Li, Kropatsch, & Steinman, 2010; Pizlo, Li, Sawada, & Steinman, 2014; Sawada, Li, & Pizlo, 2015). This Direct problem will be solved first by making a computational model that recovers depth without being given any oculomotor information or any *a priori* constraints. Having a computational model that can solve the Direct Problem of perceiving depth by using only geometrical optics prepares the way for finding out whether human beings can do this, too. Now, consider what we know about how the geometry involved in binocular disparity can be used to recover depth.

Human eyes are separated about 6.5 cm which means that the retinal images of 3D scenes will be slightly different from one another. This difference between a stereo-pair of retinal images is called "binocular disparity". Binocular disparity is one of several depth cues that the human visual system uses to perceive depth within 3D scenes. Depth perception, based on binocular disparity, has been studied for centuries. It is one of the best studied topics in visual science (see Howard & Rogers, 2012 for a review).

Binocular disparity is often decomposed into its horizontal and vertical components (Read, Phillipson, & Glennerster, 2009). Horizontal disparity plays the major role in the perception of depth when it is based on binocular disparity. It has been assumed that the visual system needs oculomotor information about the relative orientation between the two eyes to recover depth from horizontal disparity (Mayhew & Longuet-Higgins, 1982; Peek, Mayhew, & Frisby, 1984; Erkelens & van Ee, 1998). This kind of oculomotor information can be estimated from the efference copy of the oculomotor signal (Skavenski, Haddad, & Steinman, 1972; Matin, Matin, & Pearce, 1969; Skavenski, 1971; Sommer & Wurtz 2002).

Another source of oculomotor information is the 2D distribution of vertical disparity (e.g. Gillam and Lawergren, 1983; Howard & Kaneko, 1994). It has also been shown that depth perception based on horizontal disparity is affected by the vertical disparity distribution. This is often referred to as an "induced" effect. This induced effect is often explained by saying that the visual system estimates the relative eye orientations from the vertical disparity distribution. Psychophysical results also suggest that the human visual system relies on the distribution of vertical disparity, rather than on the oculomotor efference signal, whenever the information in the disparity distribution is sufficiently reliable (Mitsudo, 2007; Mitsudo, Kaneko, & Nishida, 2009; Backus, Banks, van Ee, & Crowell, 1999; Bradshaw, Glennerster, & Rogers, 1996). The visual system's speed, however, for processing the vertical disparity distribution is rather low (Ames, 1946; Caziot, Backus, & Lin, 2017; Fukuda, Kaneko, & Matsumiya, 2006; Ogle, 1938). These authors showed that the visual system needs around 500 msec for processing a change of the vertical disparity distribution. Note, however, that the human beings' intersaccadic intervals during maintained fixation are often shorter than 500 msec (e.g. Steinman, Cunitz, Timberlake, & Herman, 1967; Cunitz & Steinman, 1969), which means that the visual system must be able to use a very efficient mechanism for processing binocular disparity. This mechanism must work fast whenever saccadic eye movements occur frequently. Now that we have considered the role of vertical disparity in the perception of depth, we will consider depth perception based on horizontal disparity.

The visual system can process horizontal disparity for each point, or for each pair of points, while the visual system processes the distribution of vertical disparity. The visual system encodes horizontal disparity as *absolute* disparity first, and then converts it to *relative* disparity (Chopin, Levi, Knill, Baveli, 2016; Neri, Bridge, & Heeger, 2004; Norcia, Gerhard, & Meredith, 2017). This absolute disparity is the difference between the eccentricity angles of a point between a stereo-pair of retinal images. This relative disparity is the difference between the visual angles of two points between the retinal images (Erkelens & Collewijn, 1985a, b).[3] It has been shown that the perception of depth based on horizontal disparity primarily depends on the relative disparity (Westheimer, 1979; Erkelens and Collewijn, 1985b; Regan, Erkelens, & Collewijn, 1986; Cottereau, McKee, & Norcia, 2012). Note that the relative disparity, as well as the visual angle, is invariant against any eye movement. Potentially, this invariance of the visual angle could allow the visual system to recover depth from binocular disparity in the presence of eye movements, but note, also, that all

[3] Relative disparity can also be computed as the difference between the absolute disparities of two points (Chopin, Levi, Knill, Baveli, 2016; Schor, 2000; Westheimer & McKee, 1979). Note that this method of computing relative disparity assumes that two lines-of-sight from a stereo-pair of foveae intersect with one another at a point within a 3D scene. But note that this assumption may be violated under a natural viewing conditions. Malinov, Epelboim, Herst, & Steinman (2000) showed that the two lines may be skewed with respect to one another.

prior modeling of depth perception from binocular disparity has assumed that oculomotor information was either given or recovered first.

Models of depth perception based on binocular disparity can be categorized into two types. The first type, either implicitly or explicitly, recovers oculomotor information from binocular disparity itself *before* recovering depth (e.g. Longuet-Higgins, 1982; Mayhew & Longuet-Higgins, 1982; Peek, Mayhew, & Frisby, 1984). The second type of model implicitly assumes that the necessary oculomotor information is available. Many existing models of depth perception, based on binocular disparity, use images on a computer screen representing the left and right eyes as input to the models, rather than the retinal images in the eyes. Note that the retinal image is a two-dimensional projection of the image on a screen. This means that the retinal and the computer screen images can be transformed into one another, but doing this requires knowing both the positions and the orientations of the eyes relative to the screen. The second type also includes a model that represents 2D visual information by using a head-centric coordinate system (Erkelens & van Ee, 1998; Koenderink & van Doorn, 1976; Zhang, Cantor, & Schor, 2010). The retinal image represented in a retino-centric coordinate system can be transformed into the head-centric representation, but, once again, this transformation requires knowing the positions and orientations of the eyes relative to the head.

Our computational model recovers depth from a stereo-pair of retinal images without recovering or being given any oculomotor information. This model is based entirely on the pure geometry of optics. It does not use any *a priori* constraints. The depth recovered is represented in a head-centered coordinate system, except for a rotation around the interocular axis between the two eyes. Both the process for recovering depth and the representation of the recovered depth does not vary with eye movements.

Model

This study used a "pinhole" camera with a perfectly spherical retina as the model for our human eye. This simplified eye has the center of its optics and its center of rotation at the center of a spherical eyeball. Note that when this simplified eye rotates around its optical center, the position of the eye's optical center does not change. Lines of projection from any pair of points in a 3D scene will intersect with one another at the optical center of this simplified eye. The visual angle between these lines is the same as the distance between the projections of these points on the spherical retina of this simplified eye, and the visual angle does not change when the eye rotates. Now, consider a 3D scene composed of N points. The model developed in this study represents the retinal image of a scene as a set of $N(N-1)/2$ visual angles between pairs of the points. This representation does not tell us where the projections of the N points are on the retina relative to the fovea but they do not change when the eye rotates.

The model recovers the depth of a 3D scene from a stereo-pair of its retinal images that are represented as two sets of visual angles. The correspondence between projections of points between the stereo-pair of retinal images is taken as a given in this study. This allows our model to recover depth without using any information about the orientations of the eyes. The model recovers the 3D depth of a scene by using an optimization method. The 2D space in this optimization is characterized by the shape of a triangle formed by the arbitrary selection of any 3 points within the scene. This triangle is discussed in next section.

3D interpretations of a triangle based a stereo-pair of its retinal images

Consider 3 points P_1, P_2, and P_3 in a 3D scene and the triangle T_{123} formed by these points. The 3D scene is viewed by the eye E_L (Figure 1). The visual angles between all pairs of these points are shown and labeled as: $\angle P_1E_LP_2$, $\angle P_2E_LP_3$, and $\angle P_3E_LP_1$. The length of an edge $|P_1 - P_2|$ between P_1 and P_2 can be set to 1 without any loss of generality. This length specifies the size of T_{123}. The shape of T_{123} is controlled by two angles, namely, $\angle P_3P_1P_2$ and $\angle P_1P_2P_3$ of T_{123}. If the shape of T_{123} is given, four, or fewer than four, possible positions of E_L relative to T_{123} can be determined. This problem is referred to as the Perspective-3-Point (P3P) problem (Fischler & Bolles, 1981; Gao, Hou, Tang, & Cheng, 2003; Sawada & Minkov, 2018). The positions of E_L were computed with an algorithm used to solve the P3P problem (Fischler & Bolles, 1981; see also Sawada & Minkov, 2018).

Now, consider what happens when T_{123} is viewed by another eye E_R. The visual angles at E_R for all pairs of the points are shown and labeled as: $\angle P_1E_RP_2$, $\angle P_2E_RP_3$, and $\angle P_3E_RP_1$. Four, or fewer than four, possible positions of E_R, as well as E_L, relative to T_{123} can be computed for the given shape of T_{123}.

Recall that the shape of T_{123} can be controlled by two angles, namely, $\angle P_3P_1P_2$ and $\angle P_1P_2P_3$ of T_{123}. This means that the positions of both E_L and E_R are also controlled by $\angle P_3P_1P_2$ and $\angle P_1P_2P_3$. There are only 16 possible combinations of the positions of E_L and E_R (4 for each E_L and E_R) for a given shape of T_{123}.

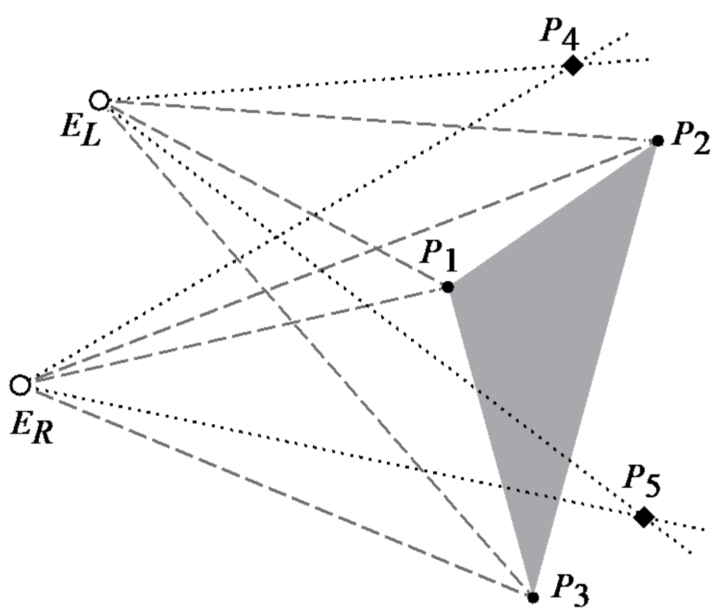

Figure 1. A stereo-pair of eyes E_L and E_R and the triangle T_{123} formed by points P_1, P_2, and P_3 and additional two points P_4 and P_5 in a 3D scene.

Recovering the Depth of a 3D Scene by Solving a 2D Optimization Problem

The shape of the triangle T_{123} and the positions of both E_L and E_R can be determined if the retinal images of an additional two points, P_4 and P_5, viewed by E_L and E_R, are given. Consider P_4 first. The visual angles between P_4 and the vertices of T_{123} at E_L and E_R are labeled as $\angle P_1E_LP_4$, $\angle P_2E_LP_4$, $\angle P_3E_LP_4$., $\angle P_1E_RP_4$, $\angle P_2E_RP_4$, and $\angle P_3E_RP_4$. If the shape of T_{123}, E_L, and E_R are given, the lines of projection to P_4 from E_L and from E_R can be written as $E_L + k_{4L}V_{4L}$ and $E_R + k_{4R}V_{4R}$ where k_{4L} and k_{4R} are free parameters and V_{4L} and V_{4R} are 3D unit vectors. The vectors V_{4L} and V_{4R} can be computed as follows:

$$(P_1 - E_L \quad P_2 - E_L \quad P_3 - E_L)^T V_{4L} = \begin{bmatrix} |P_1 - E_L| \cos \angle P_1E_LP_4 \\ |P_2 - E_L| \cos \angle P_2E_LP_4 \\ |P_3 - E_L| \cos \angle P_3E_LP_4 \end{bmatrix} \tag{1}$$

$$(P_1 - E_R \quad P_2 - E_R \quad P_3 - E_R)^T V_{4R} = \begin{bmatrix} |P_1 - E_R| \cos \angle P_1E_RP_4 \\ |P_2 - E_R| \cos \angle P_2E_RP_4 \\ |P_3 - E_R| \cos \angle P_3E_RP_4 \end{bmatrix} \tag{2}$$

where $|V_{4L}|$ and $|V_{4R}|$ are 1. Note that these two projection lines should intersect with one another at P_4 in a 3D scene, if the scene specified by the shape of T_{123}, E_L, and E_R, is a valid 3D interpretation of the stereo-pair of the retinal images of T_{123} and P_4. The distance Δ_4 between the two projection lines of P_4 can be computed as:

$$\Delta_4 = \frac{|E_L - E_R|(V_{4L} \times V_{4R})}{|V_{4L} \times V_{4R}|} \tag{3}$$

These two projection lines are skewed with respect to one another in the scene if $\Delta_4 \neq 0$ and they are not parallel to one another. The distance Δ_4 between the projection lines is the length of the shortest line segment whose endpoints are on the projection lines. These two endpoints can be written as $E_L + \acute{k}_{4L}V_{4L}$ and $E_R + \acute{k}_{4R}V_{4R}$ where $\acute{k}_{4L}$ and $\acute{k}_{4R}$ represent the distance of the endpoints from E_L and E_R. For simplicity, $\acute{k}_{4L}$ and $\acute{k}_{4R}$ will be referred to as the distance of P_4 from E_L and E_R later in this section. The distance Δ_5 between the two projection lines from E_L and from E_R to P_5 can be computed in the same way as Δ_4 (see Equations 1, 2, and 3).

Some of 16 possible combinations of the positions E_L and E_R are invalid. Note that Δ_i, $\acute{k}_{iL}$, and $\acute{k}_{iR}$ should be always positive if the 3D scene specified by the combination of E_L and E_R is a valid 3D interpretation of the stereo-pair of the retinal images of T_{123} and P_i. The combination of E_L and E_R, and the scene specified by this combination are also invalid if there is any set of 3 points (say P_{j1}, P_{j2}, and P_{j3}) that satisfy either of the following conditions:

$$\begin{cases} E_R - E_L = w_{j1L}V_{j1L} + w_{j2L}V_{j2L} + w_{j3L}V_{j3L} \\ |w_{j1L} + w_{j2L} + w_{j3L}| = |w_{j1L}| + |w_{j2L}| + |w_{j3L}| \end{cases} \tag{4}$$

or

$$\begin{cases} E_L - E_R = w_{j1R}V_{j1R} + w_{j2R}V_{j2R} + w_{j3R}V_{j3R} \\ |w_{j1R} + w_{j2R} + w_{j3R}| = |w_{j1R}| + |w_{j2R}| + |w_{j3R}| \end{cases} \tag{5}$$

where w_{j1L}, w_{j2L}, w_{j3L}, w_{j1R}, w_{j2R}, and w_{j3R} are constants, V_{j1L}, V_{j2L}, and V_{j3L} are vectors from E_L to P_{j1}, P_{j2}, and P_{j3}, and V_{j1R}, V_{j2R}, and V_{j3R} are from E_R to P_{j1}, P_{j2}, and P_{j3}. These equations show an invalid case in which some of the points P_{j1}, P_{j2}, and P_{j3} are behind the head of the observer in the scene. After eliminating these invalid combinations of the positions of E_L and E_R, the best combination can be determined from the remaining valid combinations by ascertaining that the following function is minimized:

$$\sum_{i=4}^{N_P} \frac{\Delta_i}{\sqrt{\acute{k}_{iL} + \acute{k}_{iR}}} \tag{6}$$

where N_P is the number of points in the scene, Δ_i represents the distance between the two lines of projection from E_L and E_R to the i-th point P_i (see Equation 3), and $\acute{k}_{4L}$ and $\acute{k}_{4R}$ represents the distance of P_i from E_L and E_R. Note that the number of points N_P in the scene can be more than 5 ($N_P \geq 5$). These additional points are used in the same way as P_4 and P_5 for determining the best valid combination of the positions of E_L and E_R in Equation (6).

The model recovers the depth of the 3D scene by using an optimization method. The 2D space in this optimization is characterized by $\angle P_3P_1P_2$ and $\angle P_1P_2P_3$ that are angles of T_{123} (Figure 2). The cost function that is minimized in this optimization is given in Equation (6). Once this is done, the optimization process of the depth recovery can be written as:

$$\underset{T_{123}}{\arg\min} \sum_{i=4}^{N_P} \frac{\Delta_i}{\sqrt{\acute{k}_{iL} + \acute{k}_{iR}}} \tag{7}$$

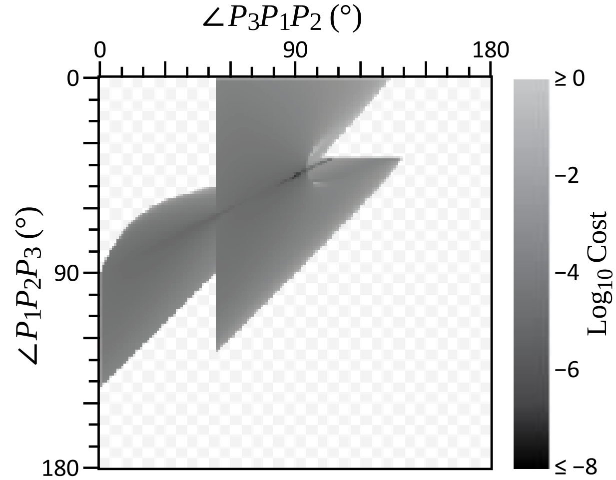

Figure 2. A 2D distribution of the cost (Equation 6) computed from a stereo-pair of the retinal images of a simple 3D scene with 5 points: $P_1 = [-20\ -20\ 57]^t$, $P_2 = [-20\ 20\ 57]^t$, $P_3 = [20\ -20\ 57]^t$, $P_4 = [20, 20, 57]^t$, $P_5 = [0, 0, 57]^t$. This scene was viewed from a stereo-pair of eyes at $[-3.3\ 0\ 0]^t$ and $[3.3\ 0\ 0]^t$. The abscissa and ordinate of these graphs represent $\angle P_3P_1P_2$ and $\angle P_1P_2P_3$ of the triangle T_{123} respectively. The grayscale levels indicate the cost computed with Equation (6). The checkered regions indicate invalid shapes of the triangle T_{123}. These shapes are invalid either because they are inconsistent with the retinal images or because 3D interpretations of the retinal images do not satisfy the condition specified by Equations (4-5). Note that these distributions are not unimodal. They have multiple local minima. The global minimum of the distribution was found by using an exhaustive search method that sampled the distribution at every 0.2° of $\angle P_3P_1P_2$ and of $\angle P_1P_2P_3$. The global minimum (1.10×10^{-14}) of the distribution was found at ($\angle P_3P_1P_2$, $\angle P_1P_2P_3$) = (90°, 45°), which represents the veridical shape of T_{123}. The minimum cost was not exactly 0 because of rounding and discretization errors.

The scale of the recovered scene is proportional to $|P_1 - P_2|$, which was set to be 1, but note that the scale cannot actually be determined no matter how many retinal images of the scene are available (Longuet-Higgins, 1981). Also, note that a line segment between E_L and E_R in the recovered 3D scene represents the interocular-axis between the stereo-pair of eyes. The scene should be scaled so that the length of the segment $|E_L - E_R|$ in the scaled scene becomes equal to the interocular distance of the observer when the interocular distance is given (around 6.5 cm for an adult human).

The position of the observer's head in the recovered 3D scene can be determined from the positions of E_L and E_R with one free parameter, namely, a rotation around the interocular axis. This means that the recovered 3D scene can be represented in a head-centric coordinate system. Finally, recall that a 3D scene, which is represented in a head-centric coordinate system, does not vary when the eye moves.

Computer Simulation

The mathematical validity and the computational robustness of the model were tested in a simulation experiment. In each trial of this experiment, a 3D scene, composed of points, was randomly-generated and a stereo-pair of its retinal images (visual angles between the points) were computed. The model was given these retinal images and used them to recover the 3D scene. The recovered 3D scene was evaluated by comparing the shapes of triangle T_{123} in the original and recovered scenes:

$$\sqrt{(\alpha_1 - \acute{\alpha}_1)^2 + (\alpha_2 - \acute{\alpha}_2)^2} \tag{8}$$

where α_1 and $\acute{\alpha}_1$ are $\angle P_3P_1P_2$ of the original and recovered scenes and α_2 and $\acute{\alpha}_2$ are $\angle P_1P_2P_3$ of the original and recovered scenes. This equation represents distance in the 2D space of the cost distribution (Figure 2) between the points representing the original and recovered scenes.

Five hundred 3D scenes were generated for each session of the experiment. The points in each scene were randomly-positioned within a range specified in the scene relative to the observer's head. The depth positions of the points from the observer were between γ and 2γ in front of the observer's head where γ is a free parameter (10, 40, and 160 cm). The head-centric eccentricity of these points was less than 45°. Eccentricity was defined as the angles between a vector along the direction of depth and vectors to the points from the observer's cyclopean eye (the midpoint between the observer's stereo-pair of eyes). The interocular distance between the stereo-pair of eyes was 6.6 cm. Note that the number of points in the scene and γ were blocked during the session.

The results of the simulation of the 3D scenes with 5 points are shown in Figure 3A. The ordinate shows the discrepancy between the original and recovered scene (Equation 8), and the abscissa shows the range of the depth positions of the points (γ). The width of the plot represents frequency (Hintze & Nelson, 1998). The 3D scene recovered by the model is never perfect. The discrepancy between the original and recovered 3D scenes can be attributed to the optimization process used to find the global minimum of the cost distribution (Equation 7).[4] The exhaustive search method used for the optimization process and the cost distribution was sampled at every 0.2° of $\angle P_3P_1P_2$ and of $\angle P_1P_2P_3$. The discrepancy between a perfectly veridical scene and the recovered scene could be more than $0.282° \approx (0.2^2 + 0.2^2)^{0.5}$ even when the cost distribution was unimodal[5]. There were also cases where our exhaustive search method produced the local minimum of the distribution rather than its global minimum. Note that a cost distribution may have multiple global minima (Kruppa, 1913/2017; Thompson, 1959). Also note that it can be difficult to know whether

[4] It was confirmed in a separate session that the cost in the distribution was virtually zero ($< 10^{-10}$) when $\angle P_3P_1P_2$ and $\angle P_1P_2P_3$ was given in a perfectly veridical 3D scene.

[5] For example, consider finding the global minimum of the following unimodal function by using the optimization method: $-e^{-(y-0.04x-0.4)^2/0.1^2} - e^{-x^2/100^2}$. Theoretically, the global minimum of this function is -2.00 at $(x, y) = (0, 0.4)$. But note that an exhaustive search method will estimate that the global minimum is -1.99 at $(x, y) = (-10, 0)$ when the equation is evaluated at every integer of x and y. The difference between the positions of this estimated global minimum and the real global minimum is substantially larger than $1.412 \approx (1^2 + 1^2)^{0.5}$.

any local minimum of the cost distribution is the actual global minimum with any numerical optimization method. This problem has not been addressed in this study.

Figure 4 shows the number of recovered 3D scenes that were nearly veridical. The ordinate shows the number of recovered 3D scenes whose discrepancy from perfectly veridical 3D scenes (Equation 8) was less than 1°. Symbols indicate the number of points in the 3D scenes and the abscissa shows the range of the depth positions of the points. The model could recover only about 60% of the 3D scenes veridically with 5 points, but it could recover about 90% with 6 points (see also Figure 3B). It is possible that the 6th point helped the model avoid local minima. Having more than 6 points only improved the model's performance a little (Figure 3B, 4).

The results of the simulation experiment showed that the model can recover a 3D scene from a stereo-pair of its 2D retinal images veridically and reliably when there are 6, or more than 6 points, in the scene.

Recall, the model uses 3 of the points in the 3D scene to define the 2D space of an optimization problem and uses the other points to compute the distribution of cost in the optimization space. Having these separate processes for recovering a 3D scene allowed us to develop a model that used a readily available algorithm that was developed to solve the P3P problem. Unfortunately, this algorithm, which is based on the P3P problem, does not resemble any known mechanism in our visual system.

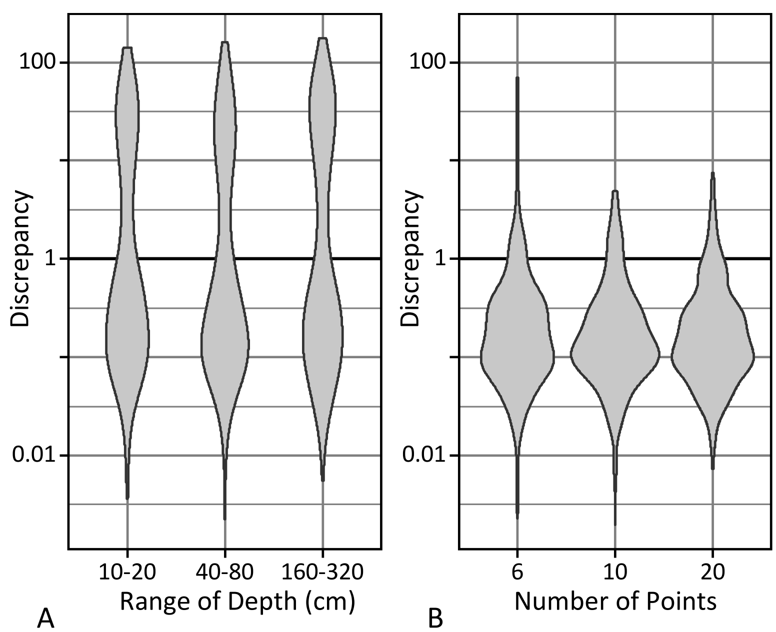

Figure 3. shows the frequency of the discrepancy (Equation 8) between the recovered 3D scenes and the perfectly veridical 3D scenes. The ordinate shows the size of the discrepancy and the width of the plot represents the frequency (Hintze & Nelson, 1998). (A) The abscissa of this graph shows the range of the depth positions of the points (γ). The number of points in the 3D scenes was 5. (B) The abscissa also shows the number of points in the 3D scene. The depth positions ranged between 40 and 80cm (γ = 40cm).

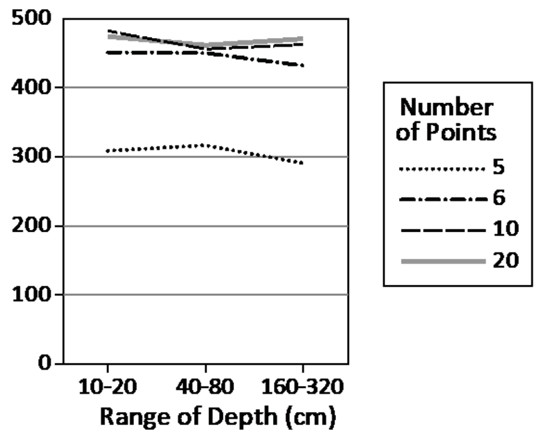

Figure 4. shows the number of recovered 3D scenes where the discrepancy from perfect veridicality was less than 1°. The abscissa shows the range of the depth positions of the points in the scene.

Discussion

The model developed in this study can recover depth in a 3D scene from a stereo-pair of retinal images without making use of the relative orientations of the eyes. The model represents the retinal image of the scene as a set of visual angles between pairs of points within the scene. The model uses only these visual angles as its input. This means that eye movements play no role in the recovery of depth. This is possible because depth is recovered within a head-centered coordinate system. Having such a model allows us to consider whether the human visual system can recover depth in a 3D scene from retinal images in the same way that our model does this (Brewer & Lambert, 2000).

This model shows that, at least from a computational perspective, the human visual system should be able to perceive depth by using only a stereo-pair of retinal images without any oculomotor information. It also shows that the perception of depth need not change when the eyes move. This can be described as a constancy of depth perception with different fixation points. Note that these properties of this visual system are consistent with our everyday life experience. Our perception of depth is reliable and it stays that way when we move our eyes (Logvinenko, Epelboim, & Steinman, 2001; Logvinenko & Steinman, 2002). These properties also allow the visual system to process stereo retinal images across eye movements which could improve the precision of depth perception (Enright, 1991; Wright, 1951).

This model is based entirely on the geometry of the optics of a schematic eye. It is not related to any known mechanisms in our visual system. This fact encourages us to revise this algorithm to make its recovery process plausible with respect to current psychophysical and neuroscientifical evidence. Once this has been done, we will have a realistic, as well as an effective, model of human stereo-depth perception.

References

Ames, A. (1946). Binocular vision as affected by relations between uniocular stimulus-patterns in commonplace environments. *American Journal of Psychology*, 59, 3, 333–357.

Backus, B., Banks, M. S., van Ee, R., & Crowell, J. A. (1999). Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vision Research*, 39, 1143–1170.

Bradshaw, M. F., Glennerster, A., & Rogers, B. J. (1996). The effect of display size on disparity scaling from differential perspective and vergence cues. *Vision Research*, 36, 9, 1255–1264.

Brewer, W. F., & Lambert, B. L. (2000). The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. *Philosophy of Science*, 68, S176–S186.

Bruner, J. S. (1973). The function of perceiving: New look retrospect. In J. S. Bruner (Ed.), Beyond the information given: Studies in the psychology of knowing (pp. 114–124). New York, NY: W. W. Norton.

Caziot, B., Backus, B. T., & Lin, E. (2017). Early dynamics of stereoscopic surface slant perception. *Journal of Vision*, 17(14):4, 1–17.

Chopin, A., Levi, D., Knill, D., & Bavelier, D. (2016). The absolute disparity anomaly and the mechanism of relative disparities. *Journal of Vision*, 16(8):2, 1–17, doi:10.1167/16.8.2.

Collewijn, H., Erkelens, C. J., & Steinman, R. M. (1995). Voluntary binocular gaze-shifts in the plane of regard: Dynamics of version and vergence. *Vision Research*, 35, 3335–3358.

Cottereau, B. R., McKee, S. P., & Norcia, A. M. (2012). Bridging the gap: global disparity processing in the human visual cortex. *Journal of Neurophysiology*, 107, 2421–2429, doi:10.1152/jn.01051.2011.

Cunitz, R. J., & Steinman, R. M. (1969). Comparison of saccadic eye movements during fixation and reading. *Vision Research*, 9, 683–693.

Enright, J. T. (1991). Exploring the third dimension with eye movements: Better than stereopsis. *Vision Research*, 31, 9, 1549–1562.

Erkelens, C. J., & Collewijn, H. (1985a). Eye movements and stereopsis during dichoptic viewing of moving random-dot stereograms. *Vision Research*, 25, 11, 1689–1700.

Erkelens, C. J., & Collewijn, H. (1985b). Motion perception during dichoptic viewing of moving random-dot stereograms. *Vision Research*, 25, 4, 583–588.

Erkelens, C. J., Steinman, R. M., & Collewijn, H. (1989). Ocular vergence under natural conditions. II. Gaze-shifts between real targets differing in distance and direction. *Proceedings of the Royal Society of London B*, 236, 441–465.

Erkelens, C. J., Van der Steen, J., Steinman, R. M., & Collewijn, H. (1989). Ocular vergence under natural conditions. I. Continuous changes of target distance along the median plane. *Proceedings of the Royal Society of London B*, 236, 417–440.

Erkelens, C. J., & van Ee, R. (1998). A computational model of depth perception based on head-centric disparity. *Vision Research*, 38, 2999–3018. http://www.ncbi.nlm.nih.gov/pubmed/9797995.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. Behavioral and Brain Science, 39, e229: 1-77, doi:10.1017/S0140525X15000965.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24, 381–395. http://refhub.elsevier.com/S0022-2496(17)30204-3/sb12

Fukuda, K., Kaneko, H., & Matsumiya, K. (2006). Vertical-size disparities are temporally integrated for slant perception. *Vision Research*, 46, 2749–2756.

Gao, X., Hou, X., Tang, J., & Cheng, H. (2003). Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 930–943. http://refhub.elsevier.com/S0022-2496(17)30204-3/sb13

Gillam, B., & Lawergren, B. (1983). The induced effect, vertical disparity, and stereoscopic theory. *Perception & Psychophysics*, 340, 121–130. http://www.ncbi.nlm.nih.gov/pubmed/6634367.

Hintze, J. L., & Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *American Statistician*, 52, 181–184.

Howard, I. P., & Rogers, B. J. (2012). *Perceiving in depth: Vol. 2. Stereoscopic vision*. New York, NY: Oxford University Press.

Howard, I. P., & Kaneko, H. (1994). Relative shear disparities and the perception of surface inclination. *Vision Research*, 34, 19, 2505–2517. https://www.ncbi.nlm.nih.gov/pubmed/7975290

Koenderink, J. J., & van Doorn, A. J. (1976). Geometry of binocular vision and a model for stereopsis. *Biological Cybernetics*, 21, 29–35. http://www.ncbi.nlm.nih.gov/pubmed/1244864.

Kruppa, E. (2017, December 25). To determine a 3D object from two perspective views with known inner orientation (G. Gallego, E. Mueggler, & P. Sturn, Trans.) Retrieved from https://arxiv.org/abs/1801.01454. (Original work published 1913 in *Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Kaiserlichen Akademie der Wissenschaften*, 122, 1939–1948)

Logvinenko, A. D., Epelboim, J., & Steinman, R. M. (2001). The role of vergence in the perception of distance: A fair test of the Bishop Berkeley's claim. *Spatial Vision*, 15, 77–97.

Logvinenko, A.D., & Steinman, R. M. (2002). Fixation on fixation impedes cognition: Reply to a Commentary by Kohly & Ono. *Spatial Vision*, 15, 387–391.

Longuet-Higgins, H. C. (1981). A computer algorithm for reconstructing a scene from two projections. *Nature*, 293, 133–135.

Longuet-Higgins, H. C. (1982). The role of the vertical dimension in stereoscopic vision. *Perception*, 11, 377–386. http://www.ncbi.nlm.nih.gov/pubmed/7182797.

Malinov, I. V., Epelboim, J., Herst, A. H., & Steinman, R. M. (2000). Characteristics of saccades and vergence in two kinds of looking tasks. *Vision Research*, 40, 2083–2090.

Matin, L., Matin, E., & Pearce, D. G. (1969). Visual perception of direction when voluntary saccades occur. I. Relation of visual direction of a fixation target extinguished before a saccade to a flash presented during the saccade. *Perception & Psychophysics*, 5, 2, 65–80.

Mayhew, J. E., & Longuet-Higgins, H. C. (1982). A computational model of binocular depth perception. *Nature*, 297, 376–378. http://www.ncbi.nlm.nih.gov/pubmed/7078648

Minkov, V., & Sawada, T. (2018). Seeing a Triangle in a 3D Scene Monocularly and Binocularly. *NRU Higher School of Economics. Series PSY Psychology*, WP BRP 91/PSY/2018. Retrieved from https://wp.hse.ru/en/prepfr_Psychology.

Mitsudo, H. (2007). Illusory depth induced by binocular torsional misalignment. *Vision Research*, 47, 1303-1314.

Mitsudo, H., Kaneko, H., & Nishida, S. (2009). Perceived depth of curved lines in the presence of cyclovergence. *Vision Research*, 49, 3, 348–361. http://doi.org/10.1016/j.visres.2008.11.004

Neri, P., Bridge, H., & Heeger, D. J. (2004). Stereoscopic processing of absolute and relative disparity in human visual cortex. *Journal of Neurophysiology*, 92, 1880–1891, doi:10.1152/jn.01042.2003.

Norcia, A. M., Gerhard, H. E., & Meredith, W. J. (2017). Development of relative disparity sensitivity in human visual cortex. *Journal of Neuroscience*, 37, 23, 5608–5619, doi:10.1523/JNEUROSCI.3570-16.2017.

Ogle, K. N. (1938). Induced size effect. I. A new phenomenon in binocular space perception associated with the relative sizes of the images of the two eyes. *Archives of Ophthalmology*, 20, 604–624.

Peek, S. A., Mayhew, J. E., & Frisby, J. P. (1984). Obtaining viewing distance and angle of gaze from vertical disparity using a Hough-type accumulator. *Image and Vision Computing*, 2, 180–190. http://doi.org/10.1016/0262-8856(84)90021-0